European Commission

Proceedings of the 2025 conference on

Big Data from Space (BiDS'25)

29 September – 3 October

Edited by P. Kempeneers, S. Lumnitz, and S. Albani

European Commission
The European Commission's science and knowledge service
Joint Research Centre

SATCEN
EUROPEAN UNION SATELLITE CENTRE

esa
→ THE EUROPEAN SPACE AGENCY

Joint Research Centre

Contact information
Pieter.Kempeneers (at) ec.europa.eu

The Joint Research Centre: EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC143704

PDF ISBN 978-92-68-31935-2 doi:10.2760/2119408 KJ-01-25-486-EN-N

Luxembourg: Publications Office of the European Union, 2025

How to cite this report: Kempeneers, P., Lumnitz, S. and Albani, S., *Proceedings of the 2025 conference on Big Data from Space (BiDS `25),* Publications Office of the European Union, Luxembourg, 2025, https://data.europa.eu/doi/10.2760/2119408, JRC143704.

Preface

In times marked by global crises, Europe's resilience–digital, societal, and environmental–has never been more vital. Being resilient means safeguarding our societies and our future in the digital space race: upholding our laws, values, and sovereignty in how data are accessed, shared, and protected. With the adoption of the European Union's Data Act (September 2025) and the world's first comprehensive AI Act, Europe has taken bold steps to ensure that innovation and competitiveness go hand in hand with trust, fairness, and fundamental rights. Together, these acts set the stage for a distinctly European way of governing the digital future, ensuring that Big Data and AI serve not just technology, but people, society, and the planet.

Big Data from Space (BiDS) is not just about observing the Earth. It is about enabling smarter, faster, and fairer decisions to meet the challenges of our time. From climate change and civil security to sustainable competitiveness and digital innovation, the insights we derive from spaceborne and terrestrial data are essential for evidence-informed action. In this sense, BiDS'25 was not only about advances in science and technology, but also about ensuring that data-driven innovation supports Europe's ability to address societal and security challenges.

Building on the success of the initiatives introduced in the last edition, BiDS'25 organised 23 satellite events, 28 demos, lightning talks, and birds-of-a-feather sessions. As in BiDS'23, code sprints were held with OSGeo and Pangeo to promote collaborative open-source software development. BiDS'25 also introduced new formats and perspectives, including an award session for the best start-up ideas. It demonstrated how scientific breakthroughs can evolve into entrepreneurial solutions and strengthen Europe's competitiveness.

Another novelty for this year's conference was the panel discussions that aligned with our mission of connecting deep tech with deep purpose. On Wednesday, October 1, "Society at Risk: Challenges and Priorities in the Space-Data Age", explored critical questions at the intersection of technology and societal well-being. On Friday, October 3, the final day of the conference, "From Raw Data to Real Decisions: Systems that Work in the Space-Data Age" focused on bridging raw data to actionable decisions through resilient infrastructure, AI, cybersecurity, and ethical frameworks–setting priorities for future research, policy, and societal impact. These conversations reflected our commitment to turning planetary-scale data into scalable solutions for global challenges.

Latvia's space sector, a partner of ESA with a rich historical heritage and modern technological strengths, brought unique value to this conference. With the support of the Investment and Development Agency of Latvia (LIAA), a B2B networking event and industrial exhibition highlighted Latvia's growing role in the European space economy.

The figures of BiDS'25 demonstrate both its reach and dynamism: two weeks before the start, 621 participants had registered. With more than 40 participants on average, the satellite events were fully booked, and 152 developers had registered for the code sprints. A total of 132 submissions were received from 34 different countries, including papers, demos, satellite, and award events. Each submission was reviewed by at least two experts from the Programme Committee. Of the 75 papers submitted, 42 were accepted as oral presentations and 21 as posters. While Earth Observation played a central role in this edition, BiDS covers the full spectrum of space domains–navigation, science and communications–and this breadth remains essential to its identity. The presentations were organised across seven thematic sessions: Towards Digital Twins: Integrating Data, Models, and Insight; Data Cubes: Advances and Applications; FAIR Workflows; Data Infrastructures & Services at Scale; GeoAI & Geospatial Intelligence; Optimising Processing from Edge to Cloud; Societal Applications: Risk, Resilience, and Resource Monitoring.

This year's BiDS'25 keynote talks captured the breadth and urgency of the challenges at the intersection of data, science, and society. Thomas Brunschwiler (IBM Research Europe, Switzerland) explored how tokens and embeddings are emerging as a new lingua franca for AI-driven Earth Observation, reshaping environmental science and data workflows through foundation models. Lynn Dudenhöfer (Senior Intelligence Professional) highlighted how the fusion of open-source intelligence and EO data is becoming indispensable for addressing rapidly evolving threats, from transnational organised crime to geopolitical instability, in a security landscape increasingly shaped by GenAI. Rosa M. Badia (Barcelona Supercomputing Center, Spain) connected the dots from Europe's world-class supercomputing infrastructure to domain-specific and FAIR application workflows and digital twins. She showed how HPC, AI, and data analytics can together deliver actionable solutions for society, with a focus on geohazards and risk mitigation. Together, these keynotes set the tone for BiDS'25: advancing science and technology while ensuring that innovation is directed towards

resilience, security, and sustainability.

Altogether, BiDS'25 offered a vibrant mix of science, technology, policy, and entrepreneurship. It stood at the intersection of climate resilience, societal security, and digital sovereignty, reminding us that responsible innovation is no longer optional. By bringing together diverse communities, BiDS'25 reaffirmed its role as the place where Europe and the world set the course for data-driven resilience and sustainability. Several contributions pointed to applications with direct impact on resilience and decision-making globally, as well as for EU Member States and institutions, showing a strong bridge between research and operations. This resonated with the European strategic context, where flagship programmes such as the EU Space Programme–covering Copernicus, Galileo, and other space infrastructures–together with key policy frameworks like the EU Strategic Compass and the EU Space Strategy for Security and Defence, supported the Union's ambition for leadership and strategic autonomy.

We express our sincere gratitude to the Programme Committee members for ensuring the excellence of this programme and these proceedings. Our local partner, the Ministry of Education and Science of Latvia, played a pivotal role in hosting BiDS'25. Thanks to their support, we gathered at the inspiring National Library of Latvia, and the University of Latvia hosted the satellite events. BiDS'25 again showcased the strong engagement of its three organising entities–SatCen, ESA, and JRC–which remain at the forefront of innovation and central to the themes of this edition. Ultimately, the outcomes of BiDS'25 extend beyond the scientific community to operational actors–including those working in security and decision-making–ensuring that innovation reaches end-users and has tangible impact.

Pieter Kempeneers, Stefanie Lumnitz,
Sergio Albani

Program Committee

Mirko Albani	European Space Agency (ESA)
Sergio Albani	EU SatCen
Conrad Albrecht	German Aerospace Institute
Anca Anghelea	European Space Agency (ESA)
Dani Arribas-Bel	The Alan Turing Institute
Vasileios Baousis	European Centre for Medium-Range Weather Forecasts (ECMWF)
Omar Barrilero	EU SatCen
Katie Baynes	NASA
Miguel Belenguer-Plomer	EU SatCen
Alessio Bozzo	EUMETSAT
Pierre-Marie Brunet	CNES
Paola De Salvo	GEO
Lorenzo Bruzzone	University of Trento
Matteo Bunino	CERN
Alessandro Burini	EUMETSAT
Michele Cecotti	European Commission, Joint Research Centre (JRC)
Laurence Chaoul	CNES
Chiara Chiarelli	European Commission, Joint Research Centre (JRC)
Massimo Ciscato	European Commission
Michele Claus	Institute for Earth Observation, Eurac Research
Mihai Datcu	DLR
Davide De Marchi	European Commission, Joint Research Centre (JRC)
Begum Demir	TU Berlin
Marco Di Tullio	European Space Agency (ESA)
Ewelina Dobrowolska	Serco Italy c/o ESA
Jean Dusart	European Commission, DG Research and Innovation
Karly De Baere	European External Action Service (EEAS)
Loïc Dutrieux	European Commission, Joint Research Centre (JRC)
Jonas Eberle	German Aerospace Center
Federico Fierli	EUMETSAT
Peter Friedl	German Aerospace Center (DLR)
Thomas Geenen	ECMWF
Rollin Gimenez	CNES
Patrick Griffiths	European Space Agency (ESA)
Thierry Ranchin	Mines Paris
Sven Tahon	European Fisheries Control Agency
Alexander Jacob	Eurac Research
Vasileios Kalogirou	EUSPA
Girts Karnitis	University of Latvia
Kaspars Karolis	Ministry of Education and Science (Latvia)
Linda Vecbiškena	Ministry of Education and Science (Latvia)
Marie-Françoise Voidrot	Open Geospatial Consortium
Tanya Walker	EARSC
Jacobo Matute	EU SatCen
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Iraklis Angelos Klampanos	University of Glasgow
Doris Klein	DLR
Alexander Kmoch	Department of Geography, University of Tartu
Manolis Koubarakis	National and Kapodistrian University of Athens
Jan Kucera	European Commission, Joint Research Centre (JRC)
Michele Lazzarini	EU SatCen

Bertrand Le Saux	European Commission DG CONNECT
Sébastien Lefèvre	Université de Bretagne Sud
Stefanie Lumnitz	European Space Agency (ESA)
Manil Maskey	NASA Marshall Space Flight Center
Gema Maza	EU SatCen
Inês Mendes	EU SatCen
Gabriele Meoni	European Space Agency (ESA)
Marco Minghini	European Commission, Joint Research Centre (JRC)
Jakub Nowosad	Adam Mickiewicz University
Tina Odaka	UMR LOPS IFREMER
Despina-Athanasia Pantazi	National and Kapodistrian University Of Athens
Ioannis Papoutsis	National Observatory of Athens
Edzer Pebesma	Inst for geoinformatics, Univ of Muenster
Fabrizio Pera	European Space Agency (ESA)
Fabrizio Pera	Serco
Salvatore Pinto	European Space Agency (ESA)
Edoardo Ramalli	European Commission, Joint Research Centre (JRC)
Connor Rhys Heeney	European Space Agency (ESA)
Sabrina Ricci	European Space Agency (ESA)
Federico Rondoni	Starion for ESA
Paula Saameno	EU SatCen
Paulo Sacramento	Solenix Deutschland GmbH
Darek Saunders	Frontex
Michael Schick	EUMETSAT
John Schnase	NASA
Céline Tison	CNES
Antonis Troumpoukis	NCSR Demokritos
Roberto Ugolotti	European Commission, Joint Research Centre (JRC)
Evelyn Uuemaa	University of Tartu
Corina Vaduva	University Politehnica of Bucharest
Linda Vecbiskena	Ministry of Education and Science
Claudia Vitolo	European Space Agency (ESA)
Wolfgang Wagner	Vienna University of Technology

Additional Reviewers

Barciauskas, Aimee
Gurung, Iksha

Table of Contents

Towards Digital Twins: Integrating Data, Models, and Insight

Data cubes: Advances and applications

FAIR workflows

Data Infrastructures & Services at Scale

GeoAI & Geospatial Intelligence

Optimizing processing from Edge to Cloud

Societal Applications: Risk, Resilience and Resource Monitoring

Risk, Disaster, and Hazard Management (Poster Session)

Water, Soil, and Ecosystems Monitoring (Poster Session)

EO Platforms, Data Access & Interoperability (Poster Session)

EO Data Processing & AI Methods (Poster Session)

EO for Humanitarian, Urban & Societal Applications (Poster Session)

TOWARDS A DIGITAL TWIN FOR DEFENCE AND SECURITY

Sergio Albani[1], Paula Saameno[1], Michele Lazzarini[1], Erik Korsbakken[2], Alberto Lorenzo[3], Maria Gorzynska[4], Andrea Masini[5], Joao Vinholi[6], Marco Chini[6], Andrea Patrono[1]

[1]European Union Satellite Centre, [2]European Defence Agency, [3]Indra Espacio S.L.U., [4]Tracasa S.A.U., [5]FlySight S.r.l., [6]Luxembourg Institute of Science & Technology

ABSTRACT

The Digital Twin Earth (DTE) developments are constantly growing, offering an innovative approach to understand and manage complex, rapidly evolving environments. While key DTE initiatives under development are addressing domains such as climate or extreme weather events, a dedicated DTE for Defence and Security (DTE4DS) will answer specific needs in these complex and sensitive domains. DTE4DS, a Joint Initiative (JI) launched by the European Defence Agency (EDA) and the European Union Satellite Centre (SatCen), addresses the management and exploitation of Earth Observation (EO) Big Data, together with a very large amount of collateral data sources for defence or security scenarios. The system is conceived to exploit synergies with existing relevant initiatives while offering unique capabilities to defence and security stakeholders.

Index Terms— Digital Twins, Earth Observation, Big Data, HPC, Security, Defence

1. INTRODUCTION

Security is an intricate subject in which a diverse number of scenarios can be triggered by different causes. Today, identifying causes of conflicts or disasters and understanding the link between these triggering events in diverse domains (e.g. climate, health, energy, food) and the impact on security is becoming more and more important not only for taking immediate responses, but also for policymaking. The evolution of security and its growing importance is highlighted in the most relevant European Union (EU) and global policies (e.g. EU Strategic Compass [1], EU Space Strategy for Security and Defence [2], EU Green Deal [3], UN Sustainable Development Agenda [4]), which are calling for advanced technological solutions to enhance current capabilities in the defence and security domains.

Digital twin is a concept widely used for design, simulation and validation of results; when applied to the modelling of the Earth and its phenomena, it is identified as a promising evolution axis to provide end users with a holistic disruptive solution that will change the access to (and the amount of) information to support decision-making processes in given scenarios. A DTE relies on the coherent exploitation of EO Big Data, including both satellite and aerial sources, and its integration of a wide variety of on-ground data sources tailored to each scenario needs. The combination of advanced AI analytics and High-Performing Computing (HPC) capacity makes it possible to generate a dynamic virtual representation of the real world that enables an interactive interface towards the user, enabling the human-triggered simulation of events and its forecasted impact. The most recent initiatives at EU level (e.g. DestinE[1]) show the potentialities of DTs, with a focus on EO data.

In defence and security, such a virtual model will mean changing the paradigm for decision-making in the field of Intelligence, Surveillance and Reconnaissance (ISR), boosting the ability to respond to diverse threats and crises in operational and strategic situations and, as final goal, also in tactical ones.

However, when addressing sensitive topics that directly affect security of citizens and societies, it is very important to consider which specific additional needs must be considered and the benefits provided for the relevant stakeholders for defence and security. The present work aims at presenting the status of the DTE4DS, an EDA-SatCen JI carried out in cooperation with industry.

1.1. Overview

The first phase of the EDA-SatCen DTE4DS JI consisted in a landscape study performed in a structured activity flow (Fig.1). The study started with a state-of-the-art analysis to assess the status of relevant technologies and to identify synergies with initiatives in the civil domain, and with the definition of a set of reference scenarios that served to identify high-level user needs. After interaction with potential users in EDA and SatCen user forums, the reviewed reference scenarios and preliminary user needs have been used to sketch a high-level system architecture, to identify the technology gaps and to define a coherent technology roadmap and development plan, paving the way for the future development phases of this DTE4DS.

[1] https://destination-earth.eu/

doi:10.2760/2119408

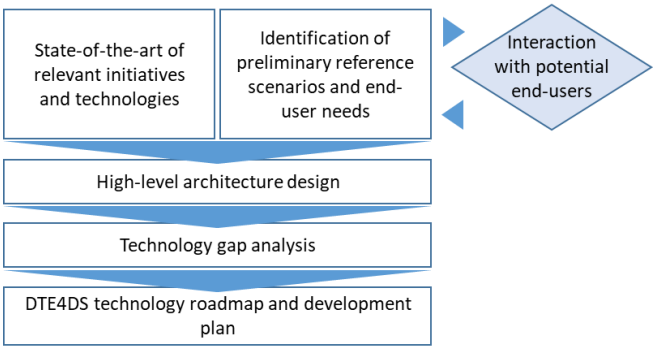


Fig. 1. DTE4DS landscape study flow

2. A DIGITAL TWIN EARTH FOR DEFENCE AND SECURITY: STATE-OF-THE-ART

2.1. DTE definition

If a Digital Twin is "*a realistic digital representation of something physical*" [5], or "*a virtual replica of a physical system whose performance it can help optimize*"[2], the DTE should be a virtual representation of the Earth that is georeferenced and connected to the world's digital knowledge archive. Such a system implements a Big Data lake and associated services that allow users to navigate through space and time, access historical data, and create future predictions. It can also be defined as digital replica of an Earth system component, structure, process, or phenomenon, obtained by merging digital modelling and real-world observational continuity – i.e. remote, in-situ, and synthetic data streams. A DTE must be seen as a living digital simulation model that updates and changes as its physical counterparts' change [6]. A DTE could also be seen as an entire ecosystem, integrating and orchestrating various DTs to allow for simulations and predictions of complex Earth scenarios.

2.2. Identified existing DTE initiatives

An exhaustive analysis of the most relevant DTE initiatives was performed during the first phases of the study (first quarter of 2025), identifying the key initiatives (Table 1) to be considered to advance in the DTE4DS development.

DestinE and EDITO projects are already functioning, however their development will continue in the next years (2030 as foreseen) to include new DTEs. In addition, there is a large and ever-growing number of local DTs initiatives in Europe (apart from manufacturing industries), forming part of digital transition strategy. The DTs are being implemented mostly at the city level, with the focus on urban planning, climate change adaptation, traffic controlling and emergency. They may be engaged in a future DTE ecosystem for downscaling simulations, predictions and training (e.g. for critical infrastructure protection).

So far, the use cases foreseen in these key initiatives do not consider directly the operational needs of users in the security/defence domain, which are, in general, more demanding in terms of reliability, accuracy and protection of information. Hence, a clear need emerges for the DTE4DS, a system specifically tailored to defence and security actors, guaranteeing maximization of synergies and no overlapping with initiatives in the civil domain. The DTE4DS would take EU capabilities for decision-making to the next level, changing the paradigm of usage of information for ISR activities.

Table 1: main DTE initiatives relevant for the DTE4DS

Initiative	Scope	Main stakeholders
DestinE[1]	EU funded initiative to develop a digital twin of our planet by 2030.	EC, ESA, EUMETSAT, ECMWF
EDITO[3]	A virtual representation of marine and coastal environments around the globe	EC, CMS, EMODnet
EDDI	Establish a robust digital twin framework tailored for defence applications, enabling enhanced simulation, planning, and predictive maintenance capabilities.	EDA, EU MoDs

3. REFERENCE SCENARIOS

Given the range of defence and security scenarios triggered by the current geopolitical situation, and the wide span of modelling specificities for each of them, it is needed to select a set of reference scenarios that serve as baseline to identify high-level user needs that drive the preliminary system design. The reference scenarios selected, based on interaction with potential users and relevant reports [7], [8], include:

1) Preparation of the battlefield
2) Forced displacement
3) Situational awareness for crises and disasters
4) Maritime surveillance

These scenarios contemplate different decision-making time spans related to the need for information (i.e. short, medium and long-term) and have served to identify end-user needs associated to each of them.

[2] Mirroring Reality. Digital Twins in Aerospace and Defence, Capgemini Research Institute (2023)

[3] https://dive.edito.eu/

doi:10.2760/2119408

4. DESIGN DRIVERS

The final technical implementation of the DTE4DS and its sharing rules should be carefully decided among relevant stakeholders. However, there are a few high-level design drivers to be considered since the early phases for the initiative, to ensure that the solution is feasible and disruptive, while aligned with user needs.

On one hand, the DTE4DS must be designed to: 1) increase the operational efficiency for decision-makers; 2) enhance the preparedness with predictive analysis; 3) provide new geospatial products for informed decision-making in the short/long term; 4) enhance the user experience 5) integrate user owned data into complex models to enhance the analysis.

On the other hand, the high-level user needs defined for each reference scenario and time span were key to understand the technology to be implemented in the DTE4DS, and the different data sources to be integrated. Therefore, the DTE4DS system shall be designed to:

1) Integrate very large amounts of EO data and collateral sources, considering high refreshment rates (up to continuous update of some sources) to ensure the provision of up-to-date and timely tailored information to different stakeholders involved in a given scenario in line with time span needs;
2) Provide forecasting capacity to assess impact of given events through advanced, yet reliable, AI models;
3) Implement a robust, secure and coherent integration of data to generate new information products and to support realistic simulations;
4) Allow a dynamic interface enabling advanced visualization (e.g. 2D/3D, AR/VR) and interaction between actors in scenario;
5) Offer the possibility to inject user-owned data to improve situational awareness capabilities.

On top of these design drivers, it will be also important to identify the standards to be implemented to guarantee usability among the security and defence communities and monitor continuously any ethics issues related to the usage of data sources with personal information.

5. DATA

While the core data will consist of Big EO Data, scenarios related to defence and security will also require data from additional sources, usually sensitive, to be integrated in the DTE4DS. This includes geolocation and positioning, on-ground imagery (e.g. video), social and environmental data and user-owned sources. Focusing solely on EO data, the variety and complexity of the geospatial big data needed to meet end-user needs across different applications is already considerable. The harmonization and adaptation of various data types (e.g. SAR, optical), along with different spatial and spectral resolutions and time series analysis of long-term archives, will present significant challenges that could be addressed by leveraging advanced AI models. Table 2 provides more details about the diversity of the data to be dealt with, along with the specific need for each time spans: near-real time, mid-term and long-term usage.

Table 2: Main data sources and usage identified accordingly to time span

Data	Near real-time usage	Mid-term usage	Long-term usage
EO data (satellite and aerial)	Quasi-real time VHR multispectral/ hyperspectral, and SAR (AoI < 100 km²)	VHR and HR multispectral/ hyperspectral and SAR (AoI 100 < 1000 km²)	Archive MR-LR multispectral/ hyperspectral and SAR (AoI > 1000 km²)
Video (in-situ)	Real-time UAV/drone feeds	Periodic aerial video updates	Archive videos
Demographic and socio-economic data	Basic overlays	Preliminary data fusion	Long-term resilience and trend analysis
Meteorological and climatological data	Realtime weather data	Short-term forecasts	Long-term climate trends
Geolocation data from mobile devices	Real-time tracking of population movements	Monitoring activity and displacement patterns	Analysis of long-term trends
Land cover maps, DEM (satellites, Lidar, etc.)	Up-to-date VHR	Up-to-date VHR	Up-to-date HR
Social media/citizens science	Real-time data about infrastructure status	Updated status on spot locations	Supportive use for long-term assessment
Security and military forces own-data sources	Real-time information for decision and actuation	Organization of short-term actions	Long-term analysis and preparedness

6. HIGH-LEVEL SYSTEM CONCEPT

The proposed high-level system architecture for the DTE4DS consists of multiple interconnected building blocks to ensure robust performance, security, scalability, and interoperability across different operational scenarios (Fig.2). The core architecture is structured into distinct layers to enable

doi:10.2760/2119408

modularity and efficient integration of very large and diverse datasets as well as advanced technologies.

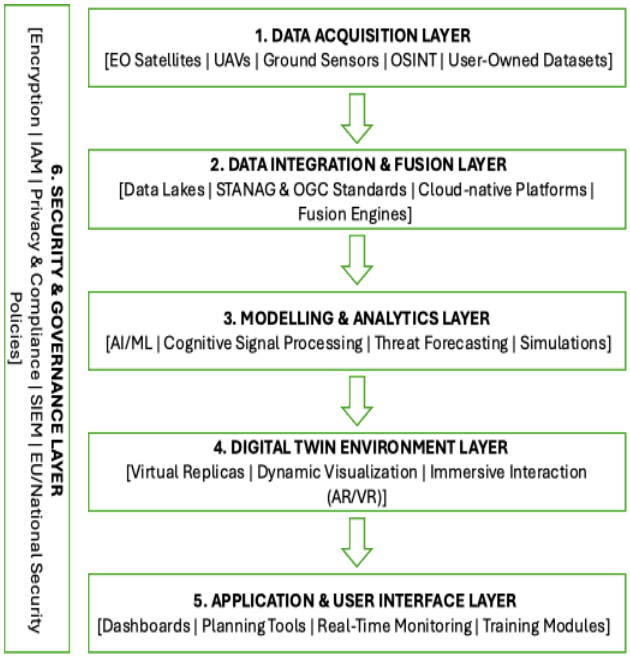


Fig. 2. High-Level System Architecture of DTE4DS

1. Data acquisition layer
This foundational layer gathers diverse data streams from multiple sources with different refreshment rates including, at least, the data listed in Table 2.

2. Data integration and fusion layer
Collected data is standardized, harmonized, and integrated through advanced fusion algorithms. This layer ensures data quality, interoperability and seamless ingestion into the digital twin environment. Technologies involved include cloud-native architectures, data lakes and data hubs compliant with STANAG and OGC standards.

3. Modelling and analytics layer
Sophisticated AI-driven analytics and simulation models provide predictive insights and scenario analysis. Techniques employed include machine learning (ML), deep learning (DL), cognitive signal processing and adaptive data fusion. Key capabilities include threat assessment, impact forecasting and decision-support simulations tailored specifically for defence and security stakeholders.

4. Digital twin environment layer
This core component hosts virtual replicas of defence and security operational environments, allowing dynamic representation, visualization and interaction. It supports immersive interfaces, enhancing situational awareness and collaborative scenario exploration.

5. Application and user interface layer
Customized interfaces and decision-support dashboards facilitate user interaction with the DT. This includes operational decision-making tools, planning modules, real-time monitoring and training environments. Advanced user experience (UX) design principles ensure usability and operational efficiency.

6. Security and Governance Layer
Given the sensitivity of defence and security applications, a transversal security layer with a robust cybersecurity framework is integrated across all layers. It comprises:

- Secure data transmission (encrypted communication protocols);
- Identity and access management (IAM);
- Data privacy and protection mechanisms;
- Security incident and event management (SIEM);
- Compliance and governance policies aligned with national and EU security regulations.

The combined architecture provides a secure, flexible and highly scalable digital ecosystem designed to consider continuous integration of new capabilities.

7. CONCLUSIONS

The present paper summarizes the status of the EDA-SatCen DTE4DS JI. The landscape study performed paved the way for the next phases that will focus on the prototyping of key technology blocks to derisk the full-scale system implementation.

REFERENCES

[1] European Union External Action Service (EEAS), “Strategic Compass: Security and Defence”, 2022.

[2] Joint communication to the European Parliament and the Council European Union, “Space Strategy for Security and Defence”, JOIN (2023) 9 final, 2023.

[3] European Commission, “The European Green Deal. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions”, COM (2019) 640, 2019.

[4] United Nations, “Transforming our world: the 2030 Agenda for Sustainable Development”, 2015.

[5] Bolton A, Butler L et all., The Gemini Principles, University of Cambridge, 2018.

[6] Naviti S., Craglia M., Mazetti P., “Digital Ecosystems for Developing Digital Twins of the Earth: The Destination Earth Case”, *Remote Sensing*, 2021.

[7] Niinistö S., Safer Together: Strengthening Europe’s Civilian and Military Preparedness and Readiness.

[8] European Commission: Directorate-General for Research and Innovation, Align, act, accelerate – Research, technology and innovation to boost European competitiveness, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2777/9106236.

doi:10.2760/2119408

ENHANCING LAND DIGITAL TWINS THROUGH THE INCORPORATION OF LAND-USE DATA IN NUMERICAL WEATHER FORECAST MODELS

Garik Gutman

NASA Headquarters, Washington DC

ABSTRACT

Although classified land-cover maps have been used in numerical weather forecast models, land-use datasets have not been incorporated at the same level, especially at a resolution of 1-10km, required in a Digital Twin design. Agricultural modifications and urbanization lead to changes in regional temperatures, roughness, albedos and flux distribution, affecting cloud and precipitation patterns, thus contributing to regional variability in weather. Improving representations of land-use distribution and dynamics is crucial for accurate weather forecasts. NASA is investing in the development of Earth System Digital Twins to better represent, predict, and investigate complex Earth system phenomena using advanced technologies, including machine learning and artificial intelligence. In particular, the NASA Land-Cover/Land-Use Change Program (LCLUC) is contributing to the development of Digital Twins by advancing the incorporation of available land-use data products based on timely updated moderate-to-high resolution satellite observations.

1. INTRODUCTION

Land-cover and land-use changes affect regional temperatures, roughness, albedos and flux distribution, leading to changes in cloud and precipitation patterns, thus contributing to regional variability in weather. Although classified land-cover maps have been used in numerical weather forecast models, land-use datasets have not been incorporated at the same level, especially at a resolution of 1-10km, required in a Digital Twin design.

In the contiguous United States, a huge portion of land surface have been altered by anthropogenic activities, such as irrigation, crops production and timber harvesting, urbanization, recreation activities. Additionally, wildland fires have been significantly impacting land cover characteristics. All these changes have direct impact on local and regional meteorology. A tool for assessing these impacts would be a Digital Twin based on an improved modeling system.

Digital Twin concept is based on developing an interactive, integrated multidomain, multiscale, digital replica of the state and temporal evolution of Earth systems. For weather forecast models, it would include dynamically integrated land-surface infrastructure and continuously assimilated timely observations of changes at land-atmosphere interface. Current challenges in numerical weather forecast modeling include relatively low horizontal resolution and low interactivity with the users. Higher resolution enables the incorporation of smaller-scale processes, described with physics, leading to higher fidelity in local information relevant for users and easier comparison with observations. Creation of an interface between model's outputs and users' inquiries is one of the advantages of DT concept compared to a common modeling approach.

2. THE LCLUC PROGRAM GOALS AND THE CURRENT OBJECTIVES

The primary goal of the NASA LCLUC program is to use satellite observations to improve our understanding of changes at land surface as an essential component of Earth System Science. The LCLUC program includes studies that detect and quantify changes in land cover and land use; examine their impact on the environment and interactions with climate and society; and model future scenarios of LCLUC impacts. The LCLUC program has been developing interdisciplinary research combining aspects of physical, social, and economic sciences, with a high level of societal relevance, using remote sensing data, methods, and tools. The LCLUC program aims to develop the capability for annual satellite-based inventories of land cover and land use to characterize and monitor changes at the Earth's surface. Land use is a human activity therefore social and economic science research plays a crucial role within the LCLUC program. It includes quantifying the impacts of changes in human behavior at various levels on land use, land-use impacts on society, and how the physical, social and economic aspects of land-use systems contribute to, are impacted by and adapt to environmental changes. LCLUC, ubiquitous worldwide, is having a significant

doi:10.2760/2119408

impact on the environment, the provision of ecosystem services, and human livelihoods at the regional, national, or global scale, often with economic and policy implications. The policy implications can, for example, be in terms of current policies that have prompted or exacerbated land-use change, or policy changes that would lead to sustainable land-use practices.

As extreme weather events become more frequent and severe, they pose significant threats to ecosystems, infrastructure, and communities. For instance, hurricanes can cause coastal erosion, deforestation, and destruction of urban infrastructure, while droughts may lead to agricultural failures, water shortages, and ecosystem degradation. Heatwaves can affect human health or lead to wildfires that devastate forests and grasslands. Similarly, floods and storms disrupt entire regions, damaging crops, displacing populations, and creating health risks. By combining advanced satellite data on land use with weather forecast models and socioeconomic indicators, LCLUC studies will capture the spatial and temporal dynamics of these events, providing insights into their impact on land use, economic stability, and societal well-being.

The goal of the current work is to contribute to Digital Twins' development for improving weather forecasts by providing near real-time data on human land uses at the highest spatial and temporal resolutions, useful in simulations of the ongoing interactive processes in the Earth's system.

3. DATA AND MODELS USED

Incorporation of land surface data in numerical models range from coarser spatial resolution sensors, such as MODIS on Terra and Aqua and VIIRS on Suomi-NPP platforms to medium-to-high resolution, such as Landsat and Sentinel-1 and -2, to very high-resolution data from commercial vendors, such as Planet Lab. Researchers also use observations from instruments on board the International Space Station, such ECOSTRESS (infrared data) and GEDI (lidar data). Models also utilize the NLCD & NLUD data (Fig. 1).

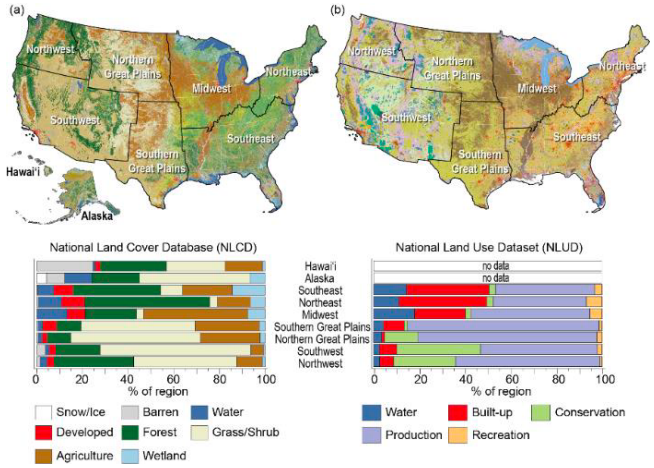


Fig. 1 Satellite-derived National Land Cover Database (NLCD) and (b) the National Land Use Dataset (NLUD) [1]

Various models have been applied in advancing the incorporation of land-use data. They include Land Surface Models, such as NOAH Multi-parameterization (MP) Heterogeneous Urban Environments (HUE) and NOAH-MP-Crop models [2], [3], [4], Community Land Model (CLM) [5], NCAR-Community Earth System Model (CESM) [6] and Weather Research and Forecasting (WRF) models, including NASA Unified WRF (NU-WRF) model [7] along with meteorological fields and re-analysis data, such as MERRA or NCEP/NCAR reanalysis.

4. TOWARDS IMPROVING WRF MODELS

The above models are being used by several NASA-funded LCLUC Program's teams to study the effect of incorporating satellite-derived dynamic boundary conditions in state-of-the art WRFs to improve the weather forecasts in urban and agriculture sectors.

4.1. Urban studies

The LCLUC ongoing studies are focused on urban infrastructure (buildings, greenspace, water features), which would account for variability at scales finer than the horizontal grid resolution of operational numerical weather prediction for US cities. These projects use WRFs with a multilayer urban canopy model to generate target surface flux fields as well as urban heat island and precipitation fields. Once validated, these fields form the targets to train a model to identify effective land-use parameters for the coarse operational models and for machine learning and deep learning algorithms component to a land surface model

doi:10.2760/2119408

to capture the sub-grid scale effects. In particular, the role of green infrastructure in coupled land-atmosphere prediction is assessed.

The remote sensing component include mapping efforts of highly detailed urban elements such as trees, paved areas, buildings, turf and natural grasses. Those are used to develop biophysical variables such as leaf area index, albedo, emissivity, building heights, and roughness lengths as input to the surface models used, such Noah-MP HUE, currently coupled with WRF models, integrates surface hydrological processes, including impervious area-to-vegetation water transfers and pavement shading and canopy interception. This, in turn, allows for a more representative urban environments and urban energy partitioning, which leads to a more realistic coupling between surface and atmosphere.

One research project focuses on appropriate spatial aggregations of remote-sensing data into the lower boundary conditions and determine model sensitivity to the various parameters. To compare simulations with in-situ field reference data values the team uses observations captured by a set of meteorological stations and then use the results to study urban heat island mitigation efforts in the downtown urban areas. The research teams construct multi-resolution datasets from a variety of platforms, suitable for estimating fractional, subpixel coverage estimates for multiple aggregations of land use. The datasets are further used by machine learning algorithms for classification and modeling. The data are then entered into boundary layer components of numerical weather forecast models. The hindcast weather simulations are constructed and compared with past and current data collected from a variety of ground-based sources, including updated stations with radiation shields and cellular data transmission, which provide the accurate representation and real-time collection of air temperature.

Another study is developing a new, high-resolution urban albedo dataset based on Landsat and Sentinel-2, separating roofs from impervious ground in the NLCD impervious surface dataset. The research team conducts and analyzes WRF simulations with the new urban albedo dataset and implement this dataset into publicly released WRF versions. The improved characterization of the albedo parameters in WRF will improve the simulation of urban meteorological variables and thus empower stakeholders and researchers to better navigate urban planning and policies.

4.2. Agricultural studies

To determine the impact of land use on vegetation-atmosphere feedback and drought development in the U.S. agricultural lands, land cover type and irrigation data are being used in numerical experiments. The NU-WRF coupled regional model is used to conduct sensitivity experiments to determine the impact of land cover type, irrigation fraction, irrigation strategy, and initial conditions on the onset and amplification of the rapid emergence and onset of land drying and vegetation stress. A prototype of Land Digital Twin is being designed to examine a range of scenarios for exploring land use impacts on short range weather conditions (atmospheric temperature, aridity, boundary layer growth, cloud development, and precipitation) and drought development. The goal is to better inform decision-making under forecasted drought by improving society drought preparedness through changes in land management strategies. The NU-WRF short-range and sub-seasonal weather forecast uses MODIS-derived land cover data and irrigation data from the U.N. Food and Agriculture Organization.

To improve weather prediction and better understand agriculture-weather interactions, NASA LCLUC researchers combine remote sensing data and machine learning techniques. They develop a suite of dynamic high-resolution annual crop and irrigation data over the continental US during the past 25 years to incorporate them into WRF/Noah MP-Crop model. The current static, outdated crop and irrigation input maps in WRF are being replaced by the suite of annual maps at 30- m field-scale historical and in-season crop types with rotation patterns, state-level crop planting and harvesting dates, the 4-km crop growing degree days, and 5-yearly field-scale 30-m irrigation area maps. Model sensitivity is being tested to quantify key factors affecting weather prediction, associated mechanisms, and uncertainty, in non-extreme and extreme conditions, including extreme precipitation, heat waves, and droughts. In developing a prototype Land Digital Twin, some researchers use Google Earth Engine, which facilitates the identification, processing, and transfer of land-use data that are further utilized in the NU-WRF model. Machine learning and Artificial Intelligence are employed to assess the impact of land surface changes on regional weather.

NASA researchers strive to integrate heterogeneous datasets to assess the implications on regional sustainability, focusing on extreme weather risks and crop production. With an interactive

doi:10.2760/2119408

interface, a Digital Twin prototype will allow stakeholders to engage in an analysis of agricultural scenarios to support sustainable land management.

5. CONCLUSIONS

The NASA LCLUC program is in a good position to contribute towards developing Earth System Digital Twins. The current paper is focused on LCLUC efforts in advancing improvements to Numerical Weather Regional Forecast Models by incorporating available land-use data on a weekly scale. The projects just have started, so it is too early to provide results of the extent the weather forecasts would improve after land-use data have been incorporated. Incorporation of dynamic spatial distribution of land use is specifically focused on providing better boundary conditions on such processes as urban structural changes and agricultural practices, such as crop rotation and irrigation.

LCLUC studies will provide indication which land-use variables are critical in models and the scale at which their impact becomes important for improving forecasts. The next step will include assessment of socio-economic impacts, which will imply the use of social science and econometric models and data. Additional efforts within the program, not described here, will be made to advance Earth System models using land-use data globally and regionally on a longer-term, decadal scale. Ultimately, Land Digital Twins based on these studies will be beneficial to stakeholders and economy.

REFERENCES

[1] Sleeter, B.M., T. Loveland, G. Domke, N. Herold, J. Wickham, and N. Wood, 2018: Land Cover and Land-Use Change. In Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II [Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 202–231. doi: 10.7930/NCA4.2018.CH5

[2] https://vlab.noaa.gov/web/emc/noah-lsm

[3] https://ral.ucar.edu/model/noah-multiparameterization-land-surface-model-noah-mp-lsm

[4] Liu X., F. Chen, M. Barlage, G. Zhou and D. Nioyugi, Noah-MP-Crop: Introducing dynamic crop growth in the Noah-MP land surface model. *J. Geoph. Research-Atmospheres*, v.121, 13,953-13,972, doi: **https://doi.org/10.1002/2016JD025597**, 2016.

[5] https://www.cesm.ucar.edu/models/clm

[6] https://www.cesm.ucar.edu/models/cesm2

[7] https://nuwrf.gsfc.nasa.gov/

BRINGING EARTH'S DIGITAL TWIN TO LIFE: ADVANCING THE DESTINE PLATFORM FOR BIG DATA EXPLORATION

Calogera Tona, Matteo Cortese, Barbara Scarda, Alexis Longuet

Serco Italia S.p.A.

ABSTRACT

This paper presents an updated overview of the DestinE Platform, developed by the European Space Agency (ESA) within the European Commission's Destination Earth (DestinE) initiative—an ambitious programme to build a high-precision digital twin of the Earth [1] [2] [3]. While serving as a unified access point to a wide range of Earth observation (EO) services and datasets, the platform also integrates advanced cloud-native processing tools, federated data access, and scalable infrastructure on EuroHPC [4].

The paper introduces key architectural and operational innovations in big data handling, including the use of Polytope [5] for multidimensional data extraction, real-time event-based workflows with Aviso, and distributed processing via JupyterLab and DASK [6]. It also explores user engagement through platform analytics, user profiles, and early case studies focused on climate resilience, environmental monitoring, and tourism applications.

Finally, the paper outlines the long-term sustainability roadmap, including eco-design, carbon impact assessment, and strategies for platform governance and maintenance. These contributions aim to position DestinE as an evolving, user-centric ecosystem for next-generation EO data exploitation and scientific collaboration.

Index Terms— Destination Earth, DestinE Platform, Big Data, Federated Access, Earth Observation, Digital Twin, Sustainability, User Engagement.

1. INTRODUCTION

The DestinE Platform, developed under the leadership of the European Space Agency (ESA), plays a central role in the European Commission's Destination Earth (DestinE) initiative. This large-scale digital infrastructure aims to build a high-precision digital twin of the Earth to support evidence-based policy making, environmental monitoring, and sustainable development.

Beyond serving as a central hub for accessing Earth observation (EO) data, models, and services, the DestinE Platform has been designed with a focus on innovation in big data processing, cloud-native architecture, and user-driven analytics [4]. It leverages distributed computing through EuroHPC, federated access layers across institutional data providers, and modular services that enable real-time or large-scale EO applications.

This paper expands on previous presentations of the platform by:

- Introducing novel technical components not covered in standard documentation;
- Highlighting how users are engaging with the platform, with early case studies;
- Clarifying the long-term sustainability and governance roadmap.

The structure of this paper is as follows:

- Section 2 describes the user journey from registration to advanced access;
- Section 3 presents the service registry and data exploitation tools;
- Section 4 introduces the platform architecture and big data handling innovations;
- Section 5 covers user engagement and community building;
- Section 6 outlines the sustainability roadmap;
- Section 7 concludes with a discussion of ongoing and future developments.

Fig. 1. DestinE Platform Homepage.

2. DESTINE PLATFORM USER JOURNEY

The DestinE Platform has been designed to offer a user-friendly and scalable pathway for accessing and leveraging high-precision EO data and services. The user journey is structured around four key steps:

1. Registration
2. Access Upgrade
3. Service Discovery via Registry
4. Support and Engagement

doi:10.2760/2119408

2.1. Registration

Users start by visiting https://platform.destine.eu and selecting the "Register" button. The registration process collects personal details, the user's institutional or organizational affiliation, and their profile category (e.g., research, public authority, NGO, citizen). Upon submission, a verification email is sent to activate the account.

Registration grants access to the basic features of the platform and allows users to explore a selection of available services and datasets.

2.2. Access Upgrade

Two access tiers are available:

- **Basic Access** (default after registration)
- **DPAD Access** (DestinE Primary and Altered Data Access), which requires approval.

DPAD Access enables users to download and utilise:

- **DestinE Primary Data**: Geospatial datasets from ECMWF-operated Digital Twins, made available via Polytope on EuroHPC infrastructure.
- **DestinE Altered Data**: Data derived from Primary Data that retain enough metadata to allow traceability, without containing significant intellectual or creative contributions from the user.

Users can request an access upgrade via the "Access Policy Upgrade" page after login. This process includes:

- Selecting the user category (e.g., academia, public authorities, SMEs)
- Reviewing and accepting the Terms & Conditions for DestinE Priority Users
- Submitting the form for manual review

2.3. Service Registry Access

The Service Registry provides a curated catalogue of operational services, publicly accessible for browsing and filtering, whereas access to the services requires an account. The services are organised into four main categories:

- Data Access management
- Data Analysis and Modelling
- Data Visualisation
- User Workflow

Services can be searched and filtered by category, data source, or tags. This modular structure supports a wide range of users — from data scientists to policy analysts — enabling discovery and integration of EO services tailored to specific needs. See Section 3 for detailed service descriptions.

2.4. Support

To ensure a responsive and inclusive user experience, several support mechanisms are in place:

- **FAQ Section** on the platform
- **Service Documentation** for each registered service
- **Dedicated Service Desk**, accessible via the "Contact Us" form
- **DestinE Platform Learning Hub** (launched in June 2025, available at https://learninghub.destine.eu designed to support user upskilling and knowledge sharing
- **Community Forum** (currently in development)

3. SERVICE REGISTRY

As of August 2025, the DestinE Platform hosts **25 operational services**, covering a diverse range of capabilities from data visualisation to advanced modelling and processing. Each service has passed a structured onboarding and evaluation process.

Below is a summary of currently available services:

- **Aviso**: Event-based data notification service for workflow automation.
- **CityNexus**: Urban digital twin modelling impacts of road and urban design changes.
- **Data Cache Management**: Supports efficient storage, handling, and user-driven requests for EO data.
- **DEA**: A no-code platform for interactive storytelling and visualisation using DestinE data.
- **DeltaTwin**: Collaborative toolbox for building and managing Digital Twin components.
- **DestinEStreamer**: Streamlines climate data access and processing.
- **DT-HEAT+**: Provides real-time intelligence on heat-related mortality.
- **Earth Data Hub**: Fast browsing, analysis, and computing on pre-processed EO datasets.
- **EDEN**: Central interface to explore and access Digital Twin data and anticipate climate-related impacts.
- **GeoAI**: Geospatial AI platform for designing and deploying AI-based EO solutions.
- **HDA**: Discovery and access to the DestinE Data Portfolio.
- **HIGHWAY**: Integrates ESA Earth Explorer datasets into DestinE with visualization and processing tools.
- **HOOK**: Serverless workflow orchestration via DestinE Data Lake.
- **Insula Code & Processing**: Visual storytelling and scalable EO data processing services.
- **Islet Compute & Storage**: Infrastructure/Platform as a Service near DestinE Data Lake.
- **miniDEA**: Lightweight visual component powered by DEA.
- **Polytope**: Efficient, federated access to EO hypercubes via API and Python client.
- **SesamEO**: Thematic discovery and download of Copernicus and other EO products.
- **STACK**: Cloud-native JupyterLab and DASK-based processing environment.
- **Tourism Square**: Analyses climate and environmental impact on tourism activity.

doi:10.2760/2119408

- **Urban Square**: Provides tools to analyse and anticipate environmental threats in urban areas.
- **Vision & VizLab**: Immersive 3D storytelling and visualization for Digital Twin data across desktop and VR/AR platforms.

Each service is accessible via the Service Registry and integrated into a common platform environment, ensuring interoperability and standardisation across user workflows.

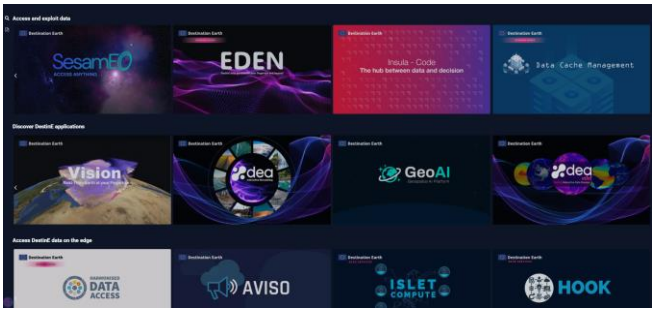


Fig. 2. DestinE Platform Service Registry.

4. BIG DATA ARCHITECTURE AND PROCESSING INNOVATIONS

One of the central technical challenges addressed by the DestinE Platform is the handling of massive EO datasets from multiple sources, generated at high temporal and spatial resolutions by Digital Twin engines. To support scalable, federated, and real-time data exploitation, several architectural innovations have been introduced:

4.1. Federated Data Access with Polytope

Polytope provides efficient, federated access to multidimensional data "hypercubes" produced by DestinE Digital Twins. It allows users to extract targeted spatial-temporal subsets via an API or a Python client, enabling selective access without the need to download full datasets. This capability is crucial for reducing data movement and supporting low-latency applications [5].

4.2. Event-Driven Processing with Aviso

The **Aviso** service implements an event-based architecture that allows users to subscribe to specific data events (e.g., new forecast availability or data ingestion) [7]. These triggers can launch downstream workflows, notifications, or custom analytics — enabling automation and real-time response.

4.3. Cloud-Native Analytics with STACK and DASK

The platform offers near-data processing capabilities through **STACK**, which integrates **JupyterLab** with **DASK**, allowing scalable computation directly in the cloud [6]. This minimises latency and supports complex operations such as time series extraction, spatial transformations, and data fusion — particularly useful for heavy geospatial workloads.

4.4. Data Processing and Modelling Services

Services like **Insula Processing**, **DeltaTwin** and **GeoAI** provide modelling, orchestration, and AI integration capabilities. These support both no-code users (via UI) and expert users (via CLI or APIs), making the platform adaptable to different skill levels.

4.5. Interoperability and API Standards

All services conform to shared platform policies and interoperability standards (e.g., STAC, OGC APIs), ensuring integration with external systems and reproducibility of results.

Collectively, these components position the DestinE Platform not only as a data access point but also as a **computational environment for big data exploitation**, enabling rapid prototyping and operational services in EO.

5. USER ENGAGEMENT AND COMMUNITY BUILDING

The DestinE Platform has demonstrated strong early user engagement across Europe and internationally. A **Public Dashboard** (available at https://platform.destine.eu/public-dashboard) provides real-time insights into platform usage, access levels, and geographic reach. As of August 2025:

- **Registered Users**: 3,783
- **Visitors (since launch)**: 58,847
- **Operational Services**: 25
- **Service Onboarding Requests**: 54

5.1. User Profile Breakdown

Analysis of user profiles shows broad participation across domains highlighted in Figure 3 below:

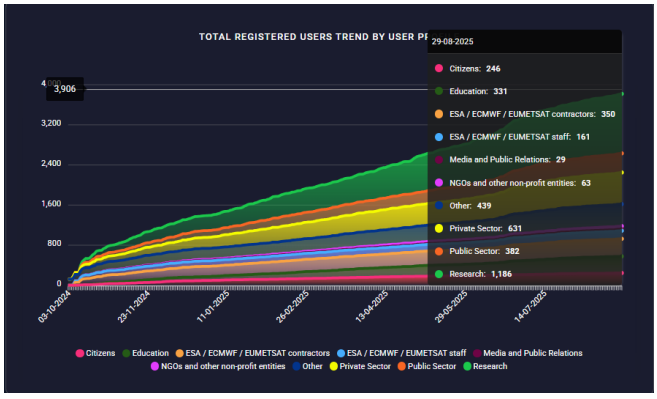


Fig. 3. Total Registered users trend by user profile.

This diversity reflects the platform's dual mission: enabling cutting-edge research and supporting operational applications such as policy-making and emergency response.

doi:10.2760/2119408

5.2. Use Case Highlights

Although still early in deployment, several compelling use cases have emerged:

- **Flood Mapping and Impact Forecasting**: Using real-time Digital Twin outputs with **DeltaTwin** and **DestinEStreamer** to model regional flood risk in Central Europe.
- **Urban Heat Island Monitoring**: Leveraging **GeoAI** and **Earth Data Hub** to study urban microclimates during extreme heat events in Southern Europe.
- **Tourism Climate Services**: Through **Tourism Square**, local governments are assessing the impact of weather variability on tourism infrastructure and services.
- **Cross-border Environmental Cooperation**: Transnational projects are using shared Digital Twin outputs to align climate adaptation strategies.

6. DESTINE PLATFORM SUSTAINABILITY ROADMAP

A strategic priority for the DestinE Platform is ensuring long-term sustainability. To this end, and following a roadmap defined in Q4 2024 and updated in 2025, the platform team has initiated a range of activities to guarantee operational continuity, environmental responsibility, and institutional governance [9] [10].

6.1. Environmental Responsibility and Eco-Design

The roadmap also integrates a clear environmental agenda, aligned with the European Green Deal and EU climate objectives:

- **Carbon Footprint Assessment**: Initial measurement frameworks are being piloted, focusing on compute- and data-intensive services.
- **Reduction Action Plan**: Early results will inform a reduction strategy for energy-intensive workflows and infrastructure.
- **Green-by-Design Principles**: Service providers are required to consider energy impact during onboarding (e.g., efficient algorithms, minimisation of data replication).
- **KPI Tracking and Reporting**: Sustainability Key Performance Indicators (KPIs) will be monitored via the platform dashboard.

7. CONCLUSIONS AND FUTURE DEVELOPMENTS

This paper has presented a revised and enriched view of the DestinE Platform — moving beyond a descriptive overview toward a more technical and strategic narrative that highlights the platform's innovations in data handling, user engagement, and sustainability [11].

7.1. Key Achievements to Date

- 3,783 registered users and 25 operational services (as of August 2025)
- Robust, modular user journey with tiered access and integrated support
- Technical services for big data federation, event-driven processing, and scalable analytics
- Clear roadmap for sustainability, eco-design, and cross-provider governance

7.2. Outlook and Next Steps

- **Forum Launch**: Targeting end of 2025 to boost community knowledge sharing and technical literacy.
- **Expansion of Onboarded Services**: A total of 54 onboarding requests were received, many of which address climate impact, agriculture, marine monitoring, and infrastructure resilience.
- **Operationalisation of Sustainability Metrics**: Including public reporting on energy consumption and carbon footprint.

With these next steps, the DestinE Platform aims to become not only a technical enabler of Digital Twin Earth initiatives, but also a sustainable, user-driven ecosystem capable of transforming how we observe, understand, and respond to planetary challenges.

REFERENCES

[1] DestinE Platform – Your gateway to a sustainable future
[2] Destination Earth
[3] I. Sanz-Morère, I. and Hintze, K.: DestinE Core Service Platform Framework, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-1370
[4] Homepage - EuroHPC JU
[5] Polytope
[6] DestinationEarth DataLake
[7] Aviso: ECMWF's data availability notification service | ECMWF
[8] C. Tona, Al. "DESP: your gateway to the Destination Earth initiative" Proc. of the 2023 conference on Big Data from Space (BiDS'23) 384 doi:10.2760/46796 6–9 November 2023
[9]C.Tona, B. Borgia, A. Longuet, B.Scarda, "AI4DESP: Unblocking Artificial Intelligence power to explore potential and challenges in Destination Earth Service Platform". ESA-ECMWF Workshop
[10] C. Tona, B.Borgia, B.Scarda, A. Longuet "Destination Earth Service Platfortm: a digital ecosystem enabling DestinE Data Exploitation", IGARSS 2024
[11] C. Tona, A. Longuet, B. Scarda, M. Cortese, E. Giuliani. "Shaping tomorrow's geospatial services: The DestinE Platform" EGU 2025.

A MULTI-AGENT SYSTEM TO ORCHESTRATE INTERACTIONS WITH DIGITAL TWINS OF EARTH

M. Tsokanaridou[1], *J. Hackstein*[2], *G. Hoxha*[2], *S.-A. Kefalidis*[1], *K. Plas*[1], *B. Demir*[2], *M. Koubarakis*[1], *M. Corsi*[3], *C. Leoni*[3], *G. Pasquali*[3], *C. Pratola*[3], *S. Tilia*[3] *and N. Longépé*[4]

[1]Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
[2]BIFOLD and Technische Universität Berlin, Germany
[3]e-GEOS S.p.A., Italy
[4]Φ-lab, ESA ESRIN, Frascati, Italy

ABSTRACT

We present a new-generation, AI-agent-powered digital assistant featuring four specialized engines for satellite imagery: search by image, search by caption, visual question answering, and knowledge graph question answering. At the core of the system is a Task Interpreter, designed as a multi-agent system, which coordinates these engines to address complex user requests for Earth observation data. The Task Interpreter comprises four agents: an Engine Routing Agent that selects the appropriate engine or rejects unmanageable requests; a Conversational Agent that handles general or out-of-scope queries; an Argument Extraction Agent that identifies image type parameters for retrieval tasks; and a Tool Feasibility Agent that assesses the applicability of tools for domain-specific queries. This multi-agent system enables seamless interaction with Digital Twins of Earth, with an emphasis on modularity and extensibility to adapt to the rapid evolution of remote sensing technologies.

***Index Terms*—** Multi-agent systems, digital assistant, digital twins, search by image, search by caption, visual question answering, knowledge graph question answering

1. INTRODUCTION

In Artificial Intelligence (AI), an *agent* is an autonomous entity capable of perceiving its environment, making decisions, and acting upon it to achieve specific goals. Multi-agent systems (MAS) is a subarea of AI studying societies of agents in cooperative or competitive settings and has a long tradition of outstanding research results. With the recent revolution of large language models (LLMs) and foundation models (FMs), the area of MAS is receiving again a lot of attention with the proposal of LLM-powered agent frameworks such as AutoGen [13], LangChain and CrewAI.

As part of these recent developments, we have seen the proposal of agent and multi-agent system architectures powered by LLMs in the Remote Sensing (RS) area [4, 9, 10, 11, 14]. *Remote Sensing ChatGPT* [4] introduces a system where ChatGPT interprets user requests and sequentially invokes specialized RS models for tasks such as object detection and land use classification. *RescueADI* [11] focuses on disaster interpretation, employing a LLM-driven agent to dynamically plan and execute multiple specialized tasks like damage assessment and rescue pathfinding. *RS-Agent* [14] extends this paradigm by integrating high-performance tools and a retrieval-augmented knowledge base to support professional geospatial analysis. *GlobeFlowGPT* [9] applies a multimodal LLM orchestrator to facilitate complex geospatial workflows, including flood forecasting and vegetation monitoring, with containerized tool integration. Similarly, *GeoLLM-Squad* [10] adopts a MAS, using an orchestrator to coordinate specialized agents for a broad range of remote sensing tasks, such as urban monitoring, climate analysis, forestry protection, and agricultural studies. Like our approach, it emphasizes modularity, extensibility, and the separation of orchestration from task-solving components.

Parallel to these developments, the emergence of *Digital Twins of Earth* (DTEs) —high-fidelity, dynamic digital representations of the Earth's systems—has created new demands for intelligent, continuous interaction with massive Earth observation (EO) datasets. DTEs require the ability to access, interpret, and integrate diverse data streams in a flexible, scalable, and context-aware manner. MAS are particularly well suited to meet these needs, enabling specialized tools to work together dynamically to support the complex data requirements of DTEs.

However, despite recent advances, there is currently *no* EO data provider that offers a digital assistant capable of guiding users in finding the EO data they seek. This is a critical functionality gap, especially as the volume of EO data made available through initiatives like *Copernicus* and *Landsat* continue to expand. Without intelligent assistance, this wealth of data remains difficult to access for both expert and non-expert users, such as journalists searching for timely EO

This work was supported by ESA project DA4DTE.

doi:10.2760/2119408

imagery of environmental disasters or policymakers monitoring climate events.

To address this challenge, we introduce the *Digital Assistant for Digital Twins of Earth (DA4DTE)*, an AI-powered multi-agent digital assistant designed to facilitate seamless interaction with EO datasets. In DA4DTE, a *Task Interpreter* operates as a multi-agent system comprising specialized agents that collaboratively interpret user requests and orchestrate the activation of appropriate search engines or tools. We distinguish between the specialised *engines* serving EO tasks, the multi-agent *Task Interpreter* with its agents—autonomous functional components responsible for specific subtasks—and the *assistant*, the overall user-facing system deployed to fulfill complex information retrieval workflows. We make the source code of our system publicly available[1].

2. MULTI-AGENT SYSTEM FOR ORCHESTRATION

DA4DTE enables a user to pose multi-modal requests, that — in addition to text— can include RS images, either uploaded or selected on the User Interface map. The assistant's toolset allows for a variety of requests including geospatial or visual queries, requests for images by describing their visual context or metadata, image search requests, and queries for explanation on image similarity results. Between the user and the DA4DTE engines lies the *Task Interpreter*: a MAS responsible for engine orchestration and the mediation between the user and individual engines. The architecture is illustrated in Figure 1, which highlights the collaborative roles of each agent module and their interactions with the user interface and underlying engine components.

To ensure future extensibility, we categorize orchestration responsibilities into two types: **core** and **assistant** tasks. *Core tasks* are permanent and fundamental to any version of the assistant, regardless of the tools or data sources integrated. In contrast, *assistant tasks* are tailored to the current implementation state and may evolve as functionalities and resources expand. Each task is assigned to a dedicated agent, forming a MAS, implemented using the AutoGen [13] framework and currently comprising the following four agents.

The first agent is the **Engine Routing Agent** (Core). This agent is a zero-shot prompted LLM that selects the most appropriate engine to activate based on the user request. It also has the capability to reject requests that fall outside the scope of all available engines.

The second agent is the **Conversational Agent** (Core). This is a fallback conversational agent designed to handle general, ambiguous, or out-of-domain queries. Although it is a capable LLM, it is specifically prompted not to respond to irrelevant requests so the assistant remains task-focused.

The third agent is the **Argument Extraction Agent** (Assistant). This is an agent dedicated to extracting key parameters required by specific tools. In the current implementation, it identifies the requested image type (e.g., Sentinel-1 or Sentinel-2) when the *Search-by-Image* engine is activated.

Finally, the fourth agent is the **Tool Feasibility Agent** (Assistant). This is a utility agent responsible for validating whether a requested operation is feasible under current system capabilities. For example, the *Search-by-Text* engine presently supports only vessel-related queries. If a user request falls outside this domain, the agent triggers a relevant explanatory message to the user.

3. ENGINES AND THEIR FUNCTIONALITIES

DA4DTE integrates four specialized engines, tailored to specific Question Answering (QA) or retrieval tasks.

The first engine is the **Knowledge Graph QA Engine** TerraQ [8]. TerraQ[2] is a QA system that is designed to process natural language requests that include spatiotemporal or metadata related criteria and satisfy the request by retrieving data from a Knowledge Graph (KG). User requests can include references to image metadata (e.g., snow percentage in an image), geoentities (e.g., the country France), administrative divisions (e.g., municipalities, regions), as well as spatiotemporal constraints.

For example, users can make requests like "Give me a hundred images of rivers near ports in France, with less than 20% snow coverage and more than 10% cloud coverage, taken in 2021". The engine then takes this request as input, translates it into a semantically equivalent SPARQL query as follows: First, relevant entities and classes are extracted from the KG. Then, relations between the retrieved entities and classes are identified, including spatial and temporal relations. At this stage, the core of the query is complete, and the expected return values are identified by a finetuned Llama 2 model. The query generator then produces the complete, executable SPARQL query. This query is subsequently enhanced by a finetuned on SPARQL Mistral-7b-v2 model, and rewritten to optimize execution efficiency by replacing GeoSPARQL functions with equivalent materialized topological predicates. In the end, the query is executed over a GraphDB endpoint, and the QA process is complete.

The second egine is the **Search-by-Image Engine**. This engine takes a query image and computes the similarity function between the query image and all archive images to find the most similar images to the query in a scalable way. This is achieved based on two main steps: i) the image description step, which characterizes the spatial and spectral information content of RS images; and ii) the image retrieval step, which evaluates the similarity among the considered hash codes and then retrieves images similar to a query image in the order of similarity. Our Search-by-Image Engine is defined based on two self-supervised methods: 1) deep unsupervised cross-

[1] https://github.com/rsim-tu-berlin/DA4DTE

[2] https://terraq.di.uoa.gr/

doi:10.2760/2119408

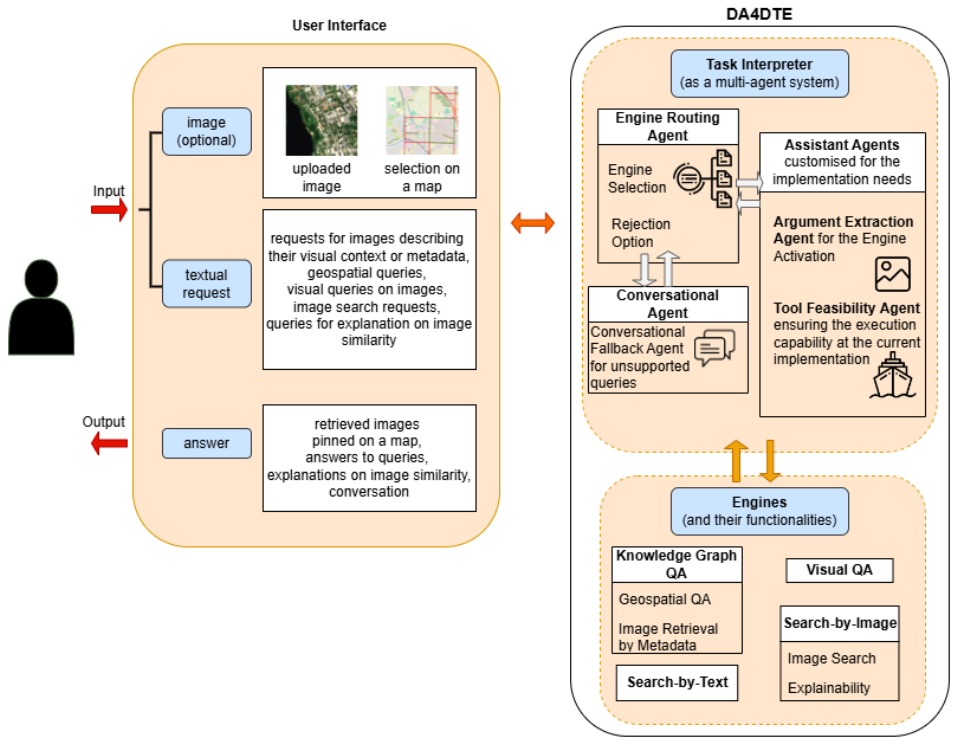


Fig. 1. High-level architecture of the digital assistant (DA4DTE), showing the user interface, multi-agent Task Interpreter, and the specialized engines (figure inspired by Figure 1 of [14]).

modal contrastive hashing (DUCH) [12]; and 2) cross-modal masked autoencoder (CM-MAE) [6]. For both methods, the image description step is composed of two modules: 1) a feature extraction module, which learns deep feature representations of RS images by exploiting visual transformers (ViT); and 2) a deep hashing module, which learns to map image representations into hash codes. The first module of the DUCH method is based on contrastive self-supervised image representation learning, while that of the CM-MAE method is based on unsupervised masked image modelling. The second module of each method employs a hashing subnetwork with binarization loss functions. Our engine has both the single-modal (also known as uni-modal) and cross-modal content-based image retrieval capability due to the consideration of the modality-specific encoders.

A key feature of the search-by-image engine is the integration of the *Explainability tools* to understand and explain the decision of the engine in retrieving a particular image given a query image. To this end, we incorporate two explainability tools: Layer-wise Relevance Propagation (LRP) [1] and BiLRP [3]. The LRP highlights areas in the input image supporting a specific class decision by generating heatmaps. Since CM-MAE is self-supervised and lacks class predictions, we train an auxiliary classification head to estimate class probabilities for each image pair. These predictions enable the generation and interpolation of class-specific LRP heatmaps, which emphasize semantically similar regions across image pairs. BiLRP, while more computationally intensive, identifies in the image pairs shared regions without needing a classification head.

The third engine is the **Search-by-Text Engine**. This engine takes a text sentence as a query and efficiently retrieves the most similar images to the query text, achieving scalable cross-modal text-image retrieval. The Search-by-Text Engine is developed by adapting the above-mentioned self-supervised DUCH [12] to be operational on text based queries. To this end, the feature extraction module is adapted to extract feature representations of image-text pairs by exploiting bidirectional transformers (e.g., BERT [2]) as text-specific encoders together with ResNet-152 [7] as image-specific encoders. The second module of each method is adapted to learn cross-modal binary hash codes for image and text modalities by simultaneously preserving semantic discrimination and modality-invariance in an end-to-end manner.

To evaluate DUCH, we constructed a vessel captioning dataset, consisting of vessel text-image pairs generated via a template-based image captioning approach. This approach consists of creating predefined sentence templates with empty slots. The slots are then filled using semantic cues from vessel bounding boxes (e.g., count, size) and contextual data from OpenStreetMap, particularly coastline proximity (i.e., vessel locations relative to harbors or coastlines). Vessel sizes, de-

doi:10.2760/2119408

rived from bounding box dimensions, were categorized into five classes (very small to very big) and mapped to two vessel types: boats (very small to medium) and ships (big and very big), reflecting typical usage and navigational context.

Finally, the fourth engine is the **Visual QA Engine**. This engine enables users to ask questions about the content of RS images in a free-form manner, extracting valuable information. It employs the LiT-4-RSVQA [5] model, which has been trained and evaluated on RSVQAxBEN[3]. The LiT-4-RSVQA architecture focuses on achieving state-of-the-art performance, while also providing rapid response times. To do so, it employs the following modules: i) a lightweight text encoder module; ii) a lightweight image encoder module; iii) a fusion module; and iv) a classification module. A RS image I and a question Q about this image are considered as input. The encoder modules produce vector representations which are subsequently passed to the fusion module. The feature fusion module consists of two linear projections and a modality combination. The projections map the two modalities with dimensions dt and dv into a common dimension df, where dt and dv denote the dimensions of the flattened output of the text and image encoder modules, respectively. The value of dv differs depending on the used lightweight transformer. The projected features are then elementwise multiplied. The classification module is defined as an MLP projection head.

4. DA4DTE IN ACTION

We now consider a use case scenario for the digital assistant. The assistant welcomes the user and asks them to pose a request. The user asks for a Sentinel-1 image from France during 2020, with snow coverage of more than 50%. Then, the Engine Routing Agent of the Task Interpreter decides that this is a request that should be fulfilled by the Knowledge Graph QA Engine which returns the appropriate image. The interaction goes on with the user asking for a similar Sentinel-2 image and then the Search-by-Image Engine is selected by the Engine Routing Agent. The term "Sentinel-2" is extracted by the Argument Extraction Agent as the modality argument, so the engine is activated and returns the appropriate image. Having selected that Sentinel-2 image, the user asks whether it presents a rural area and the answer by the Visual QA Engine is presented. Finally, the user closes the interaction with the assistant and the Engine Routing Agent of the Task Interpreter calls the Conversational Agent to answer appropriately.

5. FUTURE WORK

We plan to explore several research directions to further improve the capabilities of the system. First of all, we aim to implement an alternative Engine Routing Agent using the Function Calling paradigm in LLMs, to improve control over engine invocation compared to the current zero-shot prompting setup. We also plan to extend the assistant's capabilities to multi-step requests where multiple engines can be activated in a sequence. As the complexity of the system increases, we intend to integrate a Manager Agent to oversee and coordinate the behavior of all other agents within the Task Interpreter.

[3] https://zenodo.org/records/5084904

REFERENCES

[1] S. Bach et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015. doi: 10.1371/journal.pone.0130140.

[2] J. Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. doi: 10.18653/V1/N19-1423.

[3] O. Eberle et al. Building and interpreting deep similarity models. *PAMI*, 44(3), 2020. doi: 10.1109/TPAMI.2020.3020738.

[4] H. Guo et al. Remote Sensing ChatGPT: Solving remote sensing tasks with ChatGPT and visual models. In *IGARSS*, 2024. doi: 10.1109/IGARSS53475.2024.10640736.

[5] L. Hackel et al. LiT-4-RSVQA: Lightweight transformer-based visual question answering in remote sensing. In *IGARSS*, 2023. doi: 10.1109/IGARSS52108.2023.10281674.

[6] J. Hackstein et al. Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing. *TGRS*, 63, 2025. doi: 10.1109/TGRS.2024.3517150.

[7] K. He et al. Deep residual learning for image recognition. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.90.

[8] S. Kefalidis et al. TerraQ: Spatiotemporal question-answering on satellite image archives. In *IGARSS*, 2025.

[9] D. Kononykhin et al. From data to decisions: Streamlining geospatial operations with multimodal GlobeFlowGPT. In *ACM SIGSPATIAL*, 2024. doi: 10.1145/3678717.3691248.

[10] C. Lee et al. Multi-agent geospatial copilots for remote sensing workflows. doi: 10.48550/arXiv.2501.16254. https://arxiv.org/abs/2501.16254.

[11] Z. Liu et al. RescueADI: Adaptive disaster interpretation in remote sensing images with autonomous agents. *TGRS*, 2025. doi: 10.1109/TGRS.2025.3532594.

[12] G. Mikriukov et al. Unsupervised contrastive hashing for cross-modal retrieval in remote sensing. In *ICASSP*, 2022. doi: 10.1109/ICASSP43922.2022.9746251.

[13] Q. Wu et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. doi: 10.48550/ARXIV.2308.08155. https://arxiv.org/abs/2308.08155.

[14] W. Xu et al. RS-Agent: Automating remote sensing tasks through intelligent agents. doi: 10.48550/ARXIV.2406.07089. https://arxiv.org/abs/2406.07089.

OGC DGGS API AND ZARR: BUILDING BLOCKS FOR DIGITAL TWINS

Alexander Kmoch[1,2], Wai Tik Chan[1], Guillaume Ameline[1,2], Justus Magin[3], Tina Odaka[3], Jean-Marc Delouis[3],
Benoit Bovy[4], Anne Fouilloux[5], Evelyn Uuemaa[1,2]

[1]University of Tartu, Institute of Ecology and Earth Sciences,
Landscape Geoinformatics Lab, Tartu, Estonia
[2]Geolynx OÜ, Tartu, Estonia
[3]LOPS UMR, CNRS-IFREMER-IRD-Univ.Brest-IUEM,
Plouzane, France
[4]Georode, Liege, Belgium
[5]Simula, Oslo, Norway

ABSTRACT

The exponential growth of Earth Observation (EO) data presents significant challenges for efficient data access, processing, and analysis. Current approaches often involve disparate data formats, coordinate systems, and access patterns, limiting interoperability and scalability. Recently, the Zarr data storage format has been adopted as a unifying cloud-native foundation for various domains, including climate, EO, bio-imaging, and genomics. Secondly, Discrete Global Grid Systems (DGGS) such as HEALPIX, or ISEA-based hexagonal DGGS are being increasingly used to provide indexing beyond traditional grids, by providing equal-area pixels and location- and resolution encoding indices. Lastly, the recently published OGC API DGGS standard specifies a lightweight web service API for clients accessing data organised according to Discrete Global Grid Reference Systems (DGGRS).

We implemented a Python FastAPI service (pydggsapi) that exposes the OGC DGGS API standard with a backend utilizing Zarr archives indexed by DGGS cells. This work demonstrates a novel architecture that combines DGGS with cloud-native Zarr storage to create universal building blocks for EO data management, enabling seamless transitions between high-performance computing environments and lightweight client applications.

***Index Terms*—** DGGS, OGC API, ZARR, indexing, web service

1. INTRODUCTION

Earth Observation (EO) data volumes continue to increase exponentially, driven by new satellite missions, higher sensor resolutions, and increased temporal coverage. The Copernicus program alone generates TB of data every day. Current approaches often involve disparate data formats, coordinate systems, and access patterns, limiting interoperability and scalability across different processing environments [1].

Conventional web service standards such as Web Coverage Service (WCS), Web Coverage Processing Service (WCPS), and Web Map Service (WMS) have provided standardized access to geospatial data [11]. However, these services typically rely on traditional coordinate reference systems and raster-based approaches that present challenges when working with global datasets at multiple resolutions or when combining heterogeneous data sources.

The Discrete Global Grid System (DGGS) paradigm offers a solution to these challenges by providing a unified spatial reference framework based on hierarchical tessellation of the Earth's surface [13]. The Open Geospatial Consortium (OGC) has developed standards and best practices for DGGS implementation, including the DGGS Abstract Specification and the DGGS API. The OGC Testbed-16 Engineering Report (ER-16) and the ESA technical study on Sentinel-2 ARD handling with DGGS further elaborate on implementation approaches and use cases [12, 14].

Equal-area DGGS implementations, such as HEALPix [5], rHEALPix [4] and ISEA [15], have gained particular attention for Earth Observation applications due to their ability to maintain consistent area measurements across the globe [9], and their usability for large-scale EO analysis, including improved statistical analysis and multi-resolution data fusion.

In parallel, the scientific Python ecosystem has seen significant developments in data handling capabilities with tools like Xarray [6], which provides labelled multi-dimensional array operations. The XDGGS extension for Xarray enables direct manipulation of DGGS-indexed data within this framework [7]. Additionally, the XPublish concept provides a mechanism for exposing Xarray datasets through web service interfaces, bridging the gap between analytical environments and web-based access patterns.

In this paper, we describe universal building blocks

doi:10.2760/2119408

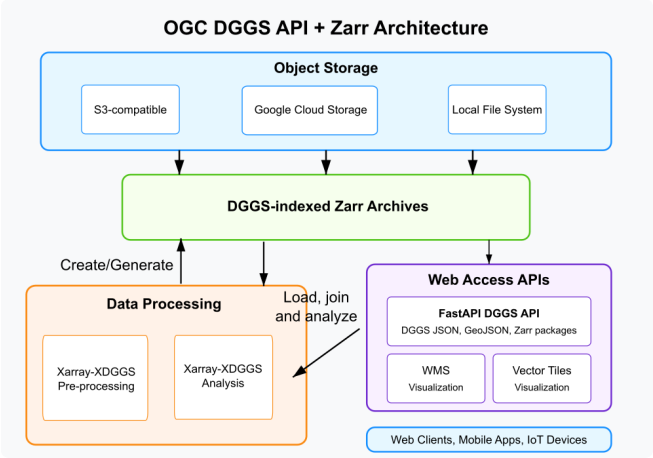


Fig. 1. Architecture

that combine DGGS with cloud-native Zarr storage to create a comprehensive framework for EO data management. This approach enables seamless transitions between high-performance computing environments and lightweight client applications while maintaining spatial alignment through DGGS indexing. By integrating these technologies, we address key challenges in the EO data processing chain, from initial data organization to final delivery and visualization.

2. METHODOLOGY

We implemented a Python FastAPI service that exposes the OGC DGGS API standard (pydggsapi [8]) with a backend utilizing Zarr arrays indexed by DGGS cells. The concept is inspired by the TiTiler and XPublish packages, which can employ a "serverless" FastAPI web service routing interface on top of cloud-native and Xarray datasets. This architecture leverages complementary technologies as shown in Fig. 1.

The Discrete Global Grid System (DGGS) serves as a universal spatial index, providing a hierarchical, multi-resolution grid system that consistently indexes geospatial data across the entire globe. This eliminates the need for reprojection when combining datasets from disparate sources and enabling immediate analysis without preprocessing steps, and extends and improves upon the grid notion, that is established in the met/ocean and climate communities.

Zarr technology functions as the cloud-optimized storage foundation, with its chunked, compressed array format enabling efficient parallel access to massive datasets stored in object storage systems while maintaining critical dimensional information and supporting selective data extraction at multiple resolutions.

As a novel composite aspect, the architecture leverages already available software, such Xarray-XDGGS, which is a package, that implements 1-D DGGS indexed arrays to work with various open-source DGGS libraries and systems. Through the Xarray DataTree model and the representation in Zarr data groups, we can aggregate data towards higher-level DGGS refinement levels. The concept corresponds to image pyramids and overviews in other cloud-native formats like COG GeoTiffs or PMTILES. Exemplary, we show use of vector tiles (MVT) to enable a visual and data access. MVT can be very efficient in-browser rendering by using webgl with MapLibre GL JS as shown in Fig. 2 a.

Ultimately, the system employs a Python FastAPI-based OGC DGGS API interface that serves as the primary access point for web clients, providing both visualization services and precise data access through standardized formats including DGGS JSON, GeoJSON, or as hybrid Zarr "packages" that maintain the original data structure while enabling efficient transfer. The web service API is meant for light-weight query and visualisation access for web- or mobile (or IoT)-based client applications.

3. IMPLEMENTATION

Our concept bridges two distinct operational scales - cloud-native big data processing, and a more refined web service-based access for lightweight clients. Zarr arrays stored in object storage serve as the unified data foundation, enabling direct access for high-performance computing and cloud-based modeling workflows. The OGC DGGS API implementation, akin to XPublish for Xarray, provides standardized, RESTful access to the same underlying data for a diverse range of web, mobile, and IoT clients and applications. This dual-scale approach ensures data consistency across use cases while optimizing for different computational and bandwidth constraints.

The data collection provider components enable access to DGGS-indexed datasets through middleware that manages connections to cloud-storage Zarr archives. Based on the OGC ER-16 report, the application also showcases an experimental connector for the Clickhouse database to provide fast on-demand aggregation and analytical queries on DGGS-indexed database tables.

The main archticture relies on pre-aggregated pyramids, where the data access middleware reads metadata from Zarr archives. During initialization, it extracts DGGS parameters from the Zarr archive's attributes, including the DGGS type (such as HEALPix or H3) and the indexing scheme used or additional index parameters (e.g., HEALPIX' nested or ring scheme), available refinement levels, and the available data variables.

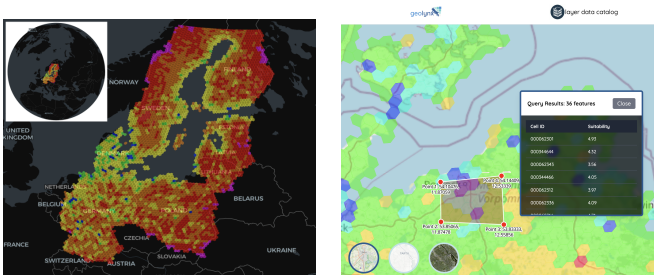
The architecture aims to abstract the underlying storage system through the Zarr application library interface, providing consistent data retrieval whether data is stored in AWS S3, Azure Blob Storage, Google Cloud Storage, or local file systems. For each API request, the middleware identifies the target collection from the URL path, loads the Zarr metadata, attaches DGGS configuration to the request context, and

doi:10.2760/2119408

queries the Zarr store instance purely on zone identifiers. To improve performance, the system maintains an in-memory cache of collection metadata, reducing storage access operations and speeding up repeated queries for common metadata.

This approach enables the FastAPI application to seamlessly serve OGC DGGS API requests while maintaining the performance benefits of cloud-native Zarr storage for Earth Observation data indexed by DGGS cells.

DGGS indexing automatically aligns diverse data products spatially and temporally, facilitating immediate analysis without further preprocessing. Via the OGC DGGS API, it is possible to provide convenient access to clients to cell-based summary queries, for examples as shown in Fig. 2 b., at a specific refinement level (i.e. higher resolution). The OGC DDGS API layer enables targeted data extraction for non-DGGS-capable clients, reducing data transfer volumes and simplifying implementation for client applications. Direct HPC access to cloud-stored Zarr archives is facilitated by high-throughput network connections and cloud-aware libraries like *fsspec* and *xarray*. These tools enable HPC applications, e.g. built on libraries like *Dask* to stream Zarr chunks directly from object storage APIs (e.g., S3), enabling efficient, highly parallelised batch processing over large areas and timescales without pre-staging the entire dataset.



(a) (b)

Fig. 2. Examples of DGGS web clients, a) MVT allows for lightweight browser visualisation; b) Cell-based summary queries

4. DISCUSSION

4.1. Universal Building Blocks for Earth Observation Data

Why universal building blocks? The combination of DGGS and Zarr creates building blocks for Earth Observation data management that address key challenges in the geospatial data pipeline. Maintaining a unified base of DGGS-indexed Zarr archives enables consistent data organization across multiple use cases. This approach supports both large-scale and local applications through a dual-access pattern: direct access to cloud-stored Zarr archives for high-performance computing and API-mediated access for web clients with bandwidth constraints.

Pydggsapi also includes the well-known OpenAPI3/ Swagger documentation and client, which provides a low-barrier entry into experimenting with the API. In the near future we will also provide training materials based on Python/Jupyter notebooks.

4.2. Standardization Requirements

The OGC DGGS working group is advancing standards for DGGS implementations, but additional components are still needed. A registry for DGGS Reference Systems similar to the EPSG codes and Proj/WKT CRS definitions is needed. Clear conventions for storing DGGS parameters in Zarr metadata fields must be established. For HEALPIX, essential parameters include the desired resolution and the indexing scheme (nested or ring), while more flexible systems like rHEALPix or ISEA-based DGGS may use additional parameters such as indexing/numbering scheme, origin, and rotation. The CF Metadata Conventions community has also picked up a discussion on grid parameter specification for NetCDF and Zarr archives [2].

The OGC DGGS API implementation requires an indexing scheme that encodes the refinement level in the zone identifier. Systems like H3 and S2 have popularized this approach, and newer DGGS references systems like IGEO7 (DGGRID/Z7 [10]) and rHEALPix can also provide this functionality. However a more accessible implementation is needed for HEALPIX (namely *nuniq* or *zuniq*). A notable requirement in the OGC DGGS API standard is subzone ordering, which currently only ISEA3H/9R appears to fully support [3]. Most DGGS implementations have a space-filling curve index and rely on associating data values with zone identifiers during data transport.

4.3. Technical Limitations and Future Work

The current implementation faces certain limitations that require further research and software engineering expertise. For example, Xarray's eager indexing of dimensions may limit its utility for very large DGGS archives, as all dimension indexes are loaded into memory. More efficient handling of multi-resolution DGGS data may require extensions to the Xarray data model or more direct Zarr-native access approaches.

The use of Zarr groups/Xarray data trees to represent refinement level aggregations requires additional experimentation to ensure interoperability and efficient chunking/sharding, which takes DGGS cell topologies and parent-child-cell boundaries into account. The OGC ER-16 report outlines the methodology for providing descriptive statistics per cell, including standard deviation, ranges, min/max values, and variance. Implementing this efficiently within the Zarr data model across refinement levels would be useful to

doi:10.2760/2119408

better convey the uncertainty of aggregated data.

Initial performance benchmarks do not indicate disadvantages compared to traditional approaches regarding data access performance. Combined with the inherent advantages of DGGS for global data analysis, these results suggest encouraging progress. There remains substantial potential for improvements in both usability and performance as the technology matures.

5. CONCLUSION

The integration of OGC DGGS API with Zarr storage represents a significant step toward universal building blocks for Earth Observation data. This approach enables a seamless continuum from big data processing to lightweight client applications while maintaining spatial alignment through DGGS indexing. We hope for the near future that this might develop towards a new paradigm for value added AI-integrated and ARD integration-ready data market APIs.

REFERENCES

[1] Bernhard Bauer-Marschallinger and Konstantin Falkner. Wasting petabytes: A survey of the Sentinel-2 UTM tiling grid and its spatial overhead. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:682–690, 2023. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2023.07.015.

[2] CF Metadata Conventions. Convention for HEALPix grid parameters, 2025. URL https://github.com/cf-convention/cf-conventions/issues/433.

[3] Ecere Corporation. DGGAL, the Discrete Global Grid Abstraction Library, 2025. URL https://dggal.org/.

[4] Robert Gibb, A Raichev, and M Speth. The rHEALPix DGGS preprint. 2016.

[5] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, apr 2005. doi: 10.1086/427976.

[6] Stephan Hoyer and Joe Hamman. xarray: N-D labeled Arrays and Datasets in Python. 5(1):10, April 2017. ISSN 2049-9647. doi: 10.5334/jors.148. Number: 1 Publisher: Ubiquity Press.

[7] A. Kmoch, B. Bovy, J. Magin, R. Abernathey, A. Coca-Castro, P. Strobl, A. Fouilloux, D. Loos, E. Uuemaa, W. T. Chan, J.-M. Delouis, and T. Odaka. Xdggs: A community-developed xarray package to support planetary dggs data cube computations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W12-2024:75–80, 2024. doi: 10.5194/isprs-archives-XLVIII-4-W12-2024-75-2024.

[8] Alexander Kmoch and Wai Tik Chan. pydggsapi: A python FastAPI OGC DGGS API implementation. [software], 2025. URL https://github.com/LandscapeGeoinformatics/pydggsapi. last accessed April 10, 2025.

[9] Alexander Kmoch, Ivan Vasilyev, Holger Virro, and Evelyn Uuemaa. Area and Shape Distortions in Open-Source Discrete Global Grid Systems. *Big Earth Data*, 2022. doi: 10.1080/20964471.2022.2094926.

[10] Alexander Kmoch, Kevin Sahr, Wai Tik Chan, and Evelyn Uuemaa. Igeo7: A new hierarchically indexed hexagonal equal-area discrete global grid system. *AGILE: GIScience Series*, accepted, May 2025. ISSN 2700-8150.

[11] OGC and Peter Baumann. *OGC WCS 2.0 Interface Standard - Core, v2.0*. The Open Geospatial Consortium (OGC), http://www.opengeospatial.org/standards/is, 2010. ISBN 09-110r3. Publication Title: WCS 2.0.

[12] Matthew B.J. J Purss, Perry R. Peterson, Peter Strobl, Clinton Dow, Zoheir A. Sabeur, Robert G. Gibb, and Jin Ben. Datacubes: A Discrete Global Grid Systems Perspective. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 54(1):63–71, March 2019. ISSN 0317-7173. doi: 10.3138/cart.54.1.2018-0017.

[13] Kevin Sahr, Denis White, and A. Jon Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003. doi: 10.1559/152304003100011090.

[14] Germain Salgues, Enrico G. Cadau, Laetitia Pessiot, Vincent Gaudissart, Silvia Enache, Ferran Gascon, Valentina Boccia, and Peter Strobl. A candidate dggs (discrete global grid system) for sentinel-2: First outcomes. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4978–4981, 2023. doi: 10.1109/IGARSS52108.2023.10281749.

[15] John P. Snyder. An Equal-Area Map Projection For Polyhedral Globes. *Cartographica*, 29(1):10–21, March 1992. ISSN 0317-7173. doi: 10.3138/27H7-8K88-4882-1752. Publisher: University of Toronto Press.

FROM SPACE TO THE EYE: EFFECTIVE VISUAL COMMUNICATION OF EARTH-OBSERVATION DERIVED URBANISATION TRENDS USING THE GLOBAL HUMAN SETTLEMENT LAYER

[1]Johannes H. Uhl, [1]Alessandra Carioli, [1]Daniele Ehrlich, [1]Thomas Kemper

[1]Joint Research Centre (JRC), European Commission, Ispra (VA), Italy

ABSTRACT

The steadily increasing amount of available Earth observation and other geospatial data provides unprecedented opportunities to measure, monitor, and understand natural, environmental, anthropogenic and social processes. However, to maximise the impact of space-based and other data on society, effective ways to extract, integrate, summarise and communicate the information contained in these sheer data volumes are needed. Herein, we focus on the latter of these components: communication of trends derived from Earth observation data by means of data visualisation. Specifically, we highlight how settlement and population dynamics, urbanisation patterns and land development processes can be visually represented to effectively communicate relevant information to a wide range of audiences, ranging from scientists to planners, policymakers and to the general public. The underlying data has been generated by the Global Human Settlement Layer (GHSL) project of the European Commission's Joint Research Centre (JRC) by leveraging and integrating vast amounts of global remote sensing data from the Landsat and Copernicus Sentinel-2 missions to derive spatio-temporal gridded data measuring the distribution and dynamics of the built environment, settlements and human population from 1975 onwards.

Index Terms— Data visualisation, data animation, GHSL, Degree of Urbanisation, visual analytics

1. INTRODUCTION

The famous cholera map of John Snow, identifying the source of a 1854 cholera outbreak in London by mapping the location of infected cases [1] is a timeless example for the importance of data visualisation to reveal spatio-temporal patterns and its impact on society. The significance of data visualisation remains unchanged, and the advances in data availability, processing capabilities, data visualisation tools and in particular digital media- and web-based dissemination channels facilitate static, dynamic and interactive data visualisation in unprecedented ways.

Visualisation is key to enable data-based insights, support evidence-based policy through data-driven decisions by stakeholders and policymakers, and an important tool for data and knowledge democratisation [2].

Earth observation data and other geospatial data are critical to monitor and understand pressing issues related to urbanisation, such as sustainable development, land take, biodiversity loss and social inequality. Detailed knowledge of the spatial characteristics and the evolution of the built environment and of human population is critical for informed urban and regional planning and policy-making, ensuring effective disaster risk management and crisis response, as also highlighted in the Agenda for Sustainable Development Indeed, accurate human settlement data aligns with the second principle of the Sustainable Development Goals (SDGs), "Leave no one behind", acknowledging gridded population data as an important resource for delivering actionable data to monitor health and infectious diseases.

The Global Human Settlement Layer (GHSL) is a project of the European Commission's Joint Research Centre (JRC), producing and providing free and open geospatial data, mapping population distributions and characterising human settlements at a global scale, at high spatial resolution of up to 10 m, and over extended time periods from 1975 to 2030 [3]. The GHSL data products integrate multi-source data, including planetary-scale Earth observation data from the Landsat and Sentinel sensors and population data from census figures in a spatio-temporal modelling framework, to produce fine-grained estimates of built-up surface, settlement age and function, building height and volume, taxonomies of the built environment, and human population, all consistently enumerated in global geospatial grids [3]. These data products have been used widely by the scientific community and beyond, since the first data release in 2016 [4].

With increasing complexity of the GHSL data ecosystem, we are currently developing effective visual-analytical methods to integrate and summarise relevant trends measured by the multivariate spatio-temporal GHSL data, fostering data usage and facilitating the access to information inherent in the data. Herein, we present selected static and animated data visualisation techniques based on GHSL data. These visualisations are automatically generated, for global, country-level, regional, and local (i.e., city-level) scope, using a combination of spatial and non-spatial visualisation techniques. They aim to represent complex datasets in comprehensive ways, making information accessible to wider audiences. These visualisations intend to facilitate faster

doi:10.2760/2119408

communication, fostering a deeper understanding of urbanisation trends and settlement dynamics to be used for outreach purposes, infographics, dashboards or other interactive data visualisation platforms.

2. DATA AND METHODS

2.1. Data sources

We use data from the Global Human Settlement Layer release R2023A [4], at spatial resolutions of 100m and 1km. For local (city and region level) we rely on the 100m data, whereas for country-level and global analyses, the aggregated 1-km GHSL datasets are used, to speed up computational processing. These datasets include: built-up surface area estimates from GHS-BUILT-S, GHS-POP (gridded estimates of residential population, derived from disaggregating census-based population counts into fine grid cells using dasymetric modelling [5], Degree or Urbanisation, a model of the rural-urban continuum (partitioning land areas into 3-7 rural-urban classes) available at 1-km resolution [6]. Moreover, we use the Copernicus GLO-30 Digital Elevation Model (CopDEM) [7], available at 30-m resolution, and global administrative boundaries from the GADM [8] dataset. Finally, the GHSL Urban Centre Database (UCDB) is used, an integrated, vector-based dataset delineating over 10,000 urban centres globally [9], according to their definition in the degree of urbanisation framework, with rich attributes on socio-demographic, environmental, and risk-related urban characteristics from various data sources.

2.2. Data processing

All GHSL raster data layers are enumerated in a global grid at 100m and 1-km resolution, respectively, in World Mollweide Equal Area projection (ESRI:54009). Urban centre boundaries from the UCDB nest within this grid. This facilitates joint processing of multiple variables. We distinguish four geographic levels for data visualisations: urban centre level, region-level, country-level, or global.
Data processing is fully automated, implemented in Python 3.9, using gdal, rasterio, geopandas, pandas, numpy, matplotlib, and seaborn python packages. Based on the user-specified geographic object of interest (e.g., an urban centre or country, specified by its name, or a region defined by a user-provided bounding box), the raster data covering the respective region are cropped and extracted to memory. In case of urban centre- or country-level visualisations, the area of interest is rasterised to the GHSL grid, and non-relevant areas are masked out. If input datasets have different spatial resolutions (e.g., CopDEM, or GHS-SMOD, which is available at 1-km resolution only), the cropped input data are resampled in memory to the highest resolution among input datasets. For spatial visualisations at country-level or at smaller extents, cropped datasets can optionally be warped into local UTM projection for mapping purposes. All relevant, gridded data are then loaded into spatio-temporal 3-d arrays, and for non-spatial visualisations, 3-d arrays are restructured to a 2-d array, and stored in tabular data frame, with columns representing the different input raster datasets, and rows representing an individual grid cell. Finally, relevant summary statistics are extracted, e.g., total built-up surface and population by degree of urbanisation class and year.

2.3. Data visualisation

Based on the spatio-temporal data cubes covering the area of interest, and/or the extracted summary statistics, different static and animated data visualisations can be produced. Herein, we discuss a selection of them, including (1) automatically produced animated GIFs of urban change per urban centre, and (2) Built-up area - population (BUPOP) plots.

2.3.1. Animated GIFs depicting urban change

Based on the spatio-temporal 3-d arrays loaded in memory, matplotlib Python library is used to render temporal slices of the data as individual maps, depicting the distributions of built-up surface, building volume, resident population, or rural-urban classes, including automatically added map elements such as scale bar, north arrow, title, and legend. Rendered images are again stored in memory and exported to an animated GIF (graphics interchange format) or MP4 movie file. Figure 1 shows examples of individual frames (i.e., epochs) of these animated maps for selected variables.

2.3.2. Built-up area – population plots (BUPOPs)

As an example of non-spatial data visualisation, we developed "BUPOPs" – built-up area – population plots. BUPOPs are bidirectional, horizontal, stacked bar charts, conveying information of four variables measured and mapped in the GHSL data collection: 1) built-up surface, 2) resident population, 3) their variation over time, 4) within classes of the rural-urban continuum, according to the Degree of Urbanisation framework. The input data is based on cross-tabulation of the gridded data and extracted summary statistics, e.g., total built-up surface area in a given year, within a given class of the Degree of Urbanisation. See Figure 2 for some examples.

3. RESULTS

We generated animated GIFs of built-up surface, building volume, resident population, and Degree of Urbanisation for each of the >10,000 urban centres of the UCDB, and produced the BUPOPs for each urban centre, and for each country globally. Such animated maps illustrate effectively the patterns of urban configuration and urban change, immediately understandable to expert and non-expert audiences, and the differences between individual cities, such

doi:10.2760/2119408

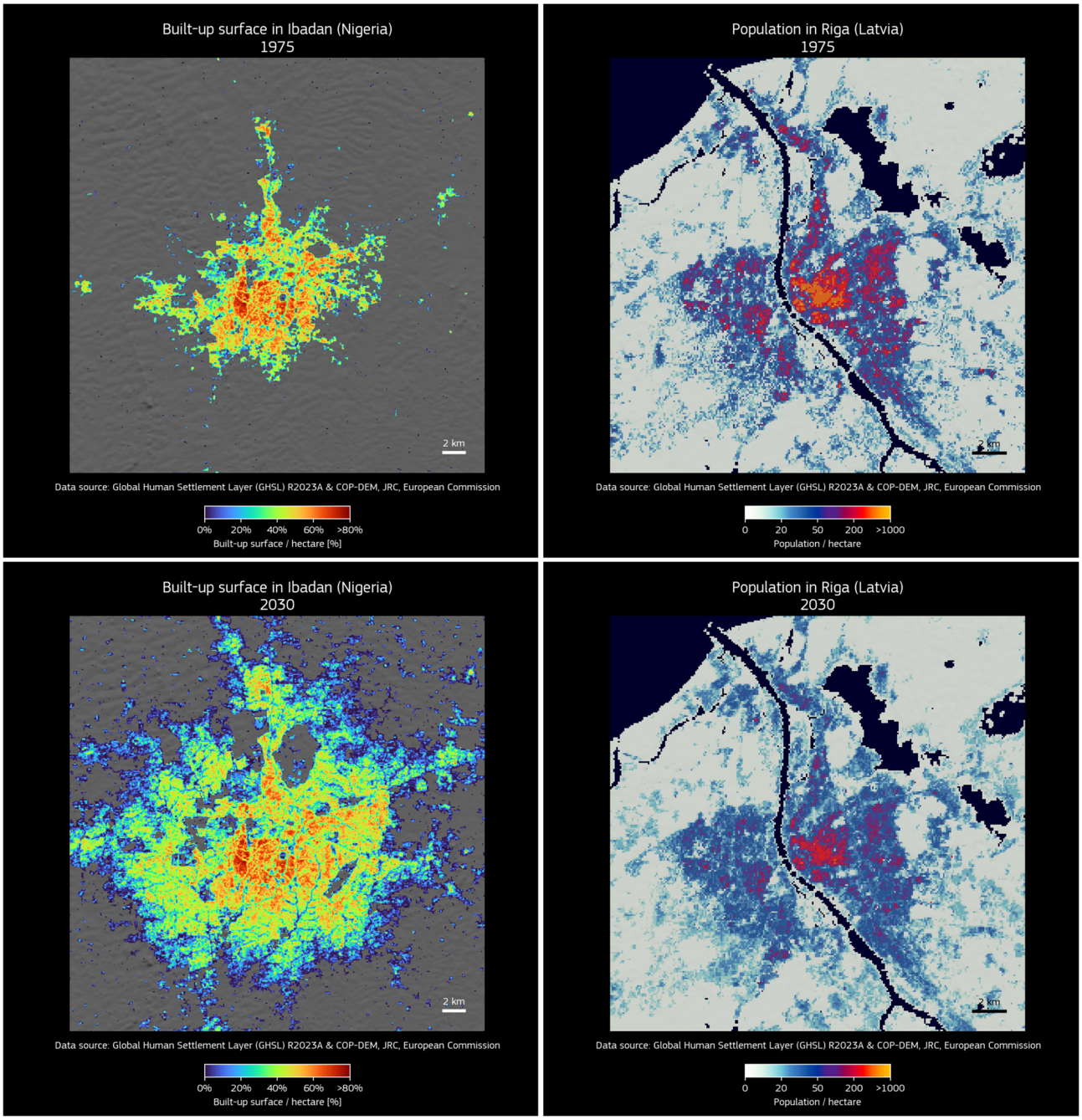


Fig. 1. Examples of stills of fully automatically produced animated maps of spatio-temporal GHSL variables at the urban-centre level. Left: Built-up surface distributions shown for the city of Ibadan (Nigeria) in 1975 and 2030, based on GHS-BUILT-S R2023A; Right: Residential population distributions shown for the city of Riga (Latvia) in 1975 and 2030, based on GHS-POP R2023A.

as the fast, concentric growth of built-up areas in Ibadan (Figure 1 left) as compared to the subtle change of the urban footprint in Riga, with declining population densities, most pronounced in the central parts of Riga, from 1975 to 2030 (Figure 1 right).

The BUPOPs shown in Figure 2 require the readers to "orient" themselves: population extends to the right, built-up surface to the left, time from bottom to top, red color indicates urban, yellow peri-urban, and green rural areas. Once the reader is familiar with this concept, the multivariate urban dynamics, including their interactions (e.g., change in built-up area in relation to change in population, which is an important metric for measuring land use efficiency using SDG indicator 11.3.1) become obvious, and the "signature" of multiple cities or regions can be compared effectively. For example, in Ibadan, population has increased much more than

doi:10.2760/2119408

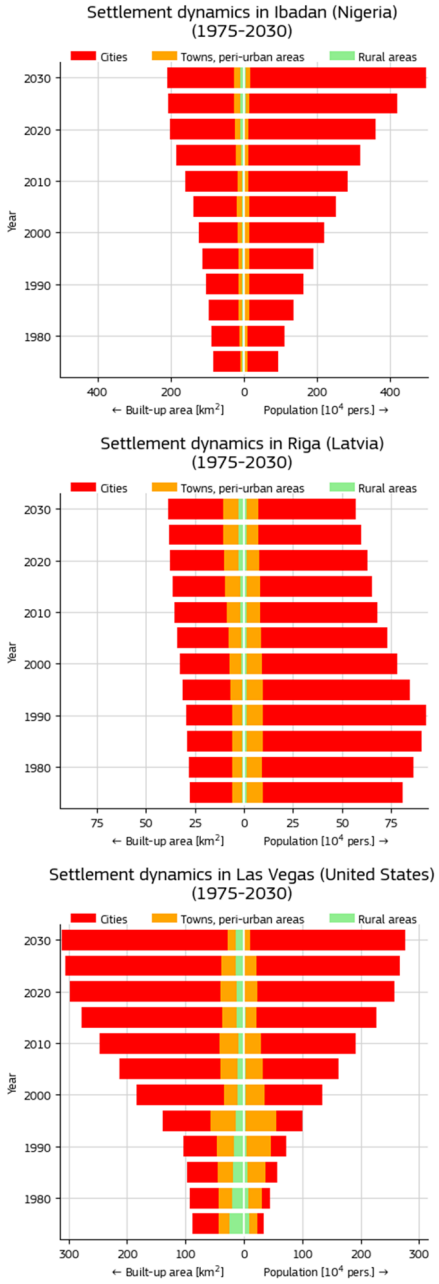


Fig. 2. BUPOPs (built-up area – population plots) effectively illustrate the settlement dynamics of a given place, region, or country, in terms of the change in built-up area and population over time and their shares across rural, peri-urban, and urban classes.

built-up area. In Riga, built-up area has experienced a very moderate growth rate, while population has been declining since 1990. In the case of Las Vegas (USA), built-up area and population have grown approximately with the same relative change rates. Moreover, the skewness of the triangular shape indicates the relationship between built-up area and population: a skew to the right indicates low built-up area per capita (e.g., Ibadan) suggesting high population density, while the slight skew to the left as observed for Las Vegas indicates the opposite, high rates of built-up area per capita, suggesting lower population densities. In all cases, growth occurs within the urban centre, while the relative shares of population and built-up area in peri-urban areas have been stable (Riga) or declining (Las Vegas) over time.

4. CONCLUSIONS AND OUTLOOK

Herein, we described selected, effective visual tools to convey information on urbanisation and settlement dynamics to expert and non-expert audiences, fostering insights from integrated, large amounts of open Earth observation and other data sources. In future work, we will develop and test further visualisation techniques and make them available to the public via a web-based tool. Importantly, the prototypes presented herein are based on locally stored data. For efficient, customised, user-centric web visualisation, we are currently exploring the use of pre-computed summary statistics and modern web-based infrastructure (e.g., Voilà dashboards [10], Google Earth Engine, cloud-optimised GeoTIFFs (COGs), cloud storage infrastructure), facilitating the real-time visualisation of large datasets such as the GHSL data. Such tools will foster data democratisation, the accessibility to space-based knowledge, and contributing to data-driven decision making.

REFERENCES

[1] K.S. McLeod, “Our sense of Snow: the myth of John Snow in medical geography“. Social science & medicine, 50(7-8), 2000.
[2] S. Bohman, “Data visualization: an untapped potential for political participation and civic engagement“. In International Conference on Electronic Government and the Information Systems Perspective (pp. 302-315). Cham: Springer Intl. Publishing, 2015.
[3] M. Pesaresi, M. Schiavina, P. Politis, S. Freire, K. Krasnodębska, K., J.H. Uhl, et al., “Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data”, International Journal of Digital Earth, 17(1), 2024.
[4] European Commission, GHSL Data Package 2023, Publications Office of the European Union, Luxembourg, doi:10.2760/098587, JRC133256, 2024.
[5] S. Freire, K. MacManus, M. Pesaresi, E. Doxsey-Whitfield, & J. Mills, “Development of new open and free multi-temporal global population grids at 250 m resolution“. Population, 250, 33, 2016.
[6] L. Dijkstra, A.J. Florczyk, S. Freire, T. Kemper, M. Melchiorri, M. Pesaresi, & M. Schiavina, “Applying the degree of urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation“. Journal of Urban Economics, 125, 103312, 2021.
[7] Copernicus DEM - Available online: https://doi.org/10.5270/ESA-c5d3d65. Last access: 22-05-2025.
[8] GADM global administrative areas. Available online: https://gadm.org, 2025. Last access: 22-05-2025
[9] European Commission, “Stats in the City: the GHSL Urban Centre Database 2025“, Publications Office of the European Union, Luxembourg, JRC139768, 2025.
[10] Voilà dashboards. Available online. https://jeodpp.jrc.ec.europa.eu/bdap/voila/. Last access: 31-08-2025.

ANALYSIS OF BIG EARTH OBSERVATION DATA CUBES WITH SATELLITE IMAGE TIME SERIES

Gilberto Camara, Felipe Souza, Felipe Carlos, Rolf Simões

National Institute for Space Research (INPE)
Avenida dos Astronautas 1758, São José dos Campos, SP

ABSTRACT

Satellite image time series methods have proven to be a valuable approach for analysing large datasets in Earth observation. Time series offer a systematic means of capturing change, leveraging the enhanced temporal resolution of satellites such as Sentinel-1 and Sentinel-2. This paper presents insights and lessons learned when applying time series analysis to big EO data, especially in the context of land use in tropical forests and statistics of tropical agriculture. We discuss the need for flexible definitions of data cubes, the benefits of the *time-first, space-later* approach, the importance of modelling events in time series, and the key differences between model validation and map accuracy. The authors also present some perspectives on possible trends in big EO data analysis.

Index Terms— Big EO analytics, satellite image time series, EO data cubes, foundational models.

1. INTRODUCTION

With free access to Earth observation (EO) massive data sets, we need new methods to measure change on our planet. One particular research area that big EO data have enabled is the analysis of time series of satellite images. Combined with EO data cubes, time series are a powerful tool for monitoring change. Using time series, experts improve their understanding of ecological patterns and processes. Instead of selecting individual images from specific dates and comparing them, researchers track change continuously.

To support research and operational work using time series, the authors have been developing the `sits` package, an end-to-end environment for land use and land cover classification [1]. Since 2020, the package has been utilised for large-scale operational land use monitoring, including the measurement of tropical deforestation and agricultural statistics. Such experience enabled package authors and users to learn several lessons related to big EO analytics, particularly in the context of time series. This paper presents some of these lessons to share experience in big EO analytics.

Thanks to IKI (Germany International Climate Initiative) for funding.

2. BRIEF DESCRIPTION OF SITS

The `sits` programming interface (API) provides a set of functions to create a workflow for land classification. The package is written in R and offers an API for both R and Python. The package leverages analysis-ready data (ARD) collections to extract and classify time series. Supported collections include AWS, Microsoft Planetary Computer, Copernicus Data Space Ecosystem (CDSE), Digital Earth Africa, Digital Earth Australia, and Brazil Data Cube. Since machine learning algorithms for time series require consistent data, `sits` creates regular data cubes from parts of ARD collections. The package also provides methods for merging data cubes from different satellites (e.g., Sentinel-1 and Sentinel-2) and ancillary data such as DEMs. Users can also apply arithmetic operations to derive new attributes from data cubes, such as vegetation indices.

Given a regular data cube and a set of ground truth data, `sits` obtains training samples containing time series for selected locations in the training area. Users can perform quality control on training samples using self-organised maps [2] and also reduce sample imbalance.

Supported machine learning algorithms include Random Forests, XGBoost, Temporal Convolution Neural Networks [3], Temporal Attention Encoders [4], and Residual Networks [5]. Data cube classification uses GPUs when available, producing probability matrices that are post-processed to remove outliers using Bayesian smoothing [6]. The package also estimates classification uncertainty to support active learning.

3. LESSONS LEARNED IN BIG EO ANALYTICS

3.1. Flexible Definition of Data Cubes

Machine learning and deep learning (ML/DL) algorithms for spatiotemporal data require conversion of ARD image collections from EO cloud services to regular data cubes. Appel and Pebesma [7] propose a definition of *data cubes* as an n-dimensional matrix of cells combining a 2D geographical location, a 1D set of temporal intervals, and a k-dimensional set of attributes. For each position in space, a data cube gen-

doi:10.2760/2119408

erates a multidimensional time series. For each time interval, users obtain a valid 2D image. In this definition, all pixels of a data cube follow the same coordinate system.

In `sits`, we have extended that data cube definition to include a further dimension related to the spatial organisation used by the ARD image collection. For example, Sentinel-2 images are organised in the MGRS tiling system, which follows the UTM grid. Thus, to process data spanning multiple UTM grid zones, EO data cubes require an additional dimension provided by the ARD tiles. This extension enables `sits` to process large-scale data, unlike systems that adopt a more restricted data cube definition.

3.2. Time-first, space-later

The *time-first, space-later* concept in satellite image classification prioritises time series analysis as the initial processing step. Then it uses spatial information after classifying all time series. The *time-first* part allows a better understanding of changes in landscapes. Time series classification is pixel-based, producing a set of class probability matrices for each pixel. This result is the input for the *space-later* part of the method. In this phase, a smoothing algorithm improves the results of the time-first classification by considering the spatial neighbourhood of each pixel. The resulting map thus combines both spatial and temporal information.

Spatial smoothing methods improve the accuracy of land classification by incorporating spatial and contextual information. The smoothing method available in `sits` uses an Empirical Bayes approach, adjusted to the specific properties of land classification. The assumption is that class probabilities at the local level should be similar. Thus, probabilities in a spatial neighbourhood of a pixel provide the baseline for comparison with those produced by the classifier for each pixel. Based on these two elements, Bayesian smoothing adjusts the probabilities of the pixels, taking into account spatial dependence [6]. Our experience is that the *space-later* part of the time series analysis significantly improves the results.

3.3. Modelling Events in Time Series

To represent change in geographical space, authors distinguish between *objects* and *events* [8]. Objects refer to entities that endure through time even while undergoing different sorts of changes. The Amazon Forest and the city of Brasilia are objects. Events occur within a well-defined period and may have distinct stages during this time. Cutting down a forest area, cultivating a crop in a season, and building a road are events. Time series analysis methods are particularly well-suited for detecting events in data cubes, as they can capture seasonal variations and abrupt changes.

Souza et al. [9] built a set of event-based training samples for measuring deforestation in the Amazon using Sentinel-2 time series. The authors considered three types of classes: (a) deforestation classes defined based on events measured by breaks in the time series, (b) natural classes with events linked to seasonal variation, and (c) stable natural classes. An example of the first case is a time series that begins with the response of a stable forest cover and is interrupted by a signal related to a forest fire. In the second case, we have seasonally flooded wetlands; their signals follow seasonal patterns. Seasonally variable classes are often confused with deforestation areas when working with single-date images. Kinnebrew et al. [10] report that Global Forest Change maps confuse wetlands with agriculture based on a single-date comparison. Using time series minimises such confusion. The authors achieved a 95% agreement with expert visual interpretation, which is much better than other automated methods [11].

3.4. Model Validation and Map Accuracy

Most works on ML models for EO use cross-validation for assessing the generalisation performance of machine learning models. Its primary purpose is to provide an unbiased estimate of a model's ability to perform on independent, unseen data, thereby helping to prevent overfitting. However, performance estimates obtained via cross-validation may not fully reflect the conditions encountered in real data. In most situations, training samples do not capture the full variation present in the entire dataset. Thus, measures of cross-validation are not reliable predictors of map accuracy. To achieve proper map accuracy, one should employ a statistically based sampling approach to compare predicted and actual pixel classes on the resulting map [12]. In real-world cases, map accuracy results differ substantially from those obtained through cross-validation.

Due to this focus on cross-validation, few works in the literature provide a statistically sound comparison of ML algorithms for EO. In Souza et al. [13], we evaluate how the temporal convolutional neural network (TCNN) [3] and the lightweight temporal self-attention (LTAE) [4] differ in their cross-validation and map accuracy. We selected a study area in Petrolina, located in the Caatinga biome, a semi-arid region characterised by land changes driven by agricultural expansion and livestock farming. These areas are complex landscapes with distinctive land use practices and climate regimes. We achieved high F1 scores for both methods during cross-validation, with values greater than 0.90.

Map accuracy measures differ significantly from cross-validation in our case study. For natural vegetation, we obtained an F1-score of 0.81 using TCNN, whereas LTAE produced an F1-score of 0.90. LTAE enabled better identification of smaller areas covered by xeric shrubland and thorny trees. Given the well-defined seasonal patterns of permanent crops and water bodies, LTAE and TCNN achieve high accuracy when classifying them. TCNN had a much lower F1-score (0.41) than LTAE (0.81) for temporary crops that do not have well-defined seasonal variations. Overall, LTAE is significantly better than TCNN for Petrolina.

doi:10.2760/2119408

The results indicate important differences between these algorithms. Convolution-based approaches, such as TCNN, are effective in areas with well-defined temporal signatures and for detecting abrupt events. In cases where the same class is associated with different temporal signatures (as in semi-arid regions), attention-based methods such as LTAE are more effective. We conclude that users need to understand how each ML algorithm works in practice to choose the best method for their problems. The work also demonstrates that cross-validation is not a reliable predictor of map accuracy.

3.5. Quality Control of Training Samples

Selecting high-quality training samples for machine learning classification of satellite images is crucial for achieving accurate results. Thus, it is beneficial to use pre-processing methods to improve the quality of samples and eliminate those that may have been incorrectly labelled or possess low discriminatory power.

When working in a large geographic region, the limitation of terms to describe nature and the variability of vegetation phenology lead to the assignment of the same label to different spectral and temporal responses. A related issue is the limitation of crisp boundaries to describe the natural world. Class definitions use idealised descriptions (e.g., "a savanna woodland has tree cover of 50% to 90%, ranging from 8 to 15 m in height"). Class boundaries are fuzzy and sometimes overlap, making it hard to distinguish between them. To improve sample quality, `sits` supports training data evaluation using a SOM-based algorithm [2]. The SOM-based method identifies potential mislabelled samples and outliers that require further investigation. Unlike methods based on confusion matrices, algorithms such as SOM allow quality estimators for individual samples. The resulting improvements show the need for further research on methods for sample quality control [14].

4. PERSPECTIVES

Research on big EO data analytics has been heavily influenced by techniques derived from Computer Vision, such as U-Net, masked autoencoders and Vision Transformers. These methods underpin most current research focused on foundational models for Earth observations [15, 16]. However, there are important shortcomings in the current generation of ML models for EO, especially those whose input consists of fixed-sized labelled patches [17]. In many cases, these patches are RGB images. Satellite images, such as Sentinel-2, have 10 spectral bands; reducing them to RGB patches leads to a significant loss of information content. As pointed out by Xiao et al. [18], foundational models for EO face several challenges, which include the differences between satellite and natural images, a shortage of large training sets, and their reliance on Computer Vision techniques.

Most current algorithms for object classification rely on the distinction between foreground (*"things"*) and background (*"stuff"*). While this design is suitable for high spatial resolution images with pixels of 3 meters or smaller, it is unsuitable for mid- to low-resolution images (pixels of 10 meters or larger). Mid and low-resolution images are continuous distributions of radiance values and are better described as fields than as collections of objects [19]. Human-sized, everyday objects depicted in natural photos differ from continuous landscapes captured in satellite images. All pixels matter when working with Sentinel-like data for land mapping and similar broad-area applications; the distinction between foreground and background is not universally applicable.

Arguably, no proper *"objects"* exist in mid to low-resolution images; image classification identifies compact regions of similar values in multidimensional spaces. While domain scientists may believe they recognise objects in a remotely sensed image, they are actually measuring fields.

A further challenge to ML models derived from Computer Vision is dealing with satellite image time series. Time series capture the evolution of geospatial fields and objects, enabling the detection of *events* such as deforestation, desertification, mudslides, and surface water loss. Event definitions are not covered in fixed hierarchies such as ImageNet because they require continuous change monitoring.

Current research on foundational models for EO focuses on combining diverse datasets, many with different resolutions and sensors, in the expectation of extracting embeddings that can be applied to various problems [15, 16]. These models make the strong assumption that spectral and temporal signatures of classes in remote sensing images are separable and mappable to a hierarchical structure such as that of ImageNet, where each term is unique and precisely defined. However, ontologies that describe the geographical world are inherently polysemic. Consider the concept of 'forest'. What counts as a forest depends on who defines a certain piece of land as one [20]. Countries use different conceptualisations of forest, based on the physical parameters of tree height and crown canopy cover [21]. The UN Food and Agriculture Administration (FAO) considers that a forest may be temporarily devoid of trees. Thus, it is unlikely that foundational models will be able to support variable definitions of 'forest' without massive improvements in data collection.

Remote sensing classification will always be task- and context-dependent. Many foundational models combine different sources, including SAR and optical images, DEMs, land use maps, and text, without considering their inherent differences. However, in the absence of suitable training data to validate the model, reported performance accuracies [22] fail to match what is currently achieved with application-centred training data sets.

Arguably, there are alternative paths to make progress in big EO analytics. One recommended approach is to place an increased emphasis on improving the collection of training

doi:10.2760/2119408

data. As stated by Roscher et al. [14]: "A shift from a model-centric view to a complementary data-centric perspective is necessary for further improvements in accuracy, generalisation ability, and real impact on end-user applications". The lack of methods for improving the extraction and evaluation of training samples remains a significant barrier to innovation in big EO analytics across various approaches. Without significant progress in this area, ML algorithms for EO will continue to require substantial effort from users when selecting datasets to achieve high-quality results.

References

[1] Rolf Simoes et al. "Satellite Image Time Series Analysis for Big Earth Observation Data". In: *Remote Sensing* 13.13 (2021), p. 2428.

[2] Lorena A. Santos et al. "Quality Control and Class Noise Reduction of Satellite Image Time Series". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 177 (2021), pp. 75–88.

[3] Charlotte Pelletier, Geoffrey I. Webb, and Francois Petitjean. "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series". In: *Remote Sensing* 11.5 (2019).

[4] Vivien Garnot and Loic Landrieu. "Lightweight Temporal Self-attention for Classifying Satellite Images Time Series". In: *Advanced Analytics and Learning on Temporal Data*. Ed. by Vincent Lemaire et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 171–181.

[5] Hassan Ismail Fawaz et al. "Deep Learning for Time Series Classification: A Review". In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963.

[6] Gilberto Camara et al. "Bayesian Inference for Post-Processing of Remote-Sensing Image Classification". In: *Remote Sensing* 16.23 (2024), p. 4572.

[7] Marius Appel and Edzer Pebesma. "On-Demand Processing of Data Cubes from Satellite Image Collections with the Gdalcubes Library". In: *Data* 4.3 (2019).

[8] Antony Galton. "Fields and Objects in Space, Time, and Space-time". In: *Spatial Cognition & Computation* 4.1 (2004), pp. 39–68.

[9] Anielli Rosane de Souza et al. "An Event-Based Approach for Training Data Selection for Mapping Deforestation". In: *Proceedings of XXI Brazilian Symposium on Remote Sensing*. Salvador: INPE, 2025.

[10] Eva Kinnebrew et al. "Biases and Limitations of Global Forest Change and Author-Generated Land Cover Maps in Detecting Deforestation in the Amazon". In: *PLOS ONE* 17.7 (2022), e0268970.

[11] Antonio Fonseca and Robert Pontius. "Comparison among Time Series Maps of Deforestation in the Amazon: How Independent Monitoring Systems Relate to Official Data". In: *XX Brazilian Symposium on Remote Sensing*. 2023.

[12] Pontus Olofsson et al. "Good Practices for Estimating Area and Assessing Accuracy of Land Change". In: *Remote Sensing of Environment* 148 (2014), pp. 42–57.

[13] Felipe Carvalho de Souza et al. "Integration of Radar and Optical Data for Identifying Tropical Forest Disturbances". In: *Proceedings of XXI Brazilian Symposium on Remote Sensing*. Salvador: INPE, 2025.

[14] Ribana Roscher et al. "Better, Not Just More: Data-Centric Machine Learning for Earth Observation". In: *IEEE Geoscience and Remote Sensing Magazine* 12.4 (2024), pp. 335–355.

[15] Daniela Szwarcman et al. *Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications*. 2025.

[16] Gabriel Tseng et al. *Lightweight, Pre-trained Transformers for Remote Sensing Timeseries*. 2024.

[17] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. "Position: Mission Critical – Satellite Data Is a Distinct Modality in Machine Learning". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024, pp. 42691–42706.

[18] Aoran Xiao et al. "Foundation Models for Remote Sensing and Earth Observation: A Survey". In: *IEEE Geoscience and Remote Sensing Magazine* (2025), pp. 2–29.

[19] Helen Couclelis. "People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS". In: *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. Ed. by A. U. Frank, I. Campari, and U. Formentini. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1992, pp. 65–77.

[20] Robin L. Chazdon et al. "When Is a Forest a Forest? Forest Concepts and Definitions in the Era of Forest and Landscape Restoration". In: *Ambio* 45.5 (2016), pp. 538–50.

[21] Alexis Comber, Peter Fisher, and Richard Wadsworth. "What Is Land Cover?" In: *Environment and Planning B: Planning and Design* 32.2 (2005), pp. 199–209.

[22] Johannes Jakubik et al. "TerraMind: Large-Scale Generative Multimodality for Earth Observation". In: *International Conference on Computer Vision 2025*. arXiv, 2025.

doi:10.2760/2119408

DEVELOPING A DATA CUBE FOR BIODIVERSITY AND CARBON DYNAMICS ASSESSMENT IN ESTONIA WITH REMOTE SENSING DATA

Evelyn Uuemaa[1], Oleksandr Borysenko[2], Jan Pisek[2], Holger Virro[1], Wai Tik Chan[1], Eveli Sisas[1], Ats Remmelg[1], Marta Jemeljanova[1], Alexander Kmoch[1]

[1]Landscape Geoinformatics Lab, Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, Tartu, Estonia

[2]Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere, Estonia

ABSTRACT

Addressing environmental challenges demands accessible and well-structured Earth observation data. This paper presents the development of a national-scale data cube for Estonia, integrating diverse remote sensing (Sentinel-1/2, Light Detection and Ranging (LiDAR)) and geospatial datasets. The aim is to provide analysis-ready data, particularly for biodiversity and carbon dynamics research, by overcoming common technical hurdles associated with Earth Observation big data. The framework emphasizes user-friendliness, offering intuitive access and visualization tools. By leveraging cloud computing and open-source standards, this work facilitates efficient data retrieval and analysis, empowering researchers and policymakers with timely environmental information for informed decision-making and sustainable development.

Index Terms— data cube, Earth Observation, biodiversity, carbon

1. INTRODUCTION

Addressing pressing global environmental challenges, including land use change and climate change, requires the availability of timely and accurate information regarding their drivers and impacts. Earth Observation (EO) data, encompassing satellite and in-situ data from diverse sources [1], has emerged as a crucial resource for monitoring these dynamics. Facilitated by free and open data policies [2] and advancements in open-source software and cloud computing [3], EO data enables more effective environmental management, informed policy assessment, and ultimately contributes to sustainable development. The capacity to process and analyse the burgeoning volumes of EO data holds immense potential for understanding complex environmental processes and informing decision-making across various scales, particularly when fusing diverse geodata and remote sensing data from disparate sources. This integration of multi-modal data, such as combining optical imagery with active sensor data like LiDAR and radar, allows for a more holistic and detailed understanding of environmental conditions and changes.

However, the effective utilization of EO big data presents significant technical hurdles [4]. Prior research has highlighted key obstacles, such as limitations in data storage, transmission, and analysis, alongside the need for developing suitable computational architectures capable of handling such immense datasets [5]. Furthermore, classical data cube implementations often fall short of providing analysis-ready data optimized for advanced analytical techniques, particularly artificial intelligence (AI) and machine learning (ML) algorithms. These methods thrive on structured, consistently formatted, and feature-rich datasets, which traditional data cubes may not inherently offer. Moreover, data cube platforms should be user-oriented, providing intuitive access, analysis tools, and customizable functionalities to cater to a diverse community of researchers, policymakers, and practitioners.

The aim of this work was to construct a comprehensive data cube at the national level for Estonia, leveraging remote sensing and geospatial data to mainly advance biodiversity and carbon dynamics research. The full potential of fusing active (LiDAR, radar) and passive remote sensing has not been fully developed and utilized yet in biodiversity and carbon modelling studies. To efficiently relate the ground measurements with remote sensing data and create spatial models, unified easily accessible data is needed. Moreover, multi-temporal (seasonal) data sets, consisting of numerous combinations of spectral bands, can hold significant potential to predict compositional vegetation classes and other environmental variables.

2. DATA AND METHODS

We used remote sensing data, including Sentinel-1, Sentinel-2 (ESA Copernicus), and high-resolution airborne LiDAR data as raw point cloud data and digital elevation model (Estonian Land and Spatial Development Board),

doi:10.2760/2119408

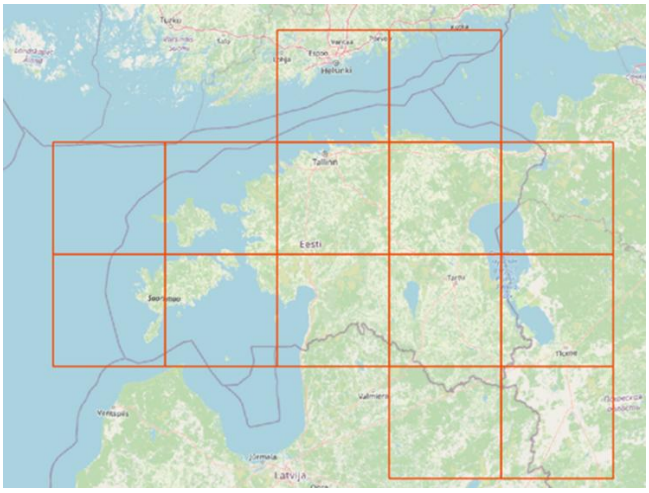


Fig. 1. The spatial sub-division of Estonia into 100×100 km processing zones.

Estonian soil map EstSoil-EH [6] for the data cube layer preparation.

We created a spatial grid to divide the area into equally sized manageable tiles for processing and export (Fig. 1). These tiles were used to create spatially aligned 10m resolution tiles in Estonian National CRS (L-Est-97, EPSG 3301). Based on the tiles, we created Cloud Optimized GeoTIFFs (COGs) (https://cogeo.org/), which ensure fast read and download operations [7].

We used the following workflow for processing Sentinel-1 and Sentinel-2 images with the Python GEE API:

1) For Sentinel-1 (ascending) and Sentinel-2 L2A images, we created seasonal (April–May, June–August, September–October) median composites for every year. For Sentinel-1, we implemented the speckle filter developed by Mullissa et al. [8]. For Sentinel-2, cloud masking using the CloudScore+ algorithm [9] was used.

2) Based on the seasonal composite images, various vegetation indices were calculated: Normalised Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Fractional Vegetation Cover (FVC), Normalised Difference Water Index (NDWI), Bare Soil Index (BSI), Normalised Difference Moisture Index (NDMI), Green Normalised Difference Vegetation Index (GNDVI), Radar Vegetation Index (RVI).

3) For biodiversity assessment, spectral species concepts and k-means clustering are commonly used to analyse gridded remote sensing data, producing 2D α- and β-diversity heterogeneity maps. We calculated α- and β-diversity heterogeneity using the biodivMapR library [10].

4) Due to infrequent cloud-related no-data pixels in seasonal composites over the past few years, we applied gap-filling using a yearly mean composite. We also generated a corresponding binary no-data mask, allowing for the identification of gap-filled areas in subsequent analyses if necessary.

In addition to the Senintel-1 and Sentinel-2 data, we added climate data from ERA5 Land monthly reanalysis data: Mean Air Temperature (2m), Total Precipitation, Minimum Air Temperature (2m), Maximum Air Temperature (2m). We also calculated different topographic indices based on the 10 m Estonian digital elevation model [11]: slope, Terrain Wetness Index, Terrain Ruggedness Index, LS-factor.

From LiDAR point cloud data [11], we calculated various indices relevant for biodiversity: ecosystem height (maximum, mean, median vegetation height), ecosystem cover (density of vegetation points within defined height layers, canopy openness), ecosystem structural complexity (coefficient of variation of vegetation height, standard deviation of vegetation height, variance of vegetation height). We used PDAL [12] for LiDAR point cloud reclassification and filtering and laspy [13] for low-level bit corrections. PDAL pipelines were used to calculate the indices.

3. IMPLEMENTATION

We utilized the high-performance cloud computing platform provided by the University of Tartu [14] to execute our data cube operations. For data storage, we employed S3-compatible object storage (buckets) to manage the Cloud Optimized GeoTIFFs (COGs). Using libraries such as rasterio, we scanned these files to extract metadata for our SpatioTemporal Asset Catalog (STAC), including asset names, extent boundaries, spectral bands, resolution, coordinate reference systems (CRS), data types, and nodata values. This information was instrumental in generating STAC JSON metadata collections, which we made accessible through a STAC-compatible web service based on pygeoapi. To enhance our service, we extended pygeoapi to enable search functionalities via the STAC-API. In QGIS, we employed the QGIS STAC API Browser plugin to connect to and query our STAC API endpoint. For high-performance computing (HPC) and scripted batch processing workflows, we utilized the Python pystac library to query the STAC metadata and select GeoTIFFs for processing. For the creation of CSW/ISO-compatible metadata, we adhered to the ISO 19115/117 and ISO 19137 XML standards for geographic metadata, which were implemented in the GeoNetwork OpenSource metadata catalog server (https://geonetwork-opensource.org/).

The core of our Data Cube Viewer (https://geokuup.ee/) is implemented in the Elixir language, using the Phoenix (https://www.phoenixframework.org/) framework. The data cube viewer component is building on the MapLibre JavaScript library to link in the WMS and WMST/Tiles layers. The COGs are visualised as WMS through GeoServer, which are directly registered from their object storage bucket locations. The Data Cube Viewer enables quick visualisations of all the layers and also timeseries queries.

doi:10.2760/2119408

In the Data Cube Viewer, datasets are also organised into collections - here datasets are referred to via the linkage to

Fig. 2. The data cube viewer that enables quickly visualize all layers and make quick queries, including timeseries.

viewing services, such as the GeoServer (https://geoserver.org) WMS. This allows users to view, compare, and query these layers. Several collections are already provided by us, including seasonal views of Sentinel-1 and Sentinel-2 indicators, terrain and topographic indices.

Several choices of the general architecture are oriented on common best practices for working with big geospatial data, such as using object storage and STAC collections to index data files, building upon the easy-to-use standards-based open-source tools like GeoServer, pygeoapi, and GeoNetwork, and relying on the Pangeo (https://pangeo.io/) ecosystem of well-integrated geospatial Python libraries for processing and workflows. However, for the user-facing Data Cube Viewer portal we adopted the Phoenix framework (built on Elixir and the BEAM virtual machine) as a pragmatic technology choice to balance performance, maintainability, with developer experience and efficiency.

We considered the Elixir, Phoenix, and LiveView combination being advantageous over separate JavaScript or Python-based alternatives, particularly in handling concurrent user connections and maintaining system stability under varying load conditions. At the same time, the developer experience with Phoenix and its LiveView capabilities allows our small team to rapidly implement responsive UI components without the complexity of maintaining separate frontend and backend codebases. In addition, we opted for the MapLibre over Leaflet or OpenLayers for its out-of-the-box rendering performance with mixed data sources, including tiles and vector data sources. Another advantage is its recent 3D globe view implementation, that interacts well with otherwise standard geospatial data sources, but does not exhibit the implementation complexities of libraries, like CesiumJS.

4. CONCLUSIONS

Our framework organizes analysis-ready spatial data in a data cube at national level, enabling efficient retrieval, storage, and extraction of spatial and temporal extents from input and project-generated datasets. The data cube is very user oriented and aims to provide easy access to the high-resolution spatial data for academic and governmental agencies. The data cube includes variables mainly relevant for biodiversity and carbon studies, but this can be also easily extended to other studies with additional data. The future work includes further processing data into multi-resolution and adding complementary datasets and processing workflows for data retrieval and analysis.

ACKOWLEDGEMENTS

This work was funded by the Estonian Research Agency (grant number PRG1764, PSG841), Estonian Ministry of Education and Research, Centre of Excellence for Sustainable Land Use (TK232), and by the European Union (ERC, WaterSmartLand, 101125476). However, views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible.

REFERENCES

[1] Giuliani, G., Masó, J., Mazzetti, P., Nativi, S. & Zabala, A. Paving the Way to Increased Interoperability of Earth Observations Data Cubes. Data 4, 113, 2019.

[2] Harris, R. & Baumann, I. Open data policies and satellite Earth observation. Space Policy 32, 44–53, 2015.

[3] Wu, B., Tian, F., Zhang, M., Zeng, H. & Zeng, Y. Cloud services with big data provide a solution for monitoring and tracking sustainable development goals. Geography and Sustainability 1, 25–32, 2020.

[4] Xu, C. et al. Cloud-based storage and computing for remote sensing big data: a technical review. International Journal of Digital Earth 15, 1417–1445, 2022.

[5] Yang, C., Huang, Q., Li, Z., Liu, K. & Hu, F. Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth 10, 13–53, 2017.

[6] Kmoch, A. et al. EstSoil-EH: A high-resolution eco-hydrological modelling parameters dataset for Estonia. Earth System Science Data 13, 2021.

[7] Chatenoux, B. et al. The Swiss data cube, analysis ready data archive using earth observations of Switzerland. Sci Data 8, 295, 2021.

[8] Mullissa, A. et al. Sentinel-1 SAR Backscatter Analysis Ready Data Preparation in Google Earth Engine. Remote Sensing 13, 1954, 2021.

[9] Pasquarella, V. J., Brown, C. F., Czerwinski, W. & Rucklidge, W. J. Comprehensive quality assessment of optical satellite imagery using weakly supervised video learning. in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2125–2135, 2023. doi:10.1109/CVPRW59228.2023.00206.

[10] Féret, J.-B. & de Boissieu, F. biodivMapR: An r package for α- and β-diversity mapping using remotely sensed images. Methods in Ecology and Evolution 11, 64–70, 2020.
[11] Estonian Land and Spatial Development Board. Elevation Data. https://geoportaal.maaamet.ee/eng/Maps-and-Data/Elevation-data/Download-Elevation-Data-p664.html, 2025.
[12] PDAL Contributors, 2025. PDAL Point Data Abstraction Library. https://doi.org/10.5281/zenodo.10884408
[13] Laspy Contributors, 2025. https://github.com/laspy/laspy
[14] University of Tartu. UT Rocket. share.neic.no. https://doi.org/10.23673/PH6N-0144

BIOCUBE: A MULTIMODAL DATASET FOR BIODIVERSITY RESEARCH

*Stylianos Stasinos**

Amazon
EU INTech
16 Rue Edward Steichen, 2540,
Luxembourg, Luxembourg

Martino Mensio, Elena Lazovik, Athanasios Trantas

Dutch Organization for Applied Scientific Research
TNO - ICT, Strategy & Policy
Anna van Buerenplein 1, 2595 DA, Den Haag,
The Netherlands

ABSTRACT

Biodiversity research requires complete and detailed information to study ecosystem dynamics at different scales. Employing data-driven methods like Machine Learning is getting traction in ecology and more specific biodiversity, offering alternative modelling pathways. For these methods to deliver accurate results there is the need for large, curated and multimodal datasets that offer granular spatial and temporal resolutions. In this work, we introduce BioCube, a multimodal, fine-grained global dataset for ecology and biodiversity research. BioCube incorporates species observations through images, audio recordings and descriptions, environmental DNA, vegetation indices, agricultural, forest, land indicators, and high-resolution climate variables. All observations are geospatially aligned under the WGS84 geodetic system, spanning from 2000 to 2020. The dataset is available at https://huggingface.co/datasets/BioDT/BioCube, the acquisition and processing code base at https://github.com/BioDT/bfm-data.

***Index Terms*—** Dataset, Multimodal, Engineering, Machine Learning, Biodiversity

1. INTRODUCTION

Biodiversity is undergoing rapid transformation due to human-induced environmental change, land-use shifts, and climate variability. Monitoring these changes at scale requires comprehensive datasets that not only capture singular modalities like species presence, but also contextual environmental information. However, most available biodiversity datasets are limited either to observational records or specific modalities such as imagery or genetic sequences, often lacking the necessary integration across environmental, spatial and temporal dimensions.

Recent advances in Digital Twins (DTs), Machine Learning (ML) and Earth Observation (EO) technologies have opened new avenues for ecological forecasting and biodiversity assessment. Yet, the full potential of these approaches is often hindered by challenges like fragmented data landscapes, inconsistent resolutions or modality gaps [1]. In response to these challenges we have engineered a multimodal dataset that provides a foundation for building scalable models that can be used for biodiversity monitoring, conservation planning and ecological forecasting at both global and local scales.

A series of ecology and biodiversity specialized datasets have recently emerged like BIOSCAN-5M [11] that contains over 5 million specimens of insects along with images, DNA barcode sequences, taxonomy, geographic information. Species distribution modeling is the focus of GeoLifeClef [4] dataset by merging 1.9 million plant and animal observations with high resolution remote sensing imagery, land cover and climate variables. In a similar direction, GeoPlant [7] provides over 5 million plant occurrence records across Europe, positivity enriched with Sentinel-2 satellite imagery and 20 years of climate time-series to support high-resolution spatial biodiversity observations. However, these datasets are far from containing enough diversified parameters to cover current needs in ecology. More specific, none of these datasets jointly integrates images, audio, eDNA, land, agriculture, conservation status, and climate variables. This gap motivated the construction of BioCube as a more diversified and holistic dataset.

The rest of the paper is organized as follows. Section 2 involves the methodology used, including data acquisition, preprocessing, and integration. In Section 3 the resulting data, its coverage, composition of modalities, and quality of the data are described. In Section 4, strengths, limitations, and open challenges regarding construction of large-scale biodiversity datasets that can be applied in ML are addressed. Finally, the prospect of the importance of BioCube to biodiversity research and ecological forecasting is discussed in Section 5.

2. METHOD

Latest ML methods like Foundation Models require large, well-curated, modality-rich datasets [2]. Accordingly, we assembled data from diverse sources, combining climate vari-

*Work performed during internship at TNO

doi:10.2760/2119408

ables, species observations, land indicators, and conservation records, as listed in Table 1. Acquisition used both API and file-based ingests to ensure scalability and spatiotemporal coverage.

2.1. Data Sources

BioCube dataset integrates data from several sources:

- **Climate Variables:** ERA5 hourly global reanalysis data, such as temperature, wind, pressure, and humidity, both in surface and atmospheric layers, obtained from the Copernicus Climate Data Store (CDS) [6].
- **Species Observations:** Images and metadata, such as taxonomy, geolocation, and timestamps collected from iNaturalist [1] and iNat2021 [5]. These datasets provide direct presence evidence.
- **Acoustic Data:** Bird vocalizations and metadata retrieved from Xeno-Canto [2] and from Xeno-Canto in GBIF [10], crucial for species monitoring where visual observation is difficult.
- **Species Descriptions and Conservation Status:** Textual records describing habitats, traits from Map of Life [9] and IUCN Red List [3], including red list index values and threat categories. The Red List is a global reference for extinction risk, with categories ranging from Extinct (EX) to Least Concern (LC).
- **Species Distribution:** Data is derived from the Living Planet Index [4], which aggregates population trends of species globally.
- **Environmental DNA (eDNA):** Genetic barcode sequences obtained from the Barcode of Life Data System (BOLD) [8].
- **Land and Vegetation Indicators:** NDVI sourced from Copernicus Land Services [5], and forest cover, land and agricultural indicators from The World Bank [6].

2.2. Acquisition Methods

API-based acquisition had a focus on the dynamic and real-time data retrieval. ERA5 climate variables were obtained through CDS using bounding boxes and temporal filters, with batch processing. Species data, including images, taxonomy, and geolocation, were collected via the iNaturalist API, while bird vocalizations were retrieved from the Xeno-Canto API, based on quality and location filters. Environmental DNA (eDNA) was sourced from the BOLD Systems API, and species descriptions together with the threat categories (e.g., IUCN Red List status) were accessed using the Map of Life API. To ensure efficiency and data integrity, we have implemented independent API modules to promote scalability and flexibility.

File-based acquisition provided an access to static and historical datasets, adding essential temporal depth and spatial coverage. The Living Planet Index (LPI) has contributed annual species distribution data from 1950 to 2020. NDVI products from Copernicus Land Services supplied vegetation indices recorded every 10 days at 1 km resolution, resampled to 0.25° grids for consistency. Land-use indicators, including arable land, irrigated areas, cropland extent, and forest cover, were sourced from the World Bank for the years 1961 to 2021. Offline datasets such as iNat2021 (2.7 million labeled images) and archived Xeno-Canto audio recordings accessed via GBIF further enriched the dataset. The complete file sizes and metadata can be found in Table 2.

2.3. Preprocessing

The obtained data could not be used in its raw format. Specific preprocessing steps detailed below needed to ensure consistency, quality and compatibility across modalities, while serving as a foundational component in the construction of structured data **batches** or **cubes** used for downstream modelling tasks. The preprocessing methods were performed during the dataset preparation phase and are crucial for generating uniform and high-quality inputs, harmonised to a 0.25° WGS84 geodesic coordinate grid, and temporally aligned to daily or monthly intervals.

- **Audio:** Silence removal, noise reduction (spectral gating), resampling, MFCC extraction.
- **Image:** Denoising, resizing, cropping.
- **Text:** Stopword and punctuation removal, stemming, lemmatisation, BERT embeddings and bag-of-words transformation.
- **eDNA:** Sequence filtering, k-mer vectorization and normalization.
- **Climate and Land Data:** Missing data interpolation, normalisation and temporal aggregation of variables such as temperature, wind, and pressure.

While end-to-end neural networks now dominate modern image, audio, and text analysis, we deliberately included traditional feature extraction methods (e.g., MFCC, TF-IDF) to ensure reproducibility and to provide baselines for researchers employing classical ML methods.

3. RESULTS

To construct unified dataset for biodiversity, we have acquired and curated multimodal data from multiple sources as a first step. Then the collected data available at Table 3 has been integrated into a structured species dataset with the following fields:

[1] https://www.inaturalist.org
[2] https://xeno-canto.org
[3] https://www.iucnredlist.org
[4] https://www.livingplanetindex.org/
[5] https://land.copernicus.eu/en/products/vegetation/normalised-difference-vegetation-index-v3-0-1km
[6] https://data.worldbank.org/indicator

doi:10.2760/2119408

Table 1. Overview of data modalities and variables included in the dataset.

Modality	Source	Variables
Surface Climate	Copernicus (ERA5)	2m temperature, 10m wind (u/v), mean sea-level pressure
Atmospheric Variables	Copernicus (ERA5)	Geopotential, temperature, humidity, wind (13 pressure levels: 50–1000 hPa)
Single-Level Variables	Copernicus (ERA5)	Land-sea mask, surface geopotential, soil type
Species Observations	iNaturalist, GBIF, Xeno-Canto	Images, audio, coordinates, timestamp, taxonomy
Descriptions	Map of Life	Text descriptions (behavior, habitat)
eDNA	BOLD Systems	DNA sequences (ATCG), taxonomic identifiers
Distribution Trends	Living Planet Index	Annual species occurrence and population trends (1950–2020)
Red List Index (RLI)	IUCN / Map of Life	Extinction risk index (0–1), categories: EX, EW, CR, EN, VU, NT, LC
NDVI	Copernicus Land (SPOT, PROBA-V)	Vegetation index values (-1 to 1), 10-day temporal resolution, 1 km spatial resolution
Agri/Forest Indicators	World Bank	Arable land, irrigated land, cropland area, forest cover, total land area

Table 2. Data Sources by File Count and Total Size

Data Source Name	Total Files	Total Size (GB)
Climate Variables	24,510	160
Species Observations	51,918	52
Acoustic Data	43,511	104.4
Species Descriptions	20,593	0.005
Environmental DNA	16,257	0.1
Species Destribution	4,922	0.03
Land Indicators	7	0.0001
Species Conversion Status	1	0.011
Vegetation Indicators	258	88

- Species Identification: Species, Phylum, Class, Order, Family, Genus
- Location and Time: Latitude, Longitude, Timestamp
- Multimodal Inputs: Image, Audio, eDNA, Description, Redlist, Distribution

Table 3. Statistics of the Species Folder Contents

Category	Count
Total number of Species	40,282
Species with eDNA, no images, no audios	15,064
Species with images, no audio, no eDNA	16,630
Species with images and audio, no eDNA	1,849
Species with audio, no images, no eDNA	2,772
Species with images and eDNA, no audio	738
Species with audio and eDNA, no images	182
Species with all modalities	273

These records are extracted from over 40,000 species folders, each containing varying combinations of modalities. To efficiently extract relevant data, we implemented a folder filtering mechanism based on a hash-table-inspired approach. Each folder is being treated as a unique bucket, and its internal CSV files (image, audio, eDNA, etc.) are scanned for timestamps. Only folders containing at least one timestamp within the target date range (2000–2020) are selected for further processing. This has minimised memory usage and accelerated BioCube's construction time by avoiding unnecessary I/O on irrelevant folders.

Image-Audio Matching: When both images and audio were available, we matched them by averaging their metadata; latitude, longitude, timestamp, and paired them to maximise spatiotemporal alignment. Additionally, species-level data such as taxonomy and distribution are matched using the closest year and location for each sample.

Efficient Storage: All data is stored in Apache Parquet format to optimize I/O operations. Tensors are serialised (as base64-encoded arrays), and each sample is assigned a `unique_id` to avoid duplication. Latitude and longitude values are rounded to a 0.25-degree resolution to align with other datasets such as climate and land-use data. If different coordinates systems were found, we transformed them to WGS84 format. The data are saved incrementally after processing each folder, enabling scalable and resilient processing.

Land indicators (agriculture, forest cover, NDVI) required additional preprocessing. Because several sources report only country-level values, we extracted country bounding boxes and interpolated these to a spatial grid to align with species-level data. NDVI at 1 km monthly resolution was harmonised separately from annual forest and agriculture statistics obtained from the World Bank and Copernicus services. Table 4 summarises the total values only for Europe.

doi:10.2760/2119408

Table 4. Summary of Environmental Indicators (Europe)

Indicator Type	Total Values	Countries
Agricultural (Arable)	2,311,390	42
Agricultural (Irrigated)	411,482	33
Cropland Area	2,276,021	38
Forest Cover	1,285,834	44
Land Area	852,248	44
NDVI (Vegetation Index)*	15,929,016	44

* NDVI values are recorded **monthly**, while all other indicators are reported **annually**.

4. CONCLUSION AND DISCUSSION

BioCube marks a big step towards progressing biodiversity research through its complete multimodal analysis of fine-grained environmental and ecological data across the globe. BioCube connects species observations including imagery, audio recordings, environmental DNA data as well as descriptive information with precise climate, land-use data, vegetation measurements and conservation metrics to fill a significant research gap between singular modality datasets. Its main strengths lie in the breadth of data types, global geospatial alignment, and open availability. These enable research in species monitoring, conservation planning, and ecological forecasting. Still, some limitations remain, like NaN and missing values, taxonomic and geographic biases, reliance on legacy feature extraction, and as well as incomplete modal or spatial representation. Overall, BioCube is a scalable step toward developing biodiversity foundation models [2], supporting hybrid experiments now, while future work should expand modalities, improve coverage, and add real-time data streams.

5. ACKNOWLEDGMENTS

This study has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057437 (BioDT project, https://doi.org/10.3030/101057437). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-10148.

REFERENCES

[1] A. Trantas, R. Plug, P. Pileggi, and E. Lazovik, "Digital twin challenges in biodiversity modelling.", Ecological Informatics, vol. 78, 2023, doi: https://doi.org/10.1016/j.ecoinf.2023.102357

[2] A. Trantas, M. Mensio, S. Stasinos, S. Gribincea, T. Khan, D. Podareanu, and A. van der Veen, "BioAnalyst: A Foundation Model for Biodiversity," arXiv preprint arXiv:2507.09080, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2507.09080

[3] C.-H. Yang, B. Feuer, Z. Jubery, Z. K. Deng, A. Nakkab, M. Z. Hasan, S. Chiranjeevi, K. Marshall, N. Baishnab, A. K. Singh, A. Singh, S. Sarkar, N. Merchant, C. Hegde, and B. Ganapathysubramanian, "Arboretum: A Large Multimodal Dataset Enabling AI for Biodiversity," arXiv preprint arXiv:2406.17720, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.17720

[4] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, and A. Joly, "The GeoLifeCLEF 2020 Dataset," arXiv preprint arXiv:2004.04192, 2020. [Online]. Available: https://arxiv.org/abs/2004.04192

[5] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha, "Benchmarking Representation Learning for Natural World Image Collections," arXiv preprint arXiv:2103.16483, 2021. [Online]. Available: https://arxiv.org/abs/2103.16483

[6] H. Hersbach et al., "ERA5 hourly data on single levels from 1940 to present," Copernicus Climate Data Store (CDS), 2023. [Online]. Available: https://cds.climate.copernicus.eu

[7] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, P. Bonnet, and A. Joly, "GeoPlant: Spatial Plant Species Prediction Dataset," arXiv preprint arXiv:2408.13928, 2024. [Online]. Available: https://arxiv.org/abs/2408.13928

[8] S. Ratnasingham and P. D. N. Hebert, "BOLD: The Barcode of Life Data System," *Molecular Ecology Notes*, vol. 7, no. 3, pp. 355–364, 2007. [Online]. Available: https://doi.org/10.1111/j.1471-8286.2007.01678.x

[9] W. Jetz, J. M. McPherson, and R. P. Guralnick, "Integrating biodiversity distribution knowledge: Toward a global map of life," *Trends in Ecology and Evolution*, vol. 27, no. 3, pp. 151–159, 2012. [Online]. Available: https://doi.org/10.1016/j.tree.2011.09.007

[10] W.-P. Vellinga, "Xeno-canto – Bird sounds from around the world," Xeno-canto Foundation for Nature Sounds, 2024. Occurrence dataset accessed via GBIF.org on 2024-09-27. [Online]. Available: https://doi.org/10.15468/qv0ksn

[11] Z. Gharaee, S. C. Lowe, Z. Gong, P. M. Arias, N. Pellegrino, A. T. Wang, J. B. Haurum, I. Zarubiieva, L. Kari, D. Steinke, G. W. Taylor, P. Fieguth, and A. X. Chang, "BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity," in *Advances in Neural Information Processing Systems*, vol. 37, pp. 36285–36313, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.12723

CONSTELLR HIVE SATELLITE MISSION: LEVERAGING BIG DATA AND DATA CUBE TECHNOLOGIES FOR THERMAL REMOTE SENSING AND ENHANCED DATA ACCESS

Daniel Spengler[1], Tobias Leismann[1],

constellr GmbH, Germany

ABSTRACT

The constellr HiVE satellite mission significantly enhances environmental monitoring through innovative thermal remote sensing capabilities. By providing global daily data at high spatial resolutions (5 to 30m), the mission effectively addresses limitations in existing satellite technologies. This paper discusses the mission specifications and the advanced data handling methodologies employed, focusing specifically on the implementation of big data architectures and data cube technology. Data cubes streamline complex spatial-temporal queries and enhance data accessibility for diverse stakeholders, including urban planners, agricultural managers, and policymakers, enabling targeted environmental interventions.

Index Terms — Big Data, Data Cube Technology, Thermal Remote Sensing, Land Surface Temperature, Environmental Monitoring

1. INTRODUCTION

The increasing frequency and intensity of climate-induced phenomena such as urban heat islands, drought, and agricultural stress highlight the necessity for timely and detailed thermal data. Current satellite technologies often compromise between spatial detail and temporal frequency, limiting their practical effectiveness. The constellr HiVE satellite constellation addresses these gaps through high-resolution thermal imaging and frequent revisits, supported by advanced big data and data cube technologies, dramatically improving environmental data handling and access.

2. CONSTELLR HIVE MISSION OVERVIEW

2.1. Satellite and Sensor Specifications

The HiVE constellation employs sophisticated thermal infrared (TIR) and visible near-infrared (VNIR) sensors. The payload for the constellation includes two primary imaging instruments: a Thermal Infrared (TIR) sensor and a Visible and Near-Infrared (VNIR) sensor. The TIR sensor, specifically designed by OHB for constellr, operates within the spectral range of 8-12 µm, capturing data in four discrete spectral bands (8.6 µm, 9.2 µm, 10.6 µm, and 11.75 µm). It features a Mercury Cadmium Telluride (MCT) detector array of 640 × 512 pixels, which is cryogenically cooled to enhance its sensitivity. The optical system of the TIR telescope comprises six refractive lenses, enabling high-precision Earth imaging from a sun-synchronous orbit at approximately 540 km altitude. For detailed technical specifications, refer to Table 1.

Accompanying the TIR sensor, the VNIR camera utilizes a commercially sourced Simerasense Multiscape100 system. This camera captures images across 10 spectral bands ranging from 400 nm to 1000 nm, corresponding closely to Sentinel-2 spectral bands to ensure interoperability. The ground sampling distance (GSD) varies between 10 m and 60 m depending on the spectral band and binning configuration, allowing comprehensive surface characterization.

The HiVE SkyBee 1 satellite was developed with ESA support under the InCubed program and was successfully launched in January 2025. Operated by constellr, SkyBee 1 enables targeted Earth surface imaging, featuring an 18.5 km swath width and nadir pointing capability of up to ±30°. Its local time of descending node (LTDN) is 10:30 AM. Figure 1 illustrates a data acquisition by SkyBee 1 conducted on May 19th 2025, demonstrating uncalibrated relative radiance differences highlighting cooler (blue) and warmer (red) areas. It shows a coastal area in Queensland, Australia, demonstrating the high resolution potential of the constellation. SkyBee 2 is scheduled for launch in June 2025, with additional satellites planned for yearly launches thereafter.

The accuracy of the Level 2 (L2) Land Surface Temperature (LST) product depends significantly on atmospheric conditions during image capture. Given a TIR instrument radiometric error of ≤2%, the absolute temperature accuracy typically ranges from 1.2K in dry arctic conditions to 2.2K in humid tropical conditions. The constellr LST algorithm generates high-resolution, surface-optimized temperature data by leveraging prior emissivity estimation, advanced atmospheric correction via real-time MODTRAN

doi:10.2760/2119408

simulations, and a flexible multi-band Equivalent Temperature approach. Constellr employs a sophisticated U-Net deep convolutional neural network for cloud masking, trained extensively on the Cloudsen12 dataset (Aybar et al., 2022) and enhanced with additional samples from challenging regions. The operational temperature range for the instruments is from -20°C to +80°C (253K to 353K). [1]

Table 1. Technical Specification of HiVE SkyBee1.

Mission parameter	VNIR	SWIR
Spectral Bands (µm)	0.44 | 0.49 | 0.56 | 0.67 | 0.71 0.74 | 0.78 | 0.84 | 0.87 | 0.95	8.6 | 9.2 | 10.6 | 11.8
LST uncertainty		<2K for mid latitudes
LST sensitivity		0.03K
Geolocation accuracy (m) wrt to Sentinel-2 reference image	<10m	1.5 pixels (42m) (CE90), for <10° off nadir angle
Band to band registration accuracy	0.2 pixels (1sigma)	0.2 pixels (1sigma)
Ground instantaneous field-of-view (m x m)	10 x 10 (Band 2/3/4/8) 20 x 20 (Band 5/6/7/9) 60 x60 (Band 1/10)	30 x 30
Signal-to-noise ratio (SNR)	10m: >1:200 | 20m: >1:170 60m: >1:500	B01: 1:530 | B02: 1:450 B03: 1:540 | B04: 1:150
Swath width / length / day (km)	20 / 1000	18.5 / 1000
Product Level for users	L2A	L2A
Metadata and Data Format	Cloud Optimized GeoTiff | STAC	Cloud Optimized GeoTiff | STAC
Data Access	API or GUI Web Platform	API or GUI Web Platform

2.2. Operations

Operating in a sun-synchronous orbit at approximately 550 km altitude, the HiVE constellation achieves daily global coverage. Skybee-1 launched on January 14th 2025 and Skybee-2 launches in June 2025., with the full constellation daily revisit operational by 2026, ensuring unprecedented temporal and spatial resolution in thermal remote sensing.

Fig. 1: show one the first images acquired by SkyBee 1 and showing the sharpness and extreme high level of detail and sensitivity of the satellite system.

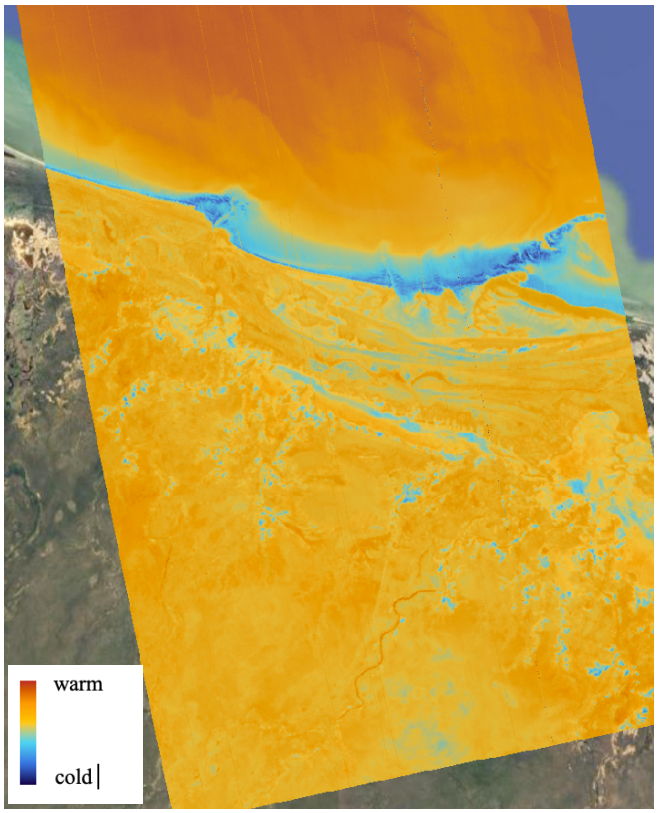


Fig. 1. HiVE SkyBee 1, LWIR thermal image, Queensland, Australia, 19.05.2025

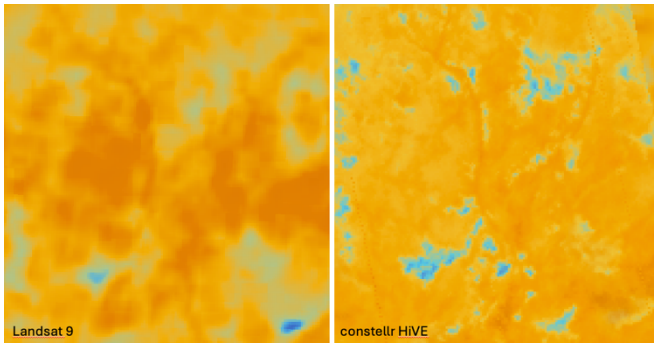


Fig. 2. Comparison of Landsat ST (20.05.2025) vs uncalibrated constellr HiVE (19.05.2025) data

doi:10.2760/2119408

3. ADVANCED DATA HANDLING INFRASTRUCTURE

3.1. Big Data Architectures

Given the massive volume and velocity of data generated by the HiVE satellites, a robust big data architecture is essential. This infrastructure uses cloud-optimized GeoTIFF and SpatioTemporal Asset Catalogs (STAC) to facilitate efficient data storage, retrieval, and integration, significantly streamlining workflows for researchers and policymakers. The use of Pangeo Stack provides a powerful and scalable environment for efficiently handling and analyzing HiVE satellite data, which often involves massive datasets with high spatial, spectral, and temporal resolution. Leveraging cloud-optimized tools within the Pangeo ecosystem, such as Xarray for multidimensional data handling, Dask for parallel and distributed computation, and Zarr for scalable storage. Users can significantly enhance their ability to perform complex analyses on HiVE datasets. The stack supports streamlined workflows for data ingestion, exploration, processing, and visualization, enabling users to rapidly interact with large volumes of satellite data without downloading entire datasets locally. By facilitating on-demand access and processing directly in the cloud, Pangeo reduces computational overhead and accelerates scientific insights derived from HiVE satellite missions.

3.2. Utilization of Data Cube Technologies

We are currently developing a multidimensional data cube by leveraging the robust capabilities of the Pangeo Stack, thereby creating an advanced data management and analytical environment tailored specifically for HiVE satellite datasets. Unlike working solely with Cloud Optimized GeoTIFFs (COGs), which are typically two-dimensional raster files, integrating a data cube framework allows us to seamlessly organize, query, and analyze satellite data across multiple dimensions, such as spatial coordinates, spectral bands, and temporal intervals simultaneously. The primary advantage of this approach is the ability to efficiently perform complex queries, rapidly extract insights from time-series analyses, and conduct large-scale parallel processing, significantly enhancing accessibility, scalability, and computational performance compared to handling individual COG files separately.

.

4. SPECIFIC USE CASES OF DATA CUBES WITH CONSTELLR DATA

4.1. Urban Heat Island Analysis

Most cities world-wide are affected by urban heat and are setting up strategies for climate adaptation and mitigations measures. At the same time, urban digital twins are becoming more and more common. This allows urban planners, real estate developers and decision makers to use thermal intelligence data cubes to analyze historical and real-time thermal data, identifying patterns and anomalies at neighborhood scales. The ability to query and visualize thermal data across different time frames assists in proactively implementing heat mitigation strategies, optimizing the placement of green, blue and grey cooling infrastructures, and enhancing urban resilience. Preliminary data from constellr's SkyBee-1 satellite reveals distinct thermal signatures associated with various urban structures and materials. For example, industrial areas and solar installations exhibit specific thermal characteristics, whereas water bodies function as natural moderating elements. Utilizing a data cube approach provides comprehensive, multi-temporal analyses of urban heat dynamics, essential for informed urban planning. This method enables city authorities to strategically position cooling infrastructures, green spaces, and reflective surfaces, effectively mitigating urban heat islands [2]. Furthermore, with the availability of constantly updated time series of thermal data, development projects can be monitored in near real time as they progress. This enables a comparable benchmarking of the thermal impact of any development as well as a clear measurement of the positive effect on heat resilience of any mitigation project. As urban energy demands for cooling are projected to rise substantially by 2050 [3], constellr's data supports proactive and sustainable urban designs.

4.2. Precision Agriculture and Water Efficiency

The high-resolution thermal data provided by constellr significantly enhances the capability to detect crop water stress by analyzing variations in land surface temperature (LST). Thermal imagery helps farmers precisely locate fields requiring targeted irrigation, optimizing water usage efficiency. Integrating these datasets into a data cube framework facilitates the systematic management and analysis of temporal and spatial trends, allowing agricultural practitioners to transition seamlessly from traditional resource-intensive methods to precision agriculture. This technique greatly mitigates drought risks, improves crop yield stability, and strengthens food security [4].

4.3. Public Health Monitoring

In the context of public health, data cubes facilitate monitoring of thermal comfort and health risks related to urban heat. Public health authorities can rapidly access relevant thermal data, identify vulnerable communities and effectively deploy targeted interventions such as cooling centers and green spaces. The escalating intensity of urban heat islands poses significant risks to public health, particularly during severe heatwaves. Constellr's thermal imaging capabilities offer crucial insights for assessing heat-related health risks. Ballester(2023)[5] indicate that heightened urban temperatures can substantially increase mortality rates, potentially causing millions of heat-related

doi:10.2760/2119408

fatalities by the end of the century. Incorporating thermal data into a data cube structure enables efficient tracking and predictive modeling of heat stress, facilitating targeted interventions such as establishing cooling centers, shaded public spaces, and urban greening initiatives. This level of detailed thermal monitoring enhances community resilience and preparedness against heat-induced health crises.

4.4. Civil Security and Defense Applications

Constellr's TIR data plays a vital role in defense and security, enabling real-time, global monitoring of critical military and civilian infrastructure. Delivered in data cube format, TIR data allow for integrated analysis across temporal and spatial dimensions. This facilitates early detection of anomalies such as overheating in industrial assets, border incursions via unusual heat signatures, or hidden underground activity through surface thermal patterns. TIR sensing also supports resilient mission planning by identifying climatic stressors like drought, flooding, or water scarcity, which is essential for training and operations. Compared to optical data, thermal observations provide earlier insight into environmental stability, aiding logistics in foreign deployments. In conflict zones, TIR enables rapid damage assessment, tracks functional changes over time, and uncovers covert or asymmetric activities not visible to optical sensors. Additionally, it serves as a discreet tool for early warning of crises—such as persistent droughts or migration movements—and for monitoring sensitive regions where camouflage or darkness limit conventional imaging.

TIR is also a key enabler for civil protection and disaster resilience, providing critical insights for early warning, response, and recovery. Thermal data cubes enable continuous spatial and temporal analysis of extreme weather events, droughts, and supply shortages, forming the backbone of national and cross-border early warning systems. TIR sensing supports disaster response by identifying heatwave- and flood-prone zones, assessing post-disaster impacts on critical infrastructure, and mapping the availability of food and energy for humanitarian aid. It also offers potential for geophysical event forecasting—such as volcanic or seismic activity—through the detection of surface thermal anomalies. On a broader scale, TIR enables economic and environmental monitoring by revealing the activity status of key industrial facilities (e.g., refineries, steel plants) and tracking climate-induced risks like wildfires and persistent droughts.

5. FUTURE PERSPECTIVES AND INTEGRATION WITH OTHER MISSIONS

The constellr HiVE mission, through its robust big data architecture and innovative data cube technologies, represents a significant advancement in thermal remote sensing. By enhancing data accessibility and usability across multiple sectors, it directly contributes to addressing critical environmental challenges and promoting informed, resilient urban and agricultural practices. In future we will combine

6. CONCLUSIONS

The constellr HiVE mission, through its robust big data architecture and innovative data cube technologies, represents a significant advancement in thermal remote sensing. By enhancing data accessibility and usability across multiple sectors, it directly contributes to addressing critical environmental challenges and promoting informed, resilient urban and agricultural practices.

REFERENCES

[1] D. Spengler, R. Benvenuto, B. Tempel, T. Leismann, C. Mittermaier, M. Bierdel, T. Menne, I.Tanouti, C. Welling and M. Gulde, "constellr High-precision Versatile Ecosphere (HiVE) Thermal Mission:Technical Overview and Application Insights", Remote Sensing Communications, 1, doi: https://doi.org/10.62880/rsc25001, 2025.
[2] Ouyang et al., Urban land surface temperature retrieval, Remote Sensing of Environment, 2024.
[3] IEA, Global Energy Review 2025.
[4] Khanal et al., Thermal remote sensing in precision agriculture, Computers and Electronics in Agriculture, 2017.
[5] Ballester et al., Heat-related mortality in Europe, Nature Medicine, 2023.

INTERPRETABLE SINGLE-LAYER REPRESENTATION OF MULTITEMPORAL VEGETATION CHANGE DYNAMICS FROM SENTINEL-2 TIME SERIES

Dirk Tiede[1], Thomas Strasser[1], Matthias Laher[2], Hannah Augustin[1], Steffen Reichel[2], Markus Kerschbaumer[2], Luke McQuade[1], Kristýna Měchurová[2], Andrea Baraldi[2] and Martin Sudmanns[1]

[1]Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

[2]Spatial Services GmbH, Salzburg, Austria

ABSTRACT

Big Earth Observation (EO) data, such as provided by the European Copernicus programme, are a great opportunity for highly frequent global monitoring of the environment. Challenges exist not only in processing big multitemporal data but also in communicating results in a meaningful and useful manner, especially for non-EO experts. Our approach uses big EO data analyses in a semantic EO data cube and communicates results using a single-layer multi-temporal representation, where colour represents different user-defined time periods and changes. The visualisation in colour-codes reduces terabytes of multi-temporal information into a single, comprehensive layer. While this approach is backed by established geovisualisation techniques, we extend it to unveil temporal processes and dynamics hidden in big EO data. The resulting layer can be used in a very simple way: It functions as an interpretable basemap, either integrated within GTIF-AT or accessed externally via WMS/STAC, to complement user or domain-specific data with a temporal perspective.

Index Terms— change detection, semantic querying, big EO data, multitemporal change indication

1. INTRODUCTION

Big Earth observation (EO) data, such as provided by the European Copernicus program, are a great opportunity for highly frequent global monitoring of the environment. Challenges exist not only in processing big multitemporal data *[1]* but also in communicating results in a meaningful and useful manner, especially for non-EO experts *[2]*. Related to establishing a Digital Twin for Austria, these publicly funded, open, and free Copernicus satellite data sets are invaluable for monitoring the environment. They also serve as valuable input to modelling approaches necessary for current global challenges, like climate change adaptation and the green transition towards an urgently needed sustainable society and its local impacts and mitigations in Austria.

ESA Green Transition Information Factories (GTIF) initiative, especially the GTIF-Austria demonstrator (GTIF-AT), wants to showcase domains and tools that support the green transition. Within the current GTIF-AT implementation many important topics for supporting and monitoring the green transition are made available to different users in an easy to grasp manner, but are often focused only on very specific domains, data sets and dates.

A key asset of the open and free Copernicus Sentinel-2 data is their temporal frequency, and GTIF-AT lacks a temporal component to a dynamic integrated view on the green transition. Generic temporal vegetation change information derived from Sentinel-2 data could enrich existing, application-specific static information layers. There is Sentinel-2 coverage at least 5 days (higher for overlapping orbits) for Austria, offering considerable change information over time, which is missing and underused in GTIF-AT as a cross-domain layer because such an approach cannot be easily integrated in a classical non-dynamic web-GIS interface.

In this study, we incorporate a temporal basemap layer into GTIF-AT, which offers general information about vegetation changes. This layer is designed for seamless combination with most thematic datasets currently available within the demonstrator. Our methodology uses all Sentinel-2 observations from 2018 onward - the first year with full dual-satellite coverage - into a reproducible and interpretable format. The result is a comprehensive multitemporal representation for Austria, which is fully automated and has the potential for global application. By providing this change detection layer, we introduce an additional temporal component to existing GTIF-AT datasets, enabling the integration of temporal dynamics across a broad range of thematic applications.

The primary challenges were twofold: first, the development and implementation of innovative, scalable, and reproducible methodologies for managing and analysing large volumes of EO data; and second, the effective communication of the resulting multitemporal insights to support spatial decision-making processes, particularly for users in Austria and beyond.

2. METHODS

2.1. Big EO Data processing

The big EO data analyses behind the multi-temporal thematic layer are conducted in a semantic EO data cube [3], where for each observation at least one nominal (i.e. categorical) interpretation is available and can be queried in the same instance. Our implementation - Sen2Cube.at, see detailed description of the implementation in [4] - is a worldwide

doi:10.2760/2119408

unique semantic EO data cube implementation available for all of Austria, where every Sentinel-2 satellite image taken since 2015 and their derived categorical information layers can be analysed in the cloud. Data cubes have the advantage that the spatial and temporal extent to be analysed can be dynamically selected using meaningful coordinates, while the images are indexed in multiple dimensions. Semantic data cubes extend this flexibility with a semantic query option that allows analyses integrating categorical information with raw data directly in the selected spatial and temporal extent.

This approach uses semantic enrichment to calculate the pixel-based percentage of vegetation versus non-vegetation observations using all Sentinel-2 images in a user defined analysis period (e.g. years or seasons). Different to index-based approaches, e.g. using NDVI only, no thresholds need to be defined since the semantic classes (here: spectral categories, see [5]) also reflect cloud-like / bare-soil-like / vegetation- and water-like categories. In contrast to machine learning / deep learning approaches, such a knowledge-based semantic enrichment approach does not rely on localized training samples and is, therefore, worldwide applicable. Scaling the approach to larger areas needs less energy consumption, which facilitates its proven transferability to all Sentinel-2 data worldwide within the ESA inCubed project SIAMaaS [6] (see also https://app.color33.io). All available imagery can be used without additional pre-processing to filter cloud-affected data. This approach allows the use of smaller cloud-free areas even in highly cloudy images, increasing the number of valid, clear observations and thereby enhancing statistical reliability. The approach can be scaled up to any region worldwide since the semantic enrichment approach does not require re-training or adaptations for other regions.

2.2. Visualisation approach

The visualisation of analytical results derived from remote sensing data plays a significant role in the communication of this information. Although effective geovisualisation is not a new area of research [7], it takes on added significance within the Copernicus Programme. This is due to the high volume and temporal resolution of data - such as that from Sentinel-2 - which makes the purposeful visualization and communication of the temporal dimension especially important alongside large-scale Earth observation data analysis workflows. The initial single-layer representation of different time steps is based on the RGB colour model and was developed together with users from different domains. It's a simple additive colour model, used to visualise the 3 different grayscale layers for each time period, each indicating the proportion of vegetation observed. The approach allows changes from three periods to be displayed on a map in a single image using different colour combinations.

Such multitemporal colour compositing is not new in remote sensing especially for visual interpretation of changes [8] [9], but we further developed this approach into a transferable technique on semantic vegetation counts for fixed time frames which can be combined individually and are afterwards still interpretable. The interpretation of the colours can be drawn from the proposed colour cube (see Fig. 2), which applies a fixed layer sequence - from oldest to newest (see Fig. 1) - to ensure consistent meaning across combinations. Since the yearly vegetation percentages are counts of cloud/snow free vegetation observations, the RGB composite is still interpretable on a yearly basis and slight colour changes can be still linked to percentage changes in vegetation. The RGB colour palette and colour cube for the interpretation does therefore not only communicate vegetation change but also changes in intensity and/or partly changed vegetation to non-vegetation and vice versa using main RGB colours and mixed colours plus their intensity.

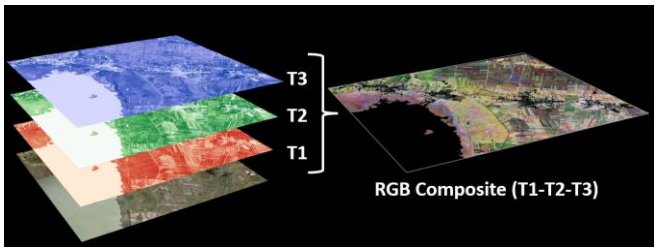


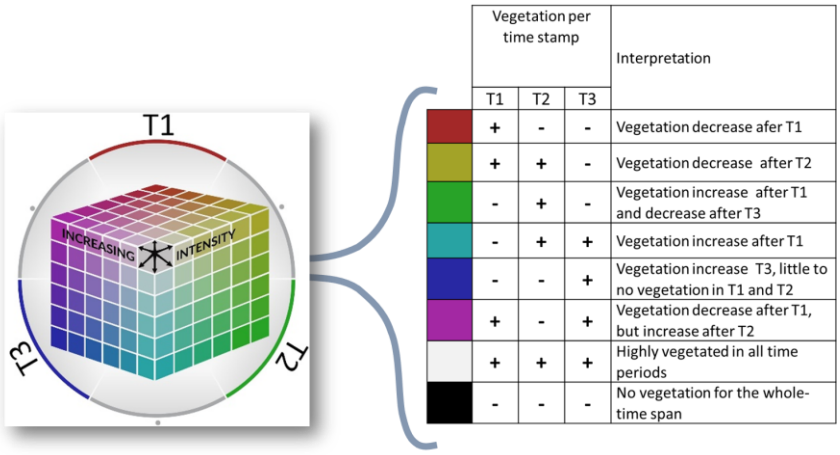
Fig. 1. An RGB colour composite created from vegetation data collected over different years. To enhance interpretation, we follow a fixed band assignment: the oldest timestamp to the red channel, the middle timestamp to the green, and the most recent timestamp to the blue channel.

IMPLEMENTATION

For the implementation a semantic query model has been developed within the semantic EO data cube and be applied to all Sentinel-2 data within Austria from 2018-2024 on a yearly basis. The first implemented layer in GTIF-AT is a long-term change combination 2018-2021-2024 (https://gtif-austria.info/narratives/vegetation-change-dynamics).

The period matches between the updating cycles of the European Copernicus Land Cover products, such as the CORINE land cover layers 2018 to 2024 (the latter is expected to be published early 2026). Eventually the single years can be combined as needed from a user perspective, e.g. Fig. 3 shows a combination of 2019-2020-2021 for evaluating specific events in a narrower time frame in the federal state of Salzburg. Fig. 4 shows a result for parts of the city of Salzburg overlaid with Urban Atlas 2018 data to stratify the result by different land use / land cover for different purposes.

doi:10.2760/2119408



Vegetation per time stamp			Interpretation
T1	T2	T3	
+	-	-	Vegetation decrease afer T1
+	+	-	Vegetation decrease after T2
-	+	-	Vegetation increase after T1 and decrease after T3
-	+	+	Vegetation increase after T1
-	-	+	Vegetation increase T3, little to no vegetation in T1 and T2
+	-	+	Vegetation decrease after T1, but increase after T2
+	+	+	Highly vegetated in all time periods
-	-	-	No vegetation for the whole-time span

Fig. 2. Proposed RGB dynamic visualisation cube for the interpretation of our RGB timely composite of different vegetation layers to directly derive and interpret change occurrence, time and duration of change, and severeness (intensity). Intense colours represent strong change, while subtle changes will be indicated by pastel colours.

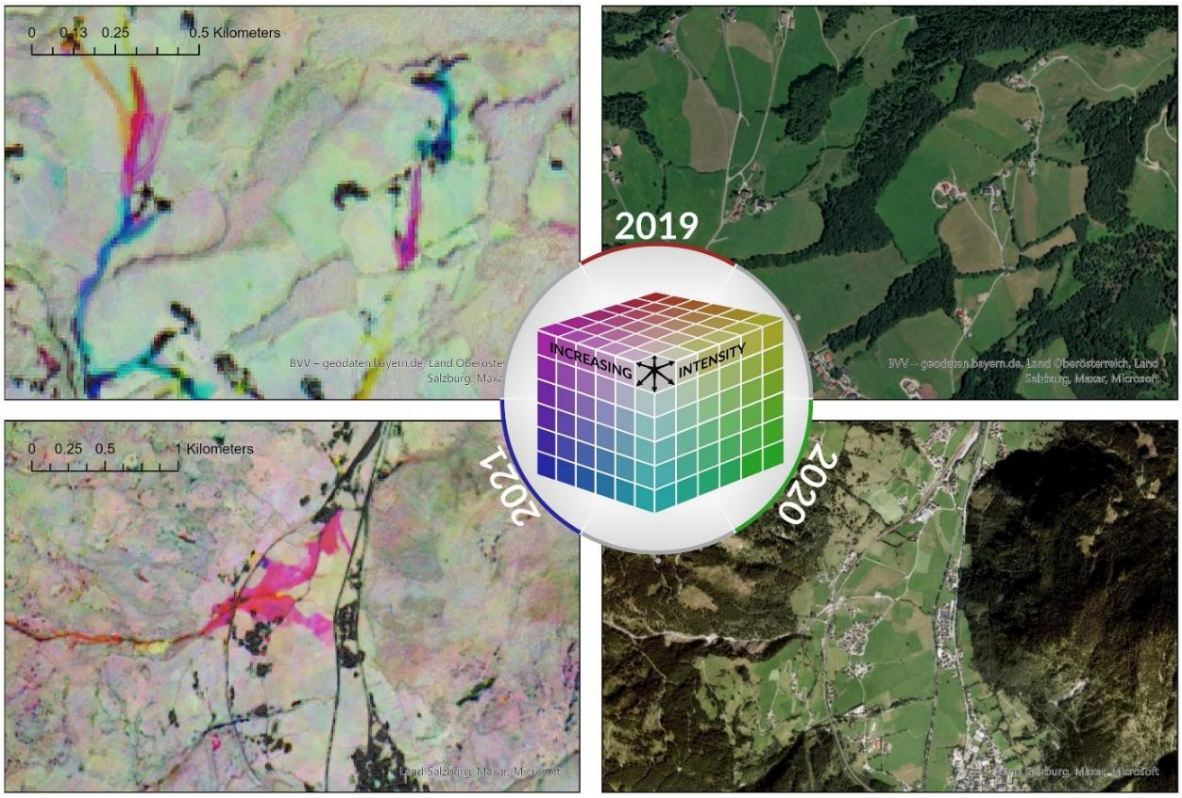


Fig. 3. Example for a one-layer representation of the changes of observed vegetation counted from every Sentinel-2 image in the years 2019, 2020 and 2021 (can be adapted to any time period (e.g. different years or seasons)). Upper left: RGB layer representing changes in road construction based on vegetation change derived from all Sentinel-2 images, the colours represent the years when the changes occurred (removal of vegetation during construction, but also vegetation regrowth of parts of the area when the roads were finished). Upper right: VHR image of the same area taken after the changes happened (>2022). Lower left: RGB representation for a mudflow taken place in Bad Hofgastein, Austria, early July 2020. Since the vegetation was removed by the mudflow the colour changes to red (not vegetated parts of 2020 and 2021), for some parts to magenta, which indicates a regrowth of vegetation already in 2021. Lower right: VHR image of the same area taken after the changes happened (>2022)

doi:10.2760/2119408

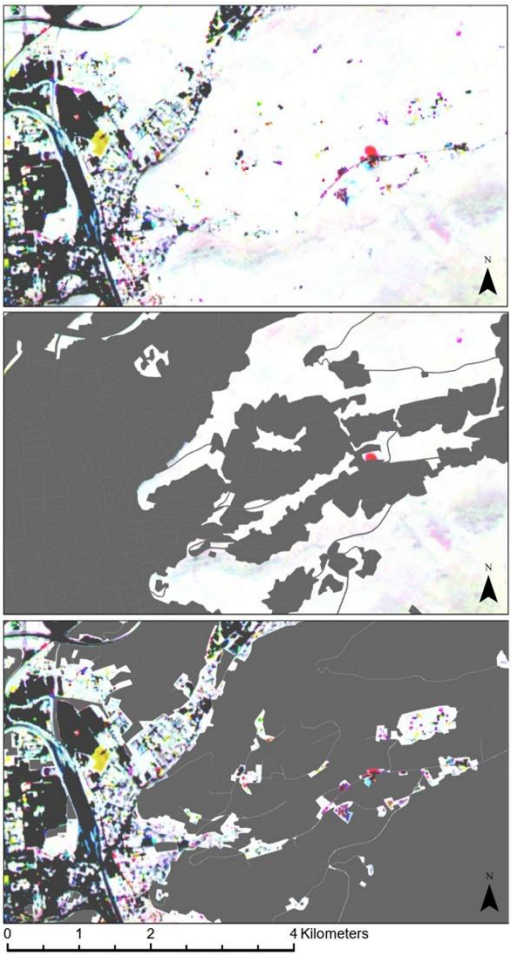


FIG. 4. Change layer combined with different land use / land cover masks taken from the Urban Atlas data (EEA, 2018) to stratify changes for different analysis purposes. Top: RGB change layer (2018–2021–2024) for a part of the city of Salzburg. Changes stratified to forest areas (middle) showing only small impact spots, combined urban fabric classes (bottom) highlighting urban densification / loss of urban vegetation.

RESULTS AND DISCUSSION

The single-layer representation of multitemporal vegetation changes is designed to improve the communication of multi-temporal analyses to a broad range of users - such as planning authorities, decision-makers, and non-EO experts. It is particularly effective for integrating change or monitoring layers with specific application topics within GTIF-AT, or for use as a WMS basemap that can be directly combined with users' sensitive internal data. This approach clearly highlights areas of change and provides insights into the intensity of those changes. Unlike conventional basemaps commonly used in GIS-based decision support systems, which typically rely on mono-temporal data like static maps or image mosaics with unclear observation dates, our method delivers time-sensitive and actionable information.

Potential application areas include, but are not limited to: comprehensive monitoring of green spaces and their dynamics across Austria, tracking land use changes related to energy production (e.g., construction of solar or wind power facilities), forest monitoring (e.g., identifying landslides, new roads, or changes in protection forests), environmental and soil protection (e.g., detecting soil sealing or agricultural deposits), and nature conservation efforts (e.g., observing vegetation shifts driven by climate change).

ACKNOWLEDGEMENTS

The research has received funding from the Austrian Research Promotion Agency (FFG) under the Austrian Space Application Programme (ASAP) within the project INTERFACE (contract no: FO999892652) and under the Digital Twin Austria Programme within the project GTIME (contract no: FO999918383).

REFERENCES

[1] M. Sudmanns *et al.*, 'Big Earth data: disruptive changes in Earth observation data management and analysis?', *Int. J. Digit. Earth*, vol. 13, no. 7, pp. 832–850, 2020. doi:10.1080/17538947.2019.1585976.

[2] G. Andrienko *et al.*, 'Geovisual analytics for spatial decision support: Setting the research agenda', *Int. J. Geogr. Inf. Sci.*, vol. 21, no. 8, pp. 839–857, 2007, doi: 10.1080/13658810701349011.

[3] H. Augustin, M. Sudmanns, D. Tiede, S. Lang, and A. Baraldi, 'Semantic Earth Observation Data Cubes', *Data*, vol. 4, no. 3, p. 102, 2019, doi: 10.3390/data4030102.

[4] M. Sudmanns, H. Augustin, L. van der Meer, A. Baraldi, and D. Tiede, 'The Austrian Semantic EO Data Cube Infrastructure', *Remote Sens.*, vol. 13, no. 23, p. 4807, 2021, doi: 10.3390/rs13234807.

[5] A. Baraldi, M. L. Humber, D. Tiede, and S. Lang, 'GEO-CEOS stage 4 validation of the Satellite Image Automatic Mapper lightweight computer program for ESA Earth observation level 2 product generation - Part 1: Theory', *Cogent Geosci.*, vol. 4, no. 1, p. 1467357, 2018, doi: 10.1080/23312041.2018.1467357.

[6] M. Sudmanns, M. Butenuth, S. Reichel, M. Kerschbaumer, A. Baraldi, and D. Tiede, 'color33 – a cloud-based service for automated semantic enrichment of optical satellite images', in *Proceedings of the 2025 conference on Big Data from Space (BiDS'25)*, Riga, Latvia, 2025.

[7] A. M. MacEachren and M.-J. Kraak, 'Research Challenges in Geovisualization', *Cartogr. Geogr. Inf. Sci.*, vol. 28, no. 1, pp. 3–12, 2001, doi: 10.1559/152304001782173970.

[8] D. Lu, Mausel ,P., Brondízio ,E., and E. Moran, 'Change detection techniques', *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004, doi: 10.1080/0143116031000139863.

[9] D. J. Hayes and S. A. Sader, 'Comparison of Change-Detection Techniques for Monitoring Tropical Forest Clearing and Vegetation Regrowth in a Time Series', *Photogramm. Eng. Remote Sens.*, vol. 67, no. 9, pp. 1067–1075, 2001.

FROM CLOUD TO CLIENT: WEB-NATIVE, IN-BROWSER EO DATACUBE EXPLORATION AND ANALYTICS USING ZARR

Nyi Nyi Nyan Lin, Martin Sudmanns, Dirk Tiede, and Hermann Klug

Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

ABSTRACT

Earth Observation (EO) analytics is experiencing a paradigm shift from cloud-centric processing toward hybrid architectures that leverage client-side computation. This research investigates the technical feasibility of fully web-native EO DataCube exploration and analytics using Zarr format, eliminating server-side dependencies for interactive analysis. Through an experimental prototype integrating Zarrita.js, OpenLayers DataTile extensions, and Web Worker-based analytics, we demonstrate sophisticated EO analysis—including Urban Heat Island mapping, spectral categorization, and zonal statistics—entirely within web browsers. Real-world validation through the SpongeCity Toolbox serving 120 municipalities shows 5.9-19 second analysis times for municipality-scale operations when cached (initial loading: 27-252 seconds), with browser cache achieving up to 43× speedup. The approach handles diverse data types (Float32/Uint8) and resolutions (10m/70m) while maintaining operational performance on both desktop and mobile devices. We developed and published the ol-zarr package for OpenLayers-based Zarr visualization [6], available as open-source software on GitHub and Zenodo. While analytical algorithms (UHI, zonal statistics) remain integrated within the SpongeCity dashboard, these implement standard algorithms demonstrating Zarr's browser-based feasibility rather than novel methods requiring separate distribution. This work advances FAIR principles in EO data access while addressing critical democratization barriers for educational institutions and resource-constrained environments where traditional cloud infrastructure is unavailable.

Index Terms— Earth Observation, Zarr, Cloud-native, Web-native analytics, Data democratization, SpongeCity

1. INTRODUCTION

The proliferation of Earth Observation data has transformed geospatial analytics, with cloud-native platforms enabling planetary-scale analysis [4, 1]. However, traditional approaches create persistent barriers through platform dependencies and infrastructure requirements that restrict access to specialized communities [3]. Recent advances in web technologies—particularly Zarrita.js for browser-based Zarr processing [13] and OpenLayers DataTile for dynamic raster rendering [10]—enable unprecedented browser-native scientific computing capabilities, with research confirming acceptable performance when properly optimized [11].

The Zarr format provides critical enabling technology through chunked access patterns aligned with browser constraints [9, 7], with NASA's adoption [8] and GeoZarr's OGC Community Standard approval demonstrating operational maturity. This research validates whether cloud-native storage formats can support fully autonomous, browser-based EO analytics without server-side computation through implementation in an operational web toolbox serving 120 municipalities.

2. METHODS

2.1. System Architecture Overview

We developed a general-purpose browser-based Earth Observation datacube system that enables direct client-side access and processing of Zarr-formatted EO data without server-side computational dependencies. The system architecture comprises two core components: (1) Interactive Exploration enabling direct visualization of datacube contents through spatial and temporal queries, and (2) In-browser Analytics providing client-side computational capabilities for EO data analysis. The technology stack integrates Zarrita.js (v0.5.0) for Zarr format handling, custom OpenLayers extensions for direct chunk visualization, Web Workers for parallel processing, and vanilla JavaScript optimization for numerical computations. This architecture eliminates traditional dependencies on tile servers, processing servers, or specialized software installations, requiring only a modern web browser for full functionality.

2.2. Zarr DataCube Architecture

2.2.1. DataCube Structure and Organization

Raster datasets are compiled into Zarr datacubes using a Python-based preprocessing pipeline leveraging Xarray for

This work was supported by the SpongeCity project within the Interreg Danube Transnational Programme.

doi:10.2760/2119408

multi-dimensional array handling, zarr-python for Zarr format writing, and Rasterio for geospatial raster processing. This pipeline is dataset-agnostic, capable of ingesting any georeferenced raster data regardless of source sensor or data type. The preprocessing generates multi-resolution pyramids and computes per-timestamp statistics, preparing data for efficient browser-based access. Our system operates on a hierarchical Zarr datacube structure designed for efficient multi-resolution access:

```
Dataset (zgroup)
+-- Zoom Levels (zgroup: 0, 1, 2, ..., n)
    |-- times (zarray)
    |-- statistics (zarray)
    |-- values (zarray: [time, band, y, x])
    |-- y (zarray)
    +-- x (zarray)
```

Each zoom level represents a pre-generated pyramid with 2× spatial resampling, enabling efficient multi-scale visualization. The datacube optimizes for interactive browser queries through specialized arrays: `times` (1D timestamps for temporal queries), `x`/`y` (coordinates for spatial indexing), `statistics` ([min, max, mean, 2nd/98th percentiles] per timestamp for dynamic visualization), and `values` ([time, band, y, x] with [1, 1, 256, 256] chunking). This chunking prioritizes spatial slice performance for tile-based rendering over pixel time series extraction—while full time series require multiple chunk fetches, point-based temporal analysis remains feasible as demonstrated in our occurrence analysis. Zarr datacubes are hosted on MinIO Object Storage with S3-compatible APIs, providing direct HTTP range requests with CORS configuration for browser access.

2.3. Interactive Exploration of EO DataCube

The interactive exploration component implements a direct Zarr-to-visualization pipeline through our custom ZarrTile class extending OpenLayers' DataTile functionality [6]. The tile loading process operates as follows: OpenLayers determines required tiles based on viewport extent and zoom level, ZarrTile computes spatial indices by intersecting tile extent with datacube's coordinate arrays, constructs multi-dimensional queries combining spatial, temporal, and band indices, and retrieves the corresponding Zarr chunks. Retrieved data undergoes processing in Web Workers for normalization and format conversion before returning to OpenLayers for rendering. We encapsulate the ZarrTile/DataTile combination within OpenLayers' WebGLTile layer, providing GPU-accelerated rendering and interactive visualization adjustments. OpenLayers' dual-layer caching—rendered tiles in memory and Zarr chunks in browser cache—persists across all zoom operations, eliminating re-fetching during navigation.

2.4. In-browser Analytics of EO DataCube

The analytics component implements comprehensive EO analysis capabilities entirely within the browser environment, processing data retrieved directly from Zarr datacubes without server-side computation. Each analysis operation requires four core inputs: analysis area(s) as polygons, dataset reference to specific Zarr datacube, spectral bands, and temporal range. Input polygon coordinates are reprojected to the dataset's CRS using Proj4.js, with geomask.js implementing the Dufour-Peyton intersection algorithm to convert polygons to gridded masks. The analysis engine constructs optimized queries by converting temporal ranges to array indices and mapping polygon extents to array indices using coordinate arrays. For multiple analysis areas, data retrieval occurs in parallel with Web Workers managing concurrent requests. Analysis computations utilize optimized vanilla JavaScript implementations with direct manipulation of typed arrays (Float32Array, Uint8Array) for memory efficiency. All analytical operations execute in dedicated Web Workers to maintain UI responsiveness through separated concerns allowing different stages to execute concurrently.

2.5. SpongeCity Toolbox Implementation

To validate our technical approach, we integrated the browser-based EO datacube system into the SpongeCity Toolbox (spongecity.zgis.at), serving 120 settlements across 12 Danube region countries. This implementation provides an ideal validation environment as primary users are non-technical stakeholders including municipal decision-makers and urban planners. We prepared three complementary datasets as Zarr datacubes: (1) ECOSTRESS Land Surface Temperature: NASA's thermal data at 70m resolution [5], (2) SIAM™ Land Surface Appearance Categories: 34-class categorical data at 10m resolution using the Satellite Image Automatic Mapper [2], and (3) SIAM™ Greenness Index: vegetation vigor indicator at 10m resolution [2]. Four analysis functions demonstrate diverse computational patterns: Urban Heat Island Analysis (temperature differentials between urban areas and rural buffers), Multi-temporal Occurrence Analysis (pixel counts meeting specified thresholds across time series), Multi-temporal Zonal Statistics (statistical summaries within user-defined polygons), and Multi-temporal Distribution Analysis (frequency distributions of categorical classes).

2.6. Performance Evaluation

We evaluated all four analysis functions across three datasets using Pécs, Hungary as our test case. Our instrumentation captures network metrics (HTTP timing, data transfer, cache hits) and processing metrics (execution time, memory usage), differentiating between data retrieval (network-bound) and processing (compute-bound) phases. Tests were conducted on iPad 8th Generation (Safari, 3GB RAM) and HP EliteBook

doi:10.2760/2119408

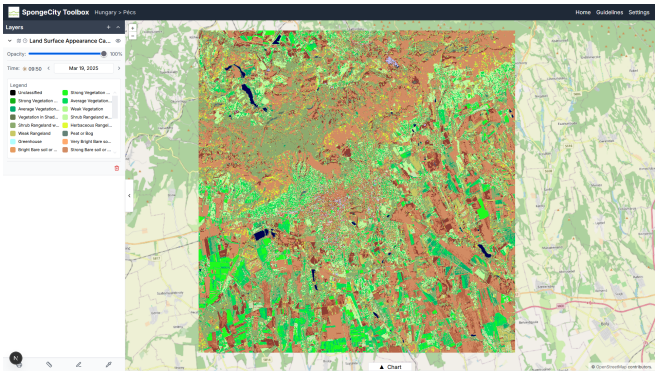


Fig. 1. SpongeCity Toolbox showing SIAM™ Land Surface Appearance Categories rendered directly from Zarr datacube storage without server-side tile generation.

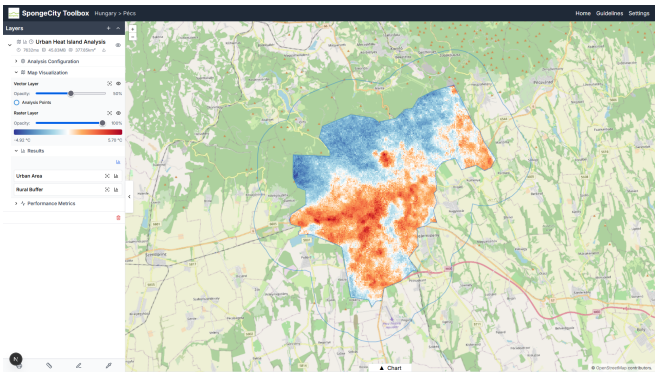


Fig. 2. Urban Heat Island analysis results rendered in the SpongeCity Toolbox, showing temperature differentials between urban areas and rural surroundings computed entirely within the browser from ECOSTRESS Zarr datacube.

840 G9 (Chrome, 32GB RAM) with 40/10 Mbps connectivity. The evaluation covered Pécs municipality (163 km²) for standard analyses and extended area (378 km²) for UHI analysis with 3km rural buffer, processing 21 ECOSTRESS thermal acquisitions (70m, Float32) and 27 SIAM™ acquisitions (10m, Uint8/Float32) from January-June 2024.

2.7. Quantitative Performance Results

Table 1 presents comprehensive performance metrics from the SpongeCity Toolbox implementation:

The evaluation reveals critical insights about browser-based EO analytics performance. Browser caching provides transformative performance improvements, achieving 2.3× to 43× speedup with the most dramatic benefits for high-frequency data access patterns. Data retrieval constitutes 87-99% of total execution time for uncached operations but drops significantly when cached, while processing times remain consistent (0.2-13.9 seconds) regardless of cache state. The 70m ECOSTRESS analysis achieves sub-3.5 second

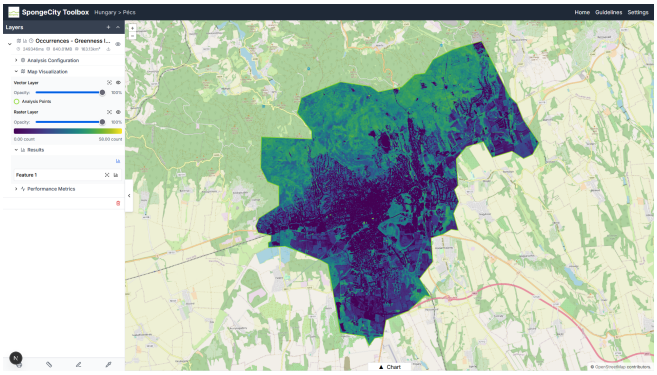


Fig. 3. Multi-temporal occurrence analysis results for SIAM™ Greenness Index displayed in the SpongeCity Toolbox, demonstrating client-side time series analysis and visualization capabilities.

performance even without caching, while 10m analyses require 27-252 seconds initially. Mobile devices demonstrate unexpected processing advantages in some cases, with both platforms maintaining stable performance requiring only 0.27-1.11 MB/s average bandwidth.

3. DISCUSSION AND CONCLUSIONS

This research demonstrates the technical feasibility of browser-native Earth Observation DataCube analytics using Zarr format, successfully implementing Urban Heat Island mapping, spectral categorization, and zonal statistics entirely within web browsers. Our work provides empirical evidence that Zarr's chunked architecture enables efficient client-side EO analytics across multiple data types and resolutions, directly addressing traditional barriers that limited EO data access to specialized communities. The browser-native approach enhances FAIR principles [12] through standardized web protocols and URL-shareable analytical configurations.

3.1. Performance and Limitations

Browser caching provides transformative performance improvements up to 43× speedup, enabling interactive experiences suitable for real-time stakeholder engagement. The system processes 3.3 million pixels with modest bandwidth requirements (0.27-1.11 MB/s average) and cross-platform compatibility. Initial loading requires 125-252 seconds for municipality-scale datasets (Pécs, Hungary: 163 km² at 10m resolution, 27 temporal acquisitions, 84.6 MB compressed), while browser memory limits restrict analysis to datasets under 400-500 MB uncompressed. Rather than competing with server-side computational throughput, our approach optimizes for zero-infrastructure deployment and immediate accessibility—enabling EO analysis where server access is unavailable or cost-prohibitive. Performance evaluation

doi:10.2760/2119408

Table 1. Performance metrics for browser-based EO analytics in SpongeCity Toolbox (Pécs, Hungary test case)

Analysis	Dataset	DType/Res	Time Steps	Area (km²)	Time (seconds)			Data (MB)		Bandwidth (MB/s)	
					Total	Retrieval	Process	Compressed	Uncompressed	Avg	Peak
iPad Performance - No Cache / Cached											
Urban Heat Island	ECOSTRESS LST	F32/70m	21	378	3.4 / 2.1	1.8 / 0.4	1.6 / 1.6	5.8	9.5	0.41 / 2.99	6.82 / 20.45
Distribution	SIAM™ Categories	U8/10m	27	163	252.0 / 5.9	249.0 / 1.9	4.0 / 4.0	43.9	84.6	0.48 / 2.90	2.40 / 9.76
Zonal Statistics	SIAM™ Greenness	F32/10m	27	163	125.2 / 120.8	123.2 / 118.9	2.0 / 1.9	513.0	338.5	1.11 / 0.91	123.97 / 134.34
Occurrence	SIAM™ Mixed	Mixed/10m	27	163	175.0 / 170.1	173.6 / 168.6	1.0 / 1.0	43.9	84.6	0.50 / 2.90	3.42 / 9.76
Desktop Performance - No Cache / Cached											
Urban Heat Island	ECOSTRESS LST	F32/70m	21	378	2.3 / 1.0	2.1 / 0.9	0.2 / 0.2	5.8	9.5	0.55 / 1.22	4.65 / 14.31
Distribution	SIAM™ Categories	U8/10m	27	163	27.7 / 19.0	13.8 / 5.7	13.9 / 13.2	43.9	84.6	0.30 / 1.61	2.72 / 13.95
Zonal Statistics	SIAM™ Greenness	F32/10m	27	163	141.6 / 16.3	138.6 / 14.7	2.9 / 1.6	513.0	338.5	1.03 / 2.24	19.32 / 2.63
Occurrence	SIAM™ Mixed	Mixed/10m	27	163	142.5 / 18.2	139.1 / 15.4	3.2 / 2.7	43.9	84.6	0.27 / 1.47	3.01 / 11.85

thus prioritizes user-centric metrics (time-to-insight, cache efficiency) over raw computational benchmarks, reflecting the system's focus on exploratory analysis and stakeholder engagement.

3.2. Conclusions and Future Work

The convergence of cloud-native storage formats with browser-based processing represents a significant advancement in Earth Observation accessibility. Our empirical validation confirms operational viability for urban planning workflows, with municipality-scale analyses completing in seconds when cached. The published ol-zarr package [6] provides reusable OpenLayers extensions for Zarr visualization, while our implementation demonstrates feasibility of complex analytics entirely within browsers. By eliminating infrastructure dependencies, browser-native EO analytics opens new possibilities for stakeholder engagement in environmental monitoring and decision-making.

As web technologies evolve and Zarr adoption expands, browser-native analytics will increasingly democratize Earth Observation data access. Future research directions include WebGPU technologies for performance improvements, temporal chunk optimization, WebAssembly kernels, and Progressive Web Application architecture for offline capabilities. The ol-zarr package will be extended to support additional Zarr v3 features and multi-dataset fusion, further advancing browser-native EO capabilities.

REFERENCES

[1] R. P. Abernathey et al. Cloud-native repositories for big scientific data. *Computing in Science & Engineering*, 23 (2):26–35, 2021. doi: 10.1109/MCSE.2021.3059437.

[2] A. Baraldi et al. Automatic spectral-rule-based preliminary classification of radiometrically calibrated spot-4/-5/irs, avhrr/msg, aatsr, aster tir and tm/etm+ satellite images using spectral angle mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 48(9):3312–3346, 2010.

[3] V. C. F. Gomes et al. An overview of platforms for big earth observation data management and analysis. *Remote Sensing*, 12(8):1253, 2020.

[4] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. doi: 10.1016/j.rse.2017.06.031.

[5] S.J. Hook, G.C. Hulley, N. Allen, R. Radocinski, H. Bao, W.R. Johnson, M. Day, B. Eng, E. Jimenez, et al. Ecostress tiled land surface temperature and emissivity daily l2 global 70 m v002, 2023.

[6] Nyi Nyi Nyan Lin. ZarrTile: OpenLayers extension for Zarr dataset visualization., August 2025. URL https://doi.org/10.5281/zenodo.17013896.

[7] A. Miles et al. Zarr: An open-source array storage format for chunked, compressed, n-dimensional arrays. *Journal of Open Source Software*, 5(50):2307, 2020.

[8] NASA IMPACT. Next-gen zarr web map visualization - dynamic client vs tiling approaches. Technical report, NASA, December 2023.

[9] D. J. Newman. Zarr storage specification version 2. Technical Report RFC-048, NASA ESDS Standards Office, 2024.

[10] OpenLayers Contributors. Openlayers v7 documentation, 2023. URL https://openlayers.org.

[11] M. Turek et al. Geospatial analysis in web browsers—comparison study on webgis process-based applications. *ISPRS International Journal of Geo-Information*, 12(9):374, 2023.

[12] M. D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

[13] Zarrita.js Contributors. Zarrita.js: A javascript toolkit for zarr. GitHub Repository and NPM Package, 2023. v0.5.0.

FAIRSENDD: FAIR WORKFLOW FOR SENTINEL-1 BASED DEFORESTATION DETECTION

Felix Cremer, Daniel Loos, Fabian Gans

Max Planck Institute for Biogeochemistry
Biogeochemical Integration
Jena Germany

Stephan Sahm

Jolin.io

Munich, Germany

ABSTRACT

The increasing volume of datasets generated by new satellite missions necessitates efficient access within cloud environments to fully harness their potential. To achieve this, algorithms must be relocated to the data, and workflows should be reproducibly offered as cloud services. In this paper, we present the development of an end-to-end FAIR (Findable, Accessible, Interoperable, and Reusable) workflow for Sentinel-1-based deforestation detection, deployed as a cloud service. This approach democratizes access to complex algorithms, ensuring that scientific results are both reproducible and applicable to new areas of interest. We discuss the challenges and benefits associated with this transition. Additionally, during the migration to the cloud we enhanced the algorithm's runtime, achieving a 27% reduction in processing time through algorithmic and implementation improvements.

***Index Terms*—** FAIR, time series analysis, Common Workflow Language (CWL), Sentinel-1,

1. INTRODUCTION

Contemporary Earth System Science increasingly depends on comprehensive data analysis workflows. As these analyses grow in complexity, understanding and reproducing a workflow based solely on the materials and methods section of a research article becomes more challenging. Publishing the source code of the data analysis is a crucial first step toward establishing a FAIR workflow. However, merely providing the workflow's code is insufficient for enabling seamless end-to-end processing by future users. To achieve this, the necessary data must be accessible, and computing resources should be co-located with the data.

In this paper we describe how we made a reproducible workflow out of a scientific analysis code. Section 2 introduces the algorithm and its data requirements. Section 3 details how we improved the code to reduce the runtime. Section 4 describes the implementation of the end-to-end FAIR workflow. The end-to-end FAIR workfow is containerized in the Common Workflow Language (CWL) adhering to the OGC Best practices for Earth Observation Application Package.

2. ALGORITHM AND DATA

In this section we describe the workflow and the necessary data for the improved and deployed Sentinel-1 based forest change detection algorithm [2]. Figure 1 illustrates the workflow of the forest change algorithm. The algorithm utilizes Sentinel-1 time series data. The Sentinel-1 data needs to be stacked as a time series so that we can apply the Recurrence Quantification Analysis (RQA) algorithm to each pixel's time series. Therefore, the algorithm currently relies on analysis ready Sentinel-1 data, where every scene of the same relative orbit is aligned to the same grid. To minimize geometric effects arising from varying relative looking angles, we conduct the time series analysis on each relative orbit independently. Currently, the analysis is performed on preprocessed Sentinel-1 data [4]. The data is organized into 15,000 x 15,000 pixel tiles within the EQUI-7 Grid [1], with each scene stored separately. This approach can result in time series with numerous missing values if all scenes of the same area are simply stacked. In Section 3 we describe how we improved the loading of the data to get time series with less missing values.

After applying the RQA algorithm on every pixel of the Sentinel-1 time series we mask forest areas with a forest/non-forest map [3] and we cluster the detected change pixel into areas of at least 30 pixels to eliminate spurious single pixel change detections. Figure 2 shows an example of the forest change algorithm for the Harz mountains in central Germany. The different coloured datasets represent the forest change for the different years between 2018 and 2022.

3. CODE ENHANCEMENT

We further improved the algorithm to make the inner loop fully allocation free, reducing the runtime of the time series analysis for a single pixel by 91% for a 100 step time series. Figure 3 compares the runtime, memory usage and allocations between the original and the improved version of the implementation. The original version exhibited a near-exponential increase in memory usage, whereas the improved version eliminates memory allocation, significantly enhancing runtime and reducing pressure on the garbage collector.

doi:10.2760/2119408

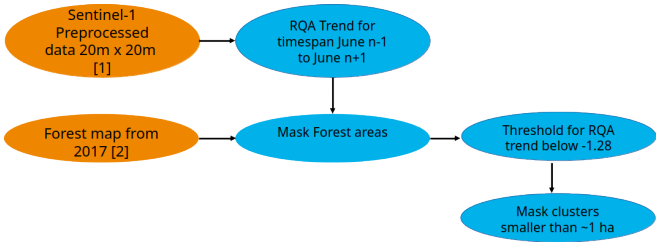


Fig. 1. Schematic workflow of the analysis. Orange ellipses specify input data and blue ellipses show the computational steps

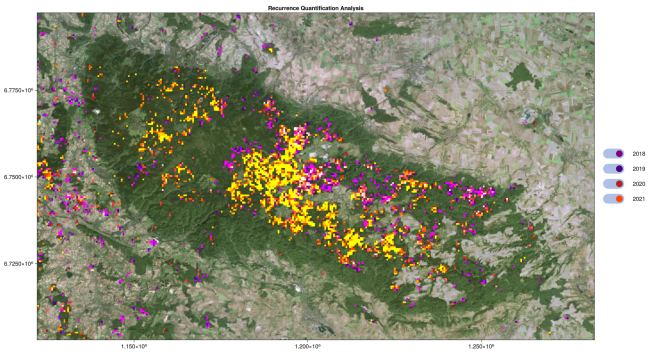


Fig. 2. Result of the forest change algorithm for the Harz mountain in Central Germany.

However, this reduction in runtime and memory footprint does not directly translate to a reduction for an entire data tile, as detailed in Table 1. The algorithmic changes for individual time series altered the input data requirements. While the previous version could handle regularly occurring missing values, the optimized version produced incorrect results under the same conditions. Consequently, we modified the data loading and preprocessing steps. Sentinel-1 data for each tile was grouped by acquisition time, allowing scenes acquired on the same date to be mosaicked together. This mosaicking process eliminated regularly occurring missing values but increased the runtime for preparing the data for an entire tile. During the recurrence analysis, we examine every time point pair, resulting in the algorithm scaling quadratically with the length of the time series. Additionally, since we do not incorporate information from neighboring pixels, the algorithm is embarrassingly parallelizable for larger areas of analysis. This characteristic provides flexibility in the workflow execution plan, allowing adaptation to the file structure and chunking schemes of the dataset at individual processing nodes. Finally, CWL workflows can be executed and orchestrated on any Kubernetes cluster using Calrissian.

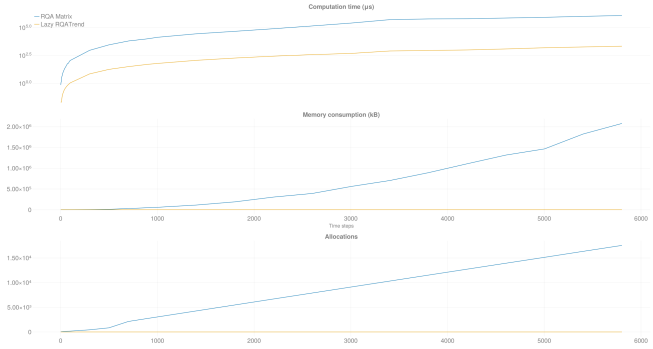


Fig. 3. Comparison of the old and the improved version of the algorithm. The old version had a near exponential increase of the memory usage while the improved version does not allocate memory.

Table 1. Improvements of the runtime and memory footprint due to the code improvements.

Indicator	v0.1	v0.2	Improved by
Duration (tile) [s]	884.03	649.04	27%
Duration (point) [s]	12.774	1.137	91%
Mem usage (tile) [GiB]	606.03	223.29	63%
Mem usage (point) [KiB]	3.75	0	100%
Mem alloc. (tile)	1.49e10	887461	99.94%
Mem alloc. (point)	8	0	100%

4. CLOUD DEPLOYMENT

A Julia library, as described in Section 3, is not sufficient on its own to ensure efficient and reproducible execution. We adhere to the principle of data locality by bringing the code to the data, rather than downloading the data for local execution. This necessitates running the code in cloud environments without direct server access, as detailed in Figure 4. We followed two different approaches to deploy the code to the cloud. For the first approach we encapsulated the analysis code in a containerized workflow following OGC Best Practice for Earth Observation Application Package (https://docs.ogc.org/bp/20-089r1.html). Hereby we utilize the Common Workflow Language (CWL) to describe the individual steps: stage in, process and stage out. The ”Stage in” process ensures uniform data access for the workflow and the user. Then we initiate the OGC API process to start the workflow via a REST API from pygeoapi. Then we use a python function to call the CWL Workflow. This nested architecture allows cloud providers to seamlessly integrate the workflow into their execution environments. The CWL employs a Docker container to host the Julia library, and we use STAC and Zarr as output formats to deliver the final results to

doi:10.2760/2119408

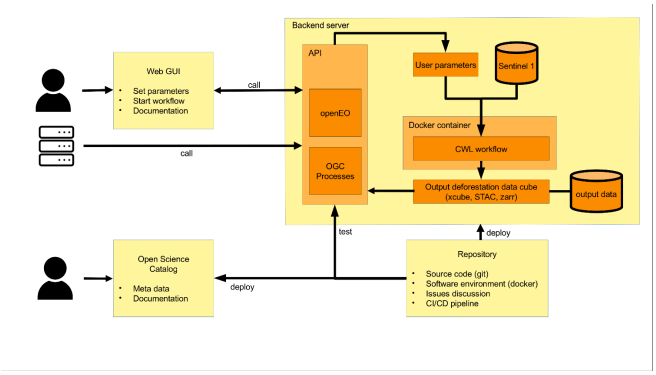


Fig. 4. Software architecture

the user.

For the second approach we compiled the function for the RQA Trend computation for a single time series into a standalone binary. This standalone binary can then be called via a C-API without having a julia runtime. This approach allows to run the RQA Trend function as an inner user defined function in an openEO backend. In this setup, the openEO backend manages data loading and postprocessing, while the Julia function is utilized solely in the core data processing stage.

The development of a FAIR workflow is influenced by the licenses and popularity of the tools and input datasets. Achieving accessibility requires permissive licenses, which is generally not a significant issue due to the widespread use of open-source software in the scientific community. While Level 2 satellite datasets are typically freely available, further processed products are often proprietary or only available at certain cloud providers, though the data can often be reproduced using open-source software.

Reproducibility is relatively straightforward to achieve through containerization, provided the algorithm is deterministic and its input data is FAIR, as is the case with our workflow. One common challenge to achieving FAIRness is availability of datasets or tools. Specific datasets may not be available on relevant cloud platforms, hindering workflow execution at scale. Similarly, some software may not be widely provided.Enhancing support for containers and CWL workflows in openEO would significantly ease the deployment of future workflows.

5. DISCUSSION AND OUTLOOK

This work demonstrates how we enhanced the algorithm and effectively brought it to the data. It highlights the necessity of evaluating algorithm improvements within the context of the entire workflow, as individual steps cannot be entirely isolated. The effort to reduce allocations in single-pixel analysis necessitated changes in data loading and preprocessing. This preprocessing step diminished the expected runtime improvements that might have been anticipated from simply extrapolating single-pixel runtime enhancements to the entire tile.

As future steps, we aim to explore deploying the algorithm across other cloud providers. Currently, this is challenged by the algorithm's reliance on stacked time series of Sentinel-1 data, which is less readily available compared to SLC or GRD data. We plan to investigate enabling the direct use of SLC data and stacking SLC data from the same burst ID to derive Sentinel-1 time series directly in radar geometry. Such a time series would be feasible since the algorithm is spatially independent.

6. CODE AVAILABILITY

The code is published free and open source under an MIT licence. The Julia package for the RQA based time series analysis is available at:

https://github.com/EarthyScience/RQADeforestation.jl

And the code for the cloud deployment is available here:

https://github.com/EarthyScience/FAIRSenDD/

7. ACKNOWLEDGEMENTS

This project was funded by the European Space Agency in the Science Result Long-Term Availability & Reusability Demonstrator Initiative. In addition, this project was supported by the ESA Network of Resources. This work uses Copernicus Sentinel data 2017-2019.

REFERENCES

[1] Bernhard Bauer-Marschallinger, Daniel Sabel, and Wolfgang Wagner. Optimisation of global grids for high-resolution remote sensing data. *Computers & Geosciences*, 72:84–93, November 2014. ISSN 0098-3004. doi: 10.1016/j.cageo.2014.07.005.

[2] Felix Cremer, Mikhail Urbazaev, Jose Cortes, John Truckenbrodt, Christiane Schmullius, and Christian Thiel. Potential of Recurrence Metrics from Sentinel-1 Time Series for Deforestation Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5233–5240, 2020. ISSN 1939-1404, 2151-1535. doi: 10.1109/JSTARS.2020.3019333.

[3] Alena Dostálová, Wolfgang Wagner, Milutin Milenković, and Markus Hollaus. Annual seasonality in Sentinel-1 signal for forest mapping and forest type classification. *International Journal of Remote Sensing*, 39(21):7738–7760, November 2018. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2018.1479788.

[4] Wolfgang Wagner, Bernhard Bauer-Marschallinger, Claudio Navacchi, Felix Reuß, Senmao Cao, Christoph Reimer, Matthias Schramm, and Christian Briese. A

Sentinel-1 Backscatter Datacube for Global Land Monitoring Applications. *Remote Sensing*, 13(22):4622, January 2021. ISSN 2072-4292. doi: 10.3390/rs13224622.

PREPARING FOR A THREE-DIMENSIONAL SENTINEL-1: TOWARDS HIGH-ACCURACY FOREST MONITORING VIA BI-STATIC SAR METHODS

Anton Kostiukhin[1], Martin Jüssi[1], Tauri Tampuu[1], Alexander Kmoch[2]

[1]KappaZeta Ltd., Kastani 42, 50410 Tartu, Estonia
[2]Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, 51003 Tartu, Estonia

ABSTRACT

Forest height and volume estimation using spaceborne SAR remains challenging in regions with uneven terrain and variability in forest structure and tree species. In this work, we investigate the potential of TanDEM-X bi-static interferometric data for hemiboreal forest height estimation in Estonia. The novelty of our study is a systematic assessment of how interferometric parameters, forest properties, and environment conditions affect forest height estimation accuracy. Combining multi-temporal TanDEM-X acquisitions with auxiliary LiDAR-derived and forest inventory data, this study is the first step towards a scalable, robust, and accurate forest height monitoring method for future multi-static SAR missions.

Preliminary results highlight the strengths and limitations of existing forest height estimation methods and provide an overlook into how forest height retrieval behaves under different conditions.

Index Terms— forest height, tandem-x, bi-static SAR, time-series analysis

1. INTRODUCTION

Sentinel-1 is a powerful data factory. No other current SAR mission produces data with systematic global coverage in such a large quantity. However, its information content is relatively limited – dual-polarization backscatter and repeat-pass interferometry data. Across-track interferometry is not feasible with Sentinel-1 due to temporal decorrelation (6 or 12 days) and short interferometric baselines (<100 m) [1]. The limited information content of Sentinel-1 sets an inherent limit to forestry applications built on Sentinel-1 reducing its scalability [2].

One of the efficient methods for forest height estimation is across-track interferometry from a bi-static SAR system, for example – TanDEM-X. Existing studies have sufficiently shown the usefulness of the TanDEM-X for forest height estimation across different forest types – boreal, temperate, and tropical forests. However, publications often omit critical details about data acquisition parameters, processing methods, or even the study areas themselves. This represents a significant research gap, especially given that numerous studies have emphasized the influence of data characteristics and processing choices on the accuracy and reliability of results.

Consequently, this work aims to fill this research gap and develop a transparent and reliable methodology for hemiboreal forest height estimation. For that the goal is to systematically quantify the effects of interferometric variables (e.g., polarization, baseline) and forest properties (e.g., height, density, tree composition) on forest height estimation using TanDEM-X data as a proxy. KappaZeta Ltd. from Estonia is developing a receive-only three-satellite constellation ("3D-SAR") to fly in formation with Sentinel-1 to enable the multi-static dimension, and therefore enrich the information provided by Sentinel-1, Fig. 1. 3D-SAR Mission Concept. The mission is expected to produce three times the amount of raw data that Sentinel-1 does for the same area. The focus of 3D-SAR is on producing a global forest height layer through single-pass across-track SAR interferometry, a proven method for high accuracy forest height monitoring.

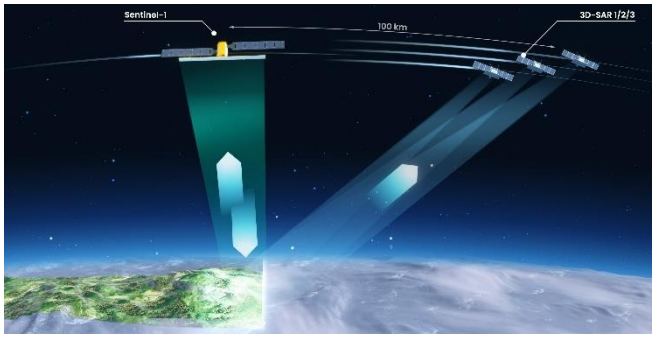


Fig. 1. 3D-SAR Mission Concept

2. TANDEM-X FOR FOREST HEIGHT ESTIMATION

TerraSAR-X and TanDEM-X are a pair of nearly identical satellites launched in 2007 and 2010 respectively. The satellites fly together in a close helix formation and acquire unique single-pass polarimetric interferometric data unaffected by temporal decorrelation [3], [4]. The data acquired by the mission was originally intended to be used for a global digital elevation model generation. However,

doi:10.2760/2119408

multiple studies have found that the X-band is also suitable for forest height estimation [4]. Since then, many researchers have attempted to perform forest height estimation with TanDEM-X data using various approaches, which are generalized here into three groups: inversion models, machine learning methods, and digital surface and terrain models (DSM-DTM) differencing techniques.

The inversion models include the Random Volume Over Ground (RVoG) model and its modifications, SINC, and C-SINC models. These models are widely used for tree height inversion in boreal (e.g. [5]), temperate (e.g. [6]), and tropical (e.g. [7]) forests. The models proved to be robust, accurate, and easy to use. In the case of the RVoG, the model also has high interpretability since it utilizes the extinction coefficient, interferometric phase, and ground-to-volume ratio [8]. However, numerous drawbacks inhibit the application of the models. For example, the RVoG requires fully polarimetric data, which is not routinely available from the TanDEM-X platform [8]. Additionally, the above-mentioned models are insensitive to terrain variations, which cause under- or over-estimation in high slope areas [6].

Among the machine learning models, Random Forest (RF) is one of the most used methods for forest height estimation [9]. Compared to the inversion models, machine learning methods are not strictly limited by the data requirements. That makes it possible to easily integrate data from multiple additional sources [10]. Moreover, RF is an interpretable method. This trait is widely used by researchers to explain the results of the modelling and define the most important features in the datasets [11], [12]. Apart from the RF, more advanced methods were used in recent studies. They include, but are not limited to, Classification and Regression Tree (CART), Gradient-Boosting Decision Tree (GBDT), Support Vector Machine (SVM), and more [13]. While these methods tend to perform slightly better, they lack the explainability of the RF method.

The DSM-DTM differencing techniques require additional high-quality DTM data for the study area. The exact approach varies from study to study. In some cases, it can be expressed as a simple DSM-DTM subtraction [13]. However, due to certain penetration capabilities of the X-band into the canopy, the results of such methods are of a lower quality [4]. Alternatively, if InSAR height is corrected to a penetration depth, the results show comparable or better accuracy to other methods [14].

In addition to the methods described, various pre-processing and data fusion techniques have positive effects on accuracy. For example, in coherence estimation, the window size cannot be set deliberately, as it might cause loss of information [8]. Given that TanDEM-X acquisitions are with high temporal resolution, temporal averaging of interferometric features can improve the results of the height estimation [11]. In case the area of interest is well covered by other data sources – terrain, climate, and optical data can be used as well [12], [13].

3. PROJECT OVERVIEW AND CURRENT STATUS

3.1. Study area

The area of interest is approximately 800 square kilometers in size, and it is split between two locations, Fig. 2. Study area. The first site is situated in the former Pikknurme forest district within Jõgeva County, Estonia. This region is well-known for extensive forest research, and it is one of the first permanent forest observational plots [15]. The second site is in the Western part of Saaremaa Island, Estonia. The areas are dominated by European Spruce (Picea abis) and Silver Birch (Betula pendula) species.

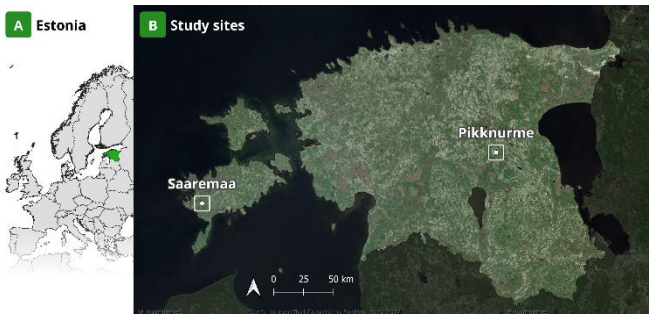


Fig. 2. Study area

3.2. Data

The primary data source used in this study is TanDEM-X. Particularly for this activity, 33 datatakes were scheduled: 18 of which cover Saaremaa and the remaining 15 cover Pikknurme area of interest. The data was acquired in VV/VH polarization, StripMap mode, from August and October 2024. There is confirmation from DLR that most of the datatakes were successful, but as of April 2025, the data has not been delivered yet. Therefore, for test purposes and pipeline development, data from the TanDEM-X Science Archive was used.

We use 5m DTM as a primary source of terrain information, which was obtained from the airborne LiDAR collected by the Estonian Land Board [16]. For comparison purposes, we also use Copernicus 30m and SRTM elevation models.

Additionally, we use tree species and forest compartments from Forest Registry [17], a canopy height map derived from airborne LiDAR data collected by Estonian Land Board. The data is available on Estonian Land Board portal [16].

3.3. Methods

So far, a major part of the work has been spent on getting familiar with the specifics of the bistatic TanDEM-X data,

doi:10.2760/2119408

data processing steps, and quality control. SNAP 8.0 is used as primary tool for interferometric processing. The processing steps are divided into three groups, depending on what data is required: backscatter, coherence, or InSAR height, Fig. 3. TanDEM-X Processing Pipeline.

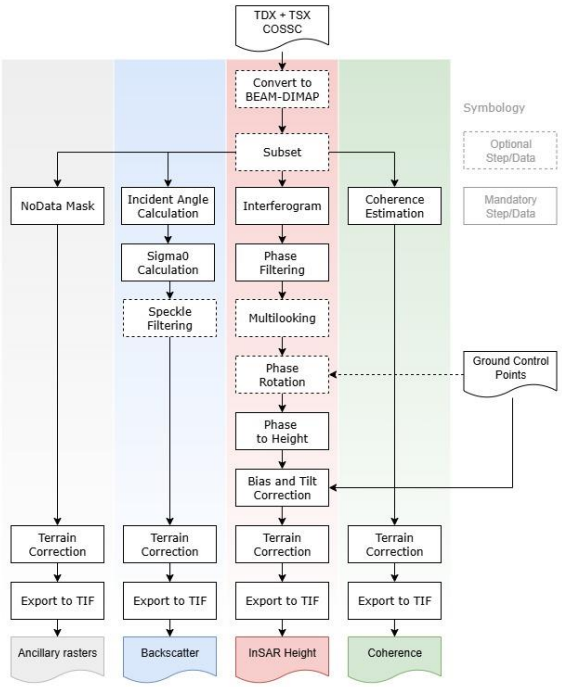


Fig. 3. TanDEM-X Processing Pipeline

Preliminary processing parameters were outlined based on the existing studies. There are different options and combinations due to the software, study area, and data differences. Thus, the final processing configuration and parameters are not stated. See Fig. 4. InSAR Height Raster for an example of the InSAR Height raster.

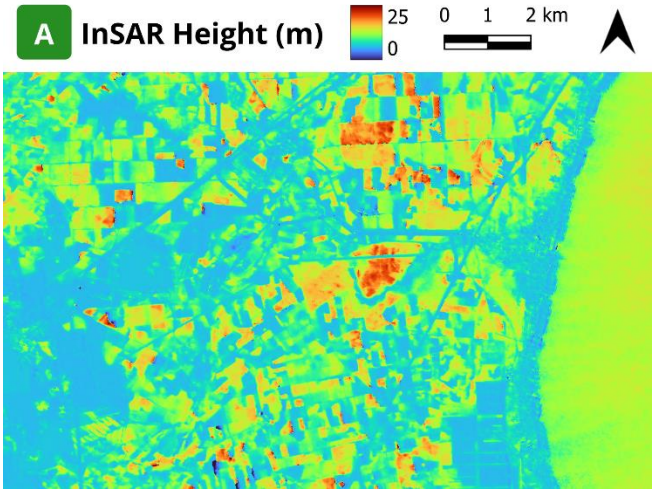


Fig. 4. InSAR Height Raster

In addition to the interferometric features, SNAP is configured to output ancillary raster data which includes no-data mask, Local Incidence Angle, Elevation, and Height of Ambiguity. Miscellaneous topographic and LiDAR data have not been thoroughly investigated and validated yet.

4. FUTURE WORK

4.1. Project outlook and 4-year plan

Forest height estimation using TanDEM-X is the first part of a 4-year doctoral project which is done by KappaZeta Ltd. in cooperation with the University of Tartu. The name of the project is “Estimation of forest height and forest volume in hemiboreal forests from multi-static synthetic aperture radar”. The project is intended to provide a better understanding of how various interferometric variables influence forest height and volume estimation. This information will be used to develop a reliable and transparent methodology to estimate forest properties.

The work is divided into three parts, which contribute to the overall goal of the project. The first two parts focus on forest height and volume estimation using Estonian hemiboreal forests as a case study. In these studies, we aim to understand the importance of various interferometric features and how to get the most information from them. Using additional data from other satellites is not our priority. Once we understand the data well enough and can provide reliable results on a local scale, we start to focus on the extended area of interest. That includes forests in Finland, Norway, and other neighboring countries. The goal of the third part is then to develop a large-scale model for forest height and volume estimation. Research progress and the results of each part will be disseminated in peer-reviewed scientific journals. The project is expected to be concluded at the end of 2028.

4.2. Near future and 1-year plan

In the near future, we are focusing exclusively on forest height estimation. We continue with the literature overview, focusing on data processing methods, forest properties, and data parameters. We start to look at the quality of the validation data and undertake additional processing steps if needed. We expect DLR to provide access to the requested data by June 2025 so that we can start data processing and methodology validation.

In addition to the requested TanDEM-X dataset, we are considering writing a proposal for extra bistatic acquisitions in the year of 2025 and 2026. We plan to have new data to be timewise as close as possible to the airborne LiDAR scanning performed by Estonian Land Board. Moreover, we are looking at the options to perform in situ data collection together with research groups from the University of Tartu.

doi:10.2760/2119408

5. CONCLUSION

In conclusion, existing forest height estimation methods from bi-static SAR data, though not without drawbacks, are powerful tools for forest monitoring. Initial findings demonstrate a variety of existing and promising new developments in the domain. Further work will focus on achieving high accuracy in forest height estimation over test areas in hemiboreal forests. The presentation will showcase the latest findings, highlighting achieved results and possible limitations.

ACKNOWLEDGEMENT

The authors would like to thank DLR for the TanDEM-X imagery used in the study under science proposal XTI_VEGE7867. LiDAR and Elevation data were provided by Estonian Land Board. Data about forest compartments and tree species were provided by Forest Registry [17].
The project is co-financed by the European Union.

REFERENCES

[1] P. Potin *et al.*, "Copernicus Sentinel-1 Constellation Mission Operations Status," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, July 2019, pp. 5385–5388. doi: 10.1109/IGARSS.2019.8898949.

[2] L. Chen *et al.*, "Improved Object-Based Estimation of Forest Aboveground Biomass by Integrating LiDAR Data from GEDI and ICESat-2 with Multi-Sensor Images in a Heterogeneous Mountainous Region," *Remote Sensing*, vol. 14, no. 12, Art. no. 12, Jan. 2022, doi: 10.3390/rs14122743.

[3] A. Moreira *et al.*, "TanDEM-X: a TerraSAR-X add-on satellite for single-pass SAR interferometry," in *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, Sept. 2004, pp. 1000–1003 vol.2. doi: 10.1109/IGARSS.2004.1368578.

[4] F. Kugler, D. Schulze, I. Hajnsek, H. Pretzsch, and K. P. Papathanassiou, "TanDEM-X Pol-InSAR Performance for Forest Height Estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6404–6422, Oct. 2014, doi: 10.1109/TGRS.2013.2296533.

[5] A. Olesk, K. Voormansik, A. Vain, M. Noorma, and J. Praks, "Seasonal Differences in Forest Height Estimation From Interferometric TanDEM-X Coherence Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 12, pp. 5565–5572, Dec. 2015, doi: 10.1109/JSTARS.2015.2501648.

[6] C. Gomez *et al.*, "Canopy Height Estimation in Mediterranean Forests of Spain with TanDEM-X Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, pp. 1–1, Feb. 2021, doi: 10.1109/JSTARS.2021.3060691.

[7] S. Yadav, H. Padalia, S. K. Sinha, R. Srinet, and P. Chauhan, "Above-ground biomass estimation of Indian tropical forests using X band Pol-InSAR and Random Forest," *Remote Sensing Applications: Society and Environment*, vol. 21, p. 100462, Jan. 2021, doi: 10.1016/j.rsase.2020.100462.

[8] A. Olesk, J. Praks, O. Antropov, K. Zalite, T. Arumäe, and K. Voormansik, "Interferometric SAR Coherence Models for Characterization of Hemiboreal Forests Using TanDEM-X Data," *Remote Sensing*, vol. 8, no. 9, Art. no. 9, Sept. 2016, doi: 10.3390/rs8090700.

[9] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[10] M. Pourshamsi, M. Garcia, M. Lavalle, and H. Balzter, "A Machine-Learning Approach to PolInSAR and LiDAR Data Fusion for Improved Tropical Forest Canopy Height Estimation Using NASA AfriSAR Campaign Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3453–3463, Oct. 2018, doi: 10.1109/JSTARS.2018.2868119.

[11] S. Antonova *et al.*, "Estimating tree height from TanDEM-X data at the northwestern Canadian treeline," *Remote Sensing of Environment*, vol. 231, p. 111251, Sept. 2019, doi: 10.1016/j.rse.2019.111251.

[12] I. Teijido-Murias, O. Antropov, C. A. López-Sánchez, M. Barrio-Anta, and J. Miettinen, "Forest Height and Volume Mapping in Northern Spain with Multi-Source Earth Observation Data: Method and Data Comparison," *Forests*, vol. 16, no. 4, Art. no. 4, Apr. 2025, doi: 10.3390/f16040563.

[13] J. Bao, N. Zhu, R. Chen, B. Cui, W. Li, and B. Yang, "Estimation of Forest Height Using Google Earth Engine Machine Learning Combined with Single-Baseline TerraSAR-X/TanDEM-X and LiDAR," *Forests*, vol. 14, no. 10, Art. no. 10, Oct. 2023, doi: 10.3390/f14101953.

[14] M. Schlund, S. Erasmi, and K. Scipal, "Comparison of Aboveground Biomass Estimation From InSAR and LiDAR Canopy Height Models in Tropical Forests," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 367–371, Mar. 2020, doi: 10.1109/LGRS.2019.2925901.

[15] A. Kiviste *et al.*, "The Estonian Permanent Forest Plots Network," in *Proceedings of an International Workshop at Beijing Forestry University*, 2012, pp. 97–108. Accessed: May 01, 2025. [Online]. Available: https://www.researchgate.net/profile/Klaus-Gadow/publication/279509791_Growth_experiment_trials_Permanent_temporal_and_interval_plots/links/559ac23c08ae99aa62ce240c/Growth-experiment-trials-Permanent-temporal-and-interval-plots.pdf#page=101

[16] Maa-amet, "Republic of Estonia Land and Spatial Development Board." Accessed: May 01, 2025. [Online]. Available: https://geoportaal.maaamet.ee/eng/spatial-data/elevation-data/download-elevation-data-p664.html

[17] "Forest Registry | Environmental Agency." Accessed: May 01, 2025. [Online]. Available: https://keskkonnaagentuur.ee/metsaregister

ENHANCING SPACE OPERATIONS WITH UNSUPERVISED ANOMALY DETECTION: THE PITIA SYSTEM

M. Tejedor, H. Jiménez, J.A. Pozo, I. Perea, J.M. Auñón

Department of Artificial Intelligence and Big Data, GMV,
Isaac Newton,11, Tres Cantos,Madrid, Spain

ABSTRACT

Anomaly detection is a problem faced daily by space mission operations control centers. These tasks involve identifying unexpected values in telemetry (TM) data originating from various onboard systems. Operators generally rely on classical systems such as the definition of nominal ranges, so the inclusion of more advanced techniques such as artificial intelligence can speed up and reduce the cost of these processes. *PitIA* was developed to address this challenge by optimizing operations and improving equipment utilization. It uses an unsupervised process to identify periods of anomalies in the data, enabling anomaly detection and a better anomaly management. We have evaluated *PitIA* using the ESA anomalies dataset published in June 2024, and in this letter we propose a realistic unattended operational deployment .

Index Terms— Anomaly detection, *PitIA*, Telemetry, Machine Learning.

1. INTRODUCTION

Anomaly detection is a critical challenge in spacecraft missions, where the reliable operation of onboard systems is essential for mission success. These missions generate large volume of TM data from various systems and subsystems, such as thermal control. The identification of anomaly behavior in this data is key to ensure operational continuity, however this data is composed of several (hundred) of individual sensors coupled in somehow between them, making manual inspection or rule-based monitoring insufficient and inefficient.

From a data perspective, TM data can be represented as multiple time series. Anomaly detection in time series is an active field of research, and the literature offers a wide range of techniques, from those based on statistical models to more complex approaches leveraging artificial intelligence. We invite to the reader to go to reference [1], where the authors stablish a taxonomy of outlier techniques based on outlier type which is used in this research.

In this context, GMV has developed *PitIA* [2], a solution capable of detecting anomalies in multivariate time series, while also identifying which variables (channels) contribute most to each anomaly. This greatly assists spacecraft operators in diagnosing the root cause of anomalous behavior. *PitIA* is based on Multivariate Statistical Process Control (MSPC), a discipline extensively studied in industrial applications [3], which we apply here to spacecraft telemetry data—specifically, the satellite anomaly database released by ESA [4]. In this work, we compare the performance of the *PitIA* solution against the algorithms proposed in the same dataset, and we further extend the study by introducing a continuous training approach tailored for real-world mission scenarios.

The paper is organized as depicted in Fig. 1, describing the main steps of the pipeline, and we finish with the result of apply this procedure to anomaly database.

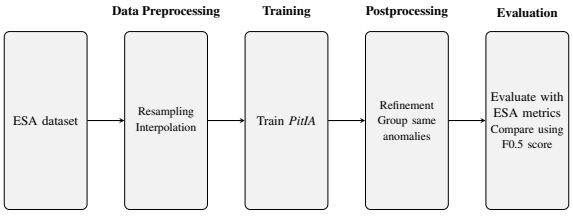


Fig. 1. Execution pipeline of the anomaly detection framework from data preprocessing to final evaluation.

2. DATASET

Due to the lack of publicly available datasets and standardized benchmarks for advanced automatic anomaly detection in space mission telemetry, the European Space Agency (ESA) has released a curated dataset of real satellite telemetry. This dataset is accompanied by a hierarchical evaluation pipeline and benchmarking results for various anomaly detection algorithms. The dataset comprises telemetry data from three different ESA missions, two of which are included in the benchmarking frame-

doi:10.2760/2119408

work. Each dataset is annotated by domain experts and cross-validated using state-of-the-art techniques.

The evaluation pipeline is tailored to the practical requirements of Satellite Operations Engineers (SOEs), featuring novel evaluation metrics and simulation of real-world operational scenarios. The benchmarking process assesses common anomaly detection algorithms to highlight areas where further improvements are needed.

3. DATA PREPROCESSING

The preprocessing pipeline begins with a resampling step that reduces the number of data points and enforces uniform sampling across TM channels—an essential requirement for time-series models. Each mission uses a specific resampling interval based on its native resolution (30 sec for Mission 1 and 18 sec for Mission 2). Following resampling, missing values are imputed using a modified zero-order hold interpolation method that preserves causality by propagating the last known value without referencing future data. This process includes timestamp alignment, value propagation, and an anomaly-preservation step that ensures annotated events are retained in the resampled dataset. Together, these steps produce a temporally consistent and complete dataset suitable for unsupervised anomaly detection [4]

4. TRAINING PITIA MODEL

PitIA performs anomaly detection through a dimensionality reduction approach based on Principal Component Analysis (PCA). Once the model is trained on nominal telemetry data, each new observation is projected onto the principal component space. Two statistical metrics are then computed to evaluate the deviation from normal behavior Squared Prediction Error (SPE), and Hotelling's T^2, being the first the main metric used to determinate anomalies. SPE measures how much of an observation is not explained by the principal components. A high SPE value suggests that the sample deviates significantly from the learned structure and may indicate a novel event or unmodeled behavior. SPE is given by:

$$\text{SPE} = \sum_{i=1}^{k} \left(x_{\text{new},i} - \hat{x}_{\text{new},i}\right)^2, \tag{1}$$

where k is the total number of observation, $x_{\text{new},i}$ is the ith observation and $\hat{x}_{\text{new},i}$ is the prediction of the observation vector from the PCA model. The number of principal components is automatically established, reaching the 90% of accumulative variance in the latent space [5]. In MSPC, an observation is considered anomalous if it exceeds the upper control limit (UCL)

$$UCL_{SPE} = \frac{\nu}{2b}\chi^2_\alpha\left(\frac{2b^2}{\nu}\right), \tag{2}$$

where ν and b are the sample variance and the sample mean of the SPE values respectively. $\chi^2_\alpha(\cdot)$ is the Chi-squared distribution at significance level α (95%). Whereas this UCL is statistically well defined, it did not give us as good results as those obtained after a post-processing. We noticed that when two or more different anomalies coincide in a given range of time, the contribution to SPE from one of them is usually greater than then rest, hidden the rest of anomalies. Figure 2 illustrates an example of SPE.

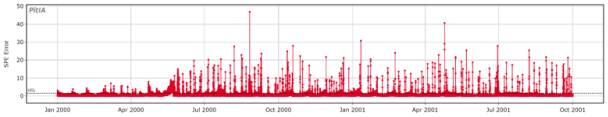


Fig. 2. SPE for Mission 2 and full set of channels. Train period: 01/01/2020 - 01/07/2020.

5. POSTPROCESSING

To enhance the reliability of the detected anomalies and reduce noise in the results, a postprocessing stage is applied.

On one hand, to refine the detection signal, the gradient of the prediction error values SPE is computed, that is, $\dot{SPE}$ to measure the rate of change over time. By using the absolute value of the gradient $\|\dot{SPE}\|$, both upward and downward changes in the error are captured. This approach highlights the magnitude of deviation between consecutive data points, offering a more sensitive anomaly indicator. In general, SPE is almost constant, so $\dot{SPE} \simeq 0$ and the thresholds for anomaly detection is just simply an inter-percentile formula, that is, $IPR = P_{95} - P_5$ and the lower and upper limits are $P_5 - 1.5IPR$ and $P_{95} + 1.5IPR$ respectively. With this configuration we obtained the best results.

On the other hand, detected anomalies that occur within short time intervals are likely to be manifestations of the same underlying issue. To account for this, temporally adjacent anomalies are grouped together if the time difference between them is below a configurable threshold. In our implementation, anomalies occurring less than 6 hours apart are considered the same one. This threshold is not a fixed or arbitrary value, but was selected based on GMV's expert input from satellite operations centers, ensuring its relevance for real-world use cases. The technique consolidates multiple small, potentially redundant anomaly events into a single larger

doi:10.2760/2119408

anomaly, thus avoiding an excessive number of consecutive alarms and facilitating clearer interpretation of the results.

6. EVALUATION

The performance of the proposed anomaly detection system was evaluated using the event-wise F0.5 score, the primary metric recommended by the ESA for benchmarking anomaly detectors. This score combines precision and recall, with a higher weight on precision, which is critical in operational settings to minimize false alarms.

An anomaly is considered successfully detected if its timestamp overlaps with a labeled anomaly event, and the system avoids raising multiple redundant alarms for the same event. The F0.5 score provides a robust balance between detection accuracy and alert relevance. In addition, two complementary metrics were used to assess the interpretability of the results: i) Subsystem-aware: Measures whether the predicted anomalies were assigned to the correct subsystem; ii) Channel aware: Evaluates whether the channels most responsible for the anomaly were correctly identified. These secondary metrics help verify the model's ability not only to detect anomalies but also to attribute them correctly, supporting actionable insights in telemetry monitoring. For brevity reasons, we restrict ourselves to F0.5 score.

7. RESULTS: FROM BENCHMARK TO OPERATIONS

Now we describe the main results, this section is divided in two: first subsection adds a new column (technique) to Table 2 and Table 3 from ESA reference [4]. This helps to measure the performance of our approach compared to other techniques. The second subsection describes a methodology to use anomaly detection models in a unattended way, ready to operations.

7.1. Benchmark

We evaluated *PitIA* on the ESA anomaly benchmark using datasets from two satellite missions, under multiple training and test configurations. Experiments were assessed using the event-wise F0.5 score, which prioritizes precision. The train/test dataset periods are the same from ESA paper: *Each mission is divided into halves of which the first half is taken as a training set and the second half as a test set. This gives 84 months of training data for Mission1 and 21 months for 16 Mission2. In both cases, the last 3 months of the training set are taken as the validation set.* A summary of quantitative results is presented in Table 1.

Configuration	PitIA	Best ESA Model	Second Best ESA Model
Mission 1 – Full channels	0.424	0.061 (Teleman-ESA P.)	0.008 (Teleman-ESA)
Mission 1 – Subset (ch. 41–46)	0.323	0.786 (Teleman-ESA P.)	0.253 (Global STD5)
Mission 2 – Full channels	0.760	0.241 (STD5)	0.100 (PCC)
Mission 2 – Subset (ch. 18–28)	0.794	0.949 (Window iForest)	0.842 (Teleman-ESA P.)

Table 1. Summary of F0.5 scores across ESA benchmark experiments

PitIA consistently achieved high F0.5 score in full-channel settings. Its performance across temporal shifts in Mission 2 demonstrates robustness, and its generalization across reduced channel subsets highlights adaptability. Analysis of undetected anomalies revealed some false positives corresponded to likely unlabelled events.

In Mission 2, *PitIA* achieved top-tier performance when using the full channel set (F0.5 = 0.760) and strong results on a reduced subsystem (F0.5 = 0.794), with excellent precision (up to 0.910). While Teleman-ESA Pruned outperformed in some cases, *PitIA*'s performance remained robust across data dimensionalities and time splits. Here we remark that Teleman-ESA Pruned is an ad-hoc version of Teleman-ESA algorithm [6] difficult to train due to hyperparameter settings and required computational resources (8.5h for trainning for full set of channels of Mission 1)

In Mission 1, a more challenging dataset due to less structured anomalies, *PitIA* still outperformed most ESA baselines and showed notable generalization in reduced input settings. Despite some missed anomalies—often attributed to weak signal changes or label inconsistencies—*PitIA* also detected events not included in the ground truth but with clear error spikes, suggesting possible unlabelled anomalies.

These findings validate *PitIA* as a practical anomaly detection tool that balances operational precision and computational efficiency, and performs well under constrained telemetry conditions.

7.2. Operations

The previous subsection shows *PitIA* performance compared to other techniques. However, this approach is still far from being ready for an operational environment, leaving several open questions: Should the model be retrained? When should retraining occur? Is it necessary to wait 84 months (i.e., 7 years of data for Mission 1) to collect sufficient training data? As we can see, these unanswered questions highlight the challenges of deploying an anomaly detection solution in real-world operations.

Unlike traditional systems that require extensive retraining or full reconfigurations, *PitIA* supports a continuous model updating approach. In our deployment scenario, the system is trained using only one month of historical data and evaluated on the immediately following month (see Fig. 3). After each cycle, the training win-

doi:10.2760/2119408

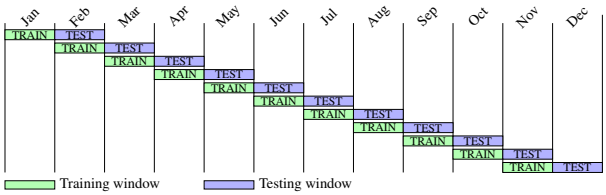


Fig. 3. Continuous model updating using monthly sliding windows.

dow slides forward by one month, allowing the model to adapt to recent patterns and preserve contextual relevance. This sliding window approach ensures that the anomaly detection model remains up-to-date without requiring full retraining from scratch.

This incremental strategy offers several operational advantages: i) no need for labeled anomalies: since *PitIA* relies on unsupervised learning, the approach is fully autonomous and does not depend on manual annotations; ii) fast model update: each training iteration takes <2 minutes, even when using up to 12 months of telemetry data, ensuring minimal computational burden; iii) scalability: this setup supports daily or incremental preprocessing, allowing integration with real-time data ingestion pipelines; and iv) robustness in operation: by always training with recent data, the system is resilient to long-term drift and evolving system behavior.

Table 2 shows the results under this configuration. Note that there is no direct comparison with other techniques, as this specific setup is not evaluated in the ESA reference. The results are consistent with those in Table 1, but in this case the evaluation is performed continuously, and only one month of data is required, so the solution can be deployed after just one month of observations.

Configuration	**PitIA**
Mission 1 – Full channels	0.424
Mission 1 – Subset (ch. 41–46)	0.332
Mission 2 – Full channels	0.882
Mission 2 – Subset (ch. 18–28)	0.880

Table 2. Summary of F0.5 scores across ESA benchmark experiments using a continuous training approach.

8. CONCLUSIONS

In this work, we presented *PitIA*, a robust and fully autonomous tool for unsupervised anomaly detection in satellite telemetry. Validated against ESA's open anomaly benchmark dataset, *PitIA* demonstrated strong performance across different missions and configurations. The system is generic, lightweight, and scalable, making it suitable for diverse operational scenarios and capable of handling large volumes of telemetry data with minimal computational overhead. Its ability to operate without human intervention, combined with mechanisms for automatic training and threshold tuning, ensures adaptability and long-term reliability.

PitIA achieved outstanding results on Mission 2 (Full channels) and performed competitively on the challenging Mission 1, showing its robustness even in noisy or weakly labeled environments. Furthermore, its design aligns with key requirements for space operations: accurate detection with low false alarm rates, ease of deployment, and autonomous retraining capabilities. As the only method evaluated that meets all operational criteria, *PitIA* is a strong candidate for integration into real-time satellite monitoring workflows.

REFERENCES

[1] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," *arXiv preprint arXiv:2002.04236*, 2020.

[2] GMV, "GMV PitIA - Sistema de Detección de Anomalías," GMV, [Online]. Available: https://www.gmv.com/es-es/productos/industria/gmv-pitia. [Accessed: Apr. 30, 2025].

[3] J. F. MacGregor, C. Jaeckle, C. Kiparissides, and M. Koutoudi, "Process monitoring and diagnosis by multiblock PLS methods," *AIChE Journal*, vol. 40, no. 5, pp. 826–838, 1994.

[4] K. Kotowski, C. Haskamp, J. Andrzejewski, B. Ruszczak, J. Nalepa, D. Lakey, P. Collins, A. Kolmas, M. Bartesaghi, J. Martinez-Heras, and G. De Canio, "European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry," *arXiv preprint arXiv:2406.17826*, 2024.

[5] Jolliffe, I. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions Of The Royal Society A: Mathematical, Physical And Engineering Sciences*. **374**, 20150202 (2016)

[6] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD '18)*, London, United Kingdom, 2018, pp. 387–395.

EOEPCA OPEN SOURCE EO DATA EXPLOITATION PLATFORM

Richard Conway[1], Chandra Taposeea-Fisher[1], James Hinton[1], Claudio Iacopino[2], Salvatore Pinto[2]
[1]Telespazio, [2]European Space Agency

ABSTRACT

Many web-based platforms provide access to satellite Earth Observation (EO) data, now often combined with cloud computing resources and applications. Users benefit from the ability to process data remotely, bypassing traditional download and storage limitations. Our vision is to enhance interoperability between these platforms to create an open network for seamless data access. To achieve this, we are establishing best practices for EO exploitation platforms based on open standards and developing a reference implementation of building blocks as free open source software. This project is sponsored by the European Space Agency (ESA), with its first iteration, EOEPCA, starting in 2018, and it's second iteration, EOEPCA+, starting in 2023, and aims to demonstrate the architecture and software in operational platforms.

Our presentation will highlight the generalized architecture, standards, best practice and open-source software components available.

Index Terms - EO, Exploitation Platform, Open Source, Interoperability

1. INTRODUCTION

The 'Exploitation Platform' concept derives from the need to access and process an ever-growing volume of data. Many web-based platforms have emerged - offering access to a wealth of satellite earth observation (EO) data. Increasingly, these are collocated with cloud computing resources and applications for exploiting the data. Rather than downloading the data, the exploitation platform offers a cloud environment with access to EO data and associated compute and tools that facilitate the analysis and processing of large data volumes.

Users are beginning to appreciate the advantages of exploitation platforms. However, the market now offers a plethora of platforms with various added value services and data access capabilities. This ever-increasing offer is rather intimidating and confusing for most users. In order to fully exploit the potential of these complementary platform resources we anticipate the need to encourage interoperation amongst the platforms, such that users of one platform may consume the services of another directly platform-to-platform.

EOEPCA+ continues the progress of the original EOEPCA initiative – but with a greater focus on the real-world use cases of platform providers. As an OSGeo Community project, EOEPCA+ has established a steering committee of Stakeholders that are committed to adopt the building blocks in their operational platforms. This ensures that EOEPCA+ efforts are focused towards needed capabilities, and that the solutions are production ready. Stakeholder engagement ranges from feature/use-case definition and adoption, through to co-design and co-development of building-blocks and features. This has led to the inclusion of several new building blocks, covering capabilities including: Datacubes, workflow federation , ML model development, automation, best practices for open science and operational resilience.

The primary users of EOEPCA+ are Platform Providers - helping them to build their platforms that provide the services needed by their users - reusing the EOEPCA+ building blocks as required. The needs of their users informs the features required of the building blocks – this includes data providers, scientists (data analysis, algorithm refinement), application developers – and ultimately policy makers that consume the value-adding information that has been derived from the platform data.

2. COMMON ARCHITECTURE

EOEPCA+, also known as EO Exploitation Platform Common Architecture, [1] is an ESA funded project with the goal to define and agree a re-usable exploitation platform architecture using standard interfaces to encourage interoperation and federation between operational exploitation platforms - facilitating easier access and more efficient exploitation of the rapidly growing body of EO and other data.

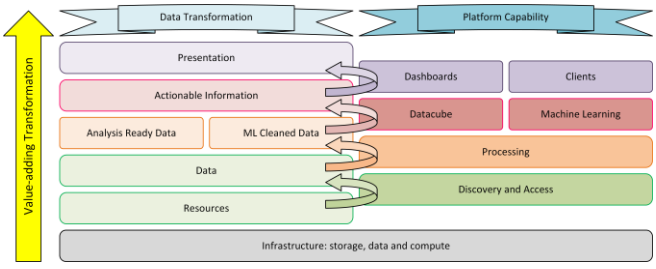


Fig. 1. Data Transformation

Interoperability through open standards is a key guiding force for the Common Architecture: platform developers are more likely to invest their efforts in standard implementations that have wide usage; off-the-shelf clients and software are more likely to be found for standards-based solutions. Whilst standardization at the service layer is a key step towards interoperability, we recognize that there are additional factors

doi:10.2760/2119408

that apply friction to full interoperability – including metadata vocabularies and data formats – in particular across different communities and domains.

The presentation of actionable information to decision makers is at the end of a potentially complex chain of data transformation, processing, interpretation and presentation. Exploitation platforms must provide the tooling and services to support these needs – **Fig. 1** illustrates the end-to-end capabilities of an information factory that transforms original data to *Actionable Information*. This is a multi-step workflow in which data may be pre-processed into *Analysis Ready Data [2]* designed for access as a *Data Cube [3]*; or prepared for input to Machine Learning model development and execution.

3. ARCHITECTURE

The System Architecture [4] is defined by a set of Building Blocks with open standard interfaces, each of which contributes to the overall capabilities of an integrated platform. A building block (BB) is defined as an open source software component that implements a specific platform capability and typically provides a service interface (REST API) – deployed to Kubernetes with its associated helm chart. Building blocks are designed to either be used on their own, or in combination as a system.

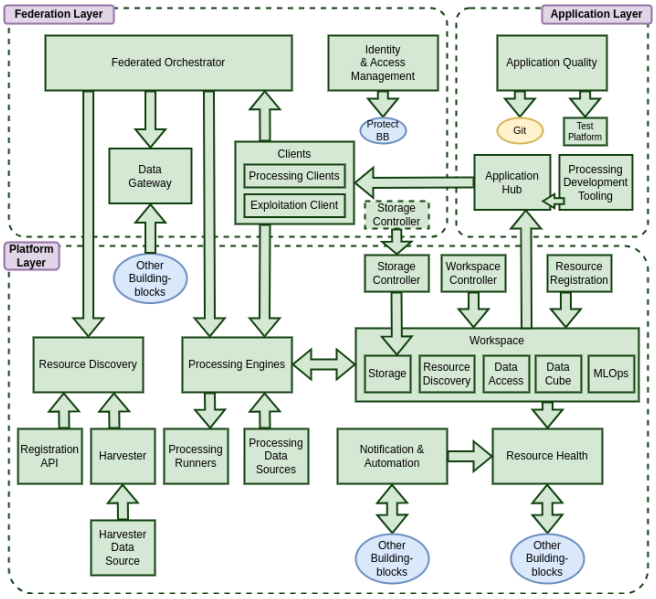


Fig. 2. Architecture of EOEPCA+

The architecture presents the building-blocks within a set of layers that attempt to reflect their notional role with a multi-platform distributed ecosystem. This layering should be regarded as illustrative as, in practice, the building-blocks and the architecture are flexible for adaption to many deployment scenarios.

The **Platform Layer** comprises capabilities for discovery of data and other resources, execution of processing workflows, and management/exploitation of added-value assets.

The **Federation Layer** comprises capabilities that operate across a set of distributed platforms, and attempt to consolidate their combined offerings towards a more homogenous consumable experience.

The **Application Layer** provides capabilities for development and publishing of applications for exploitation of platform services, and for showcasing research outcomes through information dashboards and web-enabled applications - applicable for both Platform and Federation use cases.

Fig. 3 shows the team responsible for the EOEPCA+ building block development.

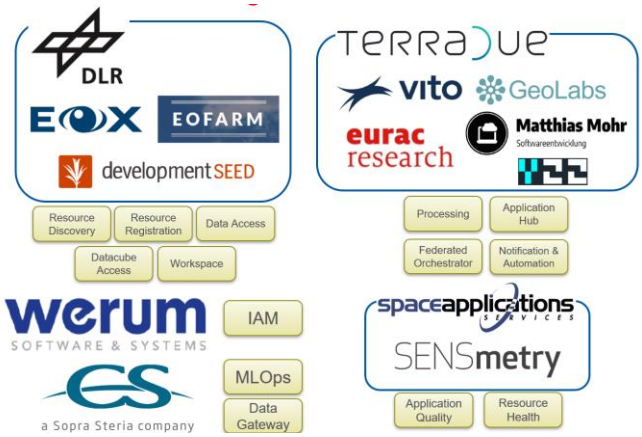


Fig 3: EOEPCA+ Building Block Development Team

3.1. Reproducible Science

To support reproducible open science, the architecture must support the capability to record the details of process and workflow execution, in order to reproduce the conditions of the original. This impacts on the following building blocks:

- Resource Discovery. Maintain records of workflow job execution details.
- Processing Engines. Report details of (sub-) workflow job executions – possibly to be consolidated by the Federated Orchestrator.
- Orchestration. Report details of orchestrated workflow job executions – by consolidation of sub-workflow jobs in Processing Engines.

To achieve all reproducibility scenarios (rerun, repeat, replicate, reproduce, reuse) the job details must be recorded regarding workflow execution at all levels. Thus, the Processing Engines must output jobs details that can be recorded as resources in the Resource Discovery. The Federated Orchestrator must similarly output aggregated job details pertaining to the overall workflow execution,

doi:10.2760/2119408

comprising the details of individual steps executed in the Processing Engines.

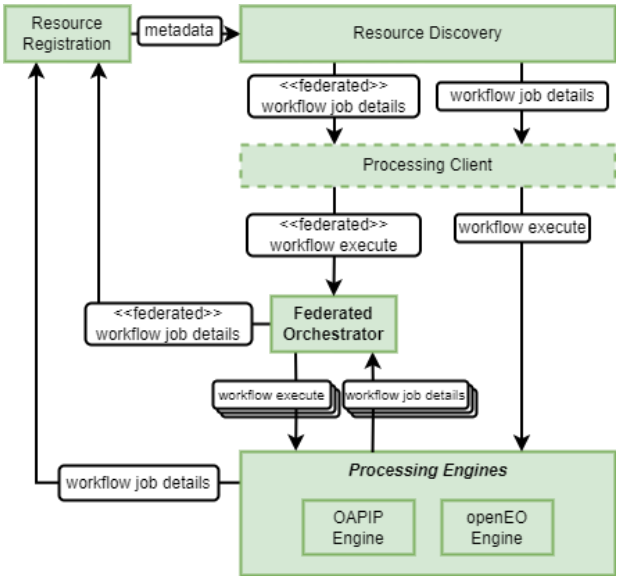


Fig. 3. Reproducible Science

3.2. User-defined Processing

The Processing Building Block is designed to provide capabilities for the hosted execution of processing workflows.

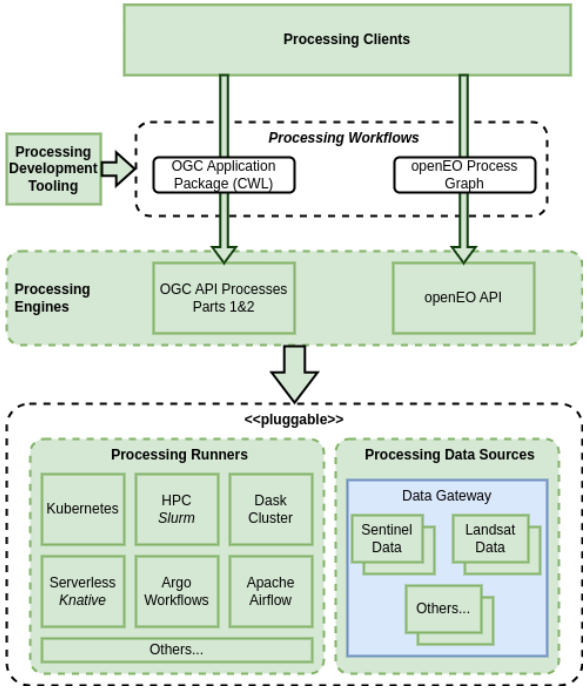


Fig. 4. Processing

As illustrated by **Fig. 4**, these workflows are defined as:

1. OGC API Processes Part 1 [5] Part 2[6] – EOEPCA+ has supported the development of the OGC Best Practice for EO Application Package [7] that formalizes how processing algorithms are packaged and described for platform integration. Particularly suited to large scale batch processing.
2. openEO API – with client-oriented semantics (Python, R and JavaScript) that abstract the API. openEO offers close Datacube integration.

Both approaches provide a portable means to submit user-defined processing for execution close to the data. To achieve this, a generic data curation approach is favored, allowing the needs of various processing workflows to be met across multiple platforms. This approach eliminates the necessity for each individual platform to develop its own unique data integration solutions.

The processing architecture is designed for an extensible set of execution engines, including Kubernetes, HPC, dask, Argo Workflows – each of which can be integrated behind the standard OGC API with support for Application Packages.

3.3. Platform Resources

The Resource Discovery building block maintains a metadata catalogue for the resources held within a platform – with OGC API Records and STAC APIs. It supports platform federation by maintaining records to resources in other platforms.

In supporting the Find capability as one of the FAIR principles, the Resource Discovery Building Block provides discovery for not only data (e.g., datasets, data cube, virtual data cube), but also workflows, job details, Jupyter Notebooks, Executable Services, Platform Services, Web Applications, Documentation, etc.

The Data Access building block, provides feature-rich and reliable interfaces to access, retrieve and visualize geospatial data assets stored in the platform, addressing human and machine users alike. Capabilities are delivered through standard service interfaces, including OGC APIs Features/Tiles/Maps - supporting data assets (incl. multidimensional data formats) persisted via common storage technologies including S3-comptabile object storage, HTTP, file system. Access to *Analysis Ready Data* is enabled through the Datacube Access Building Block.

In supporting the Reuse capability as one of the FAIR principles, the Resource Registration Building Block provides support for ingesting resources into the platform so that they can be discovered, accessed and used collaboratively.

The Workspace building block provides users/projects/teams with the capability to maintain their own resources within the platform - including data for processing, processing workflow packages ready for execution, and results output from workflow executions. The Workspace provides object storage bucket management to persist these assets – with facilities for sharing assets outside the

doi:10.2760/2119408

Workspace. Each Workspace also provides a vCluster to its members through which bespoke services can be hosted and published.

3.4. User Analysis and Exploitation

This area of the EOEPCA+ environment focuses on machine learning, interactive analysis, application best practice, event-driven behaviors and operational outcomes. The MLOps building block provides support services for training of machine learning models within the cloud platform. It also integrates within the EOEPCA+ Building Block ecosystem, with the other Building Blocks such as Processing, Workspace, and Resource Discovery.

The Application Hub building block is a comprehensive and modular platform delivering SaaS products, designed to cater to the diverse and multifaceted needs of the EO community. It is crafted to support a wide array of stakeholders, from developers and service providers integrating cutting-edge algorithms to researchers harnessing computational power, and analysts requiring clear and concise visualizations.

The Application Quality building block's aim is to support the evolution of a scientific algorithm (processing workflow) from a research project to one that can be utilised in a production environment, with tools for verifying non-functional requirements (code quality, best practice for open reproducible science, performance optimisation).

The Notification & Automation BB is designed to facilitate intra-Building-Block asynchronous communications. This means it allows different parts of the system to communicate with each other without needing to wait for responses, thereby improving efficiency and responsiveness. It supports triggers that can initiate automated behaviour. These triggers can be based on external events (e.g. events from object storage, etc) or can be scheduled to occur at certain times.

Finally, the Resource Health BB offers a generalized capability that allows all types of users to specify and schedule checks relating to their resources of interest, to visualize the outcome of the checks, and to receive notifications according to the outcome.

3.5. Platform Federation

The final area of the EOEPCA+ environment concerns platform federation, through federated workflows, abstract data access and federated user identity. The Federated Orchestrator building block allows for cross-platform workflow execution, and hybrid workflows, combining OGC API Processes, Application Packages and openEO Process Graphs.

The Data Gateway building block enables data source abstraction, with data access protocol, with the use of Python library and extensible data providers.

Finally, the IAM building blocks is key for single sign-on and external provider integration.

4. NEXT GENERATION SDI

Gomes et al [8] identifies some key attributes of next generation SDI. EOEPCA+ responds to these challenges:

- **Scalability and resilience.** Use of Kubernetes, which also provides platform-agnostic reusable building-blocks
- **Moving Code paradigm.** Use of OGC Application Packages and openEO for user-defined code
- **Job Parallelisation**. Common Workflow Language (CWL) scatter patterns combined with Kubernetes workload orchestration
- **Interoperability.** Open standard interfaces delivered through REST APIs - including OGC, STAC and openEO
- **Array Data.** Datacube services for efficient access to multi-dimensional data
- **Simplified Usage.** Programmatic clients (e.g. openEO) to abstract APIs and facilitate user update - noting that EOEPCA+ requires additional client support to cover its breadth of capabilities

5. NEXT STEPS

The EOEPCA+ architecture and reference implementation have made good progress to mature a set of building blocks for interoperable exploitation platforms. Through its *Stakeholders*, the project will continue to engage with operators of existing platforms to understand and respond to their real-world use-cases - and to support the integration of the EOEPCA+ building blocks into their platform offering – to enhance their platform capabilities and to provide feedback on the utility of the building blocks.

REFERENCES

[1] EOEPCA+ Project Team, "EOEPCA Project Portal", https://eoepca.org/.

[2] CEOS, "CEOS Analysis Ready Data", https://www.ceos.org/ard/

[3] openEO Project, "What are Datacubes?", https://openeo.org/documentation/1.0/datacubes.html

[4] EOEPCA+ Project Team, "System Architecture", https://eoepca.readthedocs.io/projects/architecture/en/latest/

[5] Open Geospatial Consortium, OGC API – Processes – Part 1: Core Specification, https://ogcapi.ogc.org/processes/.

[6] Open Geospatial Consortium, OGC API - Processes – Part 2: Deploy, Replace, Undeploy (DRU) Draft Specification, https://github.com/opengeospatial/ogcapi-processes/tree/master/extensions/deploy_replace_undeploy.

[7] Open Geospatial Consortium, OGC Best Practice for Earth Observation Application Package, 20-089r1 Version 1.0, http://www.opengis.net/doc/BP/eoap/1.0, 14th December 2021.

[8] Gomes, V.C., Queiroz, G.R., and Ferreira, K.R., 2020. An overview of platforms for big earth observation data management and analysis, https://www.mdpi.com/2072-4292/12/8/1253

doi:10.2760/2119408

METHODOLOGICAL AND COMPUTATIONAL CHALLENGES OF EVALUATING SOIL PROPERTIES MAPS DERIVED FROM HIGH RESOLUTION EARTH OBSERVATION DATA

Laura Poggio, David Rossiter, Niels Batjes, Bas Kempen

ISRIC - Worls Soil Information - Wageningen (NL)

ABSTRACT

Remote sensing data are useful covariates to aid with the production of soil properties maps. The increasingly-iner resolutions of remote sensing products aids the production increasingly-finer resolution soil maps. However, the resulting products must be evaluated for pointwise accuracy as well as how realistically they represent the soil landscape. This study presents a dedicated reproducible computational framework for integrating remote sensing into digital soil mapping workflows, and methods for the evaluation of the derived products.

Copernicus, Data integration, digital soil mapping, high performance computing, reproducibility

1. INTRODUCTION

Digital Soil mapping (DSM) at continental and global scale provides standardised global information layers based on predictive models. It is also an important tool to create soil information layers for areas for which local soil survey information is lacking. The recent availability of global and continental remote sensing-derived products, coupled with the ease of access to computational resources, has made the production of such layers easier across the globe. Therefore, it is ever more important to assess the quality of DSM-derived products.

This paper presents a computational framework to produce and evaluate soil properties maps for Europe at 20m resolution, with a key aspect being the integration of advanced remote sensing datasets.

2. DATA

2.1. Soil observations

The observations used for mapping were queried from the WoSIS database [2]. Soil observations from approximately 94k locations were available. Most of the observations belong to the LUCAS sampling [13]. The primary soil properties in table 2.1 were modelled for top soil (0-30cm).

Property	Description	Units
soc	Soil organic carbon content	g/kg
phh2o	pH in water	pH
nitrogen	Total nitrogen	g/kg
bdod	Bulk density, oven dry	kg/dm3
cfvo	Coarse fragments, volumetric	cm3/100cm3
sic	Soil inorganic carbon content	g/kg
sand	Sand	g/100g
silt	Silt	g/100g
clay	Clay	g/100g

2.2. Environmental Covariates

Sentinel 2

SCMaP [11, 20] is a specialised processing chain for detecting and analysing bare soils/surfaces on a large (continental) scale. Bare surface and soil pixels are selected using a combined NDVI and NBR index (PVIR2) calculated and applied for each individual pixel. Sentinel-2 scenes recorded between January 2018 and December 2022 in Europe were used with a cloud cover of < 80 % and a sun elevation of $>$ 20 degrees. The data products are available at a pixel size of 20 m for 10 Sentinel-2 bands (B02 to B08, B08a, B11, B12).

The products used in this work are: 1. "Bare Surface Frequency Product" provides the number of bare soil occurrences over the total number of valid observations. 2. "Reflectance Composite - Mean" represents the mean reflectance of all valid observations including vegetation, bare and other surfaces 3. "Reflectance Composite - Standard deviation", which contains the standard deviation per band for all valid observations.

Sentinel 1

The Sentinel-1 (S1) mission data were pre-processed, prepared, mosaicked and downloaded from Google Earth En-

Funding: CUP4SOIL (Framework Partnership Agreement on Copernicus User Uptake); HoliSoils (European Union's Horizon 2020 research and innovation programme grant agreement No. 101000289)

doi:10.2760/2119408

gine [10]. S1 provides SAR (Synthetic Aperture Radar) backscatter signal for VH (vertically transmitted and horizontally received) and VV (vertically transmitted and received). The polarization ratio was calculated: $(VH - VV)/(VV + VV)$ as it was proven useful for discriminating land covers with different canopy densities and it is an important parameter for soil moisture detection [9]. The overall median (2018 to 2020) was calculated and used as covariates.

Morphology

The Copernicus Digital Elevation Model [6] was used to derive the following morphological features: 1. Elevation from sea level (m) 2. Slope as the steepest slope angle, 3. Topographic wetness index [22] defined as the $ln(a/tan(b))$

Climate data

ERA5 is the fifth generation ECMWF atmospheric reanalysis of the global climate [7] combining model data with global observations. ERA5 provides aggregated values for each month for the following ERA5 climate reanalysis parameters: temperature (2m), total precipitation, and total evaporation. Monthly total precipitation values are given as monthly sums. All other parameters are provided as monthly averages.

Downscaling of climate data

The input rasters were resampled by filtering the input with an analytic kernel [4]. Each output cell is typically calculated based upon a small subset of the input cells performing convolution (i.e. a weighted sum is calculated for every raster cell). The input range is mapped to the width of the window function, so wider windows will be "sharper" (have a higher cut-off frequency). The implements FIR (finite impulse response) filtering. All of the functions are low-pass filters. The filter used were `gauss`, `box` with the radii equal to $1.5 * inputresolution$ and $3 * inputresolution$ respectively. The `gauss` filter is a gaussian kernel filter, while the `box` filter is a finite window filter.

Land cover

The Dynamic Land Cover map at 100 m resolution (CGLS-LC100) product includes continuous field layers for all basic land cover classes that provide proportional estimates for vegetation/ground cover for the land cover types.

The 2019 products [3] with the proportional estimates for vegetation/ground cover for the land cover types are provided as environmental layers in this project.

Parent material

Information from parent material was derived from the European Soil Database v2.0 [18]. Both primary (Code for dominant parent material of the STU) and secondary (Code for secondary parent material of the STU) units where rasterized in binary (presence/absence) format.

Vegetation phenology

Products to describe the phenology of the vegetative season were used [5] 1. Season amplitude, 2. productivity and 3. length.

3. METHODS

3.1. Digital Soil Mapping

The DSM approach builds on the methods described in [19]. Random Forests was used as the modelling method, using the ranger package [23], with the option `quantreg` to build Quantile Random Forests (QRF) [15]. With QRF, predictions generate a cumulative probability distribution of the soil property at each location, rather than a single average value from the ensemble of decision trees.

Recursive feature elimination [14] was used for covariates selection. Model tuning was conducted using a 10-fold cross-validation procedure applied to multiple combinations of hyper-parameters: number of decision trees (`ntree` parameter) and numbers of covariates (`mtry` parameter). Predictions accuracy was assessed using root mean squared error (RMSE) and model efficiency coefficient (MEC)[12]. The model evaluation was based on the performance metrics of the selected hyper-parameters combination. The final model was fitted with all available observations, the covariates and the hyper-parameters selected in the previous steps.

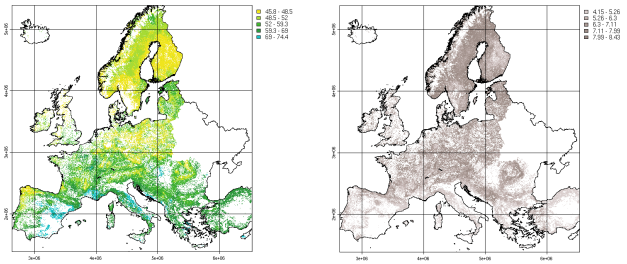
With this approach both the mean predictions and the quantiles can be obtained. The 5th and 95th quantiles were used according to the GlobalSoilMap specifications [1].

3.1.1. Maps evaluation

Maps were evaluated using the following approaches:

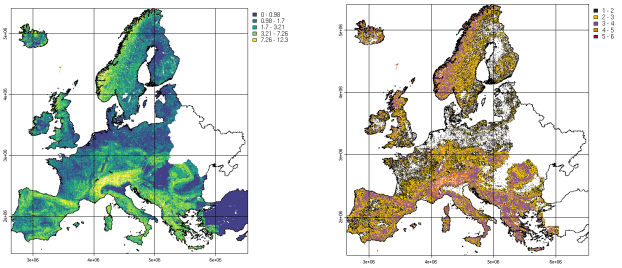
- model uncertainty $((Q95 - Q05)/Q50)$ per pixel [19]
- comparison of spatial patterns [21]
- area of applicability (AOA) of (spatial) prediction models [16]. In particular a dissimilarity index (DI) is calculated based on distances to the training data in the multidimensional predictor variable space.
- Quadmap: Variable resolution maps to better represent spatial uncertainty [17] based on quadtree algorithm recursively partitioning the map into quadrants until the uncertainty criteria are fulfilled

doi:10.2760/2119408



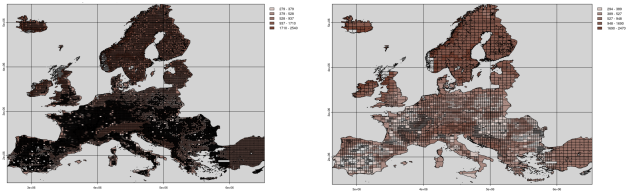
(a) Mean predictions pH (water) (b) Uncertainty for pH

Fig. 1. Predictions and their uncertainty index



(a) Dissimilarity index (b) DI by percentiles

Fig. 2. Example of AOA analysis



(a) Low uncertainty threshold (b) High uncertainty threshold

Fig. 3. Example of analysis with quadmap

4. COMPUTATIONAL FRAMEWORK

The soil properties maps were produced at 20m resolution for of Europe. This required adaption of the approach described in [8]. The methods to evaluate maps also needed to be adapted to work for continental scale products, both from a methodological and computational points of view. New approaches were implemented to tile covariates before the modelling, for the parallelisation of computation for the Area of Applicability and to compute meaningful variograms for continental extents.

All computations were performed using apptainer images containing the relevant software and packages, in particular R, python and GRASS-GIS.

5. PRELIMINARY RESULTS

The preliminary results show examples of the products for Europe (figure 1). Figures 2 and 3 show the results of the evaluation of the maps. The preliminary outputs indicate that the computational framework was able to create fine resolution soil properties maps integrating advanced earth observation data. The framework allows also the analysis of the spatial pattern of the products both in space and in the covariates space. These preliminary results indicate both the needs for in depth maps evaluation and the feasibility of adapting existing workflow to dataset with orders of magnitude more pixels and covering large geographical extents.

6. CONCLUDING REMARKS

Remote sensing provides invaluable information for soil mapping. The availability of remote sensing product at finer resolution allows to produce soil properties maps at fine resolution. This requires adapting existing workflow and methods both from a computational and conceptual point of view. What works at field or catchment level will not necessarily work for a continental or global product. Furthermore, expert knowledge (domain, users, stakeholders) is fundamental to evaluate the generated DSM products. Soil products must be created for the specific purpose and at the required scale and resolution. We need to use more options than accuracy metrics to evaluate DSM products, in particular their relationship with the landscape, model applicability and uncertainty.

REFERENCES

[1] D. Arrouays, M. G. Grundy, A. E. Hartemink, J. W. Hempel, G. B. Heuvelink, S. Y. Hong, P. Lagacherie, G. Lelyk, A. B. McBratney, N. J. McKenzie, M. d.L. Mendonca-Santos, B. Minasny, L. Montanarella, I. O. Odeh, P. A. Sanchez, J. A. Thompson, and G.-L. Zhang. Chapter three - globalsoilmap: Toward a fine-resolution global grid of soil properties. volume 125 of *Advances in Agronomy*, pages 93 – 134. Academic Press, 2014. doi: http://dx.doi.org/10.1016/B978-0-12-800137-0.00003-0.

[2] N. H. Batjes, L. Calisto, and L. M. de Sousa. Providing quality-assessed and standardised soil data to support global mapping and modelling (wosis snapshot 2023). 16(10):4735–4765. doi: 10.5194/essd-16-4735-2024.

[3] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets. Copernicus global land cover layers - collection 2. *Remote Sensing*, 12(6), 2020. doi: 10.3390/rs12061044.

[4] G. Clements. GRASS GIS: r.resamp.filter module. https://github.com/OSGeo/grass/tree/main/raster/r.resamp.filter, 2024. Accessed: 2024-12-18T21:07:30.

[5] Copernicus. Season amplitude, productivity and lenght 2017-present (raster 10 m), europe, yearly, sept. 2021, 2021.

[6] Copernicus. Copernicus dem - global and european digital elevation model, 2024. URL 10.5270/ESA-c5d3d65.

[7] Copernicus Climate Change Service (C3S). Era5: Fifth generation of ecmwf atmospheric reanalyses of the global climate. copernicus climate change service climate data store (cds), 2017.

[8] L. M. de Sousa, L. Poggio, G. Dawes, B. Kempen, and R. van den Bosch. Computational infrastructure of soilgrids 2.0. In I. N. Athanasiadis, S. P. Frysinger, G. Schimak, and W. J. Knibbe, editors, *Environmental Software Systems. Data Science in Action*, pages 24–31, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-39815-6_3.

[9] I. Gherboudj, R. Magagi, A. A. Berg, and B. Toth. Soil moisture retrieval over agricultural fields from multi-polarized and multi-angular radarsat-2 sar data. *Remote Sensing of Environment*, 115(1):33 – 43, 2011. doi: https://doi.org/10.1016/j.rse.2010.07.011.

[10] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: http://dx.doi.org/10.1016/j.rse.2017.06.031.

[11] U. Heiden, P. d'Angelo, P. Schwind, P. Karlshöfer, R. Müller, S. Zepp, M. Wiesmeier, and P. Reinartz. Soil reflectance composites—improved thresholding and performance evaluation. *Remote Sensing*, 14(18), 2022. ISSN 2072-4292. doi: 10.3390/rs14184526.

[12] P. Janssen and P. Heuberger. Calibration of process-oriented models. *Ecological Modelling*, 83(1): 55–66, 1995. doi: https://doi.org/10.1016/0304-3800(95)00084-9. Modelling Water, Carbon and Nutrient Cycles in Forests.

[13] A. Jones, O. Fernandez Ugalde, S. Scarpa, and B. Eiselt. Lucas soil 2022, eur 30331 en.

[14] Kuhn and Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28 (5):1–26, 2008. doi: 10.18637/jss.v028.i05.

[15] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006. URL http://jmlr.org/papers/v7/meinshausen06a.html.

[16] H. Meyer and E. Pebesma. Estimating the area of applicability of remote sensing-based machine learning models with limited training data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2028–2030, 2021. doi: 10.1109/IGARSS47720.2021.9553999.

[17] J. Padarian and A. McBratney. Quadmap: Variable resolution maps to better represent spatial uncertainty. *Computers Geosciences*, 181:105480, 2023. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2023.105480.

[18] P. Panagos, M. Van Liedekerke, P. Borrelli, J. Köninger, C. Ballabio, A. Orgiazzi, E. Lugato, L. Liakos, J. Hervas, A. Jones, and L. Montanarella. European soil data centre 2.0: Soil data and knowledge in support of the eu policies. *European Journal of Soil Science*, 73(6):e13315, 2022. doi: https://doi.org/10.1111/ejss.13315.

[19] L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Riberio, and D. Rossiter. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, pages 1–37, 2021. doi: 10.5194/soil-2020-65.

[20] D. Rogge, A. Bauer, J. Zeidler, A. Mueller, T. Esch, and U. Heiden. Building an exposed soil composite processor (scmap) for mapping spatial and temporal characteristics of soils with landsat imagery (1984–2014). *Remote Sensing of Environment*, 205: 1–17, 2018. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2017.11.004.

[21] D. G. Rossiter, L. Poggio, D. Beaudette, and Z. Libohova. How well does digital soil mapping represent soil geography? an investigation from the usa. *SOIL*, 8 (2):559–586, 2022. doi: 10.5194/soil-8-559-2022.

[22] R. Sorensen, U. Zinko, and J. Seibert. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences*, 10:101–112, 2006.

[23] M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. 77(1):1–17. doi: 10.18637/jss.v077.i01.

OPEN, CLOUD-OPTIMIZED, ANALYSIS-READY GLOBAL GEDI SATELLITE LIDAR DATASETS FOR LAND SURFACE APPLICATIONS

Yu-Feng HO[1], *Johannes Heisig*[2], *Milutin Milenković*[3], Leandro Parente[1], Rolf Simoes[1], Tomislav Hengl[1]

[1]OpenGeoHub, Doorwerth, the Netherlands
[2]Institute for Geoinformatics, University of Münster, Münster, Germany
[3]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

ABSTRACT

Current satellite LiDAR missions, such as GEDI and ICESat-2, provide billions of points annually that are typically not cloud-optimized and require additional quality filtering before any further analysis. In this study, we present OpenLandMap GEDI (OLM-GEDI), a new open, cloud-optimized, and global GEDI point dataset, for which we establish a spatio-temporal structure to facilitate efficient access. We show random access to OLM-GEDI achieves 20 seconds and a minute for areas around 50-thousand and 3-million km^2, respectively. The OLM-GEDI STAC catalog is further established, which can be readily loaded into a local or cloud computing environment, such as openEO. This open GEDI dataset can be beneficial to future studies to enhance their reproducibility and mitigate the complexity of handling large GEDI data volumes ($\sim$ 120 TiB) and quality filters.

***Index Terms*—** Global satellite LiDAR, GEDI, canopy, terrain, cloud-native format, Geoparquet, STAC, openEO

1. INTRODUCTION

Recent satellite LiDAR (Light Detection and Ranging) missions, such as Global Ecosystem Dynamics Investigation (GEDI) and Ice, Cloud, and Land Elevation Satellite 2 (ICESat-2), have been collecting near-global and global 3-dimensional (3D) information about the Earth's surface since early 2019 and late 2018, respectively. They utilize sensors that generate, transmit, and deflect laser pulses in several directions, which enables the sampling of 3D information, i.e., recording the backscattered laser energy distribution along the range at each sampling location, along eight and six ground tracks per every GEDI and ICESat-2 orbit [1, 2]. With pulse repetition frequencies of 242 Hz and 10 kHz, GEDI and ICESat-2 currently provide billions of points annually, which, due to their sparse, irregular, and 3D nature, are often more complex for analysis than classical satellite images.

This work has received funding from the European Union's Horizon Europe research and innovation programme (grant agreement No. 101059548)

Satellite LiDAR data have already been used for large-scale land and water surface applications. Several studies, e.g., contributed with global high-resolution (30–10 m) canopy height mapping, where sparse satellite LiDAR points were combined with multispectral images from Landsat or Sentinel-2 to achieve wall-to-wall machine-learning-based height predictions [3, 4]. In topography, satellite LiDAR data have recently been used to map a global 30 m digital terrain model again within a machine learning framework [5]. Furthermore, recent studies have also shown the potential of satellite LiDAR for monitoring the surface levels of inland and wetland waters [6, 7].

The above studies are computationally and data-intensive, requiring appropriate indexing and other optimizations for efficient data access and utilization at large scales. Furthermore, although satellite LiDAR data are openly accessible, considerable data quality filtering is still required before they can be used for specific use cases [8, 9, 10, 11]. To increase reproducibility and lower the aforementioned barriers for a broad range of users, this paper presents a global cloud-optimized, quality-filtered satellite LiDAR dataset for efficient queries, exemplified by GEDI and potentially applicable to other datasets.

2. MATERIAL AND METHODS

2.1. GEDI data

GEDI is a full-waveform satellite LiDAR that records backscattered laser energy levels (waveforms) along the laser beam's line of sight (range), near-globally, between 51.6°N and S, within a laser footprint diameter of 25 m [1]. The GEDI Level 2A product includes waveform (point-level) metrics such as ground elevation, relative height (rh) calculated between the waveform's lowest mode and different waveform energy levels (e.g., `rh98` is the relative height at the 98 percentile of the waveform energy), waveform processing algorithm information, different quality flags, and auxiliary information such as beam identifier, etc.[11]. The GEDI Level 2B product includes point-level canopy cover and vertical profile bio-

doi:10.2760/2119408

physical metrics, such as plant area index profiles, foliage height diversity, etc. [12].

2.2. Quality Filtering and Waveform Metrics Selection

Our dataset was created by combining the Level 2A and 2B GEDI datasets, then selecting only high-quality GEDI observations and finally, the most relevant metrics that can readily contribute to different land surface applications. We considered the first four years of data, i.e., acquired between March 25th, 2019 and March 15th, 2023.

We used the unique GEDI shot (point) identifier to join the Level 2A and Level 2B metrics. Then, we excluded GEDI points with a quality flag of 0 and a degradation flag greater than 0, or points with a sensitivity (the maximum canopy cover that can be penetrated) of less than 0.95. The typical sensitivity threshold values used to filter out low-quality GEDI observations in vegetation and terrain studies range from 0.90 to 0.99 [8, 9, 10, 11]. Our threshold is positioned in the middle of that range, providing not only high-quality points but making them more readily accessible to users via the cloud. To further constrain the dataset size, we made a selection of the relevant waveform metrics for land surface applications, such as RH percentile values for different processing algorithms, plant area index, different waveform energy components, etc. These metrics allow users to perform both further data filtering and modeling. A detailed list of all selected metrics and their description can be found in our GitHub repository (`github.com/Open-Earth-Monitor/GlobalEarthPoint`).

2.3. Spatio-temporal structuring

Originally GEDI data are structured in an orbit-oriented order, with each file representing a sequential time period and covering a wide spatial extent. In land surface applications, spatial blocking, global-local calibration, and local validation are commonly used for mapping and assessment. For these purposes, accessing raw GEDI data typically requires reading large volumes of data and applying subsequent filters to extract information for a specific area and time period. To enable easier access and reduce data transmission overhead, we constructed a spatio-temporally partitioned GEDI dataset optimized for efficient point queries.

Spatio-temporal structuring is carried out in two main steps: (1) streaming data from orbit-oriented raw HDF files into a spatio-temporal block structure, and (2) aggregating each partition into a single GeoParquet file (Fig 2.3). We processed each raw HDF file by filtering and selecting (mentioned in Section 2.2), using a predefined spatio-temporal block (5°×5°×year). The remaining points were then partitioned into small subsets, each stored locally according to its corresponding spatio-temporal block. Finally, these subset files–originating from multiple sources but belonging to the same block–were aggregated and converted into a single GeoParquet file.

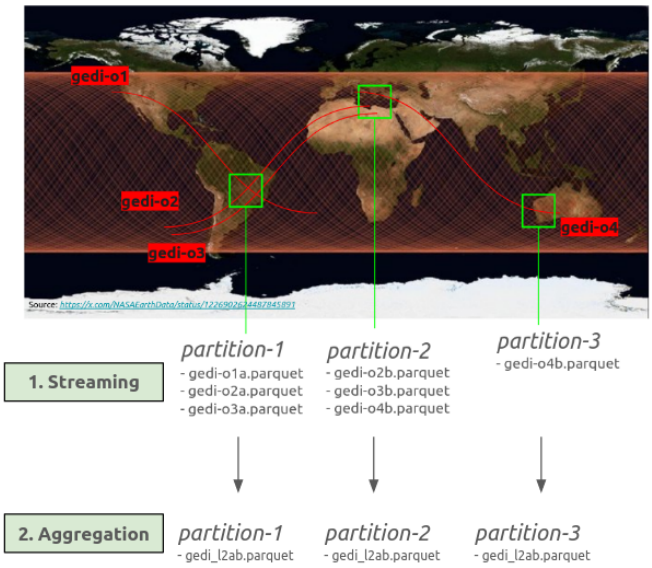


Fig. 1. Illustration of GEDI structuring achieved by partitioning and parallelization.

2.4. Data access and visualization

In order to increase interoperability, we integrated Spatio-Temporal Asset Catalogs (STAC). At the collection level, we provide collection-level metadata, including licensing information, available attributes, DOI, and a reference to STAC items for spatial querying. At the item level – representing individual partitions – we include S3 links to the corresponding GeoParquet files, along with overviews and metadata such as file size, point count, temporal coverage, and more. The STAC catalog is used to organize the partitioned data and serves as a single entry point for the dataset. Additionally, we provide an example of accessing through OpenEO [13], enabling fully cloud-based processing of the dataset.

3. RESULT

3.1. Structuring: data streaming and aggregation

Table 1 summarizes the comparison between the LC DAAC GEDI02 dataset and OLM-GEDI. The raw GEDI Level 2 dataset is composed of 2A (101.8 TB) and 2B (22.0 TB). We streamed the entire dataset using 960 CPUs across 10 servers, each equipped with 1 TB of RAM and connected to local storage servers with InfiniBand. The computing of data streaming and aggregation took 40 hours and 6 hours respectively, under full parallelization. The filtered GEDI dataset in GeoParquet format has an approximate size of 759 GB and contains 5.4 billions points.

doi:10.2760/2119408

Table 1. Summary of raw GEDI and OpenLandMap GEDI.

Product Name	**LP DAAC GEDI02**	**OLM-GEDI**
Time Coverage	03.25.2019 ∼ 03.15.2023	
Data Size	- Level 2A: 101.8 TB - Level 2B: 22.0 TB	Level 2 A&B: 759 GB
Data Format	Hierarchical Data Format (HDF)	GeoParquet

3.2. Benchmark for data access

The result (Fig 2) shows the relationship between the query area and query time for our dataset. We simulated queries of varying area sizes and locations, repeating each configuration 100 times. Each query consists of two components: (1) spatio-temporal filtering at the partition level using STAC, and (2) lazy loading at the file level using `DUCKDB`. For areas smaller than 50,000 km^2, query times fluctuate below 20 seconds. As the query area increases, the time grows accordingly. For areas larger than 3 millions km^2, query time ranges from 60 to 100 seconds, approximately the size of India.

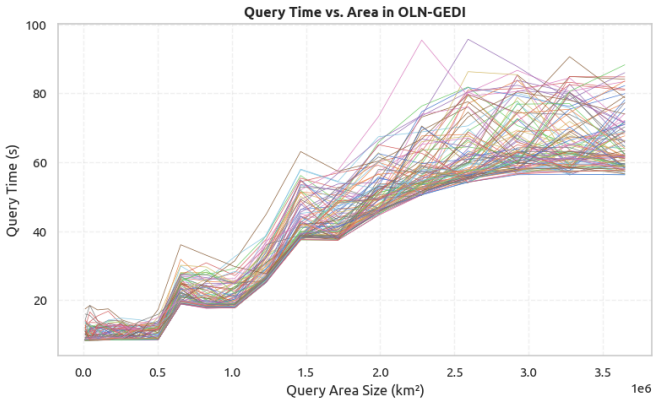


Fig. 2. Benchmark for OLM-GEDI query time versus query area size.

3.3. Use case: Streaming OLM-GEDI into an openEO Cloud-Processing Workflow

The OLM-GEDI collection can be accessed through openEO via the `load_url` process. The STAC collection (Fig 3.3) provides the metadata and easy access to the partitioned data set.

We created a notebook that demonstrates how to access and manipulate OLM-GEDI [14] in a cloud environment. Figure 3.3 illustrates the workflow of loading and overlaying OLM-GEDI and a global DEM in a given bounding box, established in openEO.

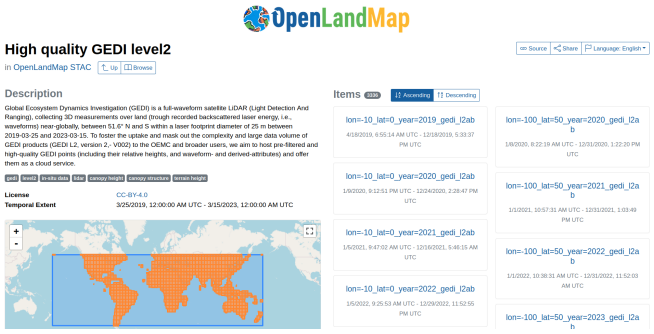


Fig. 3. STAC collection of the OLM-GEDI dataset at OpenLandMap STAC.

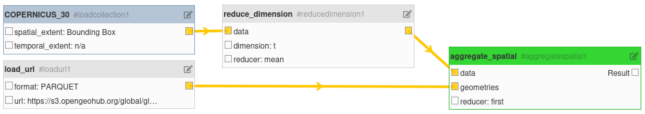


Fig. 4. openEO process graph for loading data sources and extracting raster values.

4. CONCLUSION

This study presented OpenLandMap GEDI (OLM-GEDI), a cloud-optimized, quality-filtered global GEDI point dataset. The proposed filtering methods, metric selection, and spatio-temporal structuring reduce complexity and provide an efficient means of accessing large volumes of GEDI data. The results show that this form of GEDI data partitioning–when exposed through a STAC catalog and stored in GeoParquet format–enables rapid cloud-based access and straightforward loading in a local computing environment, requiring less than 2 minutes for an area of 3 millions km^2. We also demonstrated that accessing the data through STAC on a cloud platform such as openEO is equally straightforward. OLM-GEDI can be readily applied to various land surface applications and improve research reproducibility.

5. DATA AND CODE AVAILABILITY

Data is openly available through STAC (`stac.openlandmap.org/GEDI02/collection.json`) under the Creative Commons Attribution CC-BY 4.0 license. All code to process raw data, establish STAC, benchmark is openly available in Open-Earth-Monitor GitHub (`github.com/Open-Earth-Monitor/GlobalEarthPoint`) under Apache-2.0 license.

doi:10.2760/2119408

REFERENCES

[1] Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., Luthcke, S. (2020). GEDI L2A Elevation and Height Metrics Data Global Footprint Level V001 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi: https://doi.org/10.5067/GEDI/GEDI02_A.001

[2] Magruder, L., Neumann, T., & Kurtz, N. (2021). ICESat-2 early mission synopsis and observatory performance. Earth and Space Science, 8(5), e2020EA001555. doi: https://doi.org/10.1029/2020EA001555

[3] Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., ... & Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sensing of Environment, 253, 112165. doi:https://doi.org/10.1016/j.rse.2020.112165

[4] Lang, N., Jetz, W., Schindler, K., & Wegner, J. D. (2023). A high-resolution canopy height model of the Earth. Nature Ecology & Evolution, 7(11), 1778-1789. doi: https://doi.org/10.1038/s41559-024-02371-2

[5] Ho, Y. F., Grohmann, C. H., Lindsay, J., Reuter, H. I., Parente, L., Witjes, M., & Hengl, T. (2025). Global Ensemble Digital Terrain modeling and parametrization at 30 m resolution (GEDTM30): a data fusion approach based on ICESat-2, GEDI and multisource data. doi: https://dx.doi.org/10.21203/rs.3.rs-6280607/v1

[6] Palomino-Ángel, S., Wdowinski, S., & Li, S. (2024). Wetlands water level measurements from the new generation of satellite laser altimeters: Systematic spatial-temporal evaluation of ICESat-2 and GEDI missions over the South Florida Everglades. Water Resources Research, 60(3). doi: https://doi.org/10.1029/2023WR035422

[7] Adam, M., Urbazaev, M., Dubois, C., & Schmullius, C. (2020). Accuracy assessment of GEDI terrain elevation and canopy height estimates in European temperate forests: Influence of environmental and acquisition parameters. Remote Sensing, 12(23), 3948. doi: https://doi.org/10.3390/rs12233948

[8] Milenković, M., Reiche, J., Armston, J., Neuenschwander, A., De Keersmaecker, W., Herold, M., & Verbesselt, J. (2022). Assessing Amazon rainforest regrowth with GEDI and ICESat-2 data. Science of Remote Sensing, 5, 100051. doi: https://doi.org/10.1016/j.srs.2022.100051

[9] Moudrý, V., Prošek, J., Marselis, S., Marešová, J., Šárovcová, E., Gdulová, K., ... & Wild, J. (2024). How to Find Accurate Terrain and Canopy Height GEDI Footprints in Temperate Forests and Grasslands?. Earth and Space Science, 11(10), e2024EA003709. doi: https://doi.org/10.1029/2024EA003709

[10] Pronk, M., Eleveld, M., & Ledoux, H. (2024). Assessing vertical accuracy and spatial coverage of ICESat-2 and GEDI spaceborne lidar for creating global terrain models. Remote Sensing, 16(13), 2259. doi: https://doi.org/10.3390/rs16132259

[11] Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., ... & Silva, C. (2020). The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. Science of remote sensing, 1, 100002. doi: https://doi.org/10.1016/j.srs.2020.100002

[12] Dubayah, R., Tang, H., Armston, J., Luthcke, S., Hofton, M., Blair, J. (2021). GEDI L2B Canopy Cover and Vertical Profile Metrics Data Global Footprint Level V002 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi: https://doi.org/10.5067/GEDI/GEDI02_B.002

[13] Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., ... & Reiche, J. (2021). The openeo api–harmonising the use of earth observation cloud services using virtual data cube functionalities. Remote Sensing, 13(6), 1125. doi: https://doi.org/10.3390/rs13061125

[14] Heisig, J., Ho, Y. F., & Milenković, M. (2025). Jupyter Notebook: Streaming GEDI Data into an openEO Cloud-Processing Workflow. Zenodo. doi: https://doi.org/10.5281/zenodo.15309339

BREAKING THE BOUNDARIES OF EARTH OBSERVATION – COPERNICUS DATA SPACE ECOSYSTEM AND THE CLOUD COMPUTING PARADIGM

András Zlinszky (1), Jan Musiał (2), Jurry de la Mar (3)

1: Sinergise Solutions GmbH; 2: CloudFerro S. A; 2: T-Systems GmbH.

ABSTRACT

Open satellite imagery from the Copernicus Program has the potential to address societal and environmental crises. However, data access and processing limitations, along with a lack of standards, have hindered the development of operational solutions. Copernicus Data Space Ecosystem (CDSE) marks a shift in Earth Observation (EO) by providing API (Application Programming Interface) access, overcoming the limitations of downloading and local processing. By moving data processing to the cloud, user demands for bandwidth, storage, and processing are reduced. Users can focus on data processing, since management and visualization are handled by the system. The API approach supports FAIR (Findable, Accessible, Interoperable, Reuseable) principles, enhancing the transferability of EO processing pipelines and simplifying learning. Additionally, Level 3 analysis ready data (ARD) and pre-trained models in CDSE facilitate the use of deep learning workflows. Cloud computing, integrating optimized data storage, server-side processing, and virtual machine capacity, is transforming Earth Observation.

Index Terms— Cloud Computing, Big Data, API access, code sharing, data infrastructures

1. INTRODUCTION

Free, public, global scale earth observation data has been available since the release of the Landsat archive in 2008. However, although more than 15 years have passed and the ambitions Copernicus Program has also been launched, global challenges such as deforestation monitoring and urban planning are still information starved. The European Union Common Agricultural Policy (CAP) monitoring (1) has proven that satellite data can support continental scale crop monitoring and classification, but operational applications at similar scales are rare. The reason for this is complex, but as stated in the Earth Observation for Sustainable Development Goals Compendium from 2020, "The main limitation now is not if EO data exists but where it can be stored, accessed and in a format ready to be used" (2).

The Copernicus Open Science Hub (predecessor of CDSE) has been instrumental in providing access to vast amounts of EO data, but required considerable skills to handle, did not provide instant access to most of the archive, and mainly focused on being a source for downloading data. As a result, users without access to high performance computing facilities or with limited knowledge of EO analytics coding were left behind.

Many of these limitations are solved by the cloud processing paradigm (3). If the data is stored on an accessible cloud, and API requests support data processing, the user does not need to download data or process it on their own infrastructure. The Copernicus Data Space Ecosystem, the data gateway and cloud platform of the Copernicus Program was designed according to this concept. However, CDSE goes further by offering immediate access to all Sentinel datasets, in-code access with APIs, an open codebase, an integrated processing infrastructure, and free quota to support the uptake (Fig.1) This evolution has been fostered by advancements in cloud computing, which have broken down barriers in processing capacity, data storage, and algorithm sharing. The result is substantially faster computing – eg. a benchmark test by Ray (4) showed that using API processing the download and processing time for an agricultural use case of can be reduced by a factor of more than 100.

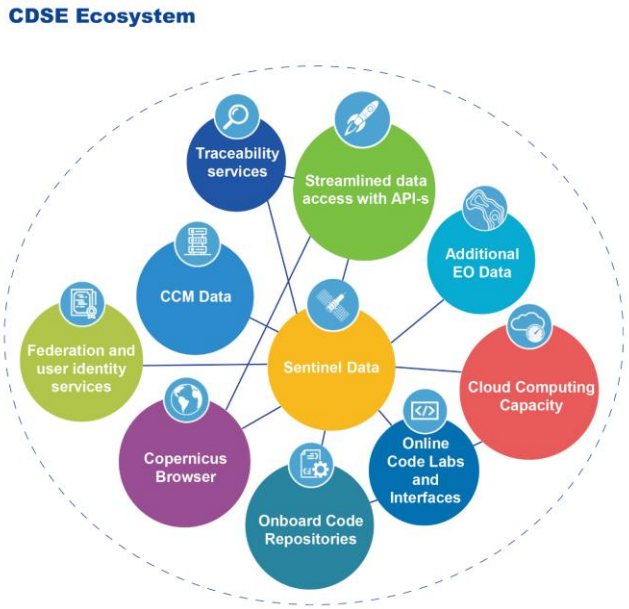


FIGURE 1: OVERVIEW OF CDSE FUNCTIONALITY

doi:10.2760/2119408

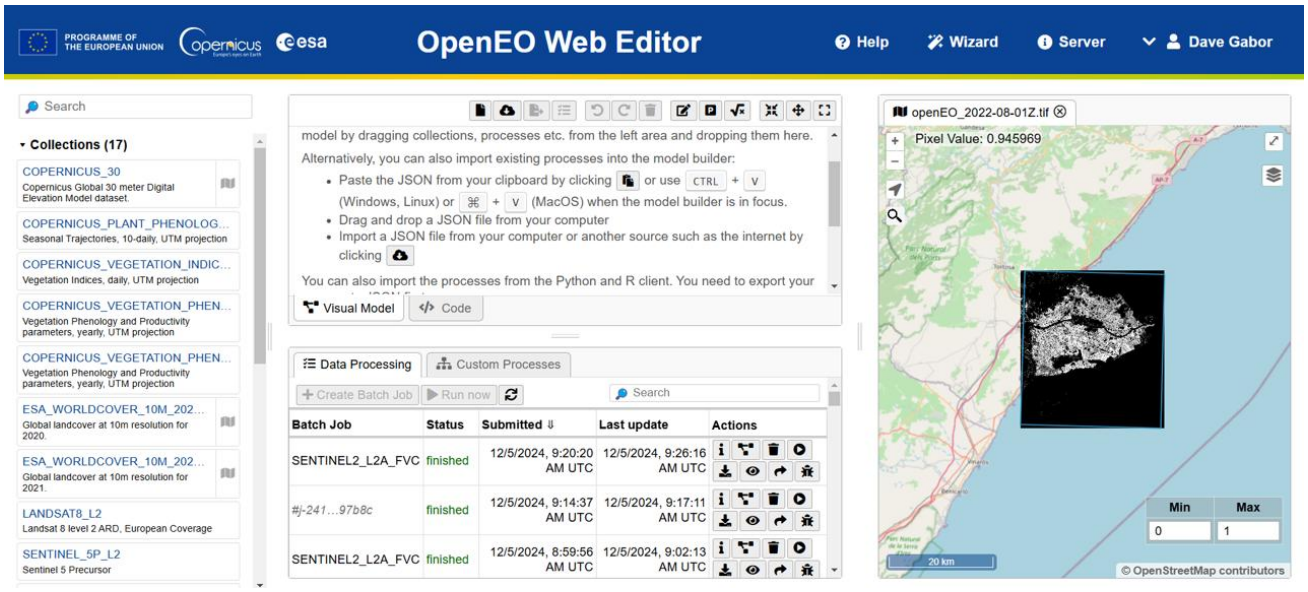


FIGURE 2: EXAMPLE OF THE OPENEO WEB EDITOR, A GRAPHICAL INTERFACE TO CDSE DATA COLLECTIONS AND PROCESSING ALGORITHMS

Here two spectral indices were calculated and a threshold applied within a Sentinel Hub Statistical API request, and compared to image tile downloading and local processing with Odata and processed with Rasterio. The full code for the benchmark is available here: https://github.com/eu-cdse/notebook-samples/blob/main/sentinelhub/EGU_notebook.ipynb

2. CLOUD COMPUTING AS A TRANSFORMATIVE SOLUTION

Cloud computing has revolutionized EO by enabling on-demand data processing through streaming-compatible data formats and optimized API requests. Transferring Terabytes of data over the web is still a cumbersome task. The legacy approach is to break the data into pre-defined chunks (tiles) and send them as a compressed archive. The new approach is to query only the data within the user's area of interest without loading the full file into the memory, do the processing on the cloud server, and transfer only the result to the user. This eliminates the need for powerful local computers, as the processing power resides in the cloud platform. DIAS (Data Information Access Services) infrastructures have already adopted this paradigm several years ago but providing it as a public service with free quota in CDSE has created the necessary critical mass for it to become an industry standard. Meanwhile, CDSE is not a commercial platform and does not compete with commercial providers. Public institutions are provided with large individual quotas, but large-scale commercial processing should still be done with commercial platforms. However, CDSE supports prototyping and learning Copernicus data and tools, and provides regional-scale processing capacity as a resource for starting.

3. SIMPLIFYING EO CODING

Nearly 80% of EARSC member companies have reported that they have difficulties in finding suitable candidates for filling positions (4). One of the main reasons behind this lack of specialists is the challenge of learning programming of the EO data processing workflows. Mastering the processing and analysis of satellite imagery requires a wide range of skills that take a long time to acquire. This is also difficult to achieve in a sequential manner: in order to carry out the simple task of requesting an image from a repository, applying a spectral index and visualising it online, a developer already has to understand big data management, data formats and projections, optical analysis, and web GIS. If we add to this the need for understanding the application domain, the requirements are nearly impossible to fulfil for someone at the early stage of their career. What is needed is a system for data access and visualisation that is easy enough to handle for domain specialists (ecologists, social scientists, agriculture experts etc.) but powerful enough to work on global scale, - and a standard-ready codebase that can be learned incrementally.

The CDSE backend, frontend, and codebase are providing exactly this, redefining the learning curve for EO coding. By handling data management in the repository and visualization in the Browser, the CDSE allows users to focus on the actual task of data processing. The integration of coding and visualization within the Copernicus Browser or openEO Web Editor (Fig.2), along with a comprehensive tutorial base (including user-contributed examples), makes EO coding

doi:10.2760/2119408

accessible to the masses. Hackathons for high school students have shown that EO data analysis is now within reach of non-specialists using these tools (5). For advanced users, CDSE offers onboarding of the custom algorithms via the openEO Algorithm Plaza, Sentinel Hub eval scripts, and JupyterLab environment (Python, R, Julia, ESA-Snap kernels supported - access date 17.06.2025). Moreover, JupyterLab offers Linux command line where users can upload and execute their own binaries. Currently, the support for onboarding custom Docker containers is not supported.

4. DATA INTEGRATION AND INTEROPERABILITY

We will never have one data repository to rule them all: different datasets require different infrastructure, and the dataset owners will always have the preference to keep data where they can control it (6). Therefore, it is necessary to operate cloud processing solutions that can integrate data from different backends while still running the processing in the cloud and not on the user's machine. OpenEO API processing and STAC (7) facilitate data integration across multiple backends, enabling seamless access and integrated processing of data from various cloud provider backends without the need for local downloads. These tools are designed specifically as a standard that can connect many data sources, and since they are open and community owned, using them does not create a competitive disadvantage for industry players. As large global datasets become interoperable, the stage is set for new, innovative applications for sustainability.

5. ADVANCING DEEP LEARNING AND AI

Deep learning for image analysis has also been available for 10 years now. Still, operational applications for earth observation are not widely seen. The main difficulty for machine learning is often the cost of data wrangling - preparing noisy, inconsistent datasets to a format compatible with the learning models.

CDSE hosts machine learning-ready data, including Sentinel-1, Sentinel-2 Level 3 mosaics and Sentinel-3 Level-2 data, significantly reducing the data preparation phase for machine learning applications. Additionally, most global datasets of the Copernicus Land Monitoring Service (CLMS) are now also served in CDSE, providing an additional layer of analysis-ready datasets, representing land cover and bio-physical quantities such as soil moisture, vegetation status, lake water quality and surface temperature (Fig. 3). These datasets are also served in a unified common structure, enabling their direct analysis with AI tools Furthermore, tools like Major Tom embeddings (8) and the compatibility of openEO with deep learning models from ONNX (9) are accelerating AI processing of Sentinel data. By providing a standard for sharing deep learning models that are compatible with API access to massive datasets, significant obstacles towards the application of AI in satellite data analysis have been removed.

6. STANDARDIZATION AND KNOWLEDGE SHARING

Transferability is always an essential issue in EO: most applications are for managing large areas, so algorithms have to work for different conditions. Standards are a key tool for transferability, but are difficult to establish in the first place. For a standard to work efficiently, it has to be compatible with a wide range of tools and datasets, well documented, actively maintained and most importantly it has to be used by a large community. The tools created in CDSE are rapidly becoming the standard for both open and commercial data and solution providers. CDSE is generating a strong mainstream with a community of more than 400 000 users, providing incentive for the adoption of the cloud computing approach, and specifically the STAC standard and the openEO and Sentinel Hub API-s by industry players and large international agencies.

In fact, according to the EARSC Survey (4), CDSE has emerged as the leading EO data platform for businesses in Europe, despite its public and free-quota nature. Its compatibility with commercial data and solution providers has fostered a thriving ecosystem where data, algorithms, and downstream products converge. Currently, Copernicus Browser has more than 500 000 visitors monthly, and more than 350 million API access requests are processed each month. A real-time interactive public dashboard enables users or prospective partners to follow the evolution of the user base and system performance here https://dashboard.dataspace.copernicus.eu/ . The next step is to grow a federated ecosystem of public and commercial actors based on this common, shared platform.

7. CONCLUSIONS

The transformative shift in Earth Observation is driven by concurrent advancements in data storage (COG/Zarr formats), API development (e.g., openEO), and virtual machine computing (e.g., Jupyter Labs). These technologies combined are providing much stronger impact than they would individually. Coupled with the open data policy of the Copernicus Program and the public virtual machine resources it offers, the new developments in cloud computing and API access have democratized access to location-based insights. The scene is set to overcome the limitations posed by the lack of expertise, computer capacity and data storage. With an internet connection and some very basic knowledge, users can start directly exploring the application of Sentinel data to their field while developing a skillset that leads towards advanced operational applications. As a result, EO data and insights are now within reach of a wider range of users, including environmental NGOs, farmers, and mainstream news sites. Above all, this leads to stronger trust and transparency (10): now, satellite data does not only enable companies and governments to monitor citizen activities but

doi:10.2760/2119408

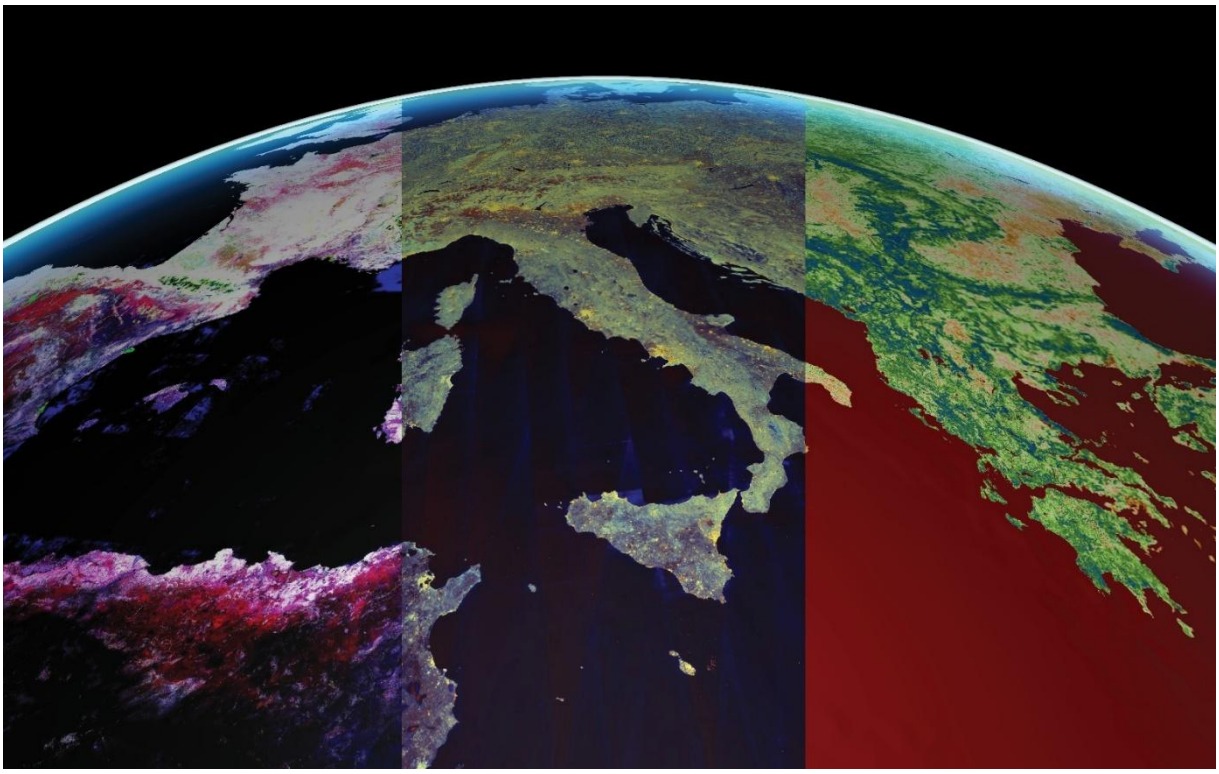


FIGURE 3: AI-READY DATA LAYERS IN CDSE: FROM LEFT TO RIGHT: AGRICULTURAL GROWTH STAGE ON SENTINEL-2 QUARTERLY MOSAICS, SENTINEL-1 MONTHLY MOSAIC, CLMS VEGETATION PRODUCTIVITY

also the other way round, as individual agents have the tools to monitor large industry and public actors and hold them accountable. The CDSE exemplifies the power of open data and collaborative innovation in shaping the future of Earth Observation. It shows how technology-aware public investment can break new ground in the application of earth observation for sustainability, supporting growth in the industry, government and non-profit sectors.

8. REFERENCES

1. **European Commission.** *Commission Implementing Regulation (EU) No 809/2014 of 17 July 2014 Laying Down Rules for the Application of Regulation (EU) No 1306/2013 of the European Parliament and of the Council with Regard to the Integrated Administration and Control System, Rural* . 1306/2013, Brussels, Belgium : s.n., 2013.

2. **O'Connor, Brian, et al.** *Earth Observation for SDG-s: Compendium of Earth Observation contributions to the SDG Targets and Indicators.* Frascati : ESA Publications, 2020.

3. *Federated and reusable processing of Earth observation data.* **Mohr, Matthias, et al.** 194, s.l. : Scientific Data, 2025, Vol. 12.

4. **European Association of Remote Sensing Companies.** *EARSC Industry Survey.* Brussels : EARSC, 2024.

5. *Advancing Students' Climate Literacy: a case from the International Summer School of Remote Sensing in Ukraine.* **Biletska, Mariia, Dovgyi, Stanislav and Babichuk, Svitlana.** Vienna : EGU General Assembly 2025, 2025. EGU25-1586.

6. *How open software, data and platforms are transforming Earth observation data science.* **Wagner, Wolfgang, et al.** Vienna : European Geosciences Union, 2025. EGU25-4155.

7. **STAC Contributors;.** Spatio Temporal Asset Catalog Specificatin. [Online] 2024. [Cited: 05 01, 2025.] https://stacspec.org/en.

8. *Global and Dense Embeddings of Earth: Major TOM floating in latent space.* **Czerkawski, Mikolaj, Kluczek, Marcin and Bojanowski, Jedrzej S.** s.l. : arxiv.org, 2024, Vol. 2412. arxiv:2412.05600v1.

9. **ONNX Developers.** ONNX Runtime. [Online] Open Neural Network Exchange, 2021. [Cited: 05 01, 2025.] https://www.onnxruntime.ai.

10. **European Association of Remote Sensing Companies.** *Sentinel Benefits Study Demonstrating the value of Sentinel data through rigorous value chain analyses and powerful user stories.* Brussels : EARSC, 2024. 4000119743/17/I-SBo.

doi:10.2760/2119408

ENABLING FAIR AND OPEN EARTH SYSTEM SCIENCE WITH EARTHCODE

Deyan Samardzhiev[1], Anca Anghelea[2], Anne Fouilloux[3]

[1]Lampata, [2]European Space Agency, [3]Simula Research Laboratory

ABSTRACT

Openly available Earth Observation (EO) data has significantly advanced climate and Earth System science; however, collaboration around the resulting downstream research datasets remains a challenge. Despite the growing focus on creating open science data catalogues across various institutions, the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) practices remains fragmented. EarthCODE aims to overcome these challenges by providing researchers in Earth system science with the tools and means to practice FAIR and Open Science. Promoting interoperability and coordination among integrated European EO platforms, EarthCODE enables scientists to discover, develop, execute, and publish data and workflows aiming to maximise reusability. This paper illustrates the strategy adopted in the federated ecosystem of EarthCODE to go beyond openness, striving to make research open, FAIR, and reproducible.

Index Terms— FAIR, Open Science, Earth Observation, Earth System Science, Cloud Computing, Reproducibility.

1. INTRODUCTION

Earth Observations are essential to advancing Earth System science and enabling innovation. Missions such as ESA's Earth Explorers and the EU's Copernicus Programme generate vast, openly accessible datasets—crucial for scientific progress. Yet these data alone are not sufficient to drive science-based action. Gaps remain in our ability to quantify global cycles, identify tipping points, and assess system resilience, all of which require large-scale coordination and knowledge exchange between researchers.

Open Science has emerged as a key enabler of sustained collaboration, supported by global programmes [1],[2] and reinforced through international policy guidance [3],[4]. This momentum comes at the time of the cloud revolution that now allows researchers to access and analyze data at scale—on demand and directly in cloud-native EO platforms.

Together, Open Science and EO platforms create the opportunity for the infrastructure required to fill the scientific knowledge gaps and address the pressing societal challenges of today – a process referred to as "Earth Action". This infrastructure for open science empowers scientists to 1. Access and process satellite and in-situ data in collaborative cloud environments, 2. Develop and publish reusable code and workflows, 3. Validate outputs and share reproducible results, and 4. Collaborate across institutional, disciplinary, and national boundaries.

ESA's vision for EO Open Science and Innovation [5] captures this opportunity, providing a structured framework to embed FAIR and Open Science practices across its Earth Observation activities. EarthCODE is part of this larger panorama of strategic initiatives. Through its ecosystem of tools and platforms, it aims to transform FAIR and Open principles from an aspiration to routine practice for Earth Science activities funded through its programme and beyond (e.g., including collaborations with EU funded research).

The EO landscape includes several major initiatives addressing open data and computational challenges. The European Open Science Cloud (EOSC) offers cross-domain infrastructure, NASA Earthdata and Copernicus Data Space provide data access, and Pangeo delivers computational environments. Yet none fully address the need for automated, portable FAIR workflows across federated EO platforms.

At a high-level, EarthCODE enables scientists to find and reuse research data, use integrated EO platforms to develop scientific workflows, and publish them by automating the "FAIRification" process. In a first stage, the key EarthCODE stakeholder groups include the activities contributing to the various ESA Science Clusters, and the Earth System Science Hub.

2. CHALLENGES

Still, on the path to implementing the vision there are numerous interlinked, complex challenges that need to be overcome. The measure–understand–predict–decide–act cycle in EO critically depends on digital research objects—data, software, workflows, models, and services—being FAIR. Despite the growing focus on creating open science data catalogues from various institutions, FAIR implementation remains fragmented.

EO datasets are often cataloged but isolated in system-specific silos with inconsistent metadata and weak links to tools and workflows. Tools like PySTAC and stactools help improve metadata generation, but require platform-specific adaptations, limiting automation and discoverability. Although open APIs and cloud services have improved data accessibility, challenges persist due to inconsistent authentication, documentation gaps, and unclear licensing.

doi:10.2760/2119408

Interoperability remains a barrier due to heterogeneous formats, metadata, and execution environments; even FAIR components remain incompatible without shared APIs, containers, or workflow engines. Reusability is limited by missing provenance, weak documentation, and absent best practices (semantic versioning, persistent identifiers), undermining reproducibility.

Despite progress on data accessibility, workflow metadata still lags; FAIR community standards only recently converged on machine-readable metadata [10] with standards for explicit provenance, standard packaging, and linking code, inputs, and configurations into discoverable, citable, executable objects.

Implementing Open Science principles in the cloud era introduces additional technical complexities. Traditional paradigms assume "one-click" retrieval, but high-resolution global datasets from cloud-optimized EO platforms are too large for cost-effective download and separating them from native infrastructure undermines reproducibility and reusability.

Moving compute next to data solves this, but requires orchestration across diverse data infrastructures, each with unique storage layouts, metadata schemas, access protocols, and authorization. Additionally, EO workflows are often tightly coupled to specific data, infrastructure, and execution environments, limiting reuse and reproducibility. Although most EO cloud platforms support STAC for data discovery, their compute interfaces remain highly heterogeneous. This fragmentation leads to a lock-in effect: code and pipelines built for one platform often require substantial modification to function on another's infrastructure or to be reused, forcing scientists to repeatedly reengineer workflows rather than advancing science.

Existing efforts only partially address the problem, for example, the EOSC provides general FAIR guidance but lacks EO-specific automation, DestinE delivers powerful compute yet prioritizes predefined operational models over open research workflows. The lack of unified FAIR automation, platform federation, and workflow portability hinders reproducible Earth System Science.

3. THE EARTHCODE ECOSYSTEM FOR FAIR AND OPEN SCIENCE

EarthCODE overcomes these challenges by providing scientists with accessible tools and guidelines to practice FAIR & Open Science. It promotes coordination among various EO cloud providers to enable portable and reproducible science across a federation of platforms by using open standards. The federated ecosystem of EarthCODE goes far beyond openness, it strives to be open, FAIR, and reproducible.

Unlike existing initiatives, EarthCODE targets automated, cross-platform FAIR workflows via: (1) true federation across EO platforms with SSO and cross-execution of workflows; (2) FAIRification automation during research (not post-hoc); (3) end-to-end reproducibility linking data, experiments, and workflows; and (4) sponsored compute to lower barriers—bridging open data and reproducible research.

EarthCODE provides a central portal which serves as the single-entry point for accessing all services. Through the portal, users engage with a federation of EO cloud platforms (Fig. 1) which provide standardized capabilities, access to EO data on the platforms, tools for developing and executing workflows, automated publication mechanisms to the EarthCODE catalog, and the ability to run published experiments, across platforms. ESA provides sponsorship to computing and storage resources on these platforms via the Network of Resources (NoR) for eligible projects. Users can also use their own local environments to access, reuse and publish to EarthCODE.

The Open Science Catalog (OSC) is EarthCODE's central platform for publishing, discovering, and accessing EO data, workflows and documentation. The OSC leverages open-source geospatial technologies like stac-browser, pycsw, PySTAC, and OpenLayers and tries to contribute back to these projects in terms of software and standardization [6].

Finally, the EarthCODE Discourse forum is a space for users to openly engage in dialog about Earth Sciences, with the ability to contact the authors of published results.

4. BUILDING EARTHCODE WITH OPEN-SOURCE

EarthCODE relies on the services provided by already mature EO Platforms and cloud computing infrastructures currently on offer on the global EO market. the initial batch of integrated platforms, EarthCODE has selected the CDSE openEO federation, Euro Data Cube and DeepESDL platforms, alongside others selected in 2025 which are not publicly announced at the time of writing this paper.

Platform providers play a key role in the EarthCODE ecosystem (Fig. 1) by either providing the FAIR Open Science environments to create scientific workflows, discover and reuse data, and publish to the catalog or by providing integrated infrastructure to run or reproduce these workflows at scale close to hosted EO data using the open, standardized metadata describing the workflows on the OSC.

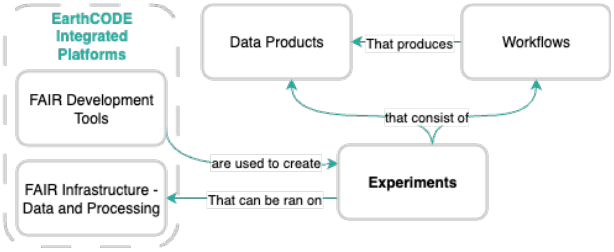


Fig. 1. FAIR Tools Platforms for development and Infrastructure Platforms as execution environments.

doi:10.2760/2119408

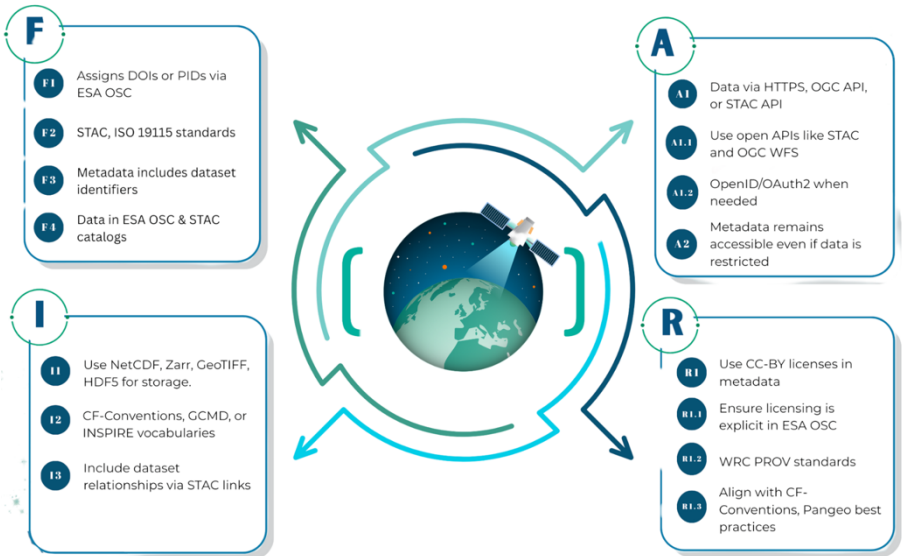


Fig. 3. EarthCODE FAIR Principles

EarthCODE significantly advances FAIRness and federation across platforms by leveraging open standards and protocols to improve cross-platform workflow execution and data reusability. This is a major step forward—particularly in addressing interoperability, a key challenge for researchers who are typically familiar with only one environment and struggle to reuse their work across others. However, further progress is still needed to ensure metadata itself meets FAIR standards consistently across platforms.

In EarthCODE, scientific output data are referred to as Products and are described in rich STAC medatada. The STAC metadata is hosted on the OSC, while the data are hosted either in the ESA Project Results Repository [9] (PRR) or in external persistent repositories. Each product includes a link to the dataset representing measured or derived variables, links to supporting documentation, and metadata capturing mission provenance, project affiliation, and thematic classification (Fig. 2). Shared taxonomies and metadata standards ensure products are findable and interoperable across platforms and domains.

Crucially, a product in EarthCODE is not an isolated artifact but is explicitly linked to an Experiment—the structured execution context that produced it. Experiments capture human-readable descriptions, machine-executable workflows, input datasets (referenced with persistent identifiers), and runtime configurations described as OGC API - Records. This complete metadata chain enables reproducibility and transparency, supporting both manual re-execution and machine-driven automation.

Workflows within EarthCODE are formalized as executable objects compatible with integrated platforms, leveraging open standards such as openEO Process Graphs, OGC API Processes (e.g., CWL, Application Packages). Platforms are then responsible for ensuring the means of execution of workflows based on the metadata defined in the experiment or workflow. Source code may be referenced, but workflows themselves must be described in an executable, platform-independent manner to guarantee cross-platform compatibility.

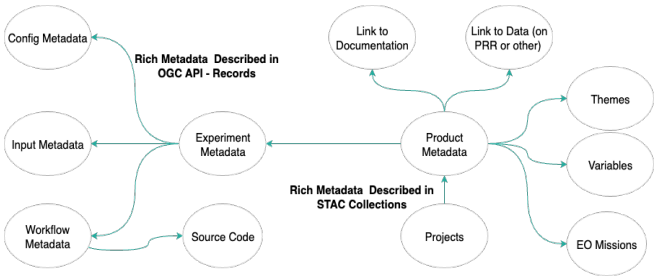


Fig. 2. EarthCODE Metadata Standards

EarthCODE implements FAIR principles for both workflows [10] and research data [11] by following community guidelines. In this context, EarthCODE adopts the broad definition of "research data" inspired by the Beijing Declaration on Research Data [12], encompassing not only datasets, but also metadata, software, methods, algorithms, and related documentation.

4.1. A FAIR Example

To demonstrate how EarthCODE implements FAIR we walk through an example experiment published in the OSC and the principles it covers [10], [11] as shown on Fig. 3.

Note that EarthCODE is under development during the time of writing of this paper. The example below generates cropland and crop type maps from input Sentinel-1 and Sentinel-2 data [13]. The experiment metadata of this example [14] is described as an OGC API Record (F2, I1), with rich metadata describing its input parameters (F2, I4),

doi:10.2760/2119408

configuration options (F2, I4, R1.3), and the infrastructures where it can be executed (F2, I1, I4), in this case, the CDSE openEO federation. It is further enriched with the EarthCODE vocabulary (I2), e.g., research themes for land.

The experiment links to the corresponding WorldCereal Workflow [16]. It is an OGC API Record (F2, F3), which represents versioned OpenEO Process Definition (I1) that reads specific inputs and configurations (I3). The workflow also links to the code which is stored on Github (F1.2). Metadata describes the infrastructure endpoints where it can be executed and thematic details (R1, I1, I2), in this case, the CDSE OpenEO Federation. Workflow components, options for execution environment infrastructure, license and provenance details are fully recorded (R1.3).

The results from the execution of the experiment are a data product ([15]) - a STAC Collection (F2, I1) enriched with EarthCODE taxonomy elements (Themes, Variables, EO Missions) (I2). It explicitly includes standardized references to the dataset and its components (F3). Qualified references to related datasets and workflows are included (I3), and standardized EO formats (COG/TIFF) ensure scalable access (A1.1). Licensing and provenance are recorded using open, standardized practices (R1.1, R1.2), linking to Sentinel data used for the analysis. All metadata aligns with community standards widely adopted in Earth Observation (R1.3) such as STAC.

All metadata including input, configuration, workflows, experiments and products on the catalog are assigned persistent, globally unique identifiers (F1, F1.1) and are indexed and searchable through the Open Science Catalog (F4), held separately from the data (A2). The catalog exposes data via open, standardized protocols such as HTTPS or STAC API (I.1, A1, A1.1, A1.2).

Together, the WorldCereal resources form a fully FAIR research object chain, where data, infrastructure, methods, and outputs remain findable (F1–F4), accessible (A1–A2), interoperable (I1–I4), and reusable (R1–R3) across platforms, infrastructures, and research communities. This metadata is automatically generated and published via integrated platforms.

5. CONCLUSION

EarthCODE is developed by implementing three workstreams, as follows: Infrastructure (WS1), FAIR Open Science (WS2), and Community Engagement (WS3), through an open competition under Best Practice procurement, renewed annually. To ensure long-term sustainability, EarthCODE is embedded in ESA's multi-year FutureEO Programme. In EarthCODE's operating model ESA provides governance, reference architecture and decision making, while platform partners commit to maintaining open-source platforms and building blocks under a shared-services agreement. The current phase focuses on initial platform integration, publication of experiments and outputs to the Open Science Catalog, and dissemination of scientific results. Phase 2 will focus on deeper interoperability using ESA's EOEPCA+ building blocks. Science stakeholders directly shape the evolution of the ecosystem as a FAIR Collaborative and Open Development Environment for Earth System Science.

REFERENCES

[1] AGU, "Open Science," AGU, 2024. [Online]. Available: https://www.agu.org/learn-about-agu/about-agu/open-science. [Accessed: 01-Jul-2024].

[2] K. Murphy, "Open-Source Science: The NASA Earth Science Perspective," The Earth Observer, vol. 33, no. 5, pp. 4–9, Sep.–Oct. 2021.

[3] European Union, "EU Open Science Policy," 2020. [Online]. Available:https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science. [Accessed: 14-Jul-2023].

[4] UNESCO, "An Introduction to the UNESCO Recommendation on Open Science," Canadian Commission for UNESCO, 2022. [Online]. Available: https://doi.org/10.54677/XOIR1696.

[5] European Space Agency (ESA), Earth Science in Action for Tomorrow's World: Earth Observation Science Strategy, ESA, Oct. 2024, 42 pp.

[6] F. Schindler, S. Pari, S. Meissl, G. Smith, E. Dobrowolska, and A. Anghelea, "Open Science Data Catalogue," Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLVIII-1/W2-2023, pp. 997–1003, 2023. doi:10.5194/isprs-archives-XLVIII-1-W2-2023-997-2023.

[7] Open Geospatial Consortium (OGC), "Open Science Persistent Demonstrator (OSPD) Initiative: Call for Participation," ver. 1.3, Nov. 2023. [Online]. Available: https://portal.ogc.org/files/?artifact_id=106421. [Accessed: 01-Jul-2024].

[8] EOEPCA, "Master System Design Document: EOEPCA.SDD.001," 2021. [Online]. Available: https://eoepca.github.io/master-system-design/current/.

[9] European Space Agency (ESA), "ESA Project Results Repository: Data Provider Guide," [Online]. Available: https://eoresults.esa.int/data_provider_guide.html.

[10] S. R. Wilkinson, et al., "Applying the FAIR Principles to Computational Workflows," Sci. Data, vol. 11, no. 1, 2025. doi:10.1038/s41597-025-04451-9.

[11] M. D. Wilkinson, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," Sci. Data, vol. 3, 160018, 2016. doi:10.1038/sdata.2016.18.

[12] CODATA, "The Beijing Declaration on Research Data," Committee on Data of the International Science Council, 2019. doi:10.5281/zenodo.3552330.

[13] K. Van Tricht et al., "WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping," Earth Syst. Sci. Data, vol. 15, no. 12, pp. 5491–5515, Dec. 2023, doi: 10.5194/essd-15-5491-2023.

[14] Open Science Catalog, "WorldCereal Experiment," [Online]. Available:https://opensciencedata.esa.int/experiments/worldcereal-experiment/record.

[15] Open Science Catalog "WorldCereal Crop Extent – Belgium" https://opensciencedata.esa.int/products/worldcereal-crop-extent-belgium/collection

[16] Open Science Catalog "ESA worldcereal global crop extent detector" https://opensciencedata.esa.int/workflows/worldcereal-workflow/record

GEO-OPEN-HACK: AN INITIATIVE ON BIG GEOSPATIAL DATA PROCESSING WITH OPEN COMPUTING INFRASTRUCTURE AND OPEN TOOLS

Milutin Milenković[1], *Giuseppe Amatulli*[2], *Tushar Sethi*[2], *Raymond Oonk*[3], *Anne Fouilloux*[4], *Tina Odaka*[5], Michele Claus[6], Valentina Premier[6], Antonio Fonseca[2], Pieter Kempeneers[7], Francesco P. Lovergine[8], Leandro Parente[9], Yu-Feng Ho[9], Johannes Heisig[10], Dainius Masiliunas[11], Wolfgang Wagner[12], Edzer Pebesma[10], Patrick Griffiths[13], Jan Verbesselt[14], Tom Hengl[9], Ian McCallum[1], Steffen Fritz[1]

[1]IIASA,Austria, [2]Yale University USA, [3]SURF, The Netherlands [4]Simula, Norway, [5]Ifremer, France, [6]Eurac, Italy, [7]JRC, Italy, [8]CNR IREA, Italy, [9]OpenGeoHub, The Netherlands, [10]Uni Münster, Germany [11]Wageningen Uni, The Netherlands, [12]TU Vienna & EODC, Austria, [13]ESA, Italy, [14]BELSPO, Belgium

ABSTRACT

To maximize the potential of big geospatial data streams, they shall be processed on high-performance computing (HPC) infrastructure, whereas local code shall be adopted to use this infrastructure efficiently. This paper presents a non-profit initiative, GEO-OPEN-HACK, that aims at lowering the technical barrier for environmental scientists through one-week hackathons on open-source HPC processing approaches, such as openEO, Pangeo, and OEMC. We present the above solutions, hackathon setup, and analysis based on the pre-event questionnaire about the participants' big data projects.

Index Terms— Geocomputation, Education, HTC

1. INTRODUCTION

A full exploration of big geospatial data, originating mainly from open satellite image archives such as those of the European Commission's Copernicus and the NASA-USGS Landsat programmes, requires both high storage and high computing capacity that no longer fit in a local workstation used for Earth Observation (EO) studies. Thus, studies typically rely on high-performance computing (HPC), which enables parallel processing across multiple processors and nodes, to dramatically accelerate their data processing [6, 5].

HPC also comes with technical complexity in its storage and computing systems. To address this, several higher-level open-source engines for geospatial workflows have been introduced recently. For example, openEO masks out code parallelization, unifies the interaction with different backends, and also supports processing user-defined functions [7]. Pangeo provides server-side computation via lazy evaluation, relying on Python tools (xarray) to handle multidimensional datasets and Dask for parallelization [1]. The Open Earth Monitor Cyberinfrastructure (OEMC) project is another open initiative that supports existing solutions with additional open tools and data.

Several hackathons, workshops, and summer schools feature Geospatial data processing with open tools. Spatial Ecology is organizing an annual event on geocomputation and machine learning (ML) for environmental applications. OpenGeoHub Foundation is also organizing yearly summer schools on data science for EO. Those events are either on the introductory or intermediate level, with a primary focus on building open geospatial workflows. At the last BiDS2023 conference, an openEO and Pangeo tutorial was given, but rather as a half-day sprint. Thus, there is a need for an event where intermediate- and advanced-level participants can learn about different HPC approaches and look at them through the prism of their use cases.

This paper presents the GEO-OPEN-HACK, a non-profit initiative focused on transferring knowledge about the advanced processing of big spatial data using open tools in an HPC environment. We outline the objectives and setup of GEO-OPEN-HACK-2024, and provide an overview of HPC processing with openEO, Pangeo, and OEMC.

2. GEO-OPEN-HACK OBJECTIVES AND SETUP

GEO-OPEN-HACK-2024 was a give-and-take, non-profit initiative introduced under the OEMC project to: (a) lower the barrier and transfer knowledge to users dealing with big geospatial data analytics, i.e., the give part, and (b) foster the co-creation of open OEMC geospatial data and tools through user feedback, i.e., the take part. The event took place from June 24 to 28, 2024, at the International Institute of Applied Systems Analysis (IIASA) in Laxenburg, Austria. The focus was on presenting HPC approaches with different levels of technical complexity so that the participants were enabled to: (a) judge whether their big data problem is suitable for

This work has received funding from the European Union's Horizon Europe research and innovation programme (grant agreement No. 101059548). We thank also SURF (www.surf.nl) for the support in using the Spider cluster.

doi:10.2760/2119408

HPC, (b) determine what would be the most appropriate HPC approach for their big data problem, and (c) process their data on HPC. This was an advanced-level hackathon, ideal for early-career researchers, scientists, and professionals who already had intermediate skills in Linux, bash, Python, and R. A basic familiarity with geoinformation processing and EO was also required. This was a one-week event featuring a keynote speech, a tutorial by teachers, a big-data story by a teacher, and a hands-on session led by participants on their own data. Here is the detailed agenda and links material.

The participants worked on three backends. They mostly worked on the Spider cluster, which is a part of the Dutch national research infrastructure, SURF. The cluster comprises 42 nodes, providing over 3,000 cores with 8 GB RAM and 80 GB local SSD scratch disk space per core. Furthermore, Spider offers PB-scale internal storage with CephFS filesystem and network uplink of 1200 Gbit/s. With openEO and Pangeo, participants also worked with the Copernicus Data Space Ecosystem (CDSE) and the European Open Science Cloud (EOSC) backends.

Fig. 1. The GEO-OPEN-HACK-2024 setup with three different HPC processing approaches and three backbends.

3. GEODATA PROCESSING APPROACHES

3.1. Direct Processing on HPC

During the direct processing on HPC, the participants had to set up an HPC environment for efficient geo-data processing. This involves organizing a logical folder structure, preparing outputs, and utilizing submission scripts, such as sbatch, to efficiently manage tasks . Proper directory and alias setup allows users to monitor jobs, handle errors, and reprocess failed tasks easily. Bash scripting automates these processes, ensuring repeatability and minimising human error. Raster processing, resampling, modelling, and landscape analysis become feasible even for continental-scale studies. Techniques such as splitting datasets, parallel resampling, and running GIS commands in batch are essential. Ultimately, such an HPC setup transforms geospatial analysis from a bottleneck into a streamlined, robust operation.

3.1.1. GDAL, pktools and GRASS GIS on HPC

In the hands-on session, multiple scripting techniques were introduced using GDAL, pktools, and GRASS. These tools provided fast, flexible, and scalable features and functions for raster-based analysis with a Bash command environment. Moreover, parallelization is accomplished by dividing a large geospatial dataset into regular tiles, conducting the operation on each tile with a dedicated CPU, and finally mosaicing them back to the original extent. A concise overview and description of the script used in the hackathon for the above steps are given on the BASH tutorial and the GRASS tutorial pages.

3.1.2. pyjeo and HPC

pyjeo is an open source library for geospatial image processing in Python [4]. It has been implemented in the Joint Research Centre of the European Commission and is distributed under the GPLv3 license. The design of pyjeo is based on a C/C++ implementation with Python bindings that are made available through the Simplified Wrapper and Interface Generator (SWIG [2]). Parallel processing is supported via multithreading using the OpenMP API in C/C++. Additionally, high-throughput computing can be achieved through a tiling mechanism when combined with a job scheduler. Unlike other libraries such as Dask, the tile-based orchestration, including the merging step, is not automated and is not part of the pyjeo library. It also relies on a job scheduler such as HTCondor to launch the individual jobs that process the different tiles. In [3], the authors show different strategies using pyjeo for the parallel processing of geospatial data in a cloud computing infrastructure.

3.2. Pangeo Approach

The Pangeo ecosystem provides an open-source, cloud-, and HPC-ready platform for scalable geoscientific and climate data analysis. Built on modern Python tools such as xarray, Dask, and Zarr, Pangeo facilitates the analysis of multi-dimensional datasets that exceed in-memory capacity. Within the GEO-OPEN-HACK event, Pangeo was introduced as a framework that supports interactive, scalable workflows, embodying FAIR (Findable, Accessible, Interoperable, Reusable) and reproducible science principles.

Pangeo involves lazy evaluation and parallel execution using Dask. This allows users to work efficiently with large volumes of data stored in cloud-optimized formats (COG and Zarr). Participants engaged in hands-on workflows implemented in Jupyter notebooks deployed on Dask clusters, exploring essential EO data processing tasks including subsetting, aggregation, statistical analysis, and visualization.

doi:10.2760/2119408

Two Pangeo deployment setups were demonstrated: one on EOSC and another on the Spider cluster. The latter leveraged Apptainer containers and the hpc-container-wrapper framework to execute Pangeo environments efficiently on SLURM. Developed in collaboration with SURF and the Netherlands eScience Center, the Spider deployment built upon open scripts and documentation contributed by Francesco Nattino, Meiert W. Grootes, and Ou Ku. Participants were instructed on launching JupyterLab with Dask via SLURM job scripts and accessing remote sessions using SSH tunneling. For users lacking container access, alternative solutions using micromamba, jupyter-forward, and dask-jobqueue were also demonstrated.

Both deployments emphasized the paradigm shift from downloading data locally to executing code where the data resides—leveraging public S3 buckets and open SpatioTemporal Asset Catalogs (STAC) for direct data access. This approach was particularly beneficial for participants constrained by limited local computational resources. The event further highlighted containerized reproducibility through the use of Docker and Binder, enabling the creation and sharing of fully executable environments.

3.3. openEO Approach

openEO is an open-source framework that standardizes and streamlines the access, processing, and analysis of EO data across heterogeneous backends. Traditional EO workflows often involve time-consuming steps, including data discovery, download, and platform-specific pre-processing. This poses challenges in scalability and interoperability, especially when integrating multiple data sources. openEO addresses these limitations by defining a unified API that abstracts the underlying data infrastructures and exposes consistent functionality through client libraries in various languages, including Python, R, and JavaScript.

During the hackathon, we showcased the capabilities and usability of the openEO framework. Our demonstration used the openEO Python client within Jupyter notebooks to connect to the CDSE backend. Participants were guided through essential steps, including library import, authentication, and backend connection. Limited memory on the CDSE Hub caused occasional kernel crashes, which were mitigated by reducing data size. Additionally, we introduced the concept of datacubes, a central feature of openEO that structures EO data in a spatiotemporal grid. This model enables server-side declarative processing, significantly streamlining complex analyses and enhancing reproducibility and cross-platform integration in EO applications.

3.4. OEMC Approach

OEMC software and data products were also presented at the hackathon to gather user feedback and foster co-development.

3.4.1. GlobalPointTools

GlobalEarthPoint is an open-source software library for accessing large geospatial point datasets stored in cloud environments such as S3. Python and R bindings enable efficient (interactive) queries and downloads of data stored in cloud-optimized formats (e.g., partitioned GeoParquet). The software supports lazy evaluation for high-performance data retrieval through technologies like Apache Arrow and Polars.

3.4.2. OEMC ML Deployment and Central app

Besides organizing and enhancing the accessibility of existing EO and in-situ data, the OEMC project is producing European-wide [5] and global datasets [6] at a 30-m spatial resolution, tailored to meet the needs of the use cases. The production of these datasets is running via an ML pipeline deployed in an in-house HPC/HTC infrastructure hosted by OpenGeoHub Foundation. With more than 1,000 CPU threads across 14 high-density nodes and 3 Petabytes of storage space, the infrastructure runs a suite of open source solutions, including `SeaweedFS` as object storage, `SLURM` as workload manager, and `Docker` as containers manager. The ML algorithms/methods rely on `scikit-learn` and utilize spatiotemporal overlay and predictions via `scikit-map`. Feature selection, hyperparameter tuning and the training of the final ML models run in a single node (96 CPU threads and 1 TB of RAM), while the prediction/inference of billions of 30-m pixels run in `SLURM`, where each chunk of input data (ex. 4000x4000 pixels) is assigned to an processing node by an embarrassing parallel strategy. Predictions of tree-based models were sped up by compiled models based on `tl2cgen` and `lleaves` libraries. The OEMC ML Deployment proved to be reliable in a production environment once all layers, produced in the context of the project, are publicly available as open data (CC-BY) at OEMC Central App, OpenLandMap STAC, EcoDataCube STAC, and Zenodo.

4. RESULTS

4.1. Hackathon Participants

The hackathon had 20 on-site and 21 online participants from 17 countries, with the majority coming from Europe. Most of them were PhD students ($\tilde{4}6\%$), followed by early-career researchers ($\tilde{2}5\%$), and postdoctoral researchers ($\tilde{1}8\%$); the rest comprised master's students, professors, and employees from the private sector. Most of them consider themselves ecologists or environmentalists, but there were also remote sensing scientists, geographers, hydrologists, and economists (Fig. 2).

The questionnaire also revealed that every fourth participant had prior experience with HPC processing. Nevertheless, most of them have also been using open-source tools and

doi:10.2760/2119408

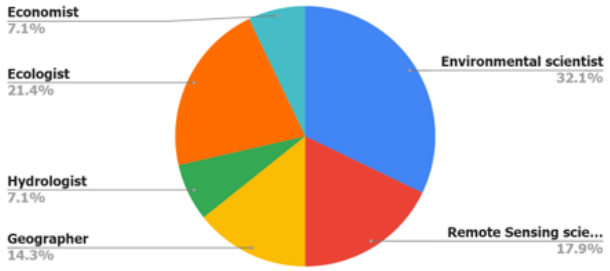


Fig. 2. The expertise of the hackathon participants.

Python modules, such as GDAL, Rasterio, NumPy, Geopandas, QGIS, GRASS, PyTorch, PKTools, Pandas, and Scikit-learn.

4.2. Perception of Big Data

Almost half of the participants (~46%) work with geospatial datasets that are stored both online and locally, whereas the rest of the participants work only with data stored online (~28%) or only with data stored locally (~26%). Their locally stored data are primarily in the 10-100 GB range, whereas only about 15% work with locally stored data on a TB scale (Fig. 3). Furthermore, the data were mainly related to the following topics: land monitoring (~20%), hydrology (~16%), agriculture (~14%), vegetation (~11%), although topics such as hazards, economics, climate, soil, urban, health, and biodiversity were also covered.

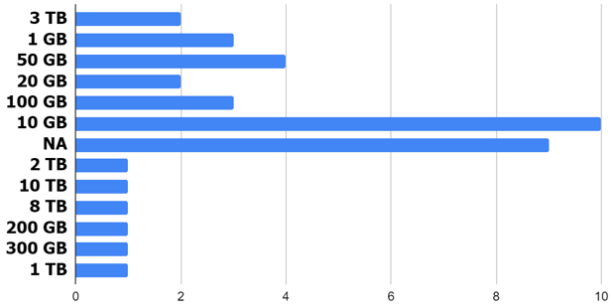


Fig. 3. What is the size of your local geospatial dataset that you work with?

5. DISCUSSION AND CONCLUSIONS

Despite the capabilities of the HPC approaches, participants encountered challenges related to environment configuration, Dask performance tuning, and conceptual adaptation from serial to parallel processing. These insights underscore the importance of streamlined deployment practices and enhanced training resources in lowering entry barriers and fostering the broader adoption of scalable open science workflows.

The pre-hackathon questionnaire, completed by participants, revealed several interesting insights. The participant structure showed that the hackathon primarily attracted researchers and PhD students, with a notably small number of participants from the private sector. However, the geospatial industry and academia should collaborate more, as open-source projects can benefit more from such partnerships. Thus, the next hackathons should focus more on engaging with industry.

The local data volumes typically processed by the participants were 10 TB or smaller. Such data can be readily handled with HPC, suggesting that the barrier is rather in adjusting code for efficient HPC processing. Therefore, presenting and interacting with different HPC processing approaches and backends is fully justified and shall remain the focus for future hackathons.

Finally, it is worth noting that, in addition to a one-time budget allocated by IIASA and the OEMC project for organizing GEO-OPEN-HACK-2024, the event received significant support from the community, including teachers and organizations such as SURF, Spatial Ecology, and OGH. Nevertheless, it is essential to secure future funding sources to ensure its long-term sustainability.

REFERENCES

[1] Abernathey, R. P., Augspurger, T., and et al. (2021). Cloud-native repositories for big scientific data. *Computing in Science Engineering*, 23(2):26–35.

[2] Beazley, D. (2003). Automated scientific software scripting with swig. *Future Generation Computer Systems*, 19(5):599–609.

[3] Kempeneers, P., Kliment, T., and et al. (2022). Parallel processing strategies for geospatial data in a cloud computing infrastructure. *Remote Sensing*, 14(2).

[4] Kempeneers, P., Pesek, O., and et al. (2019). pyjeo: A python package for the analysis of geospatial data. *ISPRS International Journal of Geo-Information*, 8(10). 10.3390/ijgi8100461.

[5] Minarik, R., Hengl, T., and et al. (2024). Soil type (world reference base) map of europe based on ensemble machine learning and multiscale eo data. *researchsquare*. Submited to PeerJ. 10.21203/rs.3.rs-5244083/v1.

[6] Parente, L., Sloat, L., and et al. (2024). Annual 30-m maps of global grassland class and extent (2000–2022) based on spatiotemporal machine learning. *Scientific data*, 11(1):1–22.

[7] Schramm, M., Pebesma, E., and et al. (2021). The openeo api–harmonising the use of earth observation cloud services using virtual data cube functionalities. *Remote Sensing*, 13(6).

doi:10.2760/2119408

RETRIEVE, TRANSFORM, DELIVER: INTEGRATING PRESERVATION AND PERFORMANCE IN THE EUMETSAT DATA LAKE

Andrea Colapicchioni, Joaquin Rodriguez Guerra, Cedric Bergeron, Guillaume Aubert

EUMETSAT

ABSTRACT

This paper presents the EUMETSAT Data Lake's (EDL) ongoing evolution and planned integration with the Long-Term Archive (LTA) to establish a sustainable, user-centric data access model. From the perspective of the EDL and its user-facing components—the Data Store and Data Tailor—we describe how this integration will enable a fluid retrieve-transform-deliver paradigm designed to optimise access, customisation, and delivery of Earth observation data. The system is being designed to address diverse user scenarios while reducing operational overhead by supporting dynamic workflows, such as on-demand generation of virtual Zarr datasets tailored to different performance or analysis needs. This evolution aims to provide a more efficient and scalable infrastructure for data access, reuse, and long-term preservation, while also strengthening business continuity through improved resilience and recovery capabilities.

Index Terms— Data Preservation, Data Lake, Cloud Integration, Multi-tier Storage, Data Repatriation, Data Customisation, Unified Data Access, Earth Observation Data

1. INTRODUCTION

Earth observation plays a central role in supporting the European Green Deal, UN Sustainable Development Goals, and climate resilience strategies. The transformation of the EUMETSAT Data Lake contributes to these societal priorities by enabling faster, more reliable, and energy-efficient access to data essential for environmental monitoring and decision-making.

EUMETSAT provides essential Earth observation data for weather forecasting, climate monitoring, and environmental studies. The EUMETSAT Data Lake serves as the ingestion, indexing, and access hub for this data. It is complemented by the Data Store [1][2], which acts as the user interface and API access point, and by the Data Tailor [3][10], which enables product transformation and customisation. Together, these services support a rapidly growing and increasingly cloud-native user base.

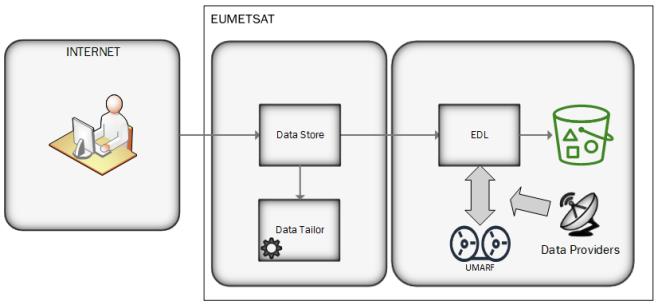


FIGURE 1 - components

Historically, data access and preservation have been handled by separate systems. The legacy UMARF (Unified EUMETSAT Meteorological Archive and Retrieval Facility) [4][5] system relies on tape-based storage, which is optimised for long-term preservation but cannot meet the performance demands of modern usage. Through the Long-Term Data Preservation & Access Management Evolution (LTDPA) initiative, the EDL will integrate with the new Long-Term Archive (LTA), leveraging it as a cost-efficient deep storage tier while enabling seamless data repatriation.

The LTDPA vision is

Efficient and affordable management of and access to large volume of satellite data with an appropriate Quality of Service (QoS)

and is driven by increasing data volumes, rising storage costs, and the need to support on-demand, customised data access workflows. The EDL will evolve to tackle these challenges by adopting a modular and incremental design, accommodating real-world workflows, and facilitating faster iteration cycles.

This vision reinforces the need to integrate the transformation layer more closely within the Data Lake architecture, not merely as a supporting tool but as a core capability. By doing so, EUMETSAT will be able to serve a broader range of use cases while maintaining sustainability, reliability, efficiency, and user flexibility.

doi:10.2760/2119408

2. OBJECTIVES AND MOTIVATION

As part of the LTDPA project, the EDL is undergoing a transformation to become a more intelligent, sustainable, and user-centric data access infrastructure, driven by several interrelated factors:

- **Exponential Growth in Data Volumes**: EUMETSAT currently manages over ten Petabytes of data, supporting many different formats. With upcoming missions such as Meteosat Third Generation, Metop Second Generation or Copernicus (Sentinels, CO2M, etc.), data volumes are expected to grow exponentially. This growth challenges current storage and access paradigms.
- **Diversifying User Expectations**: Users increasingly require cloud-native solutions, on-demand access, and integration with modern analysis tools. Emerging formats like Zarr [6][7] are designed for scalable, partial access but require flexible chunking strategies that depend on the user's intended analysis workflow.
- **Limitations of Legacy Approaches**: Maintaining multiple pre-processed versions of the same dataset to support different use cases is no longer feasible. It leads to high storage costs and conflicts with long-term preservation strategies.
- **Strong Need for Enhanced Business Continuity:** Ensuring that critical business functions relying on data access systems can continue without interruption in the face of unexpected events—such as system failures, cyberattacks, or data corruption—is essential to maintaining customer trust and operational resilience.

EUMETSAT is developing a new approach to optimise data access and meet these challenges. This approach will retrieve products from the most appropriate source (online or archival tiers), apply transformations as needed (e.g., subsetting, reformatting, reprojection), and deliver the results through a secure and manageable interface.

Integrating the Data Tailor as a native transformation layer within EDL, along with technologies like VirtualZarr and Kerchunk, will enable the on-the-fly generation of virtual data representations tailored to diverse access patterns. This approach will support performance, flexibility, and sustainability without increasing physical storage demands.

Ultimately, the LTDPA project will support:

- Innovation in data access workflows by replacing pre-generated products with dynamic, on-demand virtual datasets.
- A reduction in infrastructure duplication, leading to a more scalable and sustainable model compared to conventional satellite data distribution strategies.
- Unified access across all EUMETSAT datasets (online and archived)
- Sustainable cost models through tiered storage
- Improved usability through customisable outputs and cloud integration
- A more adaptable architecture that evolves with user needs and technology trends
- A more flexible approach to data management for business continuity.

3. METHODOLOGY

The cornerstone of the new paradigm for big dataset retrieval is a

retrieve → [transform] → deliver

workflow:

- **Retrieve**: Data will be sourced from the optimal tier, including online EDL storage or repatriated from LTA.
- **Transform:** If requested, data can be processed via the integrated Data Tailor, enabling transformations such as reprojection, reformatting, and subsetting. By using a plugin architecture, we will be able to integrate other transformation frameworks that will be available in the future.
- **Deliver**: The final product will be placed in a target location, with smart retention and quota control. Initially, the Data Store will manage this location and implement it as a standard S3-compatible bucket. However, the concept will be expanded to support delivery to alternative endpoints—such as user-provided storage in the European Weather Cloud, Destination Earth Data Bridges, or other controlled environments—allowing users to access data close to their computational resources.

doi:10.2760/2119408

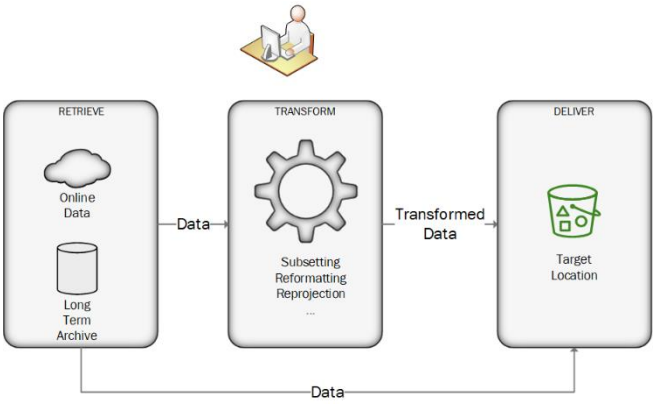


Figure 2 Retrieve-Transform-Deliver

This model will support advanced use cases such as the on-demand generation of virtual Zarr datasets. Since there is no one-size-fits-all strategy for chunking in Zarr, the ability to create tailored virtual cubes per request will be a powerful capability. Different chunking strategies will be applied dynamically, based on the user's intended access patterns (e.g. time-series analysis vs. spatial browsing), improving performance and usability in AI/ML and cloud-native workflows [6][7][8].

Additional strategic enhancements will include:

- A multi-tier storage model, with flexible movement between caching, online, and glacier tiers.
- Predictive queuing to evaluate job impacts and auto-approve low-impact user requests.
- Fast delivery to external cloud environments, supporting efficient data offloading and processing.

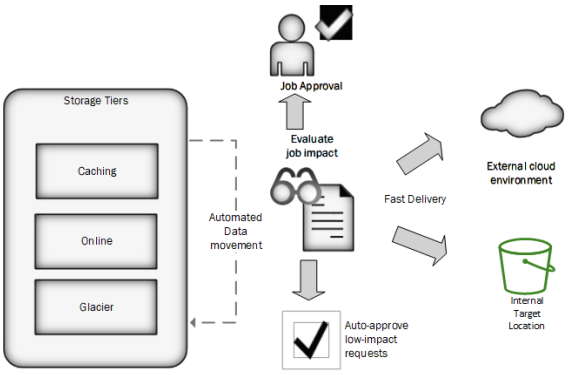


Figure 3 strategic enhancements

4. RESULTS AND EXPECTED BENEFITS

This architecture also reinforces EUMETSAT's commitment to FAIR data principles, ensuring that satellite data is Findable, Accessible, Interoperable, and Reusable. The use of open protocols and standards—such as Zarr, Kerchunk, and fsspec—facilitates integration with existing scientific tools and promotes open science.

A reliable data access service must go beyond simply making data available—it must ensure operational integrity and reliability, equitable access, and adaptability to diverse user needs. Several key principles guide our architectural and operational decisions:

- **Sustainability:** EUMETSAT must continuously balance data access performance with the cost of storage. This requires selecting between fast but costly tiers and more economical archival layers based on actual usage patterns.
- **Reliability:** Users rely on consistent service delivery. The system must be resilient to failures and designed for business continuity, particularly for critical operational and scientific use cases.
- **Quality of Service (QoS):** To prevent resource exhaustion, we enforce quotas and implement scheduling to ensure fair usage. This prevents a small subset of users from monopolising shared infrastructure.
- **Authentication:** Identifying the user allows for tiered service options. Without authentication, providing differentiated service levels or usage controls is impossible.
- **Authorisation:** Some data (e.g., Level 0 or restricted mission datasets) is not publicly available. Authorisation policies ensure that only eligible users can access these products following licensing or policy constraints.
- **Data Analytics:** Collecting usage patterns helps identify popular datasets and access methods. This data-driven approach informs resource planning, caching strategies, and the development of new features.
- **Virtualisation of the Data Endpoint**: Abstracting the physical location of the data allows us to move, replicate, or transform data behind the scenes without impacting the user. This decoupling is fundamental to supporting architectural evolution.
- **Integration with Standard Libraries**: Scientific and operational communities use established tools and workflows. Supporting libraries like fsspec [9] —a Python abstraction layer for filesystem-like access to remote data—enables users to access our services as if the data were local, without modifying their applications.

For these reasons, simply placing data in an S3 bucket with anonymous access is not sufficient. While this approach may suit simple file distribution, it falls short in delivering critical capabilities such as service monitoring,

doi:10.2760/2119408

access controls, transformation, observability, and long-term sustainability. EUMETSAT's approach, in contrast, provides a managed, intelligent, and user-aware access platform that aligns with the principles of FAIR data and ensures service continuity across missions and decades.

The EDL-centric approach will deliver:

- Unified access through a single interface for all data types, with streamlined workflows.
- Rapid delivery of customized datasets, reducing overhead and time-to-access.
- Efficient repatriation from LTA to support reprocessing campaigns, gap-filling, and user-driven data exploration.
- Cost savings through smarter tiering and reduced duplication.
- Flexibility to generate optimised data structures (e.g. Zarr cubes) tailored to each use case.

5. CONCLUSIONS AND FUTURE WORK

EDL enables a more scalable and sustainable data ecosystem by shifting from siloed services to an integrated, workflow-centric approach. The retrieve-transform-deliver paradigm supports current user needs while laying the foundation for future innovations in cloud-native data access, including dynamic data structuring and AI-ready workflows.

The project is planned to last three years. We are currently in 'phase 0', which focuses on careful requirements gathering and high-level system design. The implementation will be structured around three major iterations, each concluding with a Minimum Viable Product (MVP) that provides incremental functionality, allowing early feedback and progressive system refinement.

Future directions include:

- Extending transformation capabilities to cover more advanced use cases.
- Refining Zarr optimisation logic based on user behaviour and analytics.
- Enhancing observability and automation in the queuing and delivery pipeline.
- Delivering an FS-spec library for direct access to online data
- Preparing a clear data management approach to ensure the system meets operational and user requirements

REFERENCES

[1] EUMETSAT, "Data Store Overview," EUMETSAT User Portal, [Online]. Available: https://user.eumetsat.int/data-access/data-store. [Accessed: 13-May-2025].

[2] EUMETSAT, "Data Store Detailed Guide," EUMETSAT User Portal, [Online]. Available: https://user.eumetsat.int/resources/user-guides/data-store-detailed-guide. [Accessed: 13-May-2025].

[3] EUMETSAT, "Data Tailor Standalone Guide," EUMETSAT User Portal, [Online]. Available: https://user.eumetsat.int/resources/user-guides/data-tailor-standalone-guide. [Accessed: 13-May-2025].

[4] EUMETSAT, "Data Centre Introductory Guide," EUMETSAT User Portal, [Online]. Available: https://user.eumetsat.int/resources/user-guides/data-centre-introductory-guide. [Accessed: 13-May-2025].

[5] EUMETSAT, "Storing Data at EUMETSAT," EUMETSAT Website, [Online]. Available: https://www.eumetsat.int/storing-data. [Accessed: 13-May-2025].

[6] Zarr Developers, "Zarr-Python Documentation," [Online]. Available: https://zarr.readthedocs.io/. [Accessed: 13-May-2025].

[7] Zarr Developers, "Zarr Specification," GitHub, [Online]. Available: https://github.com/zarr-developers/zarr-specs. [Accessed: 13-May-2025].

[8] VirtualiZarr Developers, "VirtualiZarr Documentation," Read the Docs, [Online]. Available: https://virtualizarr.readthedocs.io/. [Accessed: 13-May-2025].

[9] FSSPEC Developers, "Kerchunk Documentation," GitHub Pages, [Online]. Available: https://fsspec.github.io/kerchunk/. [Accessed: 13-May-2025].

[10] EUMETSAT, "Data Tailor Web Service API," EUMETSAT User Portal, [Online]. Available: https://user.eumetsat.int/api-definitions/data-tailor-web-service-api. [Accessed: 13-May-2025].

doi:10.2760/2119408

GSSC: ESA THEMATIC EXPLOITATION PLATFORM FOR NAVIGATION DIGITAL TRANSFORMATION. ENHANCING GNSS SCIENTIFIC RESEARCH

Maria del Mar Millán [a], Raúl García [a], Pablo García[a], Enrique Saiz[a] , Sara del Rio [b], Natalia Castrillo [b], Jean-Christophe Berton [c]

[a] GMV, Isaac Newton, 11. P.T.M. 28760 Tres Cantos, Madrid, Spain
[b] Starion España S.L.U. Calle Chile 10, Edificio Madrid 92 28290 Las Rozas de Madrid, Spain
[c] European Space Operations Centre, 5 Robert-Bosch-Strasse, Darmstadt, Hessen, Germany

ABSTRACT

The GNSS Science Support Centre (GSSC) is the European Space Agency's (ESA) platform dedicated to Global Navigation Satellite Systems (GNSS) data exploitation. As part of it, GSSC Now provides centralized, cloud-native access to a rich repository of GNSS datasets, including observations from ESA missions, ground networks and third-party sources to form the ESA International GNSS Service (IGS) Global Data Centre. This platform also integrates data discovery, visualization and analysis capabilities, enabling scientists to work directly with GNSS data without the need to download or configure local tools. Through a suite of browser-based Datalabs, users can launch scientific applications on demand and perform advanced analysis in environments such as JupyterLab or Octave. This paper introduces the GSSC Now platform, highlights its architecture and data lake design, and presents practical examples of how users are engaging in GNSS science with unprecedented efficiency.

Index Terms— GNSS, open science, big data, cloud computing, scientific exploitation

1. INTRODUCTION

Global Navigation Satellite Systems (GNSS) form a cornerstone of modern geospatial science. Their ability to provide precise positioning, timing, and velocity (PVT) services has enabled a wide array of scientific and operational applications across domains such as geodesy, space weather, seismology, atmospheric science, and environmental monitoring. Researchers use GNSS data to track crustal deformation, estimate tropospheric water vapour, study ionospheric irregularities, and support real-time responses to natural hazards.

Over the past decade, the volume, diversity, and scientific relevance of GNSS data have grown significantly. This expansion has been driven not only by the increasing number of satellites and ground stations, but also by the integration of GNSS sensors into spaceborne platforms and mass-market devices. However, this growth has also introduced technical and organizational challenges. Data has historically been spread across a fragmented landscape of archives, often using non-standardized formats and metadata, limited in discoverability, and lacking tools for efficient analysis.

In response to these limitations, the European Space Agency (ESA) established the GNSS Science Support Centre (GSSC), through its Navigation Directorate and the European GNSS Evolution Programme. The GSSC aims to enhance and streamline the scientific use of GNSS data by providing a platform that supports data centralization, access, processing, visualization and analysis under a unified framework. Located at ESA's European Space Astronomy Centre (ESAC), the GSSC is a key component of ESA's digital transformation strategy, bringing together data, tools, and users in a modern, scalable environment designed to foster open science and innovation in the GNSS domain.

2. GSSC NOW: ENABLING OPEN GNSS SCIENCE

The GSSC constitutes ESA's strategic initiative to enable and enhance scientific research through the exploitation of GNSS data. Conceived as a thematic exploitation platform, the GSSC is designed to bridge the gap between vast GNSS data archives and the research community's ability to derive actionable insights from them. Its mission extends beyond simple data access, focusing instead on creating an open, integrated environment that supports the entire research lifecycle — from data discovery and pre-processing to advanced analysis, modelling, and visualization.

The GSSC provides access to a wide spectrum of GNSS-related data products. These range from Level-0 raw observables captured by ground and space-based receivers to higher-level data such as zenith tropospheric delay (ZTD), total electron content (TEC) maps, orbit and clock products, and differential code biases (DCB). These datasets are made accessible through an infrastructure that ensures responsiveness, scalability, and long-term data preservation.

At the heart of this ecosystem is GSSC Now, the interface of the platform that operates ESA's vision of open and reproducible science. GSSC Now combines a powerful metadata-driven search engine, RESTful APIs for automated data interaction, and a suite of user-facing services including graphical explorers, programmatic interfaces, and browser-executed scientific tools. This allows users to interact with

doi:10.2760/2119408

massive data archives without the need to download files or set up local processing environments.

GSSC Now also embraces the **FAIR principles** (Findable, Accessible, Interoperable, and Reusable), ensuring that GNSS data is not only available but also structured and documented in a way that facilitates discovery and reuse across scientific domains. Through standardized formats, persistent identifiers, and harmonized metadata schemas aligned with ISO and INSPIRE standards, GSSC Now provides a robust foundation for multi-disciplinary research.

Crucially, the platform supports not only traditional GNSS science such as geodesy or navigation, but also emerging domains including space weather, environmental monitoring, urban mobility, and even space science and fundamental physics. By enabling cross-correlation of GNSS signals with atmospheric, seismic, and geomagnetic phenomena, GSSC Now transforms GNSS data into a powerful sensor network for Earth and space system science.

In summary, GSSC Now stands as a catalyst for data-driven innovation, lowering the barriers to entry for scientific users, enabling collaboration across domains and institutions, and accelerating the translation and combination of raw GNSS data into meaningful scientific outcomes.

3. THE GSSC ARCHIVE AND DATA LAKE

At the core of the GNSS Science Support Centre lies its federated data archive and integrated data lake infrastructure — a foundational component that enables reliable, high-throughput access to diverse GNSS datasets. Designed for long-term preservation, rapid indexing, and on-demand availability, the GSSC archive adopts a multi-tiered architecture that blends scalability with scientific rigor.
The GSSC archive ingests and harmonizes data from both spaceborne and terrestrial sources, providing access to over 300 million individual data assets. These resources include:

- **Observation Data**: Daily and hourly RINEX-formatted observations from globally distributed ground-based GNSS receivers. These data form the backbone of positioning and geodetic studies, enabling the reconstruction of precise orbits and atmospheric states.
- **Navigation Data**: Key satellite broadcast products such as ephemerides, precise orbit and clock solutions, DCB files, and ionospheric corrections. These elements are indispensable for high-accuracy applications like real-time kinematic (RTK) positioning and PPP (Precise Point Positioning).
- **Atmospheric Products**: Derived datasets that provide critical insight into atmospheric variability, including ZTD, Slant Total Delay (STD), Integrated Water Vapor (IWV), and global/regional TEC maps.
- **Space Segment Data**: Onboard telemetry from GNSS receivers aboard ESA's Earth Observation and scientific satellites — such as **Galileo**, **Sentinel**, **MetOp**, **SWARM**, **GOCE**, and **ICESAT** — offering unique opportunities for space-based geodetic and ionospheric science.

To ensure discoverability and standardization, all resources are enriched with semantic metadata aligned to INSPIRE and ISO 19115 specifications. Each entry includes provenance, spatial and temporal granularity, processing level, sensor/platform information, and licensing constraints. These metadata elements allow for high-performance faceted search and data federation across research infrastructures.

The ingestion process itself is automated and modular. Pipelines constantly monitor and pull updates from trusted sources such as the IGS, BKG, CDDIS, IGN, ILRS, and ESA internal systems. Upon ingestion, each dataset is validated for integrity, assigned to a Resource Class (e.g., Observation, Product, Document), and catalogued within GSSC's indexing engine based on defined Science Domains — Positioning and Navigation, Geodesy, Ionosphere and Magnetosphere, and Troposphere.

GSSC also enforces use of community-standard formats, including:

- **INEX 2/3/4** for raw GNSS observables**,**
- **SP3** for satellite ephemerides**,**
- **SINEX-BIAS** for inter-frequency bias estimation**,**
- **IONEX** for ionospheric TEC maps.

This ensures **interoperability** with external GNSS processing frameworks like **RTKLIB**, **Bernese**, **GIPSY**, and ESA's own **gLAB**. The adherence to these standards also facilitates integration with federated infrastructures and European initiatives promoting FAIR and open-access data.
A significant innovation introduced by GSSC Now is its **interactive, map-based search interface**. Unlike conventional FTP or static directory structures, the GSSC Explorer allows users to query datasets based on multiple attributes, including:

- geographic bounding boxes or station locations,
- time ranges (e.g., storm periods, mission phases),
- GNSS constellations (GPS, Galileo, BeiDou, GLONASS),
- observation parameters (sampling rate, signal type, frequency),
- data quality indicators and completeness.

This interface not only accelerates data discovery but also enhances user experience, especially for multi-disciplinary researchers unfamiliar with the intricacies of GNSS data hierarchies.

In essence, the GSSC archive and data lake serve as the digital substrate upon which advanced GNSS scientific analysis is built. By consolidating heterogeneous data

doi:10.2760/2119408

sources, embedding metadata intelligence, and supporting scalable access models, ESA has laid the groundwork for a next-generation GNSS research environment that is as powerful as it is accessible

4. ARCHITECTURE AND DESIGN

GSSC Now is architected as a modular, cloud-native platform built on containerized microservices and orchestrated through Kubernetes. This design enables independent deployment, maintenance, and scaling of core services, ensuring operational resilience and flexibility in accommodating diverse scientific workflows.

The system is structured around three primary functional layers:

- **Data Lake Services**, responsible for the ingestion, storage, and classification of GNSS data. Ingested resources are validated and enriched with metadata, which is indexed for efficient discovery using Elasticsearch. The storage architecture combines high-performance local caching with scalable object storage to support both frequent access and long-term preservation.
- **Core Engine**, which coordinates processing and workflow execution. It enables users to launch analysis environments, manage personal workspaces, and run data-driven applications through GSSC Now's processing framework. This includes integration with ESA Datalabs, providing browser-based access to environments like JupyterLab and Octave.
- **User Services**, encompassing both graphical and programmatic interfaces. The GSSC Now Explorer offers faceted search and map-based navigation for dataset discovery. RESTful APIs and command-line tools provide automation options for advanced users, while session and identity management are handled via ESA's Single Sign-On (SSO) and secure JWT-based access control.

Together, these components deliver a flexible and secure platform that supports real-time interaction with GNSS data, facilitates the development of custom applications, and ensures that both new and experienced users can work seamlessly across visual and scripted interfaces.

5. BROWSER-BASED SCIENTIFIC COMPUTING

One of the most transformative features of GSSC Now is the integration of browser-based scientific computing through its **Datalabs system**. This capability fundamentally changes how researchers interact with GNSS data by eliminating the need for local software installations or high-performance computing resources. Instead, scientists can launch analytical environments directly from their web browser, bringing computation closer to the data.

GSSC Now Datalabs are based on **containerized execution environments**, managed within ESA's secure cloud infrastructure. These containers are deployed on demand and pre-configured with widely used scientific tools. Among the default options available are:

- **JupyterLab**, a powerful notebook-based environment for interactive computing in Python, widely adopted for data science, machine learning, and visualization.
- **Octave**, a numerical computation tool with compatibility for MATLAB scripts.
- **gLAB**, ESA's GNSS-Lab tool capable of full GNSS processing, from raw RINEX files to precise positioning results.

These environments are tailored to GNSS applications, and the catalogue is constantly evolving. In addition to generic tools, users have access to mission-specific and thematic notebooks, custom scripts, and pre-integrated libraries for tasks such as signal quality assessment, ionospheric analysis, or satellite orbit reconstruction.

Once a Datalab is launched, it has access to:

- the full GSSC data lake via mounted volumes,
- **persistent user storage** to retain intermediate results and notebooks across sessions,
- and optionally, **GPU resources** for acceleration of machine learning workloads or large-scale simulations.

The workflow is designed to be seamless and efficient. After logging in through ESA's SSO system, users can:

1. Search and filter data through the Explorer interface.
2. Select a dataset of interest.
3. Launch a Datalab with a single click.
4. Begin analysis immediately — no need to download data, set environment variables, or manage dependencies.

This model, in which computation is brought to the data, aligns with ESA's broader **ESA Datalabs** strategy and supports the **FAIR** and **Open Science** principles by enabling fast prototyping, reproducibility, and sharing of research workflows. Notebooks can be exported or shared via persistent URLs, supporting collaboration within the scientific community.

GSSC Datalabs represent more than just tools — they are **scientific workspaces** designed to accelerate research, lower the technological barrier to entry, and empower users to explore complex GNSS datasets without friction.

doi:10.2760/2119408

6. USE CASES: ADVANCED ANALYTICS AND MACHINE LEARNING

GSSC Now supports advanced analytics through Datalabs designed for Machine Learning (ML), citizen science, and Internet of Things (IoT)-based applications. In the **ML-Lab**, users can develop models directly in-browser using libraries like TensorFlow and scikit-learn. Typical use cases include anomaly detection in GNSS signals and prediction of atmospheric parameters such as ZTD, using GNSS and meteorological data as input.

The platform also supports large-scale participatory science. The **CAMALIOT** initiative, for example, enabled collection of over 50 million GNSS observations from smartphones during a six-month campaign. These data were processed within GSSC for ML-based forecasting of tropospheric and ionospheric conditions and are accessible through dedicated Datalabs.

In parallel, projects like **UbiSAP** illustrate the integration of low-cost IoT GNSS sensors with GSSC workflows. Using RTK techniques, researchers monitor structural displacements and landslides near critical infrastructure, with analysis and visualization performed entirely within the platform.

These examples highlight how GSSC not only enables scientific research but also extends GNSS data exploitation to public engagement and operational monitoring contexts. In this sense, the modular design and usability of GSSC Now — particularly through the Datalabs environment — provide an effective framework that can accommodate both domain experts and newcomers. The integration of scalable infrastructure and open data sources further aligns the platform with broader strategic objectives. While the cases of CAMALIOT and UbiSAP already illustrate community participation and operational relevance, documenting additional evidence of real-world adoption and uptake would further strengthen the demonstration of impact and the long-term sustainability of the approach.

7. CONCLUSIONS, LESSONS LEARNED AND FUTURE WORK

GSSC Now has established itself as a key enabler of open, data-driven science in the field of GNSS. By integrating high-quality datasets, scalable cloud infrastructure, and accessible tools into a unified platform, it addresses long-standing challenges around data accessibility, reproducibility, and usability. Researchers from diverse domains — including geodesy, ionospheric physics, meteorology, and education — can now explore, process, and share GNSS data efficiently and collaboratively.

Its open and modular architecture supports both expert users developing advanced processing pipelines, and newcomers engaging in exploratory analysis through user-friendly Datalabs. The platform exemplifies ESA's broader strategy for digital transformation and scientific innovation.

Throughout the development and deployment of GSSC Now, several lessons have been learned that can be of value to the broader Earth Observation (EO) community. Firstly, the adoption of **community standards** (RINEX 2/3/4, SP3, SINEX-BIAS, IONEX, ISO 19115, INSPIRE metadata) has proven essential to ensure **interoperability** with external infrastructures and tools (e.g. RTKLIB, Bernese, GIPSY, ESA gLAB). This approach has significantly reduced barriers for integration and collaboration across different research groups.

Another key lesson is the importance of **bringing computation to the data** through browser-based environments (JupyterLab, Octave, gLAB). This paradigm has accelerated scientific workflows, promoted reproducibility, and simplified onboarding of new users.

Looking ahead, GSSC Now will continue to evolve through several key developments:

- **Performance** improvements and **stability** of the GSSC Now platform
- **Interoperability** with other space geodetic techniques data platforms (e.g. Very Long Baseband Interference)
- **Expanded APIs** for integration with institutional workflows (e.g. in the context of the ESA Genesis mission)
- **Educational Content** to support universities and training programmes in GNSS science.

By adhering to the **FAIR principles** and contributing to ESA's **Digital Twin Earth** and **Open Science Cloud** initiatives, GSSC is building the foundation for a new generation of scientific discovery — one in which GNSS data plays a central role in understanding and monitoring the Earth.

REFERENCES

[1] Navarro, V., del Rio, S., Castillo, M., Martin, F., Ventura-Traveset, J., GSSC – Integrating Big Data, Machine Learning and Notebook Technologies for Open Science, Scientific and Fundamental Aspects of GNSS, ESA, 2019. DOI: 10.13140/RG.2.2.24797.23525

[2] Navarro, V., Alvarez, R., Pérez-López, F., Arviset, C., Ventura-Traveset, J., Furones, A.M., ESAC Science Exploitation and Preservation Platform Reference Architecture, ADASS XXVII, vol. 523, p. 285, 2019.

[3] Navarro, V., Grieco, R., Soja, B., Nugnes, M., Klopotek, G., Tagliaferro, G., See, L., Falzarano, R., Weinacker, R., Ventura-Traveset, J., Data Fusion and Machine Learning for Innovative GNSS Science Use Cases, Proc. of ION GNSS+ 2021, pp. 2656–2669.

[4] Navarro, V., et al., The GNSS Science Support Centre (GSSC): ESA's Open Platform for Scientific Exploitation of Navigation Data, Advances in Space Research, 2024. DOI: 10.1016/j.asr.2024.03.028

doi:10.2760/2119408

TOWARDS THE DEFINITION OF A BENCHMARK FOR WMTS

Michele Cecotti, Pieter Kempeneers, Edoardo Ramalli

European Commission, Joint Research Centre (JRC), Ispra, Italy

ABSTRACT

Web Map Tile Service (WMTS) is a fundamental component of a digital infrastructure for visualizing geospatial big data, supporting many applications across various domains. Despite their ubiquity and long-standing use, systematic and dedicated benchmarks are currently lacking. Practitioners will highly benefit from such a framework since it provides clear guidance on assessing the performance, scalability, and advantages of WMTS implementations under varying conditions, especially in data-intensive contexts such as those demanded by modern AI applications. This work proposes a framework to study which aspects a WMTS benchmark should consider. By empirically evaluating custom and open-source WMTS solutions, this work identifies specific factors that significantly impact test outcomes, emphasizing the need for well-defined and representative test cases.

***Index Terms*—** WMTS, benchmark, big data, data visualization, digital infrastructure

1. INTRODUCTION

In recent years, the large availability of high-resolution spatial imagery from diverse sources and formats has created a growing need for efficient systems capable of rendering and delivering such data to scientific users and the general public. Web Map Tile Services (WMTS) is a key technology that can address these challenges. Numerous applications have since been developed using various architectures and technologies to meet the rising demands of big data and AI-ready geospatial services [1] to be integrated into agentic AI systems.

As geospatial data's volume, heterogeneity, use cases, and autonomous AI agents rapidly evolve and continue to expand, pre-generating static tiles for every need and leveraging a caching mechanism is a tricky solution. Tiles must be generated dynamically in many scenarios based on specific dataset requests and styling parameters. Consequently, the performance of the WMTS itself becomes a critical factor. While previous studies have compared the performance of raster versus vector rendering [2, 3] and evaluated specific software implementations, a standardized benchmark for assessing the performance of WMTS applications remains lacking [1]. This gap can create uncertainty when selecting the best solution in a given setting. Building on prior research on benchmark development [4], this study aims to identify some characteristics that influence tile delivery performance in WMTS solutions. The findings represent an initial step toward defining a standardized benchmarking framework that practitioners can use to evaluate and select the most appropriate WMTS for their specific requirements.

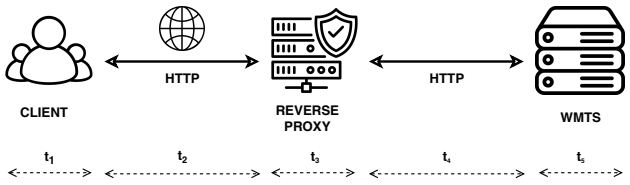


Fig. 1: A simplified client-server architecture for a Web Map Tile Service.

2. METHODOLOGY & EXPERIMENTS

The purpose of a benchmark is to provide a synthetic but informative overview of how a system performs under specific conditions. On the one hand, this can be highly valuable for practitioners, as it allows them to avoid the time-consuming process of evaluating multiple WMTS solutions every time. On the other hand, benchmarks can drive misleading or incomplete conclusions if they fail to account for critical aspects of real-world applications. This work aims to identify and analyze aspects particularly relevant to dynamic WMTS. For example, a given WMTS may exhibit excellent performance with vector data but perform significantly worse with raster inputs or even vary considerably between different types of raster formats and styles.

A first step for conducting such experiments is the definition of a generalizable digital architecture that reflects typical deployment scenarios while remaining flexible enough to accommodate a range of contexts. This ensures that the findings are broadly applicable and reflect realistic use cases.

Subsequently, isolating and evaluating individual system components is essential to understand their performance limitations under fixed hardware resource constraints. Our focus is not to identify the peak performance metrics or the saturation point of system resources of a given WMTS. Instead, we are particularly interested in how its performance varies in relation to specific variables, such as data formats, styling

doi:10.2760/2119408

Users	25^{Th}	50^{Th}	75^{Th}	95^{Th}
10	7	10	16	27
25	8	14	23	45
50	11	20	33	59
75	13	26	47	89
100	38	77	94	118

Table 1: Response Time (RT) percentiles on a minimal configuration (1 Core, 1 GB RAM) in milliseconds.

complexity, or rendering strategies.

Finally, we propose a hypothesis-driven experimental approach, formulating and empirically testing assumptions about which aspects most influence performance across different WMTS implementations. This structured methodology enables a deeper understanding of system behavior and informs the development of meaningful, context-aware benchmarks.

2.1. Overal Digital Architecture

The experimental setup and the overall architecture (see Figure 1) include three main components: the client, which sends tile requests to the WMTS; a reverse proxy, which manages the incoming and outgoing traffic; and the WMTS server, responsible for rendering and serving the tiles according to some specification detailed by the client requests. More complex and simpler data infrastructures are possible, but the architecture for this work contains the typical elements of a web service architecture.

The communication protocol is HTTP(S). We define the timing components of the request-response cycle as follows: t_1 as the client-side rendering time, t_2 as the network latency over the internet, t_3 as the time needed by the reverse proxy to handle the request and forwarding it, t_4 is the intra-network latency, and t_5 is the WMTS processing time. Therefore, the total response time observed by the client is $RT = t_1 + t_2 + t_3 + t_4 + t_5$, and it is our primary performance indicator [5]. Lower values of RT mean better performance of the WMTS and the overall digital infrastructure. Preliminary tests showed that t_2, t_3, and t_4 are negligible compared to t_5, although this assumption may not hold in more complex or distributed environments. Therefore, RT is close to t_5 in a testing environment.

A prerogative of a benchmark is isolation and reproducibility; thus, all the components we use in this work are containerized applications. In our case, the client machine is a high-performance system with 70 cores at 2.40 GHz and 1 TB of RAM, while the server machine, which in our experimental setting hosts both the reverse proxy and the WMTS, is equipped with 40 cores and 515GB of RAM.

Wor.	Rep.	Core	RAM	50^{Th}	75^{Th}	95^{Th}
1	20	1	10	7	14	31
2	10	2	20	9	15	31
20	1	20	200	7	14	33

Table 2: Response Time (RT) percentiles on different configurations in milliseconds. *Work.* stands for workers, *Rep.* for replicas. RAM values are in GB.

2.1.1. WMTS Server

In this work, we considered `GeoServer` [1] and `TileGeo` [2] as a pool of dynamic WMTS. GeoServer is written in Java, and its project has been running for many years; instead, TileGeo is an "in-house" development of the Joint Research Centre written in Python, combining the capabilities of `FastAPI` for the web service, while `GDAL` and `Mapnik` are used for handling geospatial datasets and rendering tiles. Both WMTSs support raster and vector source datasets. No particular production optimization is set for both WMTS. TileGeo is run with the *prod* starlette flag, while GeoServer, according to the documentation, comes with many optimizations by default [3].

To improve observability, the TileGeo server includes additional HTTP headers containing the WMTS processing time t_5, the CPU and RAM usage. The asynchronous architecture of *FastAPI*, built on *Starlette*, allows TileGeo to handle concurrent requests efficiently. The number of *Starlette* workers can be tuned to scale horizontally, with each worker capable of managing multiple simultaneous requests. GeoServer leverages a JAI thread pool executor. TileGeo, through the Antigrain Geometry renderer, is able to serve tiles in many different formats (JPEG, PNG, and WebP) with different setting parameters of compressions, color quantization, and quality levels. GeoServer's WMTS service primarly supports PNG, PNG8, and JPEG without the possibility of tuning specific settings. GeoServer does not natively support JPEG2000 as an input file type.

2.1.2. Client

To evaluate WMTS performance, it is essential to simulate the behavior of multiple users making simultaneous tile requests. To generate concurrent load, we use `Locust`[4], an open-source performance testing tool that supports HTTP and other protocols. Locust enables the definition of user behavior in plain Python, providing the flexibility to simulate realistic interaction patterns without being limited by a GUI

[1] https://github.com/geoserver/geoserver

[2] Available soon as OSS. The source code can already be shared based on a reasonable request.

[3] https://docs.geoserver.org/latest/en/user/production/index.html

[4] https://github.com/locustio/locust

doi:10.2760/2119408

or domain-specific language. Our Locust test suites include the modeling of user behaviour in requesting the tiles as suggested in the literature [6, 7, 8].

Locust has command-line and web-based interfaces, allowing for real-time monitoring of throughput, response times, and error rates. Thanks to its event-driven architecture (built on `gevent`), a single process can simulate thousands of concurrent users, making it highly suitable for testing systems like WMTS under high concurrency conditions. All Locust workers are deployed on the same client machine, leveraging its high-performance hardware to ensure that the generated load is not limited by client-side constraints.

Each virtual user in our test scenario initiates a single HTTP session and requests 10 tiles. This approach mimics a map browsing behavior, such as a user panning or zooming, which leads to multiple tile requests within a short time frame. This configuration is deliberate: distributing the same total number of tile requests per second across more users—each making fewer requests—would not produce equivalent conditions. Doing so would introduce additional overhead due to repeated HTTP session initialization, rather than reusing persistent connections. Consequently, such a scenario would artificially inflate network and protocol-related delays, deviating from the real-world usage pattern we aim to reproduce.

2.2. Preliniminary Experiments

Preliminary experiments indicate that client-side rendering time is negligible, as it only accounts for the completion of tile transfers—not their actual rendering in a map interface.

Following recommendations for the definition of a benchmark [4], we first tested the server's basic response capacity using a minimal HTTP service. This allowed us to validate the use of Locust for load generation and tune the number of workers appropriately. We also benchmarked the reverse proxy, an Nginx instance configured to accept up to 8K concurrent connections. The results confirm it does not represent a bottleneck in the tested configuration. To assess the capacity of the TileGeo WMTS server, we tested a single worker configured with 1 CPU core and 1 GB of RAM. As shown in Table 1, it can handle approximately 10 concurrent requests efficiently. Finally, we explored the effects of horizontal (adding more server instances, balancing the load between them) versus vertical scaling (upgrading the capabilities of the existing server). Results presented in Table 2 show no significant performance improvement. A similar investigation is also applied to GeoServer.

2.3. Data

The input datasets for our benchmark consist of Sentinel-2 Level-2A rasters over a heterogeneous test site in northern Italy, each containing only the three spectral bands (B02-Blue, B03-Green, and B04-Red) at 10 m resolution. To assess the impact of input payload volume, we prepared three input file-size categories where the covered area is different: "extra-small" counts a few megabytes, "small" hundreds of megabytes, and "large" totaling several gigabytes.

2.4. Hypotheses for Benchmark Definition

As a first step towards defining a robust and meaningful benchmark for WMTS performance, we formulated a set of hypotheses intended to identify potential factors that may introduce bias or variability in benchmark results. These factors, if unaccounted for, can lead to inconsistent or non-reproducible outcomes.

Our set of hypotheses investigates whether performance is affected by variations in the input data format, output image format, input file size, and requested zoom level. To evaluate this, we employed a range of commonly used geospatial raster formats for input, including GeoTIFF, JPEG2000 (JP2), and PNG. Each of these formats presents different characteristics in terms of file size, compression algorithms, and decoding complexity. For output formats, we tested PNG, JPEG, and WebP, as well as format variations involving different compression levels, color quantization strategies, and image quality settings. Additionally, we tested tile requests at multiple zoom levels to assess whether the scale and corresponding tile resolution influence performance, for example due to differing data volume per tile or variations in rendering complexity.

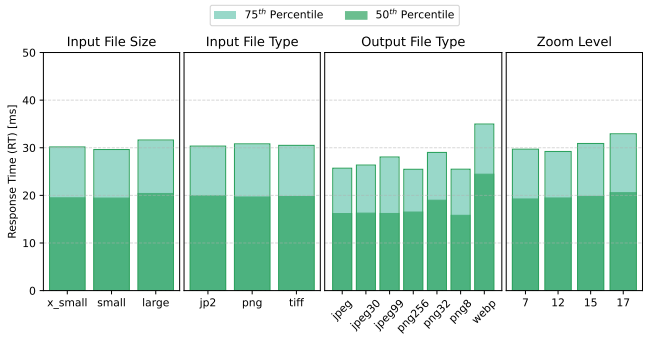
These combinations allow us to assess whether the encoding/decoding overhead, image optimization processes, or tile scale have a measurable influence on server-side performance or response time. Identifying such dependencies is crucial for defining benchmark conditions that are both fair and representative of real-world usage scenarios.

3. RESULTS

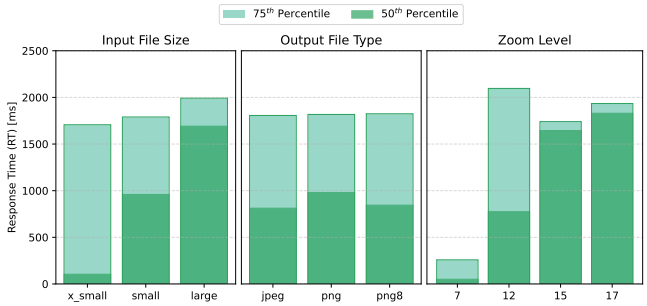
Validating the robustness of the WMTS performance measurements under different load scenarios is a key aspect of the evaluation process. For TileGeo, we conducted each experiment with both 100 and 200 concurrent simulated users, each issuing either a single-tile request or a batch of ten tiles at a time. Across all configurations, the latency distributions remained virtually unchanged between the 100 and 200 user loads and between one and ten tile batches, confirming that the results presented below are stable and reproducible regardless of moderate variations in request concurrency or tile batch size. GeoServer is more hardware resources demanding; we were able to simulate a maximum of 10 concurrent users in order not to degrade the application's performance.

Results are presented in Figure 2a for TileGeo and in Figure 2b for GeoServer. The aggregated performance measurements across our experimental conditions are summarized as follows. For input file size, "large", "small", and "extra small" tiles exhibited very similar latency distributions

doi:10.2760/2119408



(a)



(b)

Fig. 2: 50^{Th} (Dark green) and 75^{Th} (Green) percentiles of the response time (RT) per feature on Tilegeo (Figure 2a) and GeoServer (Figure 2b).

in the case of TileGeo. Instead, in the case of GeoServer, the input file size has a rather high impact with median (50th percentile) times ranging from 100 ms to 1700 ms. When varying output image format and encoding parameters, almost all the JPEG and PNG variations yield similar results for TileGeo, but in the case of WebP, the median value of 16 ms and the 95th percentile of 25 ms is significantly higher. Across zoom levels (7, 12, 15, 17), performance was again consistent for TileGeo, while this aspect has a quite relevant impact on GeoServer, degrading response time significantly as zoom levels increase. Finally, the input format (GeoTIFF, JP2, PNG) for TileGeo showed negligible differences, with all medians around 7 ms and 95th-percentiles around 19 ms.

4. CONCLUSIONS

WMTS(s) are a foundational component of digital infrastructure for visualizing geospatial big data, supporting many decision-making processes in various applications. While widely adopted, WMTS implementations vary significantly in performance, scalability, and resource efficiency, which are central aspects to be considered in cloud-native and big data applications. However, there is currently a lack of systematic and dedicated benchmarks that allow practitioners and decision-makers to evaluate and compare fairly different WMTS solutions. This is likely due to the fact that it is difficult to find WMTS with similar characteristics and scope. This contribution proposes an initial framework for identifying the elements a replicable benchmark for a WMTS should consider, thus promoting open science and interoperable solutions. We empirically evaluate open-source WMTS(s). The different behaviours of the considered WMTS presented in the results highlight that input file size, output encodings, and zoom levels are discriminatory aspects that have to be included in the benchmark definition of a WMTS. The number of WMTS servers tested can be a limit for this research. In future works, we will consider more WMTS and different system setups. Investigating other elements that may impact performance will also help better understand the additional aspects to be included in the benchmark of a WMTS.

REFERENCES

[1] Virva Svala. Web mapping tools for geospatial data visualizations–a literature review. 2025.

[2] Rostislav Netek, Jan Masopust, Frantisek Pavlicek, and Vilem Pechanec. Performance testing on vector vs. raster map tiles—comparative study on load metrics. *ISPRS International Journal of Geo-Information*, 9(2):101, 2020.

[3] Alexander Nylund. The server-side of tiled raster web maps: benchmarking and evaluating static raster tile map generation tools for geospatial data processing and visualization. 2023.

[4] Dirk Beyer, Stefan Löwe, and Philipp Wendler. Reliable benchmarking: requirements and solutions. *International Journal on Software Tools for Technology Transfer*, 21 (1):1–29, 2019.

[5] Jiri Horak, Jiri Ardielli, and Jan Ruzicka. Performance testing of web map services. *New Challenges for Intelligent Information and Database Systems*, pages 257–266, 2011.

[6] Xuefeng Guan, Bo Cheng, Aihong Song, and Huayi Wu. Modeling users' behavior for testing the performance of a web map tile service. *Transactions in GIS*, 18:109–125, 2014.

[7] Vinicius G Braga, Welder B de Oliveira, Vagner J do Sacramento Rodrigues, and Kleber V Cardoso. Understanding and modeling the behavior of web map users. *Journal of Information and Data Management*, 6(1):92–92, 2015.

[8] G Braga Vinicius, Sand L Corr, Vagner J do S Rodrigues, and Kleber V Cardoso. Characterizing user behavior on web mapping systems using real-world data. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 01056–01061. IEEE, 2018.

EXPLOITATION OF KNOWLEDGE GRAPH TECHNOLOGIES FOR GEOSPATIAL INTELLIGENCE USE CASES

Omar Barrilero, Paula Saameño, Michele Lazzarini, Miguel Angel Belenguer, Sergio Albani

European Union Satellite Centre

ABSTRACT

Geospatial Intelligence requires the exploitation and link of huge amounts of heterogeneous data. New developments in Knowledge Graph technologies enable to tackle real-life requirements and build large-scale applications, which is crucial for facing Geospatial Intelligence scenarios. In this paper, two different applications based on Knowledge Graphs technologies are presented to demonstrate the added value of Knowledge Graph for Geospatial Intelligence stakeholders. The first application integrates multiple and heterogeneous sources for the analysis of natural hazards and impact assessment, while the second application focuses on the relevant change detection for the monitoring of illegal activities. Within these use cases, Knowledge Graph demonstrates to be a useful technology for a flexible integration of knowledge, facilitating the analysis of complex relationships in the data, and an enhancement on Machine Learning algorithms and their explainability.

Index Terms— GEOINT, Knowledge Graph, Natural Hazards, Change Detection

1. INTRODUCTION

Geospatial Intelligence (GEOINT) is the collection, analysis, and visualization of geographically referenced data for the purpose of providing an understanding of a specific area or region. As the amount of geospatial data grows, the ability to effectively analyze and interpret them becomes increasingly important. The fusion of different data sources is key to GEOINT, as it allows for the consolidation, correlation, and analysis of data from disparate sources.

Knowledge graphs (KGs) are a flexible knowledge representation paradigm intended to facilitate the processing of knowledge for both humans and machines. They are widely recognized as a key enabler for a number of increasingly popular technologies including Web search, question answering, personal assistants, and (explainable) Artificial Intelligence (AI) across many sectors. However, KGs (including machine learning approaches like class expression learning) are not very popular in GEOINT domain because they present some limitations:

1. fail to scale to large graphs with billions of edges;
2. are bound to consistency in particular formalisms;
3. fail to exploit the semantics modelled into the KGs;
4. rely on a one-shot explanation paradigm if they are at all explainable.

In this paper, the results of different GEOINT use cases (developed as part of ENEXA[1] project) are presented to demonstrate the added value of KG technologies in this domain.

2. KG TECHNOLOGIES EXPLOITED

The core objective of ENEXA is to address the challenge of developing explainable Machine Learning (ML) approaches for real-world KGs with a focus on human-centred explanations. ENEXA has developed novel hybrid ML approaches that can exploit multiple representations of knowledge graphs concurrently taking advantage of different technologies presented here below. With these technologies it is possible to tackle real-life runtime requirements and build large-scale applications, which is crucial for facing GEOINT scenarios. Some of the developed technologies are introduced below.

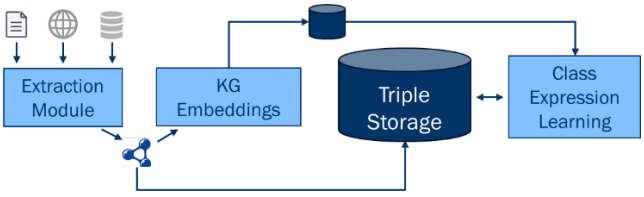


Fig. 1. Simple workflow of KG technologies considered.

2.1. Extraction modules

Extraction modules are in charge of getting information of the different sources and translate it into triples to be ingested in a KG, enabling the integration of multiple and heterogeneous data sources. For the GEOINT use cases addressed (detailed in following sections), both structured and unstructured datasets were considered. While structured data sources can be processed easily with “mappings”, unstructured datasets require a more complex processing considering Large Language Models (LLMs).

[1] https://enexa.eu/ - Project funded by the European Union

doi:10.2760/2119408

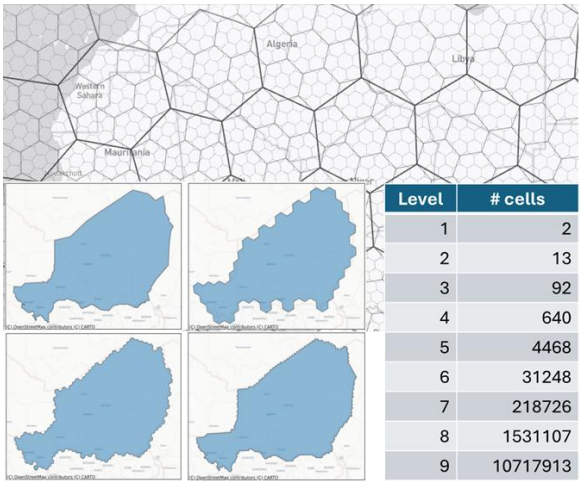
The two main challenges of the developed extraction modules are entity linking and the efficient management of geospatial data.

2.1.1. Entity linking

Entity linking is a key part of the information extraction pipeline connecting knowledge graphs to text. The methods developed in ENEXA [1] have been exploited. The entity linking enables the connection with other KG and source of information (e.g. wikidata), facilitating the exploitation of complex relationships and semantic meaning, which is one of the reasons for using KG instead of relational DB.

2.1.2. Management of spatial data

In GEOINT, the spatial relations of the features/events are critical. They can be exploited in a KG with the use of GeoSPARQL, that provides a topological ontology in RDFS/OWL for representation as well as a SPARQL query interface with a set of spatial SPARQL extension functions [2]. In order to be able to exploit the geospatial relations in triple storage not supporting GeoSPARQL, an alternative approach has been implemented. When using systems without GeoSPARQL support, the extraction module is able to generate additional triples that enable spatial queries. Within this approach, the H3 grid (Uber's Hexagonal Hierarchical Spatial Index) [3] is used. This reference grid is integrated in the KG (at a configurable level of precision (see Fig. 2)), and it is indicated with explicit triples for every spatial element in the KG, which are the H3 cells where the feature is. This way, it is later possible to make spatial queries without the GeoSPARQL functions.



Level	# cells
1	2
2	13
3	92
4	640
5	4468
6	31248
7	218726
8	1531107
9	10717913

Fig. 2. Example of H3 cells in Niger at different levels.

2.2. Triple storage

Efficient storage solutions for the KGs are indispensable for their use in real applications. In this work two different triples store have been used: Tentris [4] and GraphDB.

Tentris represents RDF knowledge graphs as sparse order-3 tensors using a novel data structure called hypertrie. It then uses tensor algebra to carry out SPARQL queries by mapping SPARQL operations to Einstein summation. By being able to compute Einstein summations efficiently, Tentris outperforms the commercial and open-source RDF storage solutions in terms of the average number of queries it can serve per second on datasets of up to 1 billion triples.

Tentris, although offer big advantages in terms of efficiency, has some limitations (work in progress) like the support for GeoSPARQL, but this issue has been solved by the use of a reference grid.

2.3. Embeddings generator

A Knowledge graph embedding (KGE) is a representation of a KG element as a vector with the objective of entity classification, link prediction or recommender systems.

DICE embeddings [5] has been used in this work to obtain the embeddings for both applications.

2.4. Class expression learning

Class expression learning aims to learn complex class expressions from the knowledge graphs. Users can harness machine learning techniques to derive meaningful insights and patterns from their data. Typically, the user can provide positive and negative examples (e.g. legal mining sites vs illegal mining sites) and the module extract "rules" to be able to classify new entities. This approach enables also the explainability of the results of a classification.

3. GEOINT USE CASE

3.1. GEOINT current challenges

The three main GEOINT challenges that can be faced using the KG technologies are:

- Data heterogeneity - In the GEOINT domain, huge amounts of data (Earth Observation (EO) and not-EO, in different formats like raster/vector or text), must be exploited together. For EO data, the collocation and transformation of raster/vector data is typically enough, but when managing the ancillary data different processing are needed, including LLM-based extraction pipelines.
- Limited use of contextual information - Some approaches in GEOINT processes just take into account the changes that are detected from EO imagery and classify them according to the changes in the reflectance (optical imagery) or back-scatter/coherence (in Synthetic-aperture radar (SAR) imagery) but, usually, to use more contextual information in order to properly classify the changes is very useful (e.g. land cover in the surroundings, distance to certain features).

doi:10.2760/2119408

This is usually made by analysts, but the automation of these processes would facilitate their job.

- Lack of Explainability - The explainability of the Machine Learning (ML) results would allow analysts to understand them and decide about their reliability.

3.2. GEOINT User stories

Different user stories have been collected from GEOINT stakeholders to assure the relevance of developed applications to real users.

1. *As a GEOINT analyst, I want to find relevant events from different sources (e.g. floods) in a specific area/time, so that I can assess their impact for relevant decision-makers (e.g. impact of events in refugee camps, impacts on critical infrastructures). For this, I want to generate queries linking OpenStreetMap data (or other geospatial databases) and event data, so that I can quickly find relevant details. Also, I want to generate queries to find infrastructure affected by specific events (e.g., floods) in a defined period / area, to support decision-making processes in relevant domains.*
2. *As a GEOINT developer, I want to find satellite images that I can use to train my models so that the labelling time is reduced.*
3. *As a GEOINT analyst, I have to monitor an area of interest during a period of several months. I want to be automatically notified when a relevant change takes place. Additionally, I want the platform to extract explanations supporting a change detected in satellite imagery, so that I can confirm most quickly their relevance.*

3.3. GEOINT applications

To address the user stories and demonstrate the added value of KG technologies, two different applications have been developed.

3.3.1. Integration of multiple and heterogeneous sources: Analysis of natural hazards and impact assessment

This application (addressing user stories 1 and 2) consists of the integration of multiple sources addressing natural hazards as well as ancillary/contextual information. The use of KG in this application aims to a) homogenize and complete the information by integrating information from different sources including data generated from EO algorithms developed by SatCen, b) detect inconsistencies (e.g. identification of duplications or contradictory data) and c) improve queries for analysts and developers, facilitating the discovery of correlations and complex links between the data.

The data integrated includes:

- Natural hazards data sources [6] containing information of past events (including affected people, severity and other event characteristics)
  - Desinventar – Sendai Framework
  - Colorado Flood Observatory
  - Copernicus Emergency Management Service
  - Wikipedia
- Ancillary data
  - OpenStreetMap for affected infrastructure
  - Socio-economic data (e.g. World Bank)
  - Land Cover
  - Satellite imagery metadata
  - ACLED (Armed Conflict Location and Event Data) [7] and migration/displacement data
- Results from EO pipelines (e.g. identification of flooding and impact assessment)

3.3.2. Relevant change detection

This application focuses on the detection of relevant changes (user story 3). The methodology has been applied, in particular, to the detection of possible illegal activities like illegal mining. It provides: a) the identification and classification of relevant changes (changes can be detected with EO, but it is needed to develop a methodology able to classify them as relevant or not) and b) the explanation of changes (the classification of the changes has to be explained in order to allow analysts a quick confirmation).

4. IMPLEMENTATION AND RESULTS

4.1. Integration of multiple and heterogeneous sources

For this application a new ontology was built based on the KnowWhereGraph ontology [8]. The entities were linked when possible with wikidata entities and other thematic dictionaries like sendai-hip for the hazard classification [9].

The extraction of the information and the ingestion in the triple store solutions of more than 1 billion triples was made in around one day, using python scripts and Tentris and GraphDB deployments in a k8s cluster.

The resulting KG has been exploited through Jupyter notebooks (see Fig. 3) with ready-to-use templates that enable final non-experts users to run SPARQL queries in a friendly way using widgets to define their requests. Some thematic examples are:

- Statistics/Evolution of past floodings in specific country/region - e.g. represent the number of floodings in Niger during the last decade;
- Identification of training datasets for new EO models - e.g. provide a list of Sentinel-1/2 images over an area affected by flooding;
- Estimation of severity of a new flooding based on historical data of previous flooding (similarity);
- Analysis of possible impact of hazards in conflicts and migration

doi:10.2760/2119408

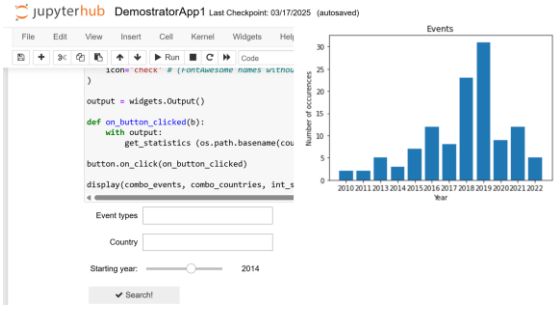


Fig. 3. Example of templates for the exploitation of KG developed.

4.2. Relevant change detection

The application consists of the integration of a monitoring pipeline in SatCen Geospatial platform (GEO-DAMP) (see Fig. 4) that is able to detect illegal mining activities based in three main models executed sequentially:

- Deforestation monitoring with HR imagery (Sentinel data)
- Contextualization of AoI: when a deforested area is detected, contextual information is extracted (an integrated in a KG) including land cover and new man-made structures (e.g. new airstrips, buildings) identified using Enhanced ML algorithms taking advantage of class expression learning
- Classification of changes as relevant or not using a trained model generated using class expressions learning obtained from positive and negative examples from known mining sites databases.

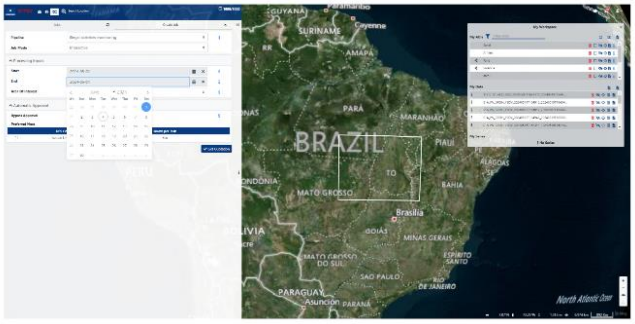


Fig. 4. Example of relevant change detection integration in SatCen Geospatial Platform (GEO-DAMP).

5. CONCLUSIONS AND WAY FORWARD

The use of KG technologies is supporting GEOINT analysts by allowing a flexible integration of huge amounts of data that facilitates the research of complex links between events by harmonizing data and combining siloed data sources. Moreover, it also enhances some ML algorithms while enabling the explainability.

The solutions developed can be optimized for specific applications. For example, the current KG for the analysis of natural hazards contains global data in order to demonstrate the scalability of the systems, but it could be adapted or even the application could create on-demand KGs with only data of specific countries or date ranges for specific tasks, allowing also to the users to decide which sources of information could be relevant to include.

Finally, although the creation of ready-to-use templates for the exploitation of the KG enables the use of the system by non-expert users, they cannot exploit its capability to its maximum extent without a minimum knowledge of SPARQL. For this reason, a future evolution will consist of supporting natural language queries. Also, an assessment of the use of other ontologies will be considered as YAGO2geo [10].

REFERENCES

[1] Fina Polat, Ilaria Tiddi, Paul Groth, and Piek Vossen. 2023. Improving Graph-to-Text Generation Using Cycle Training. In Proceedings of the 4th Conference on Language, Data and Knowledge, pages 256–261, Vienna, Austria. NOVA CLUNL, Portugal.

[2] Car, N.J.; Homburg, T.; Perry, M.; Herring, J.; Knibbe, F.; Cox, S.J.D.; Abhayaratna, J.; Bonduel, M. OGC GeoSPARQL—A Geographic Query Language for RDF Data. 2021. Available online (accessed on 30 October 2021).

[3] Uber Technologies, Inc., "H3: Hexagonal Hierarchical Spatial Indexing," Uber Technologies, Inc., 2019.

[4] Alexander Bigerl, Felix Conrads, Charlotte Behning, Mohamed Ahmed Sherif, Muhammad Saleem and Axel-Cyrille Ngonga Ngomo (2020) **Tentris – A Tensor-Based Triple Store. ** In: The Semantic Web – ISWC 2020

[5] Demir, Caglar and Ngomo, Axel-Cyrille Ngonga, "Hardware-agnostic computation for large-scale knowledge graph embeddings", Software Impacts, Elsevier 2022.

[6] SA Mazhin, M Farrokhi, M. Noroozi et al. "Worldwide disaster loss and damage databases: a systematic review". Journal of education and health promotion, 2021

[7] Raleigh, C., Kishi, R. & Linke, A. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. Humanit Soc Sci Commun 10, 74 (2023). https://doi.org/10.1057/s41599-023-01559-4

[8] Cogan Shimizu, Shirly Stephen, Adrita Barua, e. al. The KnowWhereGraph ontology. Web Semant. 84, C (Jan 2025). https://doi.org/10.1016/j.websem.2024.100842

[9] Murray, V. et al. (2020) Hazard Definition & Classification Review: Technical Report: Geneva, Switzerland, United Nations Office for Disaster Risk Reduction; Paris, France International Science Council, https://council.science/ wp-content/uploads/2020/06/UNDRR_Hazard-Report_ DIGITAL.pdf; or https://council.science/publications/hazards

[10] Karalis, Nikolaos & Mandilaras, Georgios & Koubarakis, Manolis. (2019). Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge. 10.1007/978-3-030-30796-7_12.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme within the project ENEXA under the grant No 101070305.

GRÜNBLICK - AI POWERED FOREST BIOMASS ESTIMATION SERVICE

Vytautas Jančauskas, Kalifou René Traoré, Juan Pablo Espejo Belmonte, Daniela Espinoza-Molina

Deutsches Zentrum für Luft- und Raumfahrt (DLR)
Data Science for Earth Observation Department
Münchener Str. 20, 82234 Weßling
Germany

1. ABSTRACT

Accurate forest biomass estimation is essential for climate change mitigation, biodiversity monitoring, and sustainable forest management. Recent advances in remote sensing and machine learning have opened new avenues for large-scale, high-resolution biomass mapping. In this work, we introduce Grünblick, an AI-powered service designed for scalable forest biomass estimation, leveraging multi-sensor Earth Observation (EO) data, including Sentinel-1 and Sentinel-2 imagery. The Grünblick pipeline integrates modular deep learning models, notably U-Net architectures with interchangeable feature extraction backbones, to perform pixel-wise aboveground biomass (AGB) regression. We validate our system using the public Biomassters benchmark, demonstrating significant performance gains through multi-modal sensor fusion and self-supervised pretraining strategies. Future extensions will include uncertainty quantification and global deployment capabilities.

2. INTRODUCTION

Forest biomass estimation has been a research focus for many years [1]. It is essential for various applications, including climate change mitigation, biodiversity assessment, forest management, and sustainable supply chain monitoring.

With the advent of Artificial Intelligence (AI) and Big Data from space, new methodologies have emerged, making the field of forest biomass estimation highly dynamic and interdisciplinary. This area combines remote sensing technologies, machine learning, and ecological modeling advances. Recent studies demonstrate improved AGB estimation through multi-source satellite fusion and the application of deep learning models [2, 3, 4].

Existing efforts like ESA's Climate Change Initiative (CCI) [5], FAO's Open Foris [6], and the BioMassters benchmark [7] support biomass estimation using remote sensing data, though often at lower resolutions or with less flexibility than AI-powered systems.

In this work, we introduce Grünblick, an AI-driven service designed for comprehensive forest monitoring. It targets forestry management, environmental conservation, and land-use monitoring industries. Grünblick integrates state-of-the-art deep learning architectures with software tools to facilitate global end-to-end forest biomass querying. Our system will enhance the value of Copernicus products by utilizing multispectral Sentinel-2 and Sentinel-1 Earth Observation (EO) data to generate precise forest biomass estimates.

3. GRÜNBLICK PIPELINE

Grünblick is a biomass estimation service currently under development at the German Aerospace Center (DLR). Its primary objective is to enable continuous, large-scale monitoring of biomass using a combination of multi-source satellite data, particularly multi-spectral imagery (MSI) and synthetic aperture radar (SAR). The service is being designed with adaptability and scalability in mind—capable of ingesting new data streams as they become available and integrating emerging data modalities and sources over time.

The overarching goal of Grünblick is to provide regularly updated, high-resolution biomass estimates that can support decision-making in forestry and related sectors. To achieve this, we are developing a modular processing pipeline composed of two main components: (1) a Machine Learning (ML) Model Training and Inference Service, and (2) a front-end system responsible for data caching, visualization, and delivery to end users. A high-level overview of this processing pipeline is illustrated in Figure 1.

One of the key features of Grünblick is its interactive user service, which will allow users to submit a request via a JSON file specifying an area of interest. In response, the system will generate and return a raster file containing biomass estimates for the specified region.

The ML Model Training and Inference Service underpins the core estimation functionality. This component includes the development of biomass prediction models and associated data acquisition workflows, all implemented in Python. The software is hosted on GitHub (private at the time of writing) and follows CI/CD-based development practices.

We use Google Earth Engine to obtain co-registered

doi:10.2760/2119408

Sentinel-1 and Sentinel-2 patches aligned with ground-truth AGB data, which support supervised model training. Once trained, models are deployed in an operational pipeline that includes a geographic cache for efficient access and a data delivery service enabling users to download biomass maps on demand.

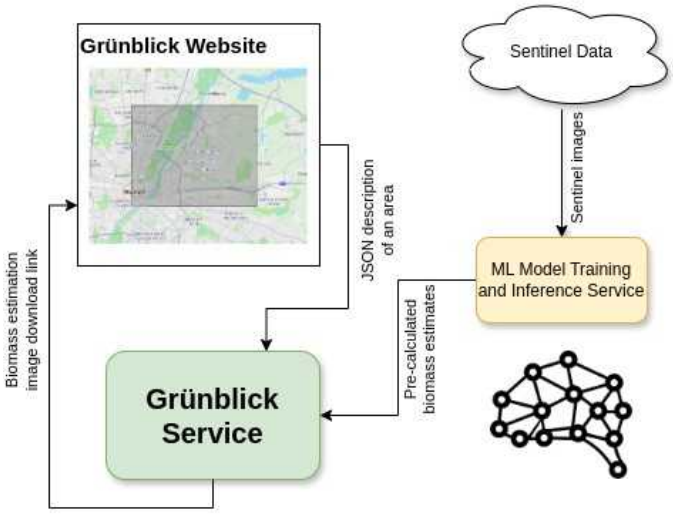


Fig. 1. Grünblick Schematic Diagram

4. DATASETS

This section outlines the dataset used for estimating above-ground biomass (AGB) in forested landscapes. The data pipeline ingests multi-source satellite imagery and aligns it with ground-truth biomass observations to train and validate machine learning models. Two primary sources are utilized: the BioMassters benchmark, and a regional dataset focused on British Columbia, Canada. These were chosen to support the Grünblick pipeline (see Fig. 1), which requires high-resolution imagery, spatial consistency, and reliable biomass reference data for supervised learning.

4.1. BioMassters

The BioMassters dataset [7] serves as a principal test bed for the experiments in this paper. it is a publicly available benchmark for above-ground biomass (AGB) estimation, derived from extensive LiDAR-based forest inventory plots across multiple regions in Finland. Collected across multiple years, between September 2016 and August 2021, the dataset offers pixel-level annotations of biomass components, including whole stem, branch, foliage, and bark, expressed in tons per hectare. Given that the AGB ground truth is not collected in one pass, the authors use a stratified sampling strategy in order to avoid any temporal bias in the geographic coverage of the territory. In practice, they identify reference areas that evenly cover the territory, for the AGB measurement campaign of each year. Then, the obtained ground-truth values are rasterized to align with the spatial resolution of the accompanying monthly satellite imagery of the past year.

Each of the 13000 data samples consists of a 2560×2560 meter satellite image patch at a 10×10 meter resolution co-registered with a 16×16 meter resolution biomass grid derived from LiDAR. The dataset encompasses approximately 8.5 million hectares of forested area and includes over 310,000 paired satellite–ground truth samples. This precise spatial alignment enables accurate pixel-level supervision and supports high-resolution model training. According to the authors, the test set is generated using 20 to 30 percent of the references areas for most of the temporal strata (2018, 2019, 2020, 2021), which are already evenly spread over the territory and without overlapping images. The remaining samples constitute the training set, which we separate into our final training and validation sets, using a 80 versus 20 percent random splitting rule.

A key strength of the BioMassters dataset lies in its rich biomass component labels, which facilitate detailed analysis of ecological structure and allow for the evaluation of model performance across a variety of biomass-related attributes. Furthermore, its standardized format and public availability make it a valuable resource for benchmarking and model comparison within the research community.

This dataset plays a crucial role in evaluating the transferability and robustness of the Grünblick system, serving as a complementary reference to our regional dataset and providing a foundational step toward the system's global scalability.

4.2. British Columbia

British Columbia (BC) is a region of interest for future experiments, due to its ecological heterogeneity and the availability of detailed forest inventory data. The province possesses a wide range of forest types and climatic zones, making it an ideal setting to assess model generalizability.

Satellite data are obtained from the Sentinel-1 and Sentinel-2 missions of the Copernicus program. Sentinel-2 provides multispectral imagery across bands including B2 (blue), B3 (green), B4 (red), B8 (near-infrared), and B11/B12 (shortwave infrared)—spectral regions frequently used in vegetation analysis. Sentinel-1 supplies synthetic aperture radar (SAR) data, offering complementary structural information and ensuring data availability under all weather conditions.

Ground-truth biomass estimates are derived from the British Columbia Vegetation Resources Inventory (VRI), maintained by the provincial government [8]. As of 2023, the dataset includes over 5.9 million delineated forest stands, covering roughly 5.5 GB in spatial data.

5. BIOMASS ESTIMATION METHODOLOGIES

The Grünblick system delivers above-ground biomass (AGB) analytics as a service for forested regions in the Northern

doi:10.2760/2119408

Hemisphere. Its analytical back-end consists of pixel-wise AGB regressors based on a modular U-Net architecture [9], adapted from semantic segmentation to perform continuous biomass regression using satellite imagery.

We employ a fixed U-Net structure while varying the encoder (feature extractor) to balance performance and efficiency. The toolbox supports backbones such as VGG, ResNet (18–50 layers), and EfficientNet-B0 [10], allowing flexible trade-offs in model complexity and accuracy. As shown in Figure 2, the choice of backbone significantly affects the quality of the predictions.

Each model currently operates in a mono-temporal setting, predicting peak AGB for a specific timestamp. Besides, the system also supports multi-sensor input fusion via channel-wise stacking of co-registered SAR and MSI data, enabling richer feature representation. Because of the current temporal restriction, our system does not perform any harmonization for the time series of the available of modalities. For the case of sensor fusion, we only consider reference areas and timestamps where both sensors are simultaneously available.

Additionally, we incorporate transfer learning using self-supervised weights pretrained on large EO datasets such as SSL4EO-S12 [11], boosting performance with limited labeled data.

Future work will integrate uncertainty quantification [12] to provide confidence intervals alongside predictions, supporting risk-aware decision-making in environmental monitoring. Additional extensions may include multi-temporal estimation, though this would increase computational demands and require some harmonization, potentially achieved by considering as target temporal consistency, the temporal availability of the most irregular sensor (Sentinel-2).

6. RESULTS

We evaluate the Grünblick system on the BioMassters dataset (Section 4.1) using two benchmarks: (1) multi-modal sensor fusion and (2) self-supervised pre-training. Model performance is reported using the coefficient of determination (R^2), a standard metric for regression.

6.1. Multi-Modal Estimation

To assess the benefit of combining EO data sources, we trained U-Net regressors with inputs from Sentinel-1 (SAR), Sentinel-2 (MSI), and their channel-wise fusion. Each configuration was evaluated using multiple encoder backbones: ResNet-18, ResNet-34, and EfficientNet-B0. As shown in Figure 2, fused inputs consistently outperformed single-sensor setups. For instance, with EfficientNet-B0, R^2 scores improved from 0.541 (S1) and 0.583 (S2) to 0.688 (fusion).

Backbone selection also affected performance. EfficientNet-B0 (5M parameters) outperformed the deeper ResNet-34 (22M, $R^2 = 0.651$) and ResNet-18 (12M, $R^2 = 0.659$), highlighting that smaller, efficient architectures can match or exceed heavier models.

These results confirm that Grünblick's modular architecture and multi-modal capability significantly enhance biomass estimation. For reference, the benchmark-leading U-TAE model [7], which uses multi-temporal imagery, achieves an R^2 of 0.765—a target future versions of Grünblick could approach by integrating temporal dynamics.

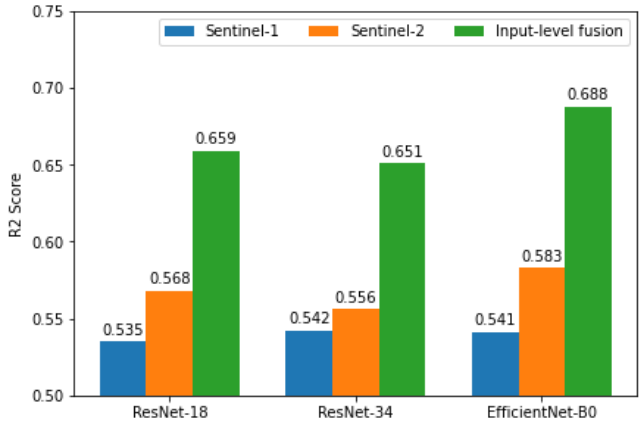


Fig. 2. Performance of U-Net regressors using different sensor inputs and backbones. Models trained for 50 epochs on the BioMassters dataset.

6.2. Effect of Self-Supervised Pre-Training

We also evaluated the impact of self-supervised pre-training (SSL) using U-Net models with ResNet-50 backbones. Initial weights came from random initialization or SSL on Sentinel-1 or Sentinel-2 data. As shown in Figure 3, sensor-consistent pre-training improved accuracy. For example, using SSL weights trained on Sentinel-1 for Sentinel-1 fine-tuning yielded $R^2 = 0.544$, compared to 0.535 from random initialization. Cross-modal pre-training (e.g., S2-to-S1) decreased performance to $R^2 = 0.524$.

In fusion scenarios, both SSL variants improved performance over baseline: Sentinel-2 pre-training resulted in $R^2 = 0.666$, Sentinel-1 yielded 0.658, and random initialization gave 0.653. These findings indicate that SSL improves transferability when the pre-training modality aligns with the downstream task.

In summary, Grünblick benefits from both input fusion and SSL-based initialization, enabling higher accuracy without added model complexity or data requirements.

7. CONCLUSION

In this work, we introduced Grünblick, an AI-driven service designed for scalable and accurate forest above-ground biomass (AGB) estimation. As a testbed, we focused on two

doi:10.2760/2119408

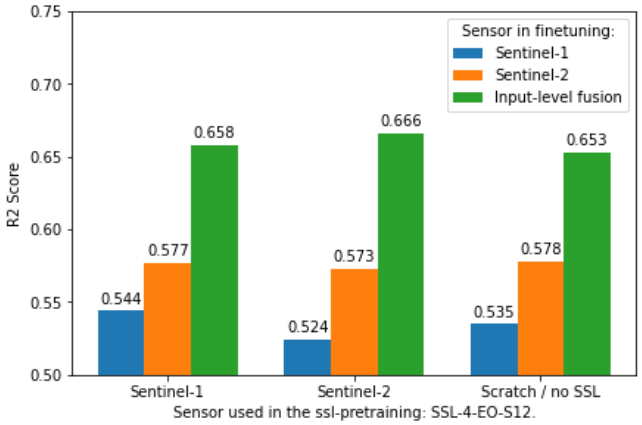


Fig. 3. R^2 for U-Net regressors (ResNet-50) trained with random or SSL-initialized weights.

regions in the Northern Hemisphere: Finland (using the public Biomassters benchmark), and British Columbia, Canada (using an in-house dataset). Both regions offered diverse conditions for evaluating AGB estimation models based on multi-sensor Earth Observation data, specifically SAR and multispectral imagery.

To address this task, we developed a modular deep learning toolbox based on U-Net architectures, enabling flexible integration of various feature extractors and sensor modalities.

Benchmark experiments demonstrate that Grünblick successfully leverages sensor fusion and self-supervised learning techniques to improve model accuracy. Specifically, fusing Sentinel-1 and Sentinel-2 data significantly enhances prediction performance compared to using individual sensors. Furthermore, fine-tuning models with pre-trained weights on EO datasets further boosts estimation accuracy, highlighting the value of transfer learning for biomass mapping.

Looking ahead, Grünblick will be extended to include uncertainty quantification in biomass predictions, offering users additional insight into the reliability of the estimates. We also plan to scale the system toward global deployment, enabling continuous, high-resolution biomass monitoring as a service for forestry, conservation, and climate research communities.

REFERENCES

[1] Taiyong Ma, Chao Zhang, Liping Ji, Zheng Zuo, Mukete Beckline, Yang Hu, Xiaojuan Li, and Xiangming Xiao. Development of forest aboveground biomass estimation, its problems and future solutions: A review. *Ecological Indicators*, 159:111653, 2024.

[2] Yaxuan Xing, Feng Wang, and Feng Xu. Above ground biomass estimation by multi-source data based on interpretable dnn model. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1894–1897, 2023.

[3] Tuomas Hame, Yrjö Rauste, Oleg Antropov, Heikki A. Ahola, and Jorma Kilpi. Improved mapping of tropical forests with optical and sar imagery, part ii: Above ground biomass estimation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(1):92–101, 2013.

[4] Autumn Nguyen and Sulagna Saha. Machine learning and multi-source remote sensing in forest carbon stock estimation: A review, 2024.

[5] European Space Agency. Climate change initiative: Biomass, n.d. Accessed: 2025-04-23.

[6] Food and Agriculture Organization of the United Nations. Open foris: Tools for forest monitoring, n.d. Accessed: 2025-04-23.

[7] Andrea Nascetti, Ritu Yadav, Kirill Brodt, Qixun Qu, Hongwei Fan, Yuri Shendryk, Isha Shah, and Christine Chung. Biomassters: A benchmark dataset for forest biomass estimation using multi-modal satellite time-series. In *Advances in Neural Information Processing Systems*, volume 36, pages 20409–20420. Curran Associates, Inc., 2023.

[8] Government of British Columbia. VRI - 2023 - Forest Vegetation Composite Rank 1 Layer (R1), 2023. Accessed: 2025-04-30.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[10] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[11] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, September 2023.

[12] Nils Lehmann, Nina Maria Gottschling, Jakob Gawlikowski, Adam J. Stewart, Stefan Depeweg, and Eric Nalisnick. Lightning uq box: Uncertainty quantification for neural networks. *Journal of Machine Learning Research*, 26(54):1–7, 2025.

doi:10.2760/2119408

AUTOMATING EARTH OBSERVATION ANALYTICS PIPELINES WITH AGENT RAVEN

Gereon Dusella[*], *Haralampos Gavriilidis*[*], *Binger Chen*[*],
Begüm Demir[*], *Volker Markl*[*,‡], *Eleni Tzirita Zacharatou*[§]

[*]BIFOLD & Technische Universität Berlin, [‡]DFKI, [§]HPI & Universität Potsdam

ABSTRACT

Efficient integration of vector databases, such as those containing administrative boundaries and land parcels, with remote sensing images is essential for various Earth Observation (EO) applications. Zonal statistics (ZS) offer a powerful tool for this purpose, but their computation remains challenging due to fragmented system interfaces, diverse preprocessing needs, and inconsistent performance across systems. Current methods optimize execution within single systems but lack support for dynamic, cross-system workflows. To address this, we present Agent Raven, the first AI-driven multi-agent system designed to autonomously manage the full lifecycle of ZS computation and deployment. Building on the Raven framework, Agent Raven dynamically selects execution backends, optimizes query pipelines, and adaptively manages workflows based on previous experiments. Our work represents a step forward in intelligent orchestration across heterogeneous systems in EO data analytics.

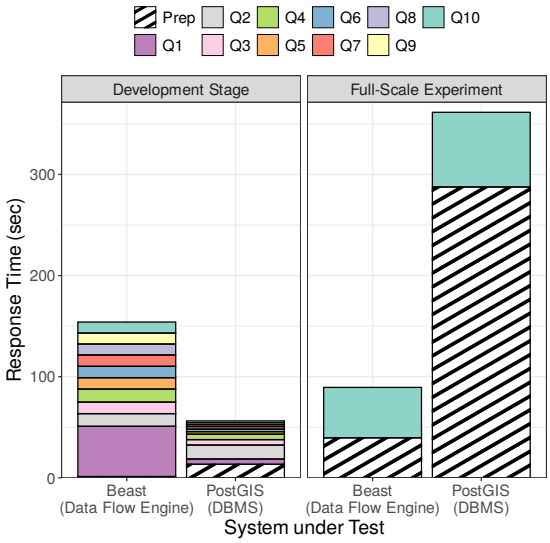


Fig. 1: Performance in different development phases

1. INTRODUCTION

The availability of remote sensing imagery has significantly increased [1, 2, 3] due to advancements in satellite technology. Programs like Copernicus [13] provide vast amounts of freely available raster data, while the volume of vector datasets (e.g., OpenStreetMap, governmental geospatial data) is also expanding. To effectively utilize these data for Earth Observation applications (e.g., climate monitoring, wildfire prediction, urban planning) [18, 17, 14, 16], efficient processing techniques are essential. A key step in these applications is the computation of Zonal Statistics (ZS), where pixel-based raster data are aggregated within defined vector-based geometries, such as city boundaries or farmland parcels. For example, to identify deforested areas, one can apply ZS on satellite images and polygons that define forest boundaries [15].

Computing ZS requires combining raster (gridded cells) and vector (geometric features) data. Geospatial systems such as PostGIS[1] and Beast [5] handle these data types, but their APIs and performance vary widely. This variability forces data scientists to navigate multiple systems, adding complexity and inefficiency. The architecture of each system also impacts its suitability for different stages of a data science project. For example, PostGIS is well-suited for development involving multiple queries on smaller datasets, while Beast is better for large-scale experiments with single-shot queries. Figure 1 illustrates this distinction.

While optimizing individual queries is well-studied in data systems engineering, optimizing the iterative process of refining an analytics pipeline is significantly more challenging. This challenge is even greater for ZS queries, as not all geospatial systems support raster-vector joins equally, forcing users to adjust queries for each system's unique API. In addition, pre-processing steps, such as rasterization, vectorization, format conversion, and coordinate reference system (CRS) alignment, are often necessary, depending on the data model of each system. As queries evolve, the optimal combination of parameters can change significantly. These challenges make it difficult for users to efficiently manage evolving ZS workflows across heterogeneous geospatial systems.

Given these challenges, emerging AI agent technologies offer promising solutions for automating complex geospatial workflows. AI agents are bridging the gap between computer scientists and other users, enabling almost anyone to accom-

[1] https://postgis.net/

doi:10.2760/2119408

plish tasks that once required years of expertise. They are now used across various fields for tasks like visual reasoning [9], code generation [11], scientific experimentation [8], and model interpretation [10]. In geospatial applications, there has been a shift from manually created processing scripts to intelligent agents that autonomously manage remote sensing data, select tools, and refine outputs for tasks like land cover mapping, change detection, or geospatial question answering [7, 12, 6]. These advancements lay the groundwork for dynamic, multi-system geospatial workflows.

We propose Agent Raven, our vision for an AI-powered assistant that supports data scientists at all stages of ZS experimentation - from initial development to continuous deployment. Agent Raven interfaces with the Raven core component [4], our framework for executing ZS queries seamlessly across multiple geospatial systems. By selecting the optimal backend and applying query optimizations based on a database of past experiments, Agent Raven learns and enhances performance over time. To the best of our knowledge, Agent Raven is the first system to offer deep, end-to-end integration of geospatial data science tasks, aiming to significantly reduce development time and operational complexity.

Our contributions are twofold. First, in Section 2, we describe how Raven integrates heterogeneous geospatial systems, providing uniform access and enabling seamless switching between them. This integration simplifies interoperability and lessens the workload for data scientists. Second, in Section 3, we propose Agent Raven, our vision for an AI-driven assistant that supports data scientists across the full lifecycle of ZS experimentation. By leveraging past experiment data, Agent Raven accelerates the transition from early-stage development to robust production pipelines.

2. PLAIN RAVEN FRAMEWORK

Today's data scientists face multiple challenges when implementing zonal statistics, due to the varying interfaces and configuration parameters exposed by existing geospatial systems, the varying pre-processing steps that these systems require, and their divergent runtime performance capabilities. In response to these challenges, Raven[2] aims to: 1) offer an easy-to-use zonal statistics interface; and 2) highlight performance differences in geospatial systems. To achieve this, Raven exposes a declarative zonal statistics interface based on a DSL that we developed. Using this DSL, Raven can transparently optimize and execute a given zonal statistics task on multiple geospatial systems. As a result, Raven provides system independence, thereby helping users avoid vendor lock-ins. Furthermore, by automating execution and providing detailed performance results, Raven simplifies selecting the most efficient system for a given workload. In the following, we give a brief overview of Raven's components.

[2] https://github.com/polydbms/RaVeN

2.1. Architecture Overview

Figure 2 presents Raven's architecture. Raven accepts a ZS task expressed in its DSL (the query) and relies on its `Pipeline Planner` for optimization. Combined with a `Capabilities` file specifying any system limitations, the planner identifies any necessary pre-processing steps, such as format or CRS conversions, and builds a Pipeline representation that it passes to the `Execution Interface`. This system-developer-provided interface includes a `IR` (Internal Representation) `Converter` and a `GSS` (Geospatial System) `Connector`. The IR Converter translates Raven's abstract syntax tree (AST) into system-specific code using parameterized templates, and the GSS-Connector enables execution on the underlying systems and result retrieval. Additionally, Raven stores execution metrics, e.g., runtime and resource consumption for each step, in its experiment database, which is accessible to other systems. The current systems supported by Raven are PostGIS, Beast, Apache Sedona[3], HeavyDB[4], and RasDaMan[5].

2.2. Zonal Statistics Parameters

To simplify Zonal Statistics (ZS) queries across different geospatial systems, Raven provides a domain-specific language (DSL, Listing **1**) that abstracts system-specific syntax and allows users to define and tune ZS queries in a structured way. We have identified four key operator classes that a tunable ZS query consists of: *Dataset operators* (L. 2–4) specify the raster and vector datasets used for analysis. *Aggregation operators* (L. 6–7) define how pixel values within vector-defined zones are processed, including grouping, filtering, and computing summary statistics. *System operators* (L. 9) determine which geospatial system executes the query. *Execution Parameter operators* (L. 11–12) allow fine-tuning of execution, such as raster tile size adjustment, vector simplification, and CRS alignment.

2.3. Zonal Statistics Pipelines and Optimizations

The AST generated by Raven's `Pipeline Planner` (cf. Figure 2) encapsulates the end-to-end processing of a ZS task. This includes pre-processing operations, such as changing format to support loading into the given system, aligning CRSs, filtering the datasets, as well as the join and aggregation operations. Raven then allows a data scientist or AI agent to configure each of these parameters individually, enabling fine-grained control over the pipeline execution.

We can categorize these parameters into three groups. First, they can reduce the number of processed pixels and vector features as early as possible. Second, they can tune the

[3] https://sedona.apache.org/
[4] https://heavy.ai
[5] https://rasdaman.org/

doi:10.2760/2119408

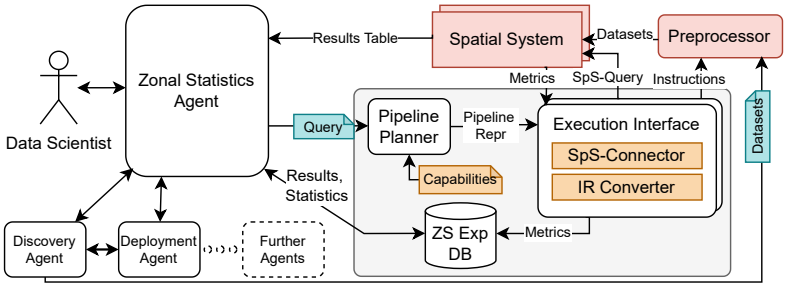


Fig. 2: Agent Raven Architecture

```
# Datasets definition
zs_result = ZSGen(
    raster="/data/vegetation_idx",
    vector="/data/plots") \
# Aggregation operations
  .group("oid") \
  .summarize({"avg": ZSGen.AVG}) \
# Systems
  .system([ZSSystem.Beast]) \
# Parameter settings (optional)
  .raster_tile_size("auto") \
  .vector_simplify([0, 200])
```

Listing 1: ZS Query in Raven's DSL

partitioning of raster and vector data to improve query execution. Third, they can minimize computational overhead by avoiding expensive operations when possible. These tuning techniques leverage the available ZS parameter operators and incorporate methods from existing research. Together, they can speed up ZS queries significantly [4].

2.4. Benchmarking Mode

The performance of ZS tasks in different geospatial systems can vary significantly depending on the data and workload. To uncover this, Raven features a dedicated benchmarking mode. This mode allows users to execute multiple pipelines and produce detailed performance plots, e.g., breakdown performance of different pipeline stages, facilitating easy comparison of different systems and parameter combinations. As a result, users can gain insights into potential bottlenecks and enhance system performance by fine-tuning available parameters. Overall, Raven's integrated benchmarking component provides valuable tools for optimizing zonal statistics tasks across diverse geospatial systems.

3. AGENT RAVEN: AUTONOMOUS ZONAL STATISTICS BEYOND RAVEN

Calculating the results of a ZS task is only one part of broader multi-stage geospatial data science pipelines. To cover other parts of the pipelines, we propose Agent Raven, a multi-agent extension of the plain Raven framework. Agent Raven allows users to describe analytical goals in natural language, while internal agents automatically discover datasets, construct pipelines, invoke tools like plain Raven, and manage execution from development to deployment. It can also retrieve external resources, such as data catalogs or tool manuals, to assist its reasoning during task planning and execution.

3.1. Background on AI Agents for Geospatial Workflows

An AI agent is a system that perceives its environment, reasons about goals, and acts autonomously to fulfill user requests. These agents typically leverage large language models and tools to interpret user intent, plan tasks, access external systems, and manage workflows adaptively. While traditional single-agent systems struggle with scalability, specialization, and responsiveness when workflows become complex, multi-agent systems organize multiple specialized agents under an orchestrator that coordinates their collaboration [6]. Each agent focuses on a smaller set of capabilities, such as dataset discovery, ZS, or deployment, while the orchestrator handles planning, task assignment, and execution monitoring. Multi-agent frameworks can more easily scale across domains, integrate heterogeneous tools, recover from errors, and provide faster intermediate feedback.

Agent Raven uses a multi-agent framework consisting of three agents: the `Discovery Agent`, which identifies and retrieves relevant input datasets; the `Zonal Statistics Agent`, which constructs ZS pipelines and selects appropriate geospatial systems by invoking the core Raven system as a tool; and the `Deployment Agent`, which manages downstream tasks such as continuous monitoring and scheduled deployment. A centralized memory allows the system to improve decision-making across tasks. This shared memory will replace the database used in the plain version of Raven.

3.2. Workflow Example of Agent Raven

We envision Agent Raven as a deeply integrated multi-agent system, where specialized AI agents collaborate to automate the end-to-end ZS workflow. Instead of requiring users to manually script queries, Agent Raven allows users to simply express their goals in natural language. The agents then automatically handle dataset retrieval, pipeline construction, execution, and continuous pipeline deployment.

Consider a data scientist interested in monitoring a specific geospatial area over an extended period, such as tracking the percentage of trees in a given region to observe deforestation. They would interact with Agent Raven by specifying the task and suggesting relevant criteria for suitable datasets. In response, the orchestrator in Agent Raven schedules dataset retrieval to its Dataset Discovery Agent, which searches for and returns a selection of candidate datasets, complete with metadata. To minimize perceived latency and improve user experience, Agent Raven proactively initiates multiple parallel actions. While the Dataset Discovery Agent retrieves

doi:10.2760/2119408

datasets, the ZS Agent begins preparing preliminary pipeline templates based on the user's task description. When candidate datasets are identified, the ZS Agent automatically generates the pipeline representation and invokes the underlying plain Raven engine as a tool to execute the ZS operations. This process includes selecting an efficient geospatial system capable of handling the candidate datasets. To provide early feedback and save resources, Agent Raven initially executes the pipeline on a small geospatial subset, quickly producing preliminary results. In cases where datasets are particularly large, Agent Raven may suggest applying approximate query processing to reduce the dataset size and speed up the query, while trading accuracy. Additionally, Agent Raven leverages its shared memory, which records all past task steps, parameters, and outcomes, to predict optimal configurations based on prior experience, further reducing the need for user intervention. If a dataset appears highly promising, Agent Raven can even pre-run partial queries while awaiting final user confirmation, further improving the perceived latency.

Once the user is satisfied with the preliminary results, Agent Raven will switch over to full-scale experiment mode. It again analyzes all parameters, considering available resources and time, selecting the best system and execution strategy. The final results will be passed to the continuous Deployment Agent, which manages ongoing deployments and regularly updates Agent Raven on its operations. This allows Agent Raven to adapt any parameters if necessary.

4. OUTLOOK

This paper presents Agent Raven, an AI-powered multi-agent system that automates ZS-based EO analytics pipelines across diverse geospatial systems. By adaptively selecting execution systems, optimizing queries, and managing workflows based on historical performance, Agent Raven enhances both the efficiency and accessibility of EO applications.

Looking ahead, we plan to extend Agent Raven with real-time data streaming support and integrate additional geospatial backends. Moreover, we plan to incorporate fault-tolerant execution strategies within the multi-agent framework to ensure robust execution in dynamic environments.

REFERENCES

[1] European Space Agency. Copernicus data space ecosystem, 2024. URL https://dataspace.copernicus.eu/.

[2] Ahmet Kerem Aksoy, Pavel Dushev, Eleni Tzirita Zacharatou, Holmer Hemsen, Marcela Charfuelan, Jorge-Arnulfo Quiané-Ruiz, Begüm Demir, and Volker Markl. Satellite image search in AgoraEO. *PVLDB*, 15(12):3646–3649, 2022.

[3] Arne de Wall, Björn Deiseroth, Eleni Tzirita Zacharatou, Jorge-Arnulfo Quiané-Ruiz, Begüm Demir, and Volker Markl. Agora-EO: A Unified Ecosystem for Earth Observation – A Vision for Boosting EO Data Literacy –. In *Proc. Big Data from Space (BiDS)*, 2021.

[4] Gereon Dusella, Haralampos Gavriilidis, Laert Nuhu, Volker Markl, and Eleni Tzirita Zacharatou. Multi-Backend Zonal Statistics Execution with Raven. In *SIGMOD/PODS*, pages 532–535. ACM, 2024. doi: 10.1145/3626246.3654730.

[5] Ahmed Eldawy et al. Beast: Scalable Exploratory Analytics on Spatio-temporal Data. In *CIKM*, pages 3796–3807. ACM, 2021.

[6] Chaehong Lee et al. Multi-agent geospatial copilots for remote sensing workflows. *ArXiv*, abs/2501.16254, 2025.

[7] Chenyang Liu et al. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *CoRR*, abs/2403.19646, 2024.

[8] Chris Lu et al. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292, 2024.

[9] Dídac Surís et al. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, pages 11854–11864. IEEE, 2023.

[10] Tamar Rott Shaham et al. A multimodal automated interpretability agent. In *ICML*. OpenReview.net, 2024.

[11] Tanmay Gupta et al. Codenav: Beyond tool-use to using real-world codebases with LLM agents. *CoRR*, abs/2406.12276, 2024.

[12] Wenjia Xu et al. Rs-agent: Automating remote sensing tasks through intelligent agents. *ArXiv*, abs/2406.07089, 2024.

[13] European Commission. Copernicus programme. https://www.copernicus.eu/en, 2025.

[14] Stefanie Holzwarth and et al. Earth Observation Based Monitoring of Forests in Germany: A Review. *Remote Sensing*, 12 (21):3570, January 2020. doi: 10.3390/rs12213570.

[15] Parag Kadam, Nicholas Magnan, and Puneet Dwivedi. A spatial dependence approach to assessing the impacts of Sustainable Forestry Initiative's Fiber Sourcing certification on forestry Best Management Practices in Georgia, United States. *Forest Policy and Economics*, 157:103071, 2023. doi: 10.1016/j.forpol.2023.103071.

[16] Paul J. Pinter, Jr., Jerry L. Hatfield, James S. Schepers, Edward M. Barnes, M. Susan Moran, Craig S.T. Daughtry, and Dan R. Upchurch. Remote Sensing for Crop Management. *Photogrammetric Engineering & Remote Sensing*, 69(6):647–664, 2003. doi: 10.14358/PERS.69.6.647.

[17] Jerry C. Ritchie, Paul V. Zimba, and James H. Everitt. Remote Sensing Techniques to Assess Water Quality. *Photogrammetric Engineering & Remote Sensing*, 69(6):695–704, June 2003. doi: 10.14358/PERS.69.6.695.

[18] Kali E Sawaya, Leif G Olmanson, Nathan J Heinert, Patrick L Brezonik, and Marvin E Bauer. Extending satellite remote sensing to local scales: Land and water resource monitoring using high-resolution imagery. *Remote Sensing of Environment*, 88(1):144–156, 2003. doi: 10.1016/j.rse.2003.04.006.

doi:10.2760/2119408

An Interoperable Data Economy to Enable GeoAI via Spatial Tokenizers (DGGS)

Michael Jendryke[1,2], João Manuel[1,2], Ludovic Augé[3], Emmanuel Mondon[4], Gino Caspari[1,2]*

[1]GeoInsight, Kaiserstr. 25, 59505 Bad Sassendorf, DE; [2]GeoInsight, Forstweg 65A, 3012 Bern, CH; [3]Dataionics, 1 rue Jean-Paul Laurens, 31400 Toulouse, FR, [4]Space Cooperative Europe

Abstract

This abstract introduces a unifying spatial framework based on Discrete Global Grid Systems (DGGS) to facilitate an interoperable data economy and scalable GeoAI workflows. It proposes a spatial tokenizer, a DGGS-based abstraction layer that partitions the Earth's surface into hierarchical, equal-area zones, providing a shared reference for spatial data fusion, indexing, and AI modeling. This approach supports FAIR principles, Data Spaces, and addresses challenges of data harmonization, explainability, and reusability. The DGGS acts as an analog-to-digital converter for geospatial data, transforming the continuous Earth surface into discrete, computable units. By adopting DGGS, data producers and consumers can decouple analytics from native formats, enabling applications like climate resilience assessment and risk modeling. The concept is demonstrated through a DGGS API and storage layer.

***Index Terms*—** DGGS, GeoAI, Spatial Tokenizer, Data Economy, Interoperability

1. Introduction

The exponential growth of spatial data has opened new possibilities for analytics, forecasting, and decision-making [1]. However, technical fragmentation - across coordinate systems, data models, and storage architectures - still limits effective integration of Earth Observation (EO) with AI-driven analytics [2]. This abstract proposes a unifying spatial framework based on Discrete Global Grid Systems (DGGS)[3] to underpin an interoperable data economy and enable scalable GeoAI workflows.

We introduce the concept of a spatial tokenizer: a DGGS-based abstraction layer that systematically partitions the Earth's surface into hierarchical, equal-area zones, providing a shared reference for spatial data fusion, indexing, and AI modeling [Figure 1]. This approach aligns with FAIR principles[4], supports Data Spaces [5], and addresses core challenges of data harmonization. Here, the term 'tokenizer' is borrowed from the broader AI community, where 'tokenization' refers to the process of breaking raw data into smaller, discrete units.

Just as digital systems require discrete units to process and represent information, AI requires data to be digitized into consistent, structured tokens. In contrast, geographic space is inherently continuous and heterogeneous. A DGGS acts as the analog-to-digital converter (ADC) for geospatial data, transforming the continuous Earth surface into discrete, computable units [6]. It digitizes geography in the same way a microphone digitizes sound - enabling AI to understand, index, and model spatial phenomena.

By adopting a DGGS as an underlying indexing and storage mechanism, data producers and consumers can decouple analytics from native formats and projections, enabling policy-relevant applications such as climate resilience assessment, dynamic risk modeling, and near-real-time monitoring. The tokenizer concept is demonstrated through a DGGS API and storage layer within the context of the Open Geospatial Consortium (OGC) [7].

This work aligns with the European Commission's Interoperable Europe initiative, which promotes cross-border and cross-sector interoperability as a foundation for public sector innovation and digital sovereignty [8]. It also supports objectives of the EC's SIMPL (Smart middleware platform) framework, which aims to establish trusted mechanisms for data sharing and reuse in accordance with the European Data Strategy [9].

2. Background and Related Work

Recent efforts such as the Copernicus Data Space Ecosystem, Destination Earth (DestinE), and ESA's Digital Twin Earth highlight the need for unified spatial infrastructures that can accommodate diverse data streams and analytical workloads [cf. 10]. However, these initiatives still rely heavily on conventional GIS paradigms, where spatial joins, reprojections, and data fusion are computationally expensive and semantically inconsistent.

DGGSs offer a promising alternative, standardizing spatial representation into discrete, indexable units [3]. This inherently supports hierarchical aggregation, consistent downscaling, and AI-ready feature extraction. Prior implementations (e.g., ISEA3H, H3, rHEALPix) have demonstrated technical feasibility, yet lack unified APIs, integration with modern cloud-native architectures, or adoption as analytical primitives for AI [cf. 11].

Our work builds on ongoing standardization efforts in the OGC and extends them with a programmable interface for data ingestion, transformation, and retrieval. The DGGS-based tokenizer enables alignment of datasets by spatial keys rather than geometry, offering substantial performance and interoperability benefits.

doi:10.2760/2119408

Fig. 1. Hexagonal grid over Europe, a common spatial data space for GeoAI

3. GeoPlegma: A Unified Interface for DGGS

While the DGGS concept is powerful, its practical use remains limited by fragmented implementations and complex, often theoretical, tooling. To address this, we introduce GeoPlegma: a lightweight and user-friendly software interface designed to aggregate and unify access to diverse grid systems, including DGGRID, DGGAL, HEALPix, H3, S2, and others.

GeoPlegma abstracts the complexity of each underlying DGGS implementation and exposes a common interface for spatial indexing, data ingestion, and retrieval. It lowers the entry barrier for developers, analysts, and institutions looking to adopt DGGS for scalable spatial analysis. Rather than promoting a single grid, GeoPlegma emphasizes interoperability and composability, making it possible to discretize and quantize, i.e. tokenize, space and give the ability to analyze across different grids for specific use-cases.

This approach enables experimentation, comparison, and co-existence of multiple DGGS types within a single processing framework, which is providing practical interoperability for AI and EO workflows. It also supports the emergence of a modular ecosystem of spatial tools, aligned with the broader goals of the Interoperable Europe initiative and the upcoming Digital Europe Programme [12].

4. Toward a Unified Spatial Data Space

The vision of a federated, interoperable data space is gaining traction across domains, from EO to civil security to digital governance. However, the technical realization of such a space still faces substantial challenges. A key example is the International Hydrographic Organization (IHO), where national hydrographic offices manage independent data lakes, yet seek a common spatial framework for seamless integration and analysis. A DGGS can serve as this harmonizing layer, enabling consistent representation and cross-border interoperability.

Importantly, a unified spatial data space must go beyond conventional GIS. EO and geospatial datasets are often siloed and incompatible with structured data workflows. DGGS zones act as keys that allow data to be reshaped, indexed, and queried like database records. This process enables users to view and access their data lake through the lens of a chosen grid geometry that best suits their use case.

Through spatial tokenization, data becomes accessible not only by location or geometry, but as rows in a scalable, schema based on Zone IDs. This supports the design of modular, queryable data systems that conform to Data Space principles and enable composable spatial analytics at scale.

doi:10.2760/2119408

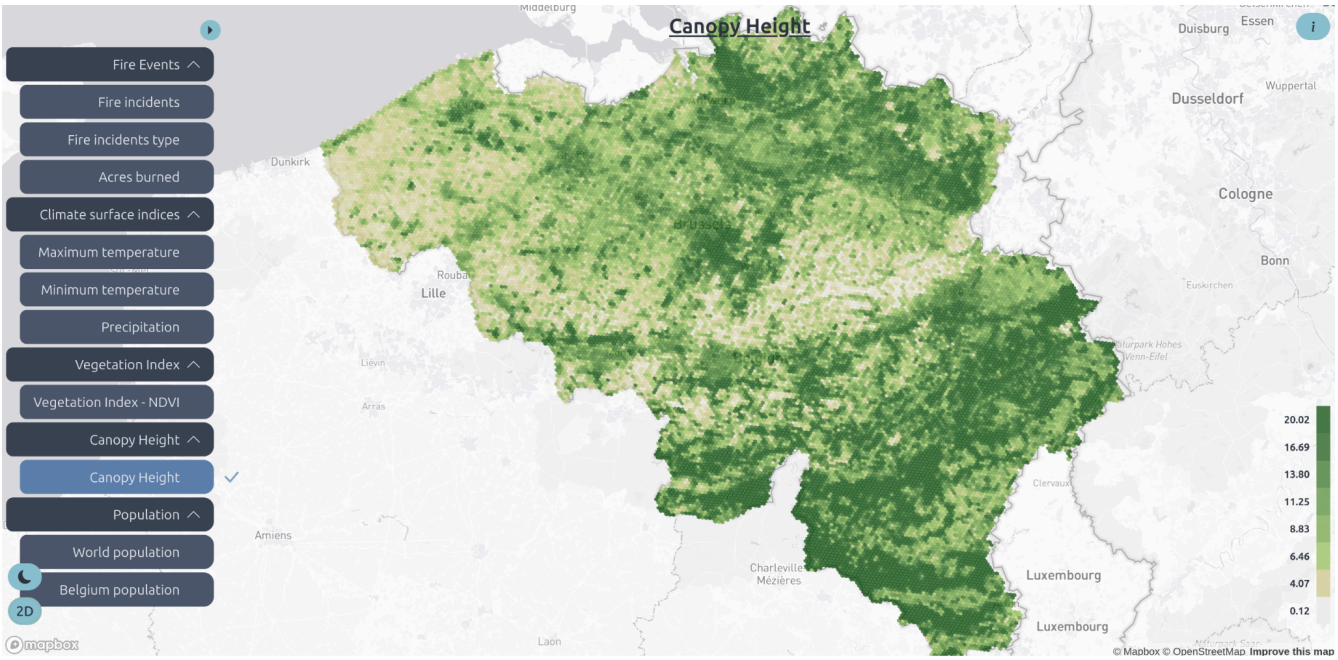


Fig. 2. Example data of canopy height in Belgium at https://demo.geoinsight.ai

5. Implementation and Demonstration

We are actively developing components to operationalize the DGGS tokenizer concept. The DGGS API hosted at https://dggs.geoinsight.ai/ is part of the OGC pilot initiative and provides endpoints for spatial queries, zone inspection, and hierarchical operations. It enables real-time DGGS-based interactions through a standard interface. GeoPlegma (https://github.com/GeoPlegma), written in Rust, implements multiple DGGS variations from scratch while also leveraging existing open libraries. The codebase is modular and designed for performance, with an emphasis on type safety and extensibility. It has been released under the MIT or Apache 2.0 license. A visualization tool is in progress, expected to be deployed at https://dggs.io, enabling users to explore DGGS coverage and interact with geographic information and insights through an intuitive web interface. Additionally, integration with the STAC (Spatio Temporal Asset Catalogs) ecosystem is planned to enhance access and metadata exchange. Previews and system diagrams will be made available via https://demo.geoinsight.ai, where users can evaluate the proposed system architecture and workflows in action (Figure 2).

6. Use Cases and Applications

The primary use case of a DGGS is to enable data interoperability across heterogeneous sources, unlocking advanced analytical capabilities. By assigning spatial units to each data point - regardless of origin, projection, or resolution - DGGS provides a universal reference frame for integration. A typical use case involves combining EO data with ancillary sources such as demographic statistics, IoT sensor streams, climate variables, and social media signals [cf 13, 14, 15]. Once harmonized through DGGS, this fused data can support real-time monitoring, automated index generation, advanced spatial statistics, and machine learning workflows, providing a foundation for operational GeoAI. In the maritime domain, the International Hydrographic Organization (IHO) illustrates the potential of DGGS to create a pan-European spatial data space. Here, national hydrographic datasets can be consistently partitioned and queried through shared grid zones, promoting cross-border interoperability and unified maritime analytics. DGGS zones also offer strong policy relevance. Because the zones are discrete, stable, and reproducible, users can query the same data slice across systems and over time. This consistency ensures transparency, reproducibility, and comparability, which are key enablers for regulation, reporting, and evidence-based policy.

One forward-looking GeoAI scenario envisions exploratory queries: users could select zones of interest and let the system automatically search for similar spatial patterns elsewhere in the dataset. This supports unsupervised discovery, anomaly detection, and feature-based spatial search at scale. In the long term, we envision GeoAI architectures that directly leverage the structure of DGGS. The fixed topology and relations between zones across the hierarchy make DGGS a natural substrate for Graph Neural Networks (GNNs) [16], where each zone acts as a node and

doi:10.2760/2119408

adjacency relations define the graph. Such models can learn from neighborhood context, perform hierarchical reasoning, and generalize spatial knowledge across scales. To ensure broad participation and respect organizational boundaries, federated learning approaches are planned, where institutions contribute local models trained using a DGGS without sharing the data itself. This enables collaborative AI training across jurisdictions, preserves data sovereignty, and supports scalable GeoAI development across Europe and beyond.

7. Perspectives and broader adoption

By solving the AI readiness of the geospatial data thanks to DGGS, the next frontier of Geo related AI , in particular with Large Models, will then be the data access per se. Not only providing a smooth and powerful access, the challenge will be to go beyond the bottleneck of sourcing input material, i.e. geospatial data. In other words, tearing down the AI Geodata wall so that AI can actually deliver its potential at the full scale. With respect to that concern, strategies like federated data access and mutualization of resources are being explored and should be examined to scale up and feed AI applications accordingly.

8. Conclusions

The proposed spatial tokenizer based on Discrete Global Grid Systems (DGGS) offers a unifying framework to address technical fragmentation in geospatial data integration, enabling scalable GeoAI workflows. By transforming the continuous Earth surface into discrete, computable units, DGGS acts as an analog-to-digital converter for geographic data, facilitating consistent data harmonization, indexing, and AI modeling. This approach supports FAIR principles and Data Spaces. GeoPlegma, a user-friendly software interface, simplifies access to diverse grid systems, promoting interoperability and lowering the entry barrier for DGGS adoption. The implementation of the DGGS API and associated tools, aims to create a unified spatial data space based on zones. This creates a basis for operational GeoAI. Our long-term visions include GeoAI architectures that leverage the structure of DGGS for Graph Neural Networks and federated learning approaches to ensure data sovereignty and collaborative AI training. The DGGS approach promises transparency, reproducibility, and comparability, essential for evidence-based policy and regulation.

REFERENCES

[1] Pei, T., Song, C., Guo, S., Shu, H., Liu, Y., Du, Y., ... & Zhou, C. (2020). Big geodata mining: Objective, connotations and research issues. *Journal of Geographical Sciences*, *30*, 251-266.

[2] Zou, L., Song, Y., & Cervone, G. (2024). Geospatial big data: theory, methods, and applications. *Annals of GIS*, *30*(4), 411-415.

[3] Sahr, K., White, D., & Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. Cartography and Geographic Information Science, 30(2), 121–134. https://doi.org/10.1559/152304003100011090

[4] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., ... & Schultes, E. (2020). FAIR principles: interpretations and implementation considerations. *Data intelligence*, *2*(1-2), 10-29.

[5] Bacco, M., Kocian, A., Chessa, S., Crivello, A., & Barsocchi, P. (2024). What are data spaces? Systematic survey and future outlook. *Data in Brief*, *57*, 110969.

[6] Mechenich, M. F., & Žliobaitė, I. (2023). Eco-ISEA3H, a machine learning ready spatial database for ecometric and species distribution modeling. *Scientific data*, *10*(1), 77.

[7] Purss, M. B. J., Gibb, R., Samavati, F., Peterson, P., & Ben, J. (2016). The OGC® Discrete Global Grid System core standard: A framework for rapid geospatial integration. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 3610–3613). IGARSS 2016 - 2016 IEEE International Geoscience and Remote Sensing Symposium. IEEE. https://doi.org/10.1109/igarss.2016.7729935

[8] European Commission, “Interoperable Europe Portal,”, Publications Office of the European Union, 2025. Available: https://interoperable-europe.ec.europa.eu/

[9]European Commission, “Simpl: Cloud-to-edge federations empowering EU data spaces,”, Publications Office of the European Union, 2025. Available: https://digital-strategy.ec.europa.eu/en/policies/simpl

[10] Destination Earth, Destination Earth (DestinE), a European Commission flagship initiative to support the green transformation, Publications Office of the European Union, 2025. Available: https://digital-strategy.ec.europa.eu/en/policies/destination-earth

[11] Bowater, D., & Stefanakis, E. (2018). The rHEALPix Discrete Global Grid System: considerations for Canada. Geomatica, 72(1), 27–37. https://doi.org/10.1139/geomat-2018-0008

[12] The Digital Europe Programme, Publications Office of the European Union, 2025. Available: https://digital-strategy.ec.europa.eu/en/activities/digital-programme

[13] Jendryke, M., Balz, T., McClure, S. C., & Liao, M. (2017). Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. Computers, Environment and Urban Systems, 62, 99–112. https://doi.org/10.1016/j.compenvurbsys.2016.10.004

[14] Jendryke, M., & McClure, S. C. (2021). Spatial prediction of sparse events using a discrete global grid system; a case study of hate crimes in the USA. International Journal of Digital Earth, 14(6), 789–805. https://doi.org/10.1080/17538947.2021.1886356

[15] Caspari, G., dos Santos Manuel, J., Gago-Silva, A., & Jendryke, M. (2024). Employing discrete global grid systems for reproducible data obfuscation. Scientific Data, 11(1). https://doi.org/10.1038/s41597-024-03354-5

[16] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, *32*(1), 4-24.

LLM-READY SPATIO-TEMPORAL DATA: ENABLING AGENTIC GEOAI WITH STAC

C. Chiarelli[*], *I. Trandafir*[§], *P. Kempeneers*[*]

[*]European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy, [§]Unisystems

ABSTRACT

AI-ready data is an evolving concept, with its definition depending on the specific AI application. The adoption of structured metadata and standardized APIs can significantly improve data findability and accessibility for Large Language Models (LLMs). This paper concentrates on the Spatio-Temporal Asset Catalog (STAC) standard, examining its strengths in supporting LLM spatio-temporal data retrieval. For this purpose, a prototype was developed to demonstrate how an LLM agent can retrieve, and process geospatial data exposed by a STAC catalog. The study reveals some key challenges such as excessive token use, imprecise spatial queries, and redundant endpoints, and shows possible targeted mitigations. The results demonstrate STAC's foundational strengths and the Model Context Protocol (MCP)'s enabling role for agentic geospatial artificial intelligence (GeoAI) workflows.

***Index Terms*—** STAC, LLM agent, GeoAI, MCP, LLM-ready spatio-temporal data

1. INTRODUCTION

The concept of AI-ready data is still evolving, and there is currently no universal method that can preemptively prepare all data for every AI application. Gartner has noted that AI readiness highly depends on the specific AI use case [1]. Consequently, rather than addressing AI-ready data in general—which spans disparate domains from structured time-series datasets to unstructured multimedia—this paper narrows its scope to LLM-ready spatio-temporal data, described through the STAC standard.

Retrieving the data is the first fundamental step in any data processing and analysis workflow, whether traditional or AI-based. It is therefore essential for geospatial artificial intelligence (GeoAI), where geospatial studies and artificial intelligence intersect [2]. A typical application is rapid conversational access to spatio-temporal assets that can improve environmental monitoring and crisis management. This can foster better situational awareness, faster anomaly detection, and more effective resource coordination.

This paper focuses on spatio-temporal datasets and demonstrates how the Spatio Temporal Asset Catalog (STAC) standard [3] with its rich metadata and formalized API specifications establishes a strong foundation for agentic geospatial intelligence. To this end, a working prototype was implemented, and it allowed to highlight challenges and mitigation strategies.

2. FROM STAC TO LLM-READY SPATIO-TEMPORAL DATA

STAC has become a key standard for organizing and sharing geospatial data, especially for Earth Observation and remote sensing. It provides metadata with extensive descriptive fields that enable AI models to parse and contextualize imagery and sensor data. Domain-specific extensions [4] further enhance this metadata, supporting specialized AI tasks with improved accuracy. In addition, STAC metadata and API specification [5] defined via the JSON schema [6] and the OpenAPI [7] standard respectively, provide a machine-readable interface that can be used as a basis for automated geospatial intelligence.

STAC's extensibility is another major strength. Community-driven extensions can add new metadata fields without breaking core specifications, ensuring backward compatibility while adapting to evolving needs. Its alignment with Open Geospatial Consortium (OGC) standards [8] and support for cloud-native formats like Cloud Optimized GeoTIFF (COG) [9] and Zarr [10] also streamline integration with machine learning and geospatial analytics pipelines.

However, despite these native strengths, achieving full LLM-readiness requires additional capabilities that STAC alone does not provide. This requires additional components which include conversational tool descriptors that define human-readable functions, dynamic controls like pagination and rate-limiting to support interactive, dialog-based usage, and contextual metadata summaries providing aggregated information [11].

3. LLMS, AGENTS, AND TOOL CHAINING

LLMs are neural networks trained on massive text corpora to generate human-like text. They can summarize information, answer questions, translate text, or carry on a conversation by predicting the next word in a sequence. However, these models inherently lack the capacity to interact with external tools or retain memory across interactions and their knowledge is frozen at training time. Moreover, every LLM has a finite "context window" meaning it can only consider a limited amount of text at once. These constraints – static knowledge, no stateful memory, and limited context – restrict an LLM's ability to operate autonomously or reliably over extended tasks.

doi:10.2760/2119408

Agentic systems build on LLMs by adding autonomy, planning, and external interaction. In technical terms, an AI agent uses a language model as its reasoning core but can formulate plans and take actions beyond the original training data. Key to agentic systems is access to tools and external resources. Tools are executable components—such as operations, algorithms, or services—that perform specific tasks when invoked. Resources refer to structured datasets or content repositories that provide information upon request. Unlike a stand-alone LLM, an agent can break a complex task into subtasks and can orchestrate multiple tools in sequence (a process sometimes called tool chaining) where the output of one tool serves as the input for the next, creating complex processing pipelines that can adapt to diverse tasks and data scenarios [12].

Central to this new paradigm is Model Context Protocol (MCP), an open standard recently introduced by Anthropic [13]. MCP enables secure and standardized communication between LLM agents and external systems, eliminating the need for custom integrations with each new tool or resource. By providing a unified interface for tool and data discovery, MCP empowers agents to dynamically expand their capabilities and operate in increasingly complex and interconnected environments.

The evolution from traditional LLMs to intelligent agentic systems marks a significant step toward enabling autonomous task execution. These advancements lay the groundwork for agentic geospatial intelligence, where AI-driven agents autonomously manage geospatial data retrieval, processing, and analysis with contextual awareness.

4. GEOAI PROTOTYPE

A prototype was developed to demonstrate how geospatial data, structured with the STAC standard, can be made LLM-ready and integrated into an agentic intelligent system (see Fig. 1).

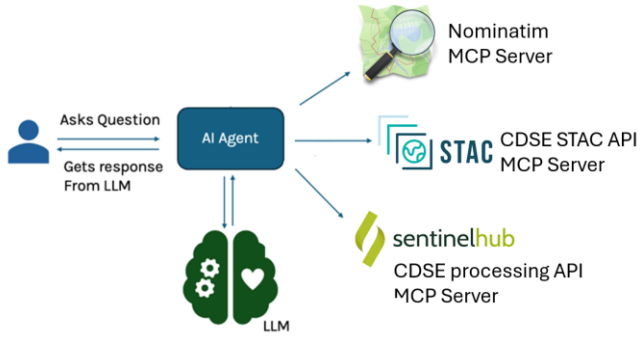


Fig. 1. Agentic architecture accessing three MCP servers (on the right)

It consists of an LLM agent that interacts with three MCP servers. The servers were implemented using Anthropic MCP Python SDK [14]. A first MCP server (Nominatim) is responsible for geocoding and is used to convert names of locations on earth to GeoJSON polygons. It relies on OpenStreetMap's Nominatim service [15]. Central for this study is the STAC MCP server that links the AI agent to the STAC APIs provided by the Copernicus Data Space Ecosystem (CDSE) [16]. This STAC endpoint was chosen because of its openness and matureness. This server allows the agent to discover and search the available geospatial data in the CDSE catalog. The third MCP server is responsible for image processing, also based on a CDSE implementation (Sentinel Hub) [17]. In this study, it is used as an example to complete the agentic processing chain.

Standard API endpoints are exposed as MCP servers, creating the corresponding tools and resources. In Fig. 2 a simple example is shown of how to construct an MCP tool by wrapping an API function call.

The AI agent itself was built with the mcp-agent framework [18]. It orchestrates tool invocation and response handling hereby interacting with the three MCP servers and the LLM based on LLM Claude 3.5 Sonnet [19]. Finally, a simple user interface was built with Streamlit [20] to directly interact with the user and visualize the results.

```python
@mcp.tool()
async def list_collections(ctx: Context) -> Dict:
    """
    List all available collections in the CDSE STAC catalog.
    This tool retrieves information about all available collections
    in the Copernicus Data Space Ecosystem (CDSE)
    STAC catalog. Collections are logical groups of items that share
    common properties and represent specific datasets like Sentinel-1,
    Sentinel-2, etc.
    Returns:
        A dictionary containing information about all available collections,
        including their IDs, titles, descriptions, spatial and temporal
        extents, and other metadata.
    """
    stac_client = ctx.request_context.lifespan_context.stac_client
    return await stac_client.get_collections()
```

Fig. 2. Example of MCP tool wrapping getCollections method of STAC APIs.

An end-to-end pipeline from the text query to the final geospatial insight is demonstrated using the following query: "Analyze vegetation over the Ispra (Italy) area in the most suitable day of spring 2024". This query was intentionally phrased ambiguous, to test the interpretation capability of the LLM in the context of GeoAI. Vegetation can be analyzed with remote sensing imagery, typically by calculating the normalized difference vegetation index (NDVI) [21]. This requires the LLM to select an optical sensor with the appropriate spectral bands, for instance the multi-spectral instrument (MSI) onboard on the Copernicus Sentinel-2 satellites. In this case, cloud cover is an important factor when selecting the optimal acquisition date of optical Earth Observation imagery. This request enchained three steps shown in Fig. 3.

doi:10.2760/2119408

I'll help you analyze vegetation in Ispra, Italy. Let me break this down into steps:

1. First, I'll geocode Ispra to get its coordinates
2. Then, I'll search for Sentinel-2 L2A imagery (best for vegetation analysis) during Spring 2024 (March 20 - June 20)
3. Finally, I'll process the imagery for vegetation analysis

Fig. 3. Agentic workflow explained by the LLM to the user.

Information recorded in a log file provided insight into how the LLM agent reaches its final answer. It was shown how the request was translated into a chain of subsequent calls to tools and resources.

1. **Geocoding**: convert the input location (in this case Ispra) to a polygon by calling the Nominatim MCP server.
2. **Search data:**
   a. Query the list of existing Earth Observation collections available in the CDSE STAC catalog by calling the *getCollections* method of the STAC APIs**.** This retrieves information on the available collections and identifies those that include spectral bands suitable for calculating the Normalized Difference Vegetation Index (NDVI). Among them, the Sentinel-2 L2A collection is identified, providing the red (B04) and near-infrared (B08) bands.
   b. Query the catalog by calling the *search* method of the CDSE STAC APIs**.** This selects the images in the retrieved collection (Sentinel-2 L2A) with the lowest cloud coverage. The LLM agent formulates the query parameters (coordinates, date range and others) according to the STAC specifications and passes them to the STAC MCP server.
3. **Image processing**: call CDSE Sentinel Hub processing APIs to compute NDVI on the extracted items. The LLM agent formulates the query parameters according to the Sentinel Hub specifications and passes them to the Sentinel Hub MCP server, which returns the image with the calculated index.

This workflow shows how an LLM agent can autonomously decide and chain steps leveraging MCP-compatible services, with the LLM providing the logic and natural language instructions. The prototype demonstrated that the LLM agent successfully completed its workflow like a human programmer: retrieved the location geometry, then queried the STAC catalog for relevant data, and finally generated a meaningful image along with an explanation of how to interpret it, as shown in Fig.4. No modifications or fine-tuning were applied to the LLM. The selection of criteria, such as identifying the lowest cloud coverage or choosing the appropriate spectral bands, was based solely on the input prompt and the provided context.

Further analysis and interpretation of the resulting NDVI image are essential to fully address the user's query. However, these steps would require additional specialized tools or cooperating AI agents and are therefore left for future research. The primary aim of the prototype in this study was to assess the STAC readiness for integration with LLMs, with the data processing components included mainly to complete the end-to-end workflow.

NDVI image color interpretation using bands B04 (red) and B08 (NIR):
- Dark areas indicate dense vegetation (because healthy vegetation strongly absorbs red light (B04) which is shown in the red channel)
- Brighter/lighter areas indicate less vegetation or non-vegetated surfaces
- The darkest areas in the image represent the densest vegetation cover, as these areas have the highest red light absorption
- Water bodies appear bright due to their higher reflectance in the red band compared to vegetated areas
- Urban areas and bare soil typically appear in brighter tones

This image from May 28, 2024 (14.1% cloud cover) shows the vegetation patterns around Ispra during peak spring growth, where the darkest areas in the image indicate the healthiest and densest vegetation cover.

Fig. 4. Result image of NDVI calculation over Ispra area in Spring 2024 and related LLM description.

5. DISCUSSION

The geospatial AI prototype that was implemented in this study has revealed some key challenges for making LLM ready data leveraging on STAC. Initial experiments highlighted several obstacles. One major issue was excessive token usage, with conversations exceeding 70,000 tokens and

doi:10.2760/2119408

frequently reaching the LLM's token limit. Another challenge was that the LLM struggled to handle structured geospatial queries directly. When given a place name (e.g., 'Ispra'), it would attempt to estimate the location by retrieving its coordinates and then constructing a bounding box around it. This approach led to imprecise spatial representations and introduced unnecessary steps into the workflow. Additionally, the use of one-to-one endpoint mappings, resulted in inefficient and redundant API calls, emphasizing the need for optimized endpoints.

To overcome these challenges, several targeted optimizations were implemented. The use of the STAC API extension, *Sort* and *Fields*, significantly reduced token usage by limiting metadata requests to essential fields and ordering results by relevance. This reduced the conversation size from approximately 70,000 tokens to just 4,000. Local caching of geometries and images helped avoid redundant data transfers by storing geometry definitions and image payloads locally, thus eliminating the need to transmit large JSON or binary streams to the LLM. Furthermore, the integration of authoritative polygon geocoding through the Nominatim tool simplified preprocessing and reduced query complexity by directly retrieving accurate GeoJSON polygons.

These optimizations substantially improved both the efficiency and reliability of the agentic workflow. Future enhancements may include the implementation of adaptive pagination strategies and the use of embedding-based pre-filtering of assets to further streamline and refine query handling.

6. CONCLUSION

This paper articulates an agentic approach to LLM-driven spatio-temporal data retrieval and analysis leveraging on STAC. A prototype was implemented to demonstrate how an LLM agent can retrieve, and process geospatial data exposed by a STAC catalog, using MCP-enabled tools. The study reveals some key challenges for making LLM ready spatio-temporal data, suggesting mitigation strategies and highlighting STAC's strengths.

In this prototype, relatively simple queries were used to validate the feasibility of the approach. Further investigation is required to evaluate performance with more complex and diverse queries, which will provide deeper insights into robustness and scalability. Future work should explore adaptive interaction strategies, multi-agent orchestration, and enhanced tooling for image analysis, visualization and batch processing. These advancements aim to extend agentic geospatial workflows to broader and more complex application scenarios.

REFERENCES

[1] Gartner, "What is AI-Ready Data?", 2024. Accessed: 23/05/2025 [Online]. Available: https://www.gartner.com/en/articles/ai-ready-data

[2] Gao, Song, Yingjie Hu, and Wenwen Li. "Introduction to geospatial artificial intelligence (GeoAI)." Handbook of geospatial artificial intelligence. CRC Press, 2023. 3-16.

[3] STAC, Accessed: 23/05/2025 [Online]. Available: https://stacspec.org/

[4] STAC extensions, Accessed:23/05/2025 [Online]. Available: https://stac-extensions.github.io/

[5] STAC specifications, Accessed:23/05/2025 [Online]. Available: https://stacspec.org/en/about/stac-spec/

[6] JSON Schema, Accessed:23/05/2025 [Online]. Available: https://json-schema.org/

[7] OpenAPI standard, Accessed:23/05/2025 [Online]. Available: https://www.openapis.org/what-is-openapi

[8] OGC standards, Accessed:23/05/2025 [Online]. Available: https://www.ogc.org/standards/

[9] Cloud Optimized GeoTIFF (COG), Accessed:23/05/2025 [Online]. Available: https://cogeo.org/

[10] Zarr, Accessed:23/05/2025 [Online]. Available: https://zarr.dev/

[11] Is Your API AI-ready? Guidelines and Best Practices, Accessed:23/05/2025 [Online]. Available: https://www.blobr.io/guide-build-ai-copilot/api-ai-ready-guidelines-best-practices

[12] Prashant Sharma, "LLM to RAG to Agents to MCP: What the Shift!?", 2025. Accessed:23/05/2025 [Online]. Available: https://www.linkedin.com/pulse/llm-rag-agents-mcp-what-shift-prashant-sharma-axrfc/

[13] Anthropic, "Introducing the Model Context Protocol (MCP)", 2024. Accessed:23/05/2025 [Online]. Available: https://www.anthropic.com/news/model-context-protocol [Online Accessed:23/05/2025].

[14] MCP Python SDK, Accessed:23/05/2025 [Online]. Available: https://github.com/modelcontextprotocol/python-sdk

[15] OpenStreetMap's Nominatim service, Accessed:23/05/2025 [Online]. Available: https://nominatim.openstreetmap.org

[16] CDSE STAC API, Accessed:23/05/2025 [Online]. Available: https://stac.dataspace.copernicus.eu/v1/api.html

[17] CDSE Sentinel Hub API, Accessed:23/05/2025 [Online]. Available: https://documentation.dataspace.copernicus.eu/APIs/SentinelHub/Process.html

[18] MCP Agent, Accessed:23/05/2025 [Online]. Available: https://github.com/lastmile-ai/mcp-agent

[19] Claude 3.5 Sonnet, Accessed:23/05/2025 [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[20] Streamlit, Accessed:23/05/2025 [Online]. Available: https://streamlit.io/

[21] Lozano-Tello, A.; Siesto, G.; Fernández-Sellers, M.; Caballero-Mancera, A. Evaluation of the Use of the 12 Bands vs. NDVI from Sentinel-2 Images for Crop Identification. Sensors 2023, 23, 7132. https://doi.org/10.3390/s23167132

doi:10.2760/2119408

CATEGORICAL TIME-SERIES BASED ON SEMANTIC EARTH OBSERVATION WORKFLOWS IN LAND COVER MONITORING OF SEMI-ARID AREAS

Nimrod Kibet[1], Martin Sudmanns[1], Andreas Braun[2]

[1] University of Salzburg, Department of Geoinformatics Z_GIS

[2] University of Tübingen, Department of Geosciences

ABSTRACT

Earth observation (EO) data plays a key role in the analysis and monitoring of land cover dynamics. With increased availability of long time series of EO data, it is essential to develop effective workflows across studies to evaluate the land cover changes. In contrast to approaches relying on analysing the reflectance values directly, we developed a workflow for long EO time series analysis based on semantic categories and evaluated its effectiveness in monitoring and quantifying long-term land cover dynamics in semi-arid regions. The analysis aimed to answer the question, "How effective are a semantic EO workflows in monitoring and quantifying long-term land cover dynamics, while ensuring automation, reproducibility and scalability?" A semantic EO workflow was used to analyze the data, relying on the Satellite Image Automatic Mapper (SIAM™) to semantically process the datasets and the *Semantique* python library for semantic querying. The semantic EO workflow is complementary to existing approaches but provides additional insights into spatio-temporal land cover dynamics and confirms the reproducibility of results, automation and scalability of the approach.

Index Terms— Earth Observation, Land Use Land Cover, Semantic Enrichment, Semantic Querying.

1. INTRODUCTION

Land use land cover change (LULCC) is a global phenomenon that has significant environmental impacts including vegetation changes and biodiversity loss [1]. Over the past four decades, LULCC has intensified and is closely linked to accelerating global environmental crises [2]. Arid and semi-arid regions - covering about 41% of the Earth's land surface and supporting roughly 2.5 billion people - are particularly vulnerable, with vegetation highly sensitive to climate change and water scarcity [3].

Meanwhile, the increasing availability of long-term remotely sensed Earth Observation (EO) data offers both opportunities and challenges for innovative data retrieval, processing, and analysis. As the volume of EO data grows, traditional methods of managing it have become inefficient and unable to meet the demand for timely insights [4]. In turn, automated workflows have emerged as solutions. While many of them rely on analysis of reflectance values or spectral indices such as the normalized difference vegetation index (NDVI), a semantic approach based on semantic enrichment (SE) of spectral categories can provide additional insights, because categorical analysis can express queries such as "how many", "how often" etc. SE refers to the process of assigning concepts from global ontologies to semantic types in local ontologies to create a terminology knowledge base (TKB) [5]. One of the approaches used to produce a semi-symbolic layer from EO data is using the Satellite Image Automatic Mapper (SIAM™) software [6]. It implements a fully automated physical-model-based decision tree that assigns a multi-spectral color name (category) to EO data without training samples. Those categories can be semantically queried in EO data cube instances using semantic EO data cubes.

Effectiveness in a semantic EO workflow is measured by analyzing attributes of 1) automation – reducing manual interpretation by mapping data to semantic concepts, 2) reproducibility – ensuring transparent and shareable workflows through explicit ontologies and mappings, and 3) scalability – enabling efficient large-scale, multi-temporal analyses via EO data cubes and standardized interfaces. Given there has been limited research on the effectiveness of semantic EO workflows in analyzing long-term land cover dynamics, this study aimed to apply a semantic EO workflow using a semantic querying and analysis approach for categorical time-series applied to Landsat 8 and Sentinel-2 data.

2. METHODS

2.1. Trend in EO Data Generation and Management

The Landsat and Copernicus Sentinel programs have been central to EO satellite technology development and have realized a rich repository of satellite imagery. Currently, satellite observations provide wide area coverage and long-term sources for EO data and are crucial for monitoring and identifying insights on earth changes and anthropogenic influence [7]. Improvements in the storage and processing of large EO data do help with the monitoring and analysis.

2.2. Semantic EO Data Cube

A data cube is a multi-dimensional array that arranges data in a way that makes data storage, access, and analysis easier than file-based storage and access [8]. EO data cubes contain raster data that is organized along multiple dimensions that can be directly accessed. The structure improves data accessibility, allowing users to retrieve specific subsets through spatial or temporal queries [9]. Semantic EO data cubes move beyond data storage and provision and offer basic, interoperable spectral categories as building blocks of image-derived information within the cube. Users can create more expressive, thorough rulesets and queries as a result, and semantic analyses can be integrated into basic rule-sets in domain language [8]. Certain semantic content-based queries covering a user-defined area of interest (AOI) in each temporal extent are possible given

doi:10.2760/2119408

semantic enrichment that includes clouds, vegetation, water, and "other" categories.

2.3. Semantic Enrichment with SIAM

SE of EO images refers to “interpreted content of EO imagery (i.e., mapping data to symbols that represent stable concepts)” [10]. SIAM™ is a fully automated program that works without training data or human input [11]. It is designed for:

1. Converting multi-sensor, multi-spectral (MS) reflectance data into a structured set of color names at different levels of detail, forming a hierarchy of color vocabularies.
2. Identifying connected regions (super pixels or segments) within the color-mapped image.
3. Assessing the accuracy of image compression using vector quantization (VQ) by measuring the root-mean-square error (RMSE).

SIAM™ emulates a pre-attentive phase of human vision, performing the first classification of satellite data autonomously without the need for user guidance or training [6]. It is the first, necessary step towards semantic analysis, which requires semantic querying on user-side to generate meaningful results.

2.4. Semantic Querying with Semantique Library

The *Semantique* Python package allows implementation of a structured framework for semantic querying in EO data cubes. There are three main components of the framework and include the real-world, image, and mapping domains. *Semantique* makes a clear separation between the image domain and the real-world domain. In this study, it is used for the analysis of the long categorical time series.

2.5. Study Area

The research focuses on Kanthuni area, Makueni County, Kenya. The area was selected for this study as it lies in the arid and semi-arid zone of the Eastern region of the country which is characterized by frequent draught and minimal rainfall. Due to the rainfall shortage in the region, sand dams have been significantly constructed between 2011 and 2014 as a solution for water harvesting and management. In this context, our method supports the research on the impact of these sand dams on their environment [12].

2.6. Data

The primary datasets for this study were Landsat 8, and Sentinel-2, which provide multi-temporal coverage of the study area over a 10-year period for the former and 8 years for the latter. ERA5 data were also used as complimentary to the categorical time series analysis. ERA5 is the 5th generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate from 1940 to the present [13].

2.7. Data Processing and Analysis

Semantique was utilized for data preparation and entity mapping for vegetation cover. Next, the framework’s filtering capabilities were used to filter cloud cover. Query recipes were then employed for analysis to compute metrics such as the percentage of vegetation cover and seasonal changes. The *reduce* function was also crucial in the analysis as it allowed reduction by time and space before query recipes were executed in the defined data cube for the area of interest (AOI). When reducing by time, the *reduce* verb aggregates data across different timestamps for each spatial location. This means getting a single value for each spatial coordinate, summarizing the number of observations across all timestamps. Conversely, reducing by space aggregates data across all spatial locations for each timestamp. It results in a single value for each timestamp, summarizing how many locations reported the presence of the concept (vegetation).

The analysis for this study included vegetation analysis, seasonality analysis, greenness index (GI), vegetation intensity analysis. A total of 1,137 scenes were processed for the area of interest: sentinel-2 (519) and Landsat 8 (608). To examine spatial distribution of vegetation class observations, reduction by time approach was applied which checks the frequency at which vegetation class and subclasses were recorded at a particular pixel on the AOI expressed as a percentage. Seasonality analysis was done using monthly mean vegetation percentages where an additive model was adopted, since it assumes that the observed data is the sum of three elements: residuals, trend, and seasonality. GI - a numerical value calculated using satellite imagery to measure the amount of vegetation present in an area, was derived from SIAM™. Further, a semantic querying approach was used to evaluate vegetation intensity over time, indicating potential vegetation changes and classifying outputs in three vegetation categories (strong, average, and weak).

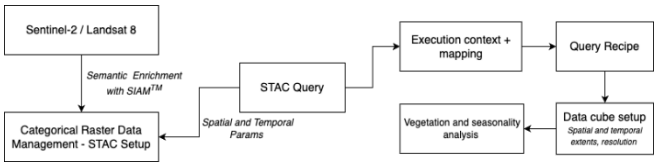


Fig 1. Data processing and querying workflow

3. RESULTS

As indicated in fig 2 and 3, the highest vegetation observation frequency was 42.6% and 36.3% for Sentinel-2 and Landsat 8 SIAM™ outputs.

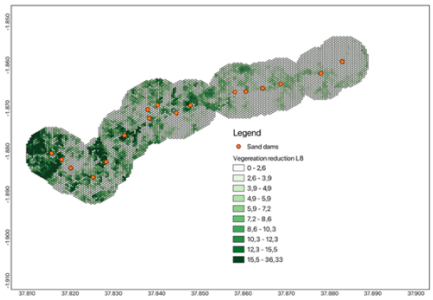


Fig 2. Landsat frequency of vegetation observation as a percentage (2013-2023)

doi:10.2760/2119408

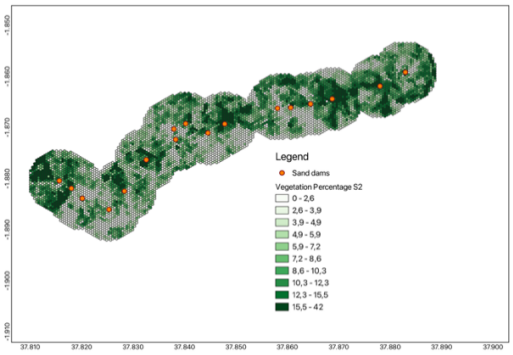


Fig 3. Sentinel frequency of vegetation observation as a percentage (2016-2023)

The analysis of intensity subclasses—weak, average, and strong—focuses on the frequency of observation of each subclass as a percentage over the dataset temporal range, as highlighted in fig 4. The average vegetation subclass had the highest frequency of observation compared to the other subclasses.

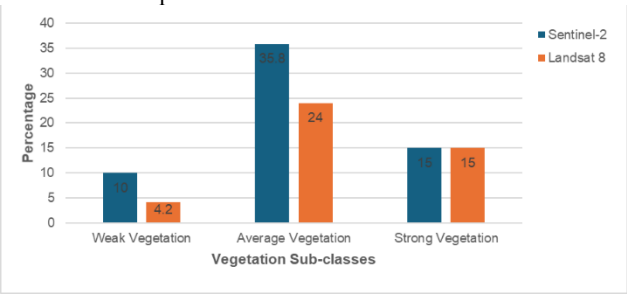


Fig 4. Maximum percentage of vegetation frequency observation

The seasonal decomposition of vegetation class observation frequency in fig 5 reveals distinct patterns across the original raw data, trend, seasonal, and residual components. Both Sentinel-2 and Landsat 8 datasets exhibit periodic peaks in the raw data, with notable increases in 2018, 2020, 2022 and 2023. In this case, the vegetation entity consisting of several vegetation spectral categories was reduced over space and the result is the relative area and not the vegetation intensity. The relative area covered by vegetation can be sensitive to precipitation or irrigation.

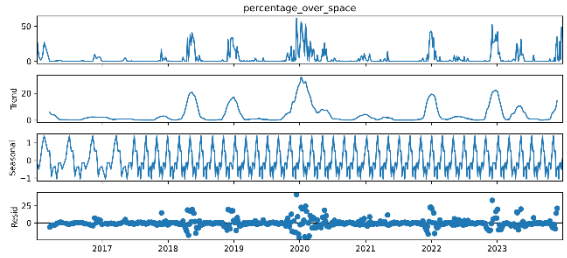


Fig 5. Seasonal decomposition of vegetation percentage over space (2016-2023)

The seasonal vegetation observation frequency data, grouped into wet and dry seasons and overlaid with seasonal precipitation trends highlight the interplay between vegetation dynamics and rainfall patterns. Sentinel-2 (fig 6) and Landsat 8 (fig 7) outputs consistently show that wet season vegetation observation frequencies peak during periods of increased precipitation.

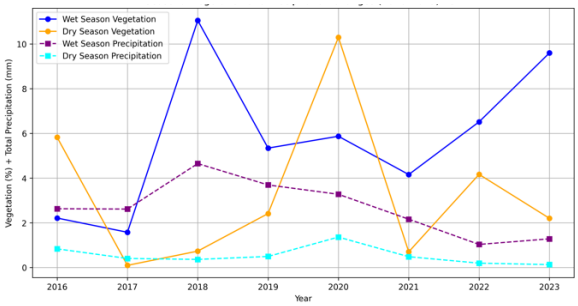


Fig 6. Overlay of vegetation observation frequency and precipitation data (Sentinel 2016-2023)

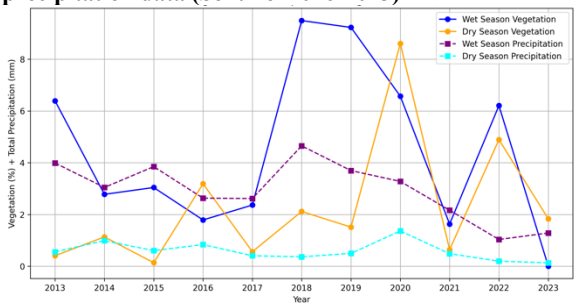


Fig 7. Overlay of vegetation observation frequency and precipitation data (Landsat 2013-2023)

Fig 8 and 9 show that in contrast to vegetation percentage indicating the covered area, the greenness index is more sensitive to vegetation health, biomass, and water content which are important considerations as the area of interest is a semi-arid area [14]. The index captures this seasonal variability peaking during the wet season (March to May and October to December) and dropping during the dry season (January to February and June to September)

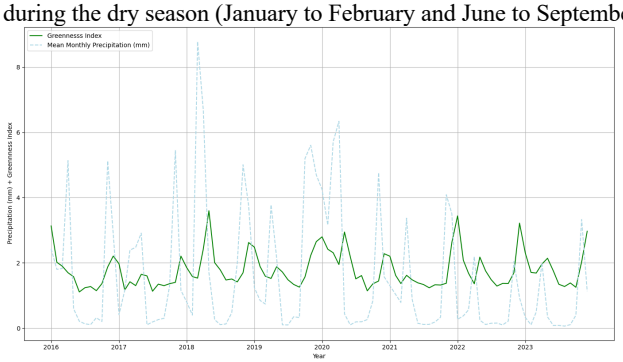


Fig 8. Mean monthly greenness and precipitation data (Sentinel 2016-2023)

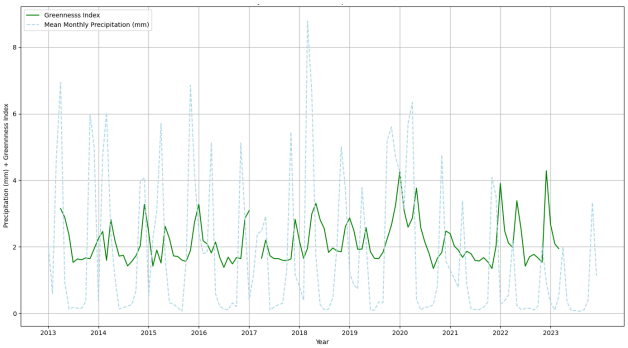


Fig 9. Mean monthly greenness and precipitation data (Landsat 2013-2023)

doi:10.2760/2119408

4. DISCUSSION AND CONCLUSIONS

The study evaluated the effectiveness of a semantic EO workflow in monitoring and quantifying long-term land cover dynamics. Vegetation frequency observations recorded in both Sentinel-2 and Landsat 8 had a maximum record of 42.6% and 36% respectively, providing an overview of the entire dataset. The spatial distributions of the frequency of vegetation class observations had higher percentages recorded in areas around sand dams, indicating the water resilience has an impact on vegetation cover. The distribution was also more pronounced in Sentinel-2 than Landsat 8; this could be attributed to the spatial resolution of Landsat 8 (30m) compared to Sentinel-2 (10m). The consistency between the outputs from Sentinel-2 and Landsat 8 SIAM-derived data confirms the reliability and transferability of the approach across different data products. This not only underlines the confidence of the results but also highlights the flexibility of semantically enriched data cubes for multi-sensor analyses, which is crucial for achieving comprehensive temporal coverage in studies requiring long-term environmental monitoring. In this study, the semantic EO workflow was effective as it autonomously transformed the raw spectral values into valuable land cover classes without human interpretation, by means of ontology-mappings. The reproducibility was also met because the query recipes can be reapplied for different temporal datasets and retain the same results without rewriting scripts for analysis. Lastly, scalability is also feasible for potential future applications since the workflow, underpinned by EO data cubes and standardized interfaces, is built for handling progressively more enormous and complex datasets more effectively. Future studies should focus on *Semantique*'s library query expressiveness in supporting complex spatiotemporal patterns and uncertainty quantification for more nuanced detection of land cover changes.

5. OUTLOOK

This work showed semantic-based EO workflows provide nuanced insights into landscape evolution at different temporal scales. This is particularly relevant in areas of high seasonal variation and strong phenological dynamics, where traditional and static methods of LULC classification fall short. Once these dynamics have been assessed at both annual and long-term scales, their association with water retention and conservation efforts like use of sand dams can be examined in detail to attribute landscape changes to the date and location of sand dam construction. This can be achieved using geostatistical methods and space-for-time substitution, but also requires a multidisciplinary perspective including groundwater changes, topography, climate trends and land surface temperature changes. Earth observation methods can provide data on all of these aspects and help to assess the actual impact of sand dams on their environments.

6. ACKNOWLEDGEMENTS

This work was partially funded by the Dr. Erich Ritter-Stiftung Foundation, under project number T0021/43246/2023.

7. REFERENCES

[1] Seyam MMH, Haque MR, Rahman MM. Identifying the land use land cover (LULC) changes using remote sensing and GIS approach: A case study at Bhaluka in Mymensingh, Bangladesh. Case Studies in Chemical and Environmental Engineering. 2023;7:100293-.

[2] Wahdatyar R, Khokhar MF, Ahmad S, Rahil MU, Stanikzai MA, Khan JA, et al. Exploring the dynamics and future projections of land use land cover changes by exploiting geospatial techniques; A case study of the Kabul River Basin. Heliyon. 2024;10(20):e39020-e.

[3] Mahesh GM, Chauhan N, Goyani M. Enhanced Multistage Content Based Image Retrieval International Journal of Computer Science and Mobile Computing Enhanced Multistage Content Based Image Retrieval. 2013.

[4] Wulder MA, Hermosilla T, Stinson G, Gougeon FA, White JC, Hill DA, et al. Satellite-based time series land cover and change information to map forest area consistent with national and international reporting requirements. Forestry: An International Journal of Forest Research. 2020;93(3):331-43.

[5] Jiang S, Feng X, Zhang B, Shi J. Semantic enrichment for BIM: Enabling technologies and applications. Advanced Engineering Informatics. 2023;56:101961-.

[6] Baraldi A. Satellite Image Automatic Mapper™ (SIAM™) - A Turnkey Software Executable for Automatic Near Real-Time Multi-Sensor Multi-Resolution Spectral Rule-Based Preliminary Classification of Spaceborne Multi-Spectral Images. Recent Patents on Space Technology. 2011;1(2):81-106.

[7] Zhao Q, Yu L, Du Z, Peng D, Hao P, Zhang Y, et al. An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. Remote Sensing. 2022;14(8):1863-.

[8] Augustin H, Sudmanns M, Tiede D, Baraldi A. A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict. GI_Forum. 2018;1:214-27.

[9] Kopp S, Becker P, Doshi A, Wright DJ, Zhang K, Xu H. Achieving the Full Vision of Earth Observation Data Cubes. Data. 2019;4(3):94-.

[10] Sudmanns M, Augustin H, van der Meer L, Baraldi A, Tiede D. The Austrian Semantic EO Data Cube Infrastructure. Remote Sensing. 2021;13(23):4807-.

[11] Baraldi A, Puzzolo V, Blonda P, Bruzzone L, Tarantino C. Automatic Spectral Rule-Based Preliminary Mapping of Calibrated Landsat TM and ETM+ Images. IEEE Transactions on Geoscience and Remote Sensing. 2006;44(9):2563-86.

[12] Walper C, Braun A, Hochschild V. A Satellite-Based Framework to Investigate the Impact of Sand Dams on Landscapes in Semi-arid Regions. In: Naddeo V, Choo K-H, Ksibi M, editors. Water-Energy-Nexus in the Ecological Transition: Natural-Based Solutions, Advanced Technologies and Best Practices for Environmental Sustainability. Cham: Springer International Publishing; 2022. p. 287-90.

[13] Soci C, Hersbach H, Simmons A, Poli P, Bell B, Berrisford P, et al. The ERA5 global reanalysis from 1940 to 2022. Quarterly Journal of the Royal Meteorological Society. 2024;150(764):4014-48.

[14] Elmogy M, El-Bakry HM, Alkhawlani M, El Bakry H. Text-based, Content-based, and Semantic-based Image Retrievals: A Survey. 2015.

REAL-TIME BLIND DEFOCUS DEBLURRING FOR EARTH OBSERVATION: THE IMAGIN-E MISSION APPROACH

Alejandro D. Mousist

Thales Alenia Space, Tres Cantos, Spain

ABSTRACT

This work addresses mechanical defocus in Earth observation images from the IMAGIN-e mission aboard the International Space Station (ISS), proposing a blind deblurring approach adapted to space-based edge computing constraints. Leveraging Sentinel-2 data, our method estimates the defocus kernel and trains a restoration model within a Generative Adversarial Network (GAN) framework, effectively operating without reference images.

On Sentinel-2 images with synthetic degradation, SSIM improved by 72.47% and PSNR by 25.00%, confirming the model's ability to recover lost details when the original clean image is known. On IMAGIN-e, where no reference images exist, **perceptual quality metrics indicate a substantial enhancement**, with **NIQE improving by 60.66%** and **BRISQUE by 48.38%**, **validating real-world onboard restoration**. The approach is currently deployed aboard the IMAGIN-e mission, demonstrating its practical application in an operational space environment.

By efficiently handling high-resolution images under edge computing constraints, the method enables applications such as water body segmentation and contour detection while maintaining processing viability despite resource limitations.

***Index Terms*—** GenAI, defocus noise, remote sensing, edge computing

1. INTRODUCTION AND STATE-OF-THE-ART

The IMAGIN-e mission (**I**SS **M**ounted **A**ccessible **G**lobal **I**maging **N**od-**e**) is a space edge computing initiative hosted aboard the ISS. IMAGIN-e operates as a functional demonstration payload with real-world applications for Earth observation. Its primary objective is to evaluate the capabilities and operating modes of onboard edge computing by processing Earth observation data directly within the payload. An optical sensor was integrated to capture images that fuel onboard applications. However, the captured images exhibit significant mechanical defocus characterized by wide dispersion and smoothing (see Fig. 1), complicating precise interpretation and hindering the extraction of meaningful insights.

In this context, missions like Sentinel-2 from the Copernicus program -which provide multispectral images with higher spatial resolution (GSD) and additional spectral bands— could serve as a reference to estimate the defocus kernel when contrasted with IMAGIN-e RGB images. Nonetheless, IMAGIN-e images are not georeferenced at origin and include uncertainties (e.g., the sensor's final orientation due to its uncharacterized mechanical and thermoelastic misalignments), posing a significant challenge for restoration in the absence of sharp reference images.

Recent studies, such as Popika and Lelechenko [1], have used synthetic distortions to train models for satellite image restoration in post-processing. Our approach builds on this idea, adapting it for onboard edge computing to enable real-time correction within the IMAGIN-e payload (see Section 3).

Fig. 1: Captured image from the IMAGIN-e payload without further processing, showing significant mechanical blur.

Traditional deblurring approaches, such as the Wiener filter [2] or Richardson-Lucy deconvolution [3], rely on known blur kernel characteristics, which limits their performance for complex, non-uniform blurs observed in space-based imagery. The advent of deep learning has enabled robust alternative strategies. Early methods employed Convolutional Neural Networks (CNNs) to learn the mapping between blurred and sharp images [4, 5], while GAN-based approaches like DeblurGAN [6, 7] addressed blind deblurring when the blur kernel is unknown. More recently, transformer-based architectures have emerged as promising candidates for image restoration tasks. For instance, DeblurDiNAT[8] presents a compact model that leverages dilated neighborhood attention mechanisms to achieve robust generalization and high perceptual fidelity, even in out-of-domain settings . In parallel, MIMO-Uformer [9] integrates a U-shaped structure with window-based attention (W-MSA), enabling efficient capture of both local and global dependencies with a computational footprint suitable for resource-constrained environments.

Despite these advances, most state-of-the-art approaches assume access to paired blurred-sharp images or mandate substantial computational resources, rendering them incompatible with the onboard processing constraints of the IMAGIN-e mission.

1.1. Contribution of This Work

Our research contributes a blind deblurring methodology for satellite imagery without reference images that leverages Sentinel-2 data to characterize the defocus kernel. We adapt MIMO-Unet++[10] for space-based edge computing, optimizing computational efficiency while preserving restoration quality. Quantitative and qualitative analysis validates our approach, showing significant improvements in structural similarity and edge preservation. Additionally, we pro-

doi:10.2760/2119408

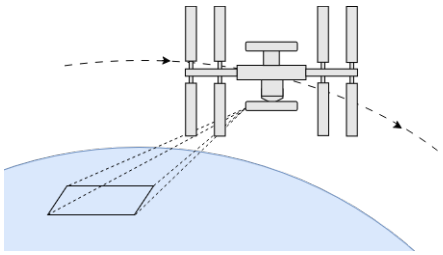


Fig. 2: Illustration of the payload orientation on the Bartolomeo platform, showing its backward tilt relative to the ISS trajectory.

vide insights into deep learning-based image enhancement for space-based observation systems with limited resources.

This study introduces a generative AI framework for defocus correction within the constraints of the IMAGIN-e mission, enhancing onboard edge computing for Earth observation and enabling the effective utilization of otherwise compromised instruments. The ability to perform defocus correction onboard enables the utilization of restored images in time-sensitive applications including disaster response, where immediate image analysis can save lives, precision agriculture requiring real-time crop monitoring, and environmental hazard detection where processing delays could compromise critical decision-making.

2. PROBLEM CHARACTERIZATION

2.1. Platform and Payload Orientation

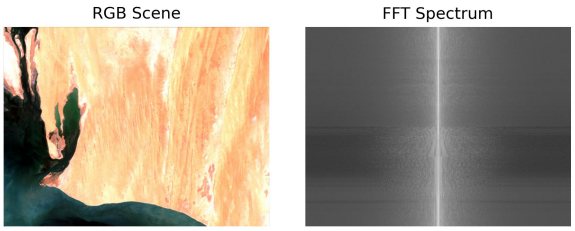
The payload is hosted on an external platform for payload hosting, mounted on the Columbus module of the ISS. Although its nominal alignment is Earth-facing, the imaging system is not perfectly oriented in the nadir direction; rather, it is directed a few degrees backward relative to the ISS trajectory (see Fig. 2). This orientation results in a non-perpendicular incidence angle compared to a purely nadir-pointing configuration, potentially affecting the observation geometry and data acquisition characteristics. Moreover, the payload was installed using a robotic arm, so the exact sensor orientation relative to nadir was not known *a priori*.

2.2. Sensor Data Characteristics

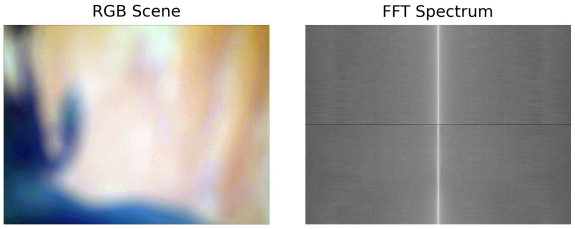
The sensor acquires RGB images compressed in JPEG format at a resolution of 2048×1536 pixels. The Ground Sample Distance (GSD) ranges from 37.5m to 41m, depending on altitude variations, ISS pitch fluctuations, and terrain elevation changes. The captured images exhibit significant optical defocus noise, likely due to mechanical miscalibration, while some images also display minor shot noise, though its intensity is considerably lower than that of the defocus blur. Figure 3 provides a spectral comparison between an IMAGIN-e capture and its corresponding Sentinel-2 scene, highlighting the frequency-domain effects of these noise sources.

2.3. Onboard Deblurring Process

The deblurring process is designed to be executed onboard without dedicated acceleration hardware as a critical step in the postprocessing stage of the capture pipeline. It takes place immediately after image acquisition, ensuring that restoration is completed before the images are passed on for further analysis. Third-party applications, which request image captures and process them upon availability,



(a) Sentinel-2: Scene and FFT spectrum



(b) IMAGIN-e: Scene and FFT spectrum

Fig. 3: Comparison of Sentinel-2 and IMAGIN-e images along with their frequency spectra. The Sentinel-2 scene, composed of RGB bands downscaled to a 40m GSD, and its corresponding frequency spectrum are presented in (a). The IMAGIN-e scene and its respective frequency spectrum are shown in (b), illustrating the effects of defocus and alterations in the frequency domain.

rely on this preprocessing step to enhance data quality and optimize downstream computational tasks.

Given the constraints of onboard execution without specialized hardware, the deblurring model must operate efficiently within the platform's limited computational resources. To meet this challenge, the MIMO-Unet++ model was selected for its high efficiency in generative processing, enabling real-time deblurring with minimal hardware requirements. By integrating this model into the capture pipeline, image restoration is performed onboard without compromising system performance, ensuring that the processed images maintain the necessary fidelity for further analysis.

3. METHODOLOGY: DEBLURRING WITHOUT REFERENCE IMAGES

3.1. Model Architecture and Training Strategy

To enhance structural features critical for georeferencing, we extracted 1024×1024 pixel patches from Sentinel-2 imagery and downscaled them to 256×256 pixels. This size reduction simplified the learning process by focusing the model on sharpening primary edge structures rather than on subtle textures. A batch size of 4 patches was chosen to balance computational efficiency with training stability. We used a MultiStepLR schedule with an initial learning rate of 1e-4, reducing it every 500 iterations by a factor of 0.5 over 3000 iterations to progressively refine the model's ability to produce spatially coherent reconstructions.

Initially, only defocused images—accompanied by tentative geolocation from the ISS's position and attitude data were available, making it extremely difficult to align these images with established ground references due to severe defocus and unknown noise characteristics. To tackle this, we first trained an early version of the MIMO-Unet++ model using RGB images generated from Sentinel2

doi:10.2760/2119408

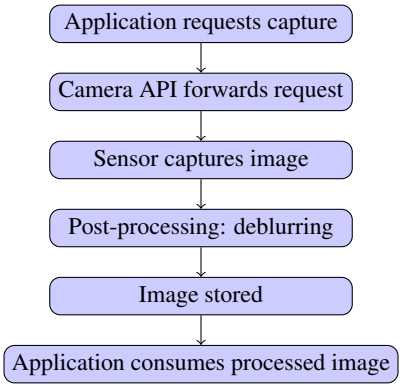


Fig. 4: The diagram illustrates the position of the deblurring process within the image processing chain. An application requests an image from the camera API, which then communicates with the sensor for acquisition. The raw image undergoes a post-processing stage, including deblurring, before being stored for later consumption by the application.

products and augmented with various noise types (Gaussian, defocus, shot, motion, and spin blur). The outputs of this model allowed us to correlate the images relative to their Sentinel-2 counterparts, leading to improved noise characterization and the creation of more realistic synthetic training data.

Subsequently, we used these synthetic images to train a refined MIMO-Unet++ model within a GAN framework, with the model serving as the generator, *i.e.*, the component responsible for producing deblurred images from blurred inputs. A multi-scale discriminator—designed to evaluate the generator's outputs at multiple image resolutions simultaneously, capturing both global structures and fine-grained details—was employed, inspired by Pix2pixHD[11], enhanced with self-attention mechanisms[12] and spectral normalization, ensuring effective extraction of features across all resolutions and promoting superior image reconstruction.

The overall loss function combined the standard adversarial loss with an L1 loss and an FFT-domain loss—as proposed in the original MIMO-Unet++ framework—as well as a perceptual loss computed using a VGG16 [13] model pre-trained on Sentinel-2 images. This comprehensive training strategy yielded a robust generator capable of delivering deblurred images with enhanced visual fidelity and structural accuracy, which is crucial for Earth observation tasks in edge computing environments.

3.2. Edge Implementation

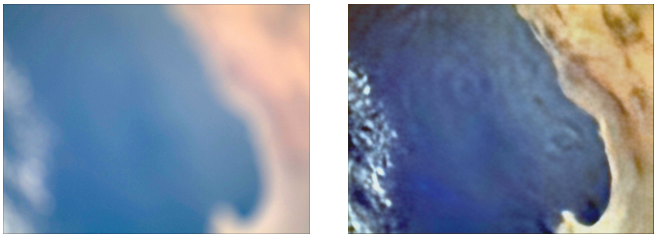
For deployment in the IMAGIN-e mission, the model must operate onboard a hosted payload on the ISS, sharing computational resources with other processes and without dedicated acceleration hardware. Therefore, it is imperative to maintain low latency to ensure seamless integration into the image post-processing pipeline (see Fig. 4). The system constraints summarized in Table 1 require that processing speed and resource usage be carefully managed to meet the rigorous demands of edge computing environments.

4. RESULTS AND DISCUSSION

The proposed deblurring approach significantly enhances image clarity and structural reconstruction. Initial models trained on

Table 1: Problem conditions

Parameter	Value
Acceleration HW	Not present
Available RAM memory	300 MB
Virtual memory	2 GB
Available CPU	3 cores (shared)



(a) Raw image (b) Deblurred image

Fig. 5: Initial deblurring effectively sharpened main borders but produced low quality images and ringing effect on some captures. Left image (Fig.5a) shows the output of the sensor, while right image (Fig.5b) shows the deblurred scene with the initial model.

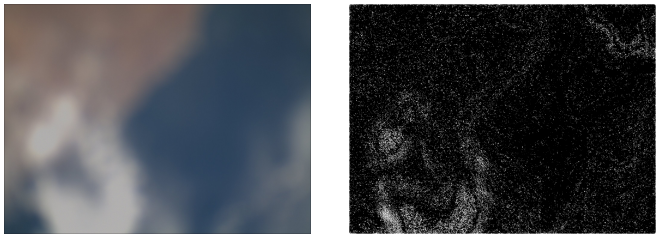
Sentinel-2 imagery were able to improve the sharpness of IMAGIN-e data (see Fig. 5), enabling subsequent georeferencing and a more comprehensive characterization of noise type, effective resolution, and spectral sensitivity. In addition, the application of a Sobel edge detection filter confirmed that, despite some undetected boundaries, the edges of critical objects and terrains were more clearly delineated (See Fig. 6). These improvements are paramount for subsequent object detection and segmentation tasks in onboard applications.

Quantitative evaluation demonstrates a substantial enhancement in image quality across multiple metrics (see Table 2). On Sentinel-2 images, SSIM improved by 72.47% and PSNR increased by 25.00%, calculated by comparing noisy synthetic images with reference images in the initial state and processed images with the same references in the final state. In contrast, for IMAGIN-e, image perceptual quality improved significantly, with NIQE showing a 60.66% enhancement and BRISQUE improving by 48.38%. Since these metrics evaluate image quality without requiring clean reference images, they are particularly valuable for real-world applications where reference-free assessment is necessary, as is the case for IMAGIN-e.

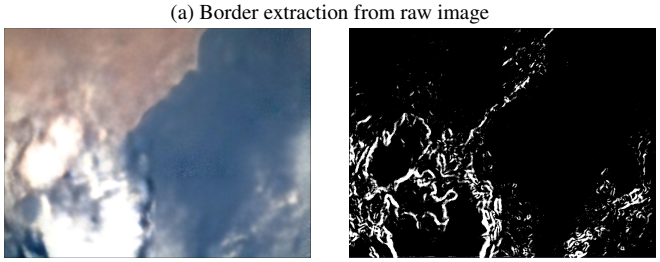
From a computational standpoint, the deblurring process operates within the edge computing constraints outlined in Table 1. Under these conditions, the model successfully processes a 2048x1536 pixel image in approximately 5 minutes, demonstrating its ability to handle high-resolution inputs despite resource limitations. Peak memory consumption reaches 600 MB, exceeding the available RAM and requiring the use of virtual memory. While this contributes to an extended processing time, the results highlight the model's adaptability in constrained environments and underscore the role of efficient memory management in optimizing performance.

Occasional ringing artifacts were observed, probably due to scaling operations during patch processing (see Fig. 7). Moreover, the effective Ground Sample Distance (GSD) varied between 37.4 m and 41 m, reflecting the dynamic imaging conditions of the ISS and underscoring the need for adaptive processing workflows.

doi:10.2760/2119408



(a) Border extraction from raw image



(b) Border extraction from deblurred image

Fig. 6: Edge detection using a Sobel filter from both the raw image (6a) and the deblurred version of it (6b)

.

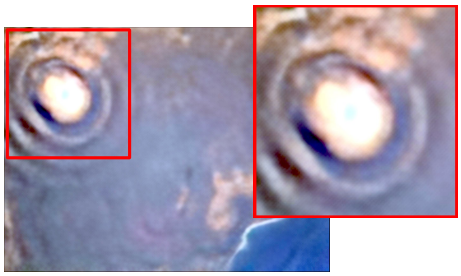


Fig. 7: Ringing effect on the images

5. CONCLUSIONS AND FUTURE WORK

Despite the inherent complexity of blind deblurring, our results demonstrate that incorporating Sentinel-2 imagery enables effective iterative processing. This strategy refined image synthesis techniques and achieved acceptable outcomes without sharp reference images. The final model executes efficiently onboard during post-processing, ensuring IMAGIN-e compatibility and maximizing instrument utilization.

Moreover, the restored images prove valuable for specific applications, such as water body segmentation and coarse contour detection for map generation. However, it is important to note that while these results are promising for certain contexts, the current resolution is insufficient for detecting small objects or for the fine segmentation of closely related classes. This limitation reflects the trade-off between processing speed and image quality inherent in edge computing scenarios.

Further research could focus on leveraging enhanced onboard computational resources to deploy more powerful models that process image patches at their original resolution. By eliminating the need for downscaling and subsequent upscaling, this approach would likely yield images with increased realism and detail. Such improvements could enhance the deblurring performance while expanding the applicability of processed imagery, especially in tasks that require the detection of small objects or fine-grained segmentation.

Table 2: Image quality metrics for Sentinel-2 synthetic validation images and IMAGIN-e real ones

Dataset	Metric	Original	Deblurred	Δ%
Sentinel-2	SSIM	0.4442	0.7662	**+72.47%**
(Synthetic)	PSNR	24.0127 dB	30.0159 dB	**+25.00%**
IMAGIN-e	NIQE	21.9257	8.6263	**+60.66%**
(Real)	BRISQUE	110.8351	57.2149	**+48.38%**

REFERENCES

[1] Viacheslav Popika and Lidia Lelechenko. Machine learning models for eos sat-1 satellite image enhancing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 1095–1098. IEEE, 2024.

[2] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series: With engineering applications. 1949.

[3] William Hadley Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

[4] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[5] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[6] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[7] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[8] Hanzhou Liu, Binghan Li, Chengkai Liu, and Mi Lu. Deblurdinat: A lightweight and effective transformer for image deblurring. *arXiv e-prints*, pages arXiv–2403, 2024.

[9] Jian Zhang, Baoping Cheng, Tengying Zhang, Yongsheng Zhao, Tao Fu, Zijian Wu, and Xiaoming Tao. Mimo-uformer: A transformer-based image deblurring network for vehicle surveillance scenarios. *Journal of Imaging*, 10(11):274, 2024.

[10] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring, 2021. URL https://arxiv.org/abs/2108.05054.

[11] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[12] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

doi:10.2760/2119408

SENTINEL-1C DATA PROCESSING WITH A SCALABLE SCIENCE-MISSION FRAMEWORK

Richard Hofmeister[1], Knut Bernhardt[1] and Faisal Rafi[1]

1 – Werum Software & Systems AG, Lüneburg, Germany

ABSTRACT

The data processing orchestration used in the Sentinel-1C Copernicus Production Service has been developed from a science-mission data processing framework in the ESA Earth Explorer programme. The methodology of the framework including specific implementations for the Sentinel-1C mission and the application setup as deployed in the public cloud is presented. Results from the commissioning phase show the performance and scalability of the system. Stability and flexibility of the data processing framework, as inherited from the operational science-mission context proves to serve well for its application for a larger-scale operational Sentinel mission.

Index Terms— Copernicus, Earth Explorer, systematic production, cloud computing

1. INTRODUCTION

The Copernicus Programme of the European Commission is the largest Earth Observation programme on Earth and consists of a variety of missions and operational initiatives to collect environmental data and exploit valuable information gathered in the different subservices. The Sentinel-1 mission is the first in the series of Sentinel satellites developed by the European Space Agency (ESA) for the Copernicus Programme. It consists of a constellation of synthetic aperture radar (SAR) imaging satellites designed to provide continuous, all-weather, day-and-night imagery of the Earth's surface. Its constellation consists of two satellite units, the older Sentinel-1A satellite launched already in 2014 and the newer Sentinel-1C, which has been launched on 5 December 2024 and replaced the retired 1B unit. Sentinel-1's SAR imagery supports a wide range of applications, including land and sea monitoring, natural disaster mapping, sea ice observations, and ship detection. This data is widely used for environmental monitoring, emergency response, and maritime safety.

The ground segment of the Copernicus Space Components are procured as industrial subservices, which are interlinked closely and coordinated centrally with common procedures. The payload data processing is delegated to the Production Services, which are renewed periodically to enable competitive service evolution. The computational environment for the software elements of the Production Services are operated in European cloud infrastructures with their inherent benefits of optimized performance, scalability, portability, political resilience and compliance.

With the launch of Sentinel-1C, a new production system was put into operation and its setup and features are presented in this paper. The software system is based on a processing framework "Olib", which has been developed and used successfully for scientific missions in ESA's Earth Explorer programme such as Swarm, EarthCARE, Biomass ([1], [2], [3], [4], [5]) as well as the new Generic Processing Orchestration System for Earth Explorer missions [8]. The framework matured and additional components have been added in the past years, such as a data management setup with AI-supported operations for a Copernicus Long-Term Archive [7].

2. METHODS AND TECHNOLOGIES

The Sentinel-1C data processing builds on a setup of software facilities implemented using the Olib processing framework. The Production Service delivers Sentinel-1C data processing by a dedicated DevSecOps team, along with comprehensive service management, reporting, and security oversight. For the methodology presented in the following, the focus is on the data processing facilities.

The major use cases of the Olib have been the payload data processing facilities, i.e. the Core Processing Facilities (CPF) as part of the Earth Explorers' PDGS and later the Copernicus Production Service (CPS), for which the OLib-software has been qualified since 2024. The Olib is the central element of the systematic processing function implementing all or most of a mission's production model. Processing starts with the availability of raw data from the acquisition facility and usually covers Level-0 (raw time ordered data), followed by Level-1 (engineering), Level-2 (geophysical) product generation. The actual processing algorithms are implemented in processor software elements, which are integrated into the CPS production workflows as external components. The Olib provides the processing management layer in charge of the processing orchestration, which allows to properly execute the various processors in line with the Sentinel-1-specific production model.

The processing orchestration is fully configurable and is able to handle re-use of existing baseline workflows from the Copernicus and Earth Explorer context. To do this, the Olib is able to handle production rules and timers for making

doi:10.2760/2119408

decisions on processing steps to be performed according to the availability of data or time events, as well as processors in charge of executing the elementary processing steps of the production model. The mission specific configuration has been adopted in the orchestration workflows for the systematic processing of the Sentinel-1C products.

2.1. CPS processing facility

The CPS uses a software system consisting of multiple components as depicted in Figure 1. It is a distributed system with management entities containing the high-level business logic and so-called agents, which carry out the low-level business logic (e.g. individual production or dissemination steps) using a typical micro-service approach. The micro-services are deployed in a dedicated Kubernetes cluster platform, where agents are running on Kubernetes pods. Deployment of the Cluster and on-demand scaling is easy and helps to manage different load situations and system recovery. Although orchestration frameworks such as Apache Airflow, Spring Cloud Data Flow, Pegasus, and Prefect support a range of workflows, this PDGS-oriented framework leverages a combination of cloud native concepts with efficient local caching, reliable long-term request queue handling, and effective priority handling for all mission's specific workflows based on different timeliness requirements e.g. NRT, NRT-PT and FAST-24 - computed dynamically on runtime. A generic workflow language, such as CWL, is not used here, since the processing steps are defined based on ESA's Task Table ICD. The Olib software is based on widely used and well supported COTS and standard protocols and the framework is published under an open source license. The system is implemented in Java Open JDK 11, the database system used is PostgreSQL with PostGIS extension and Linux is used for the operating system for the involved machines.

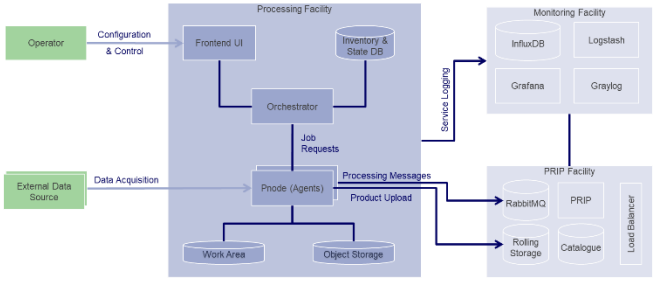


Fig 1: Processing facility as part of the Copernicus Production Service of the Sentinel-1C mission

The Sentinel-1C Production Service is grouped into three functional elements (see Figure 1): the Processing Facility, the Production Interface Point (PRIP) Facility and the Monitoring Facility. These facilities constitute the automatic software system as part of the service elements of the Production Service. External Clients can access the Monitoring facility through the monitoring dashboard and the PRIP service via the respective interfaces such as subscription to new publications, data queries and data downloads.

2.1.1. Processing Orchestrator and Nodes

The Processing Orchestrator is the core component of CPS software where the business logic is located. The orchestrator provides requests, workflows, resource control and general service functionality. The data harvesting and processing tasks as part of the production system workflows are distributed to a configurable number of processing nodes and harvesting nodes. Processing Messages are handled by a message queue within the Processing Facility which are consumed by the PRIP facility. Measurements and log messages are provided to the Monitoring Facility to keep track of the status of the processing jobs.

The software provides horizontal and vertical scaling possibilities leading to flexible sizing in order to support any kind of (ad-hoc) scaling needs. The scalability is also used to fine-tune the performance of the system when bottlenecks are identified, usually by adding components, however, automatic scaling based on system metric, as supported by the framework, is not applied in the CPS, because resources were provisioned to anticipate predictable, consistent, and systematic production workload. The orchestrator makes use of multiple production nodes to perform the CPU and I/O intensive work like higher level data production, transfer of data from/to external entities, and trace calculations as well as routine quality control. Each node contains a single instance of the CPS Processing Node Handler, which is connected to the CPS Orchestrator. It delegates processing tasks, data transfers, and traceability calculations to its associated agents. Intermediate products are cached locally for efficient access as soon as jobs can be merged in case of the same node resources selection (e.g. between Level-1 and Level-2 steps).

2.2. CPS Monitoring

The monitoring facility in Fig. 1 is supplied with operational information from all service elements, and it builds the monitoring subsystem in the technical setup of the Production Service. It is responsible for metrics gathering and consolidation as well as logging.

The monitoring information data is visualized in specific dashboards using Grafana, which provides all means to navigate through historical and current time windows. The raw values used for the visualization and key performance indicator calculation, i.e. communication status with external interfaces, components availability, production status, and resource usage, are collected automatically by the Monitoring and Logging component. Two example screenshots are provided in Fig. 3. Additionally, alarms are configured on critical thresholds to inform the operators proactively with push notifications.

doi:10.2760/2119408

Fig 3: Informative Monitoring Dashboards for the Sentinel-1C operators, visualizing the downlink information (upper panel) and the resource usage (lower panel)

3. SENTINEL-1C PRODUCTION CHAIN

The Sentinel-1C payload data production is implemented for all instrument acquisition modes: the StripMap (SM), Interferometric Wide swath (IW), Extra Wide swath (EW), and Wave (WV) mode, while a generic data processor is used for the generation of Level-1 and Level-2 data.

The Level-0 production process starts with collecting the raw input data known as Channel Access Data Unit (CADU). The CADU data is either downloaded from the ground stations utilizing CADU interface Point (CADIP) or via the European Data Relay System (EDRS). Subsequently, the raw data then undergoes different collective and consolidative steps depending on downlink chunking, gaps and polarization to ensure that the data is complete and properly consolidated. Once aggregated, the L0 raw data is assembled and segmented into overlapping L0 slices along the azimuth direction - these L0 sliced products serve as the foundation for the higher level processing steps.

For subsequent L1 processing, the Level-1 Instrument Processing Facility (IPF) is utilized. Several processing steps such as calibration, Doppler Centroid (DC) estimation, and terrain height correction are performed by the L1 IPF. Level-1 Single Look Complex (SLC) and Level-1 Ground Range Detected (GRD) products are systematically produced based on the Level-0 products. The EW Mode constitutes an exception to SLC workflow because it is triggered over specific geographical areas only.

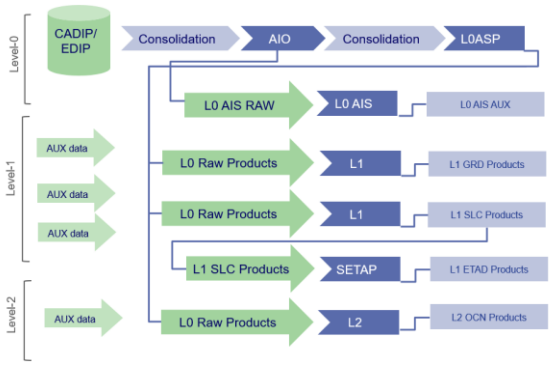


Fig 4: Simplified Sentinel-1C production model focusing on processors without refinement for the different acquisition modes

For Level-1 Extended Timing Annotation Dataset (ETAD), Sentinel-1 Extended Timing Annotation Processor (SETAP) is utilized. The SETAP processor uses Sentinel-1 Level-1 SLC Annotation products and additional Auxiliary data to generate ETAD products. These ETAD products provide improved geometric accuracy for Sentinel-1 Level-1 SLC products.

Level-2 OCN processing is systematically triggered for WV mode, while for other modes (IW, EW, SM), it is only triggered when the products footprints are located over the sea.

As a part of SAR Post Processing (SPP) chain, Orbit Baseline and TOPSAR Synchronization (OBS), products are generated systematically by a dedicated IPF. The ERRMAT processing chain systematically generates Error Matrices files. These files contain detailed error characterization which is essential for accurate calibration of Sentinel-1 SAR data.

The orchestration framework supports both event-driven and scheduled workflows, while CPS relies on events from the messaging system (RabbitMQ). Once an event occurs, the corresponding workflow is initiated. However, some workflow-steps may wait until all required dependencies are met before proceeding. For example, a Datatake that is split across different downlink stations needs to be consolidated. The CPS creates a request and waits until all segments corresponding to a Datatake become available, then it proceeds with execution of the request.

3.1. New AIS Processor

The integration of an Automatic Identification System (AIS) for vessels with Sentinel-1's radar capabilities improves maritime safety by helping ships avoid collisions, supporting efforts to detect illegal activities, and mitigating pollution risks.

The Sentinel-1C mission includes the use of the AIS processor for ship tracking, enhancing their capabilities for

doi:10.2760/2119408

maritime monitoring. The satellite carries an antenna that captures AIS signals transmitted by ships, which include details such as the vessel's identity, location, and direction. The AIS processor on the Sentinel-1C satellite allows for precise tracking of ships and is intended to be used for the upcoming Sentinel-1D unit as well. AIS imagery is made available to European Maritime Safety Agency (EMSA), enabling EMSA to utilize this satellite AIS data in conjunction with terrestrial data to enhance the safety of various maritime operations.

3.2. Resource usage during Commissioning

The initial stages of the Sentinel-1C commissioning were designed to test the nominal functioning of the Satellite's instruments under minimal load. As the payload increased during later stages, the production capacities were scaled up to anticipate the increased load accordingly. During the final phase of Sentinel-1C commissioning, the system load was ramped up to nominal operational capacity - approximately 800 GB of raw data was processed to produce ~5 TB across Level-0, Level-1, and Level-2 products on daily-basis. The production was carried out on a Kubernetes based cluster consisting of over 625 CPU cores and 3 TB of RAM, distributed across different machine types tailored to meet varying processing demands. Fig. 5 depicts the number of products produced by Sentinel-1C during a single week of the commissioning phase.

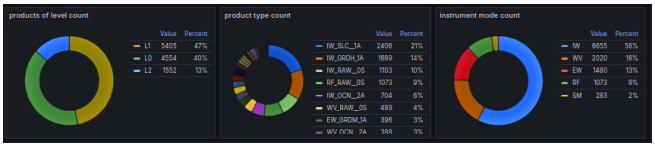


Fig 5: Production counts for 7 days during commissioning accumulated by processing level (left), product type (middle), and instrument mode (right)

4. SUMMARY AND CONCLUSIONS

A generic processing framework for scientific Earth observation missions' processing in the ground segments is used in the Copernicus Programme for the operational payload data processing of the Sentinel-1C mission. The system concept and software implementation has been matured with the different use cases and similar setups will build the data processing backbone of some of the near-future science missions in the Earth Explorer programme.

A modified instance of this processing facility is used by the Authors' company to operate one of the Copernicus Long-Term Archive Services in a public-cloud environment since year 2020 [7]. These multi-purpose use cases of the framework enhance the stability and variety of the system's modules.

The software framework is published as Generic Processing Orchestration System in a Space CODEV repository [9]. Although the software had a science focus originally, the Sentinel-1C application as well as the Copernicus Long-Term Archive Service demonstrate the successful knowledge transfer from explorational science activities to large scale, operational programmes of the society.

REFERENCES

[1] E. Friis-Christensen, H. Lühr, D. Knudsen, R. Haagmans, Swarm – An Earth Observation Mission investigating Geospace, Advances in Space Research, Volume 41, Issue 1, Pages 210-216, 2008

[2] Wehr, T., Kubota, T., Tzeremes, G., Wallace, K., Nakatsuka, H., Ohno, Y., Koopman, R., Rusli, S., Kikuchi, M., Eisinger, M., Tanaka, T., Taga, M., Deghaye, P., Tomita, E., and Bernaerts, D.: The EarthCARE Mission – Science and System Overview, doi:10.5194/egusphere-2022-1476, 2023

[3] Shaun Quegan, Thuy Le Toan, Jerome Chave, Jorgen Dall, Jean-François Exbrayat, Dinh Ho Tong Minh, Mark Lomas, Mauro Mariotti D'Alessandro, Philippe Paillou, Kostas Papathanassiou, Fabio Rocca, Sassan Saatchi, Klaus Scipal, Hank Shugart, T. Luke Smallman, Maciej J. Soja, Stefano Tebaldini, Lars Ulander, Ludovic Villard, Mathew Williams, The European Space Agency BIOMASS mission: Measuring forest above-ground biomass from space, Remote Sensing of Environment, Volume 227, Pages 44-60, 2019

[4] Pruin, B. C. Caspar, C. Stella, N. Junike and A. Strecker: EarthCARE Processing Facility and EarthCARE L2 Testbed - a synergetic setup to support scientific algorithm development, IAC-18,B1,IP,3,x43883, 2018

[5] Pruin, B., N. Junike and A. Strecker: Selective data processing in DIAS for localized time series analysis - a specific use case for a generic dias processing suite, Proc. of the 2019 conference on Big Data from Space (BiDS'19), Pages 193-196, 2019

[6] Moreno, J., R. Colombo, A. Damm, Y. Goulas, E. Middleton, F. Miglietta, G. Mohammed, M. Möttus, P. North, U. Rascher, C. van der Tol, M. Drusch, Quantitative global mapping of terrestrial vegetation photosynthesis: The Fluorescence Explorer (FLEX) Mission, IGARSS 2017 (IEEE International Geoscience and Remote Sensing Symposium), Fort Worth, Texas, USA, 2017

[7] Hofmeister, R., T. Keßler, A. Strecker, A. Petersen, Q. Saalfeld, K. Bernhardt, Operational strategies for a continuously growing public-cloud archive, Proc. of the PV conference 2023, 2023

[8] Hofmeister, R., Bernhardt, K., Strecker A., Keßler, T. and Caspar, C.: Processing Framework For Scientific Earth Observation Missions Proc. of the 2023 conference on Big Data from Space (BiDS'23), Pages 309-312, 2023

[9] http://www.space-codev.org

OPERATIONAL PROCESSING OF DLR'S SENTINEL-1 NORMALIZED RADAR BACKSCATTER PRODUCT

Jonas Eberle, John Truckenbrodt, Mario Winkler

German Aerospace Center (DLR), German Remote Sensing Data Center

ABSTRACT

Operationalizing algorithms in the Earth Observation (EO) domain requires scalable and maintainable workflows to manage the variety and large data volume of satellite missions. This paper presents the processing system for DLR's Sentinel-1 Normalized Radar Backscatter (NRB) product, designed for operational deployment and data production. High-Performance Computing (HPC) resources are utilized to efficiently process the large data volumes of Sentinel-1 on DLR's high performance data analytics platform terrabyte. The workflow is modeled using Business Process Model and Notation (BPMN), chosen for its clarity, standardization, and suitability for managing complex processing chains. The combination of BPMN and HPC enables automation, robustness, and traceability in the product generation process. This approach supports consistent and high-quality production of the NRB product, demonstrating the importance of structured workflows in transitioning EO algorithms from research to operational environments. It is based on open source software developed in the framework of ESA's Exploitation Platform Common Architecture (EOEPCA).

***Index Terms*—**Workflow, BPMN, HPC, Analysis-Ready-Data, Sentinel-1, EOEPCA

1. INTRODUCTION

In the field of Earth Observation (EO), the transition of algorithms from research to operational use presents significant challenges. These include not only algorithmic robustness and data quality but also the development of scalable, maintainable, and automated workflows capable of supporting continuous production. The Copernicus Sentinel-1 mission provides Synthetic Aperture Radar (SAR) data on a global scale. To maximize the scientific value of this data, derived products such as the Normalized Radar Backscatter (NRB) [1] must be generated routinely and reliably.

The NRB product represents a radiometrically calibrated, terrain-corrected SAR backscatter dataset, suitable for long-term monitoring and geophysical analysis. Ensuring the operational availability of such a product involves more than implementing the core algorithm, it requires a comprehensive processing system that addresses data ingestion, scheduling, execution, error handling, and data publication.

To meet these demands, workflows play a central role. In this context, workflows are not merely sequences of tasks but structured representations of the entire processing pipeline, enabling transparency, reproducibility, and scalability. Business Process Model and Notation (BPMN) was selected as the workflow language due to its standardized notation and suitability for modeling complex, conditional logic in a readable format. BPMN supports collaboration between domain experts, software engineers, and operators, improving communication and maintainability throughout the system lifecycle.

The implementation on DLR's high performance data analytics platform terrabyte [2] leverages High-Performance Computing (HPC) infrastructure to address the computational load imposed by the large data volumes of Sentinel-1. The use of HPC resources ensures timely product generation, while BPMN enables clear orchestration of parallel processing steps and quality control procedures.

This paper describes the design and implementation of the operational processing system for the Sentinel-1 NRB product, with a focus on the integration of hybrid processing environments (Cloud and HPC) within BPMN-based workflows. The goal is to demonstrate how modern workflow technologies support the operationalization of EO algorithms, enabling reliable and scalable production systems aligned with the growing demands of satellite-based Earth monitoring and considering the different perspectives of algorithm developers, processing environments, workflow orchestration, and operations.

2. THE SENTINEL-1 NRB PRODUCT

The Sentinel-1 NRB product was originally defined by University of Jena for ESA. Together with the product definition, a prototype processor was implemented to generate test datasets for product demonstration. This work has been continued at DLR by operationalizing the prototype processor and refining the product definition. The main measurement is radiometrically terrain corrected (RTC) gamma naught backscatter (γ_T^0) per acquired polarization. Additionally, several ancillary layers support the backscatter interpretation: a multi-layer data mask containing layover and shadow masks as well as water body mask, the ellipsoidal and local incident angles, the local contributing area, a ratio for converting to sigma naught backscatter, and a layer delineating the extent of the source products.

doi:10.2760/2119408

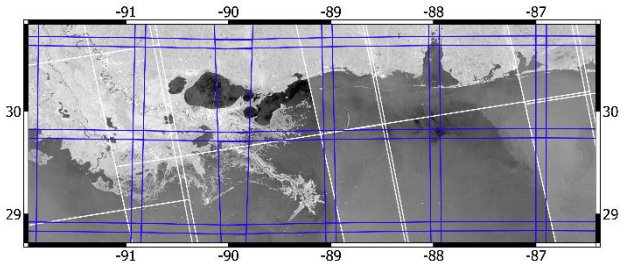


Fig. 1. Sentinel-1 NRB product in MRGS tiling scheme. Blue: MGRS tiles; White: Sentinel-1 GRD footprints.

2.1. Software

The open-source Analysis Ready Data (ARD) processor s1ard [3] orchestrates the ARD generation. It queries source scenes from a database, invokes a core SAR processor for generating geocoded SAR backscatter and ancillary products, and converts the output into the final NRB structure including file format conversion, re-gridding and metadata generation. As core SAR processor ESA SNAP [4] is currently used because of its open-source availability and comprehensive functionality. However, the modular design facilitates easy replacement of the core processor by other software solutions. pyroSAR [5] is used for reading SAR product metadata, configuring and running SNAP workflows, as well as downloading needed ancillary for processing like digital elevation models and orbit data.

2.2. Processing steps

The key processing steps are as follows:

- **Configuration Setup:** Generates a configuration file specifying parameters such as scene selection, output paths, and metadata.
- **Job Order Creation:** Defines the area of interest and acquisition parameters (e.g. time range, mode) and creates a vector geometry file.
- **Job List Generation:** Queries the database to identify MGRS tiles and matching Sentinel-1 scenes, storing scene-to-tile associations in job lists.
- **SAR Processing:** Executes SAR core processing for each scene using the defined configuration.
- **ARD Conversion:** Converts processed scenes into NRB tiles through mosaicking, cropping, and metadata generation.
- **STAC Registration:** Ingests the final product metadata into a STAC-compliant database for API-based access.

2.3. Data input and output

Sentinel-1 GRD products serve as input. The whole archive is stored on terrabyte for immediate availability. Careful curation of this product ensures that always the latest processed GRD products are available to avoid duplication. The ARD data output is gridded – identical to Sentinel-2 data – in tiles of 110x110 km size aligned to the Military Grid Reference System (MGRS) with a pixel spacing of 10 m (see Fig. 1). The binary data of the product is stored in cloud-optimized GeoTIFFs (COGs). In addition, comprehensive metadata has been defined in JSON files compliant to the SpatioTemporal Asset Catalog (STAC) specification.

The data volume of the final NRB product is about 2.3 times that of the input GRD product resulting in 9 PB. Data volume is increased by adding several ancillary layers and COG overviews as well as through the MGRS grid overlap. A decrease in volume is achieved by applying compression. The full archive is currently being processed and gradually made available to users. An example of data cube analysis capabilities was presented at the last BiDS [6].

3. PROCESSING ENVIRONMENT

The terrabyte EO exploitation platform is used to produce the Sentinel-1 NRB product. Terrabyte is a hybrid high performance data analytics platform based on on-premise High-Performance Computing (HPC) and on-premise cloud-like infrastructures. Both, HPC and cloud, are connected internally with high-speed network to the Data Science Storage (DSS), where all data – EO data, intermediate data, and the final product – is stored. DSS Management Servers are available for high-speed data access.

3.1. Algorithm packaging

Packaging is an important activity to have a trusted source of dependencies necessary for the algorithms used in a processing campaign. Today, Docker containers are used often to package an algorithm with a defined set of software necessary to execute the algorithm. However, containers cannot be used in all environments or need to be converted to other container formats (e.g., HPC systems do not allow for Docker containers, but they can be converted to Charliecloud or Singularity containers). To be interoperable with different infrastructures, Micromamba [7] has been selected to conduct the algorithm packaging with all software dependencies, which can be used in both, HPC- and cloud-based infrastructures. Micromamba environments can be either used individually (e.g., as base environment for an HPC job) or installed within containers (e.g., Docker container for cloud-based processing).

3.2. terrabyte HPC

The terrabyte HPC system, with 44,000 virtual CPU cores and 287 TB RAM, supports high-throughput processing of compute-intensive jobs. For the Sentinel-1 NRB processing, two HPC job types are executed: (1) SAR scene processing using ESA SNAP software (16 cores, 45 GB RAM, 60–90 min runtime) and (2) ARD conversion using the Python-based s1ard (4 cores, 40 GB RAM, 5–10 min runtime per tile). At full capacity, up to 2,750 SAR jobs can run in

doi:10.2760/2119408

parallel. Jobs are submitted via SLURM CLI or REST API, accessible only within secure networks. Accordingly, workflow steps involving job submission must run within these environments (e.g., HPC login-nodes).

3.3. terrabyte Cloud

While HPC is used for CPU and RAM-intensive processing tasks, the cloud environment can be best used for low-level tasks, such as input data discovery, creation and registration of metadata as well as data validation. terrabyte Cloud has a maximum of 3,000 CPU cores and is used for all terrabyte web services (e.g., metadata catalogue, visualization services, support forum, workflow orchestration). For the Sentinel-1 NRB processing it is mainly used for workflow orchestration.

3.4. terrabyte Data Science Storage Management Nodes

The terrabyte Data Science Storage Management Nodes provide the best performant access to the large Data Science Storage system. These nodes can be used for input and output data validation and integrity checks (e.g., checksums) as well as data transfers from the temporary processing folder to the central location for data publishing.

4. PROCESS ORCHESTRATION

Operational product generations require the coordination of a sequence of processing steps (see Chapter 2.2). This set of tasks must be designed and integrated into structured and repeatable processes to ensure the reliability, maintainability and scalability of the whole processing system. In this context, workflow definitions based on the Business Process Model and Notation (BPMN) specification provide a structured and transparent approach to process definition and orchestration. By providing a clear and visual framework, BPMN models enhance the visibility of the process during the development as well as the operational phase, bridging the gap between stakeholder requirements, technical implementation and operational tasks. They serve as the backbone of operational production, enabling the automated execution of workflows in a controlled and traceable manner.

4.1. BPMN models

A BPMN model is a static representation of the workflow as a BPMN 2.0 process definition. It defines the sequence of activities (service tasks), events and (conditional) sequence flows to be executed during the workflow and is usually stored as an XML file. An overview and the specification of all possible BPMN modelling elements can be found in [8]. The workflow for Sentinel-1 NRB has been separated into four models to achieve a better overview, organization and error handling for the operator:

- Data discovery and job preparations for a user-defined area and time of interest
- Grouping of the processing for each Sentinel-1 data take found in data discovery step
- Geocoding of each Sentinel-1 GRD scene
- ARD conversion and data publishing for each MRGS tile of data take

4.2. Automated product validation

An important step in the workflow is the automated product validation for both SAR processing and ARD conversion to the final output format. Diverse errors can happen during the processing, e.g. issues with access to the data storage, execution in SNAP, creation of metadata, failed downloads of auxiliary data, processing timeout.

After the HPC job is finished either in a failed or completed state, the generated output is validated and the log files are scanned for typical errors. If there is a “well known error” (e.g., failed downloads or access to storage is not available), the processing job will be automatically cleaned and retried, which is modelled in the BPMN diagram. Only for specific or unknown errors or too many retries, the processing job will be directed with a conditional sequence flow to a human user task in the BPMN model. An operator or expert of the algorithm needs to manually check the error and restart or finish the execution of the job.

4.3. System architecture

As a central BPMN-workflow orchestration tool, the Open Source BPMN engines Camunda 7 and Flowable 7 can be used on terrabyte. Currently, the NRB processing workflow is operated with Camunda but with the upcoming release 2.0 of ESA’s Exploitation Platform Common Architecture (EOEPCA) components, the workflow will be migrated to EOEPCA’s Resource Registration building block, which uses Flowable as workflow engine.

A BPMN workflow engine supports the execution of service tasks as internal, synchronous invocation of code deployed in the workflow engine itself or as external, asynchronous tasks provided to an external worker. This external worker process is an entity that is independent of the workflow engine. It does not need to run in the same process, on the same machine or even in the same cluster. When the workflow engine encounters a service task that is configured to be externally handled, it provides this unit of work as a job, which can be polled and acquired by the external worker. The worker then executes the job and sends the result back to the engine. The benefit of this pattern is that the entities performing the actual work are independent of the workflow engine which allows the worker to be deployed, scaled out or maintained independently.

In the NRB processing system the workers are distributed between the terrabyte HPC, Cloud, and DSS Management Node environments. CPU intensive workflow tasks like SAR processing and ARD conversion are handled by workers deployed on the Login-Node which submit corresponding

doi:10.2760/2119408

HPC jobs to the cluster and manage their states. In contrast, the more data and access related tasks like data validation, data transfer and metadata registration are deployed on both, Cloud and DSS Management Nodes, to benefit from fast access to the platform storage.

4.4. Operations

For the operations of the product generation, dashboards have been developed to show different aspects of the processing campaign: (1) the HPC resources available and used (see Fig. 2), (2) the amount of scenes processed, (3) the amount of errors, (4) the HPC jobs and their current state (queued, running, finished), (5) the storage capacity for the intermediate processing location as well as the final output location, and (6) the amount of ARD tiles generated. In case of errors, the operator can use the Camunda web interface to analyze the error and inspect the log files.

Fig. 2. Monitoring of HPC resources used.

5. CHALLENGES AND OUTLOOK

Producing Sentinel-1 NRB products operationally involves key challenges in data handling, resource management, and data format limitations. Large data volumes, especially during parallel processing, require stringent cleanup procedures to avoid storage overload. Temporary files must be removed immediately after each processing to maintain system stability. Job scheduling also poses difficulties: Fast ARD jobs (5–15 minutes) must be prioritized over longer geocoding tasks (60–90 minutes) to optimize HPC utilization. This demands adaptive orchestration strategies that respond to workload fluctuations. Furthermore, the reliance on GRD input data simplifies access—since the full archive is online via DLR's terrabyte platform—but introduces border inconsistencies between scenes. SLC data avoids this issue, yet its larger size and limited online availability present significant processing and access challenges. These issues underline the need for flexible, automated systems capable of balancing efficiency, quality, and operational readiness.

Future improvements to the NRB production system focus on enhancing data quality and ensuring operational sustainability. A shift from GRD to SLC input is planned to increase data quality, but this requires handling significantly larger datasets and addressing the limited availability of SLC data, which needs to be retrieved from DLR's long-term archive. Additionally, with the end of open-source support for Camunda 7 by late 2025, workflow orchestration will migrate to the Harvester component of ESA's EOEPCA+ framework. Harvester supports workflow modeling and execution while maintaining compliance with open-source and European data standards. These transitions aim to boost the system's quality, scalability, and interoperability, ensuring the S1-NRB production remains reliable and adaptable in evolving computing environments.

6. CONCLUSIONS

The Sentinel-1 NRB product enables consistent, analysis-ready backscatter data for scientific and monitoring applications. Its operational production relies on a robust system that combines HPC capabilities with BPMN-based workflow orchestration. This paper detailed the system's architecture and addressed challenges related to data volume, job prioritization, and input data limitations. Planned transitions to SLC input and open-source orchestration tools like EOEPCA's Harvester will further enhance product consistency and long-term maintainability. The S1-NRB system demonstrates how research-grade EO algorithms can be successfully operationalized for large-scale, high-throughput production, supporting long-term environmental monitoring on a continental scale.

REFERENCES

[1] Truckenbrodt, John, M. Wolsza, A. Valentino, C. Albinet, A. Wendleder, J. Eberle and K. Molch (2023). The ESA Sentinel-1 Normalized Radar Backscatter Product. PV2023, 2023-05-02 - 2023-05-04, Geneva, Switzerland.

[2] Eberle, Jonas, M. Schwinger, and J. Zeidler (2023). Challenges in the development of the EO Exploitation Platform terrabyte. In: Big Data from Space, Pages 97-100. Publications Office of the European Union. Big Data from Space 2023, 2023-11-06 - 2023-11-10, Wien, Austria. doi: 10.2760/46796.

[3] https://github.com/SAR-ARD/s1ard

[4] https://step.esa.int/main/toolboxes/snap/

[5] https://github.com/johntruckenbrodt/pyroSAR

[6] https://github.com/DLR-terrabyte/demo-bids23

[7] QuantumStack (2025). Mamba documentation website. https://mamba.readthedocs.io/en/latest/ (accessed 2025-05-02).

[8] Object Management Group, OMG, "Business Process Model and Notation (BPMN) Specification", 2013, https://www.omg.org/spec/BPMN/2.0.2/PDF

doi:10.2760/2119408

CHALLENGE AND SOLUTION FOR ALGORITHM OPTIMISATION FOR GLOBAL PROCESSING OF FIRE BURNED AREA AT SCALE AND AT MINIMUM COSTS

Hannes Neuschmidt[1], Martin Böttcher[1], Thomas Storm[1], Ekhi Roteta[2], Andreas Hangler[3], Carsten Brockmann [1]

[1]Brockmann Consult GmbH
[2]University of the Basque Country
[3]Cloudflight GmbH

ABSTRACT

Global mapping of burned areas (BA) at high spatial resolution is critical for understanding fire impacts on ecosystems and emissions. This paper analyzes the technical obstacles in processing global Sentinel-2 20m imagery for BA detection and introduces an optimized, incremental time series algorithm designed to minimize computational costs. The new approach, implemented within the Calvalus processing system on the Copernicus Data Space Ecosystem, eliminates redundant input/output operations by avoiding the storage of intermediate products and leveraging a sliding window strategy. This results in substantial reductions in memory usage and processing time, enabling scalable and cost-effective production of global high-resolution BA datasets.

Index Terms— *Burned Area mapping, Sentinel-2, High-resolution remote sensing, High volume data throughput, Algorithm optimization*

1. INTRODUCTION

Fire has a profound impact on ecosystems, atmospheric chemistry, CO2 emissions as well as human settlements and air quality. Burned area (BA) estimation from satellite data has been performed since the 70s [1]. Large scale BA products from medium resolution optical satellites are available but have been shown to be unable to reliably detect small fires (< 100 ha). In [2], [3] an approach for large scale BA mapping using Sentinel-2 images with 20m resolution has been developed and applied to create a data base covering all of sub-Saharan Africa.

However, when applied globally, the processing of high resolution imagery on very large scales poses significant computational challenges, due to the very large data sets involved. The processing chain must be extremely efficient and well adapted to the underlying computer infrastructure to make global scale processing possible and cost effective. This paper analyses the computational challenges involved in global processing of high resolution burned area, using Sentinel-2 20m data with the algorithm proposed in [2], [3] and presents a new formulation of the algorithm as an incremental time series algorithm.

Processing is done using the Calvalus processing system, developed by Brockmann Consult GmbH, Germany (https://www.brockmann-consult.de/calvalus/) which provides a framework for big data earth observation processing based on Apache Hadoop [4]. For the global computation of burned area, Calvalus is deployed on the Copernicus Data Space Ecosystem (CDSE) [5].

2. ALGORITHM

The algorithm is presented here in an abbreviated manner, focusing on aspects that are relevant to the technical challenges associated with very large-scale processing. For a more in-depth discussion of the algorithm, see [2], [3].

Fundamentally, burned area is determined in two phases. In the first phase, a per-pixel confidence level of burned area is computed for all (unordered) pairs of distinct Sentinel-2 MSI L2A products from the same platform with an observation date no longer than 40 days apart. Pairs consist of one later observation, termed *reference* and one earlier observation, termed *predecessor*. The second phase encompasses all remaining steps in which the confidence levels are filtered and aggregated into monthly products.

The determination of burned area between a pair of Sentinel-2 observations in phase one takes into account the differences between the Mid-Infrared Burned Index (MIRBI) and Normalized Burned Ratio 2 (NBR2) spectral indices and the NIR band B8A of the reference and predecessor images. Pixels classified as burned are only considered if they form a patch larger than 30 ha and are close to an active fire, sourced from the VIIRS sensor. The scene-classification (SCL) mask provided by the Sen2Cor algorithm is used to exclude clouds, no data and water pixels from the analysis. This first step is called *pre*.

doi:10.2760/2119408

The second phase comprises a sequence of filters and aggregations. The steps of the second phase are termed *post*, *fuse* and *tile*. In the first phase, the same observation serves as a reference for multiple pairs of observations, visualized in Fig. 1. The post step classifies for each reference and each pixel if the pixel has been observed, if it was burned and selects the confidence level from the pre step with the least temporal distance to the reference where the pixel has been observed in both elements of the pair. Both pre and post steps operate only on observations from the same platform, either Sentinel-2A or Sentinel-2B. In the second step of the second phase, fuse, pixels classified as burned by one platform that have been classified as observed but not burned by the other platform are discarded. The final step of the second phase, tile, temporally aggregates the filtered burned area confidence levels and classifications into a monthly product.

Further steps, not discussed here, spatially aggregate and resample the results into 5-by-5-degree tiles and a global product with reduced resolution. Each of the steps described in this paper operates on *input products* (Sentinel-2 MSI L2A) or on intermediate products produced by a prior step which have the same extent and UTM projection as the inputs.

2.1. Original Implementation

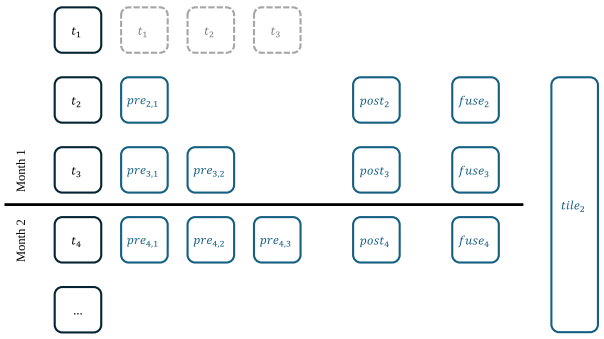


Fig. 1. Processing flow of the original implementation for a single platform. Each box represents an intermediate or output product that is written to disk and can be inspected. Products depend on inputs to their left. Additionally, fuse and tile steps depend on inputs from the other platform.

The baseline implementation, which has been used to generate products for south-Saharan Africa for the years 2016 [2] and 2019 [3] computes each step atomically, reading all necessary inputs and writing a single output. This approach allows for maximum parallelization on the algorithm step level, as each computation is only dependent on its required inputs to be present on the distributed file system. The local state in the main memory on any specific node is only relevant to a single step and held only for a short time. Fig. 1 shows the structure of the algorithm for a single platform.

2.2. Performance considerations for global processing

While the baseline implementation allows for extremely parallel execution, the number of parallelizable steps far exceeds the number of available computing notes. There are two main contributions to redundant input and output (I/O) operations. First, intermediate results are created on disk which are not part of the delivered product. This is particularly relevant in the case of the pre step, which creates an output for each pair of inputs. Second, each step must read all input products required for its computation. However, the required inputs for many steps overlap significantly. For the pre step, each input product acts as a reference product for four to eight predecessors and as a predecessor for four to eight other reference products. Therefore, the same product may be read up to 16 times, significantly driving up resource costs. The fuse step reads a time series of all post step intermediate products in a time window of 80 days. The fuse step for the next observation date of the same platform reads the same time series shifted by a few days, which changes the required inputs only at the edges of the time window. Significant time is spent on writing to disk and reading inputs.

To make the global production cost effective, a new computational approach has been developed to more efficiently compute the same burned area product for very large-scale processing. The new implementation is designed to avoid the most significant computational inefficiencies described above. The new approach seeks to minimize I/O operations, keep memory usage low to accommodate multiple instances on one machine and optimize CPU usage.

2.3. Incremental time series approach

The new implementation is based on an incremental time series analysis approach. Fig. 3 shows the set of intermediate products held in main memory at a single instant in time. The algorithm maintains two cursors into the time series of Sentinel-2 observation dates of the current granules, which are the inputs to the algorithm, to define the processing window. The two cursors can be interpreted in relation to the input products as well as the intermediate products produced by the post step, as described in section 2.1.

In the new implementation, these intermediate products are never written to disk. We name the cursor indicating the start of the processing window the *time frame cursor* and the cursor at the end the *product cursor*.

doi:10.2760/2119408

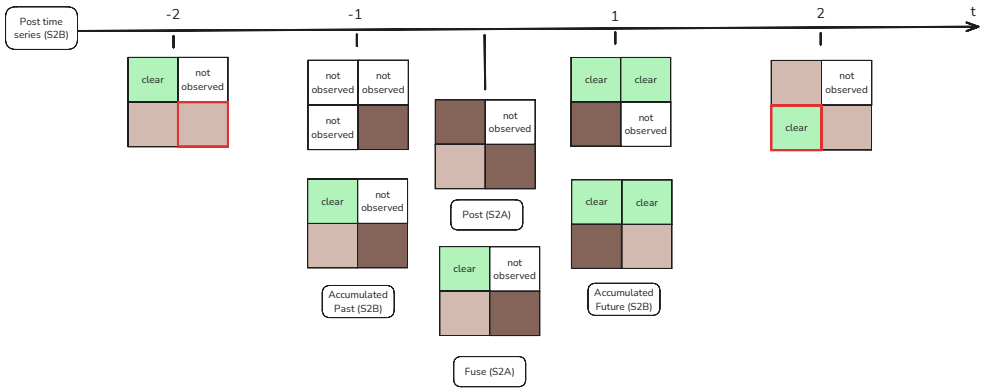


Fig. 2. Fuse step. To filter a target observation of S2A, the observations of S2B 40 days before and after the target observation are aggregated separately. Dark pixels represent a high confidence burned area, light pixels low confidence burned area. White pixels are not observed. A thick red border marks pixels that are overwritten in the aggregation process.

In addition to the intermediate products in between the cursors, the algorithm maintains accumulated state products for the fuse (one per platform) and tile steps. These state products have a similar memory footprint to a single intermediate product of their respective step and serve to reduce the number of intermediate products in memory.

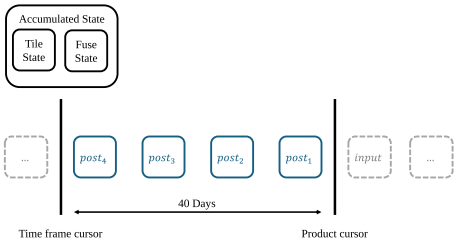


Fig. 3. Processing flow of the new implementation

In relation to the input products, any product older than the time frame cursor as well as its derivative products from the post and fuse steps is completely processed. The data from these inputs only affect future results though the accumulated states described below. With regard to the post step intermediate products, the cursors define the range of post step intermediate products that are required to compute the future fuse step intermediate products. The product following the time frame cursor is the next product to be filtered in the fuse step. The time frame cursor advances, after a fuse step has been performed and the result is integrated into the algorithm state. The product cursor advances when a new product must be opened to generate a new post step intermediate product. The products with time stamps in between the two cursors are, by definition, the minimal set of products needed to advance the algorithm. This sliding window approach minimizes product reads and writes to the achievable minimum, as every input product is read exactly once, and no intermediate products are written or read. Only the final products required for delivery are written to disk.

In the following, the dependencies between the various algorithm steps are analyzed, to find the minimal set of products required at any point in the algorithm. These dependencies are expressed as the (intermediate) input products read by each step in the original implementation of the burned are algorithm. Trivially, each pre step requires two input products, predecessor and reference, when it is performed. The post step requires the results of all pre steps with the same reference product, as it selects the most recent classification and confidence levels for one reference. The fuse step, visualized in Fig. 2, requires all products from the other platform, e.g. Sentinel-2B for a main input computed from Sentinel-2A observations, in a time window of ±40 days. The tile step requires all results computed by the fuse step belonging to observations in its target month, plus all fuse result products from 40 days before the beginning of the month, to make sure the same burned area is not counted twice, in two separate months.

It could be concluded that the maximum number of products that the algorithm must keep in memory encompasses 40 days of fuse result products before the beginning of the month and 40 days of post result products after the end of the month in addition to fuse and post result intermediate products inside the month. This maximum occurs at the end of the month, when the earliest intermediate product required by the tile step has a time stamp 40 days before the beginning of the month, and the latest intermediate product required by the fuse step, 40 days after the most recent observation, are furthest apart. Fortunately, we can reduce the number of open products significantly, by taking advantage of the fact that most intermediate products can be aggregated in time immediately, as only the aggregated products are required to compute the later algorithm steps. This strategy can be applied to the post, fuse and tile steps.

doi:10.2760/2119408

This process yields an advantage, as the algorithm can discard those intermediate products that affect further computation only in the form of the aggregated product.

The post step selects the confidence level of that pre step computation which has the smallest temporal distance between reference and predecessor. This means, it is enough to keep track of the most recent confidence level for each pixel, when computing the pre steps for a single reference in order of largest temporal distance to smallest. In this case, only two intermediate results need to be saved, the most recently computed one and the aggregation for the current reference.

The fuse step filters the intermediate product for a single reference observation produced by the post step, by aggregating future and past time series of post step products of the other platform, to confirm or reject burned area observations. The post step intermediate product to be filtered is referred to as the target. Fig. 2 shows the aggregation step for a short time series. The fuse step prioritizes information of the post step results from observation dates closest to the target reference date. Information from the latest observation before the target has the highest priority in the past aggregation and information from the observation immediately following the target has the highest priority in the future aggregation. For the fuse step, this process is subject to two caveats: Once a post step result has been integrated into the aggregated state, the fuse step cannot be computed for any observation time stamp before the newly integrated post step result because of the overwritten information. Secondly, the algorithm cannot use the same strategy to keep an accumulated state into the future. The future aggregation must be computed backwards in time, starting from the latest observation and finishing with the observation immediately following the target. Fig. 2 shows information that is overwritten in the aggregation process with a thick, red border. The pixel overwritten in the future aggregation is relevant for the next fuse step computation. However, it is lost in the aggregation for the current fuse step. This means that the algorithm needs to keep access to the product at t_2. Conversely, the product at t_{-2} can be safely discarded, because the product overwriting its information has higher priority for all future fuse step computations.

The tile step proceeds in a similar way to the past comparison of the fuse step and keeps an accumulated state of the most up to date information for the currently unfinished month. Taking advantage of aggregated state products for the post, fuse and tile steps, the algorithm needs to keep track of post result intermediate products for 40 days to compute the future aggregation of the fuse step. The aggregated products eliminate the need to access intermediate products behind the time frame cursor. The past aggregation of the fuse step is replaced by an aggregated product. Besides accumulated states for tile and fuse with a memory requirement similar to a single intermediate product of their kind, no products outside of the 40-day time frame need to be kept in memory.

2.4. Reduction in IO operations for sub-Saharan Africa

The global SFD dataset is based on 3,060,962 Sentinel-2 L2A input products. The 2019 sub-Saharan Africa product has 882,642 inputs of the same type. In the 2019 production, using the original implementation, the pre step was executed 2,426,459 times, reading two products each. The number of input product reads is reduced by a factor of approximately 5.5, from reading two products per observation pair to reading each input only once. Further, all reads and writes of intermediate products from the pre, post, fuse and tile steps have been eliminated, including the 2.4 million pre step intermediates (sub-Saharan Africa), read once, and 327,741 post intermediates, read multiple times each in the fuse step.

3. CONCLUSIONS

In this paper, we discuss the challenges posed by global production of a high resolution burned area dataset as well as an optimized implementation scheme for the existing burned area algorithm. The new implementation minimizes input and output operations by eliminating all intermediate products that were created by the original implementation and drastically reducing the number of reads of the input products. At the cost of losing the intermediate information for later inspection, the new implementation considerably reduces the computational requirements.

REFERENCES

[1] E. Chuvieco *et al.*, “Historical background and current developments for mapping burned area from satellite Earth observation,” *Remote Sensing of Environment*, vol. 225, pp. 45–64, May 2019, doi: 10.1016/j.rse.2019.02.013.

[2] E. Roteta, A. Bastarrika, M. Padilla, T. Storm, and E. Chuvieco, “Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa,” *Remote Sensing of Environment*, vol. 222, pp. 1–17, Mar. 2019, doi: 10.1016/j.rse.2018.12.011.

[3] E. Chuvieco *et al.*, “Building a small fire database for Sub-Saharan Africa from Sentinel-2 high-resolution images,” *Science of The Total Environment*, vol. 845, p. 157139, Nov. 2022, doi: 10.1016/j.scitotenv.2022.157139.

[4] Apache Software Foundatation, *Apache Hadoop*. Accessed: May 02, 2025. [Online]. Available: https://hadoop.apache.org/

[5] G. Milcinski, J. Bojanowski, D. Clarijs, and J. de la Mar, “Copernicus Data Space Ecosystem-Platform That Enables Federated Earth Observation Services and Applications,” in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2024, pp. 875–877. Accessed: Apr. 11, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10642308/

CLOUD-NATIVE DATA SERVICES AT EUMETSAT: A PORTFOLIO APPROACH FOR SCALABLE USER ACCESS FOR A DIVERSE USER COMMUNITY

Daniel Lee, Michael Schick

EUMETSAT

ABSTRACT

EUMETSAT, Europe's organisation for the exploitation of meteorological satellites, has transformed its approach to data access from predominantly real-time satellite broadcasts to a versatile, cloud-native data access portfolio. This shift allows for the service portfolio to be managed according to cost and service level targets. The services are available using a combination of public clouds, private clouds, and on-premises environments, including the European Weather Cloud (EWC). Leveraging an extendable, service-based architecture has allowed EUMETSAT to rapidly adopt and support interfaces such as the WMO Information System 2 (WIS2), backed by the robustness of the cloud-native approach. Additionally, EUMETSAT actively enhances its data readiness for Artificial Intelligence (AI) and Machine Learning (ML) workflows guided by adherence to FAIR principles, especially in data accessibility and reusability.

Index Terms — cloud-native, FAIR, data access, services portfolio, AI/ML readiness

1. INTRODUCTION

Historically, EUMETSAT provided satellite data primarily through real-time dissemination systems, requiring users to independently manage local archives if they needed fast access to time series. Historical products could be retrieved from a tape archive service, but this was not an online service and response times were high. Increasingly diverse user needs and growing satellite data volumes brought this model to its limits, prompting the development of a modern, centralized, cloud-native data service portfolio. This portfolio is more flexible and is compatible with a multitude of workflows and use cases. This paper describes EUMETSAT's transformation, emphasizing its alignment with the BiDS 2025 theme of scalable digital infrastructures optimized for diverse applications. The new architecture has served EUMETSAT well as user needs have evolved and technology opportunities have presented themselves.

2. CENTRALIZED PROCESSING, DECENTRALIZED USER CONTROL

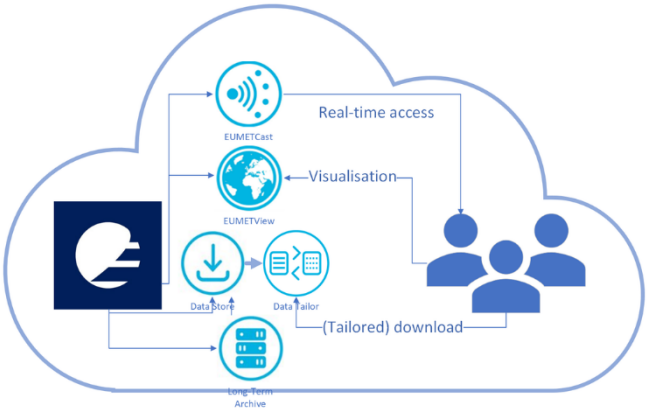


Fig. 1. EUMETSAT's data access services portfolio.

EUMETSAT's legacy is as a real-time data provider for nowcasting and numerical weather prediction. Additionally, its data products were used for monitoring the climate. This led to a bifurcation of data access early in the organisation's history; data was made available to users with the lowest timeliness possible in order to maximise its use in making weather-related decisions, often to preserve life and property, in real-time, and additionally, data was available from a long-term, tape-based archive.

doi:10.2760/2119408

This setup was cost-effective and fulfilled the needs of EUMETSAT's users very well. However, the addition of new missions, the continuous increase of data volumes, user demands, and the need for online data access changed the environment, exposing gaps in EUMETSAT's data access strategy. As more missions were added to EUMETSAT's portfolio and the use of satellite data products became more feasible for a wider range of users, our potential community expanded greatly. In order to allow more users to reap the benefits of our data products it was necessary to provide data to users in such a way that didn't require them to maintain antennae for real-time data reception, operate their own archives to access time series on short notice, or write their own software in order to visualize our products or use them in their own processing chains.

EUMETSAT's data access strategy was therefore reworked to leverage a portfolio that takes the burden associated with data management and processing off users' shoulders by centralizing a variety of data services:

- **EUMETCast**: A real-time, high-availability multicast service delivering time-critical satellite data globally, suitable for users who require immediate, operationally relevant data.
- **Data Store**: Provides online access to extensive historical and near-real-time datasets, enabling access and integration of data products into user workflows.
- **EUMETView**: Offers interactive visualization of satellite products directly using standardized web interfaces, removing the need for custom visualization software.
- **Data Tailor**: Allows users to transform datasets into custom formats tailored specifically to their unique processing toolchains and specific use case requirements.

All these data services can be used in the cloud of users' choice, be it a public cloud, on-premises, or a hybrid. Furthermore, EUMETSAT further enhances this portfolio by offering hosted processing in the form of managed, data-proximate processing environment, significantly reducing infrastructure demands previously placed on individual users. Several environments offer computing services tailored specifically to the needs of different user communities, sometimes as part of a federative offering with other organisations. This lets users maintain decentralized control by freely deploying and interacting with these services across various environments, including public clouds and local on-premises setups.

3. THE ADVANTAGES OF SERVICE ORIENTATION: WMO INFORMATION SYSTEM 2

EUMETSAT's adoption of a service-based architecture benefits not only our users directly, but also our internal development and maintenance teams, by providing flexibility. This flexibility is evidenced by the rapid implementation of EUMETSAT's interface to WMO Information System 2 (WIS2). The existing architecture required only minimal, targeted extensions to support WIS2, avoiding major rewrites or additional service deployments.

Previous architectures would likely have required significant development work or the development of a new bespoke service to fulfil the required interfaces. However, the portfolio approach allowed us to extend services surgically. In this case, all that was required was the addition of a pub/sub service to the Data Store announcing the availability of new data products. Users – including WIS2 infrastructure – can then access the products of their interest from the Data Store directly. No changes to our access management technology stack were necessary.

This extendibility significantly enhances EUMETSAT's ability to serve the WMO user community, increasing data visibility and accessibility while incurring low costs and minimal

doi:10.2760/2119408

operational disruption, and validates the flexibility and future-readiness of the chosen architecture.

4. MULTI-MISSION ARCHITECTURE AND SCALABLE ACCESS

Previously, each satellite mission required bespoke adjustments to infrastructure and data services, increasing both complexity and user-side costs. Data formats varied significantly, and users often developed unique tools to handle data visualization and processing. In contrast, the new portfolio allows users to tailor data directly to their workflows with minimal effort, significantly reducing barriers to adoption. These customisations can be executed on EUMETSAT computing assets using the Data Tailor Web Service, which interfaces with products stored in the Data Store, or after downloading from the Data Store using a local Data Tailor installation.

The Data Store itself is engineered to be data-agnostic, and its functionality is limited to data discovery and download. Under the previous architecture, data products accessed from the tape archive could be customized before download via a monolithic service. This rich functionality proved cumbersome because the provision of data collections to users was tightly coupled with any customisations that we wanted to offer on those data. Decoupling the Store and Tailor functionality has made the portfolio more flexible; new data collections can be added asynchronously with customisations, and the Tailor service can be upgraded independently of the storage and provisioning services. Further details on the Data Store can be found in "Retrieve, transform, deliver: Integrating preservation and performance in the EUMETSAT Data Lake" by Colapicchioni et al., which is also presented at BiDS 2025.

EUMETView also represents a significant step forward in the user experience. The service profits significantly from the use of standard data formats in EUMETSAT's newer generations of satellites. Previously, mission data was provided in bespoke formats; users had to understand the bespoke formats that were used and develop their own software if they wanted to visualize our data. EUMETView has made this situation much simpler; for heritage missions, users can simply use OGC services to consume our data in the form of a Web Map Service. Certain products are also available as Features or Coverages.

The use of netCDF as the native format for new-generation satellites not only simplifies the development of visualization and processing software, but also allows EUMETSAT to more rapidly develop and deploy visualisations that are available to users as a service on EUMETView. EUMETView also allows users to more easily share views of our data with colleagues.

All of these services are available both via a graphical user interface as well as via REST APIs, which makes it easy to interface them with downstream automated processing. This allows us to offer a high-quality user experience, whether our users are humans or machines.

Furthermore, as mentioned previously, real-time data provision used to be available only to users who operated satellite antennae to receive products downlinked from space. This service, EUMETCast, has been expanded to offer reception not only via satellite but also via terrestrial networks. On a per-user basis this allows the EUMETCast services (both Satellite and Terrestrial) to disseminate more than a petabyte of data to users per day, located worldwide.

5. FAIR PRINCIPLES AND FUTURE DIRECTIONS

EUMETSAT's commitment to FAIR (Findable, Accessible, Interoperable, and Reusable data) principles drives continuous enhancements in data discoverability, accessibility, and reusability. Expanding the availability of comprehensive time-series datasets improves their utility for climate research, training, and various operational purposes. Future initiatives include allowing data discovery using SpatioTemporal Asset Catalog

doi:10.2760/2119408

(STAC), which we anticipate will significantly enhancing catalog navigation and interoperability. Such developments ensure that EUMETSAT's satellite data holdings become increasingly user-friendly and widely usable.

6. AI/ML READINESS AND INNOVATIVE DATA PROVISION

EUMETSAT is also exploring innovative approaches to data provision such as data cubes and formats such as zarr to facilitate AI/ML workflows. This should allow data to not only be discovered and utilized, but also allow our data to be used at massive scale. Concurrently, first experiments with generative AI chatbots are being trialed, aimed at simplifying interactions with our data services by enhancing data discoverability, streamlining API usage, and making it easier to navigate our comprehensive documentation. Although these technologies remain exploratory, they demonstrate our forward-looking perspective to preparing our data services for our users' evolving requirements and the needs of our diverse and growing user community.

7. OUTCOMES

The data services have produced an excellent performance track record in operations. On a daily basis, more than 500 TB are delivered to in real-time to users on in 150 countries using EUMETCast. EUMETView serves up more than 2 million images per day, and over 400 TB per month are delivered to users by the Data Store and Data Tailor. These services are reachable in the cloud of users' choice, using APIs, CLIs, and GUIs. And the technology stack has proven reliable and adaptable as new needs arise.

8. CONCLUSIONS

The portfolio approach adopted by EUMETSAT is a significant step forward in the way users can access and exploit our data. The new portfolio offers a richer set of capabilities, improving the user experience while saving costs on all sides simultaneously.

Because EUMETSAT remains dedicated to serving national meteorological and hydrological services, an important consideration in the portfolio's design was not only ease of use, but also flexibility. For this reason, we have worked hard to achieve a situation where our services can be used with equal ease in on-premises computing environments, as well as public and hybrid clouds. This includes hosted processing cloud offerings that EUMETSAT offers for specific user communities. Thus the portfolio allows us to harness the advantages of centralized processing and storage, whilst allowing users to be flexible in where they work and consume these services and data products.

Furthermore, by building a cloud-native architecture, the services are highly scalable and interoperable. This lets us respond rapidly to technological advancements and new user needs. An example of this is the speed and ease with which it was possible to extend the data services to fulfil the needs of WIS2. It is also demonstrated in our ability to experiment with emerging technologies and use cases. Currently known examples are catalogue services such as STAC, as well as new data interfaces such as data cubes and zarr. Our focus is not only on human, but also on machine consumption, which means that our data needs to be available at scale, and these efforts have already proved worthwhile through the ease with which experimental generative AI applications can work with our data services.

EUMETSAT remains committed to providing life- and property-saving data services to our users, and the use of a cloud-native architecture has helped us pursue this goal with flexibility, low costs, and fast learning cycles. Similar organisations might do well to use this template also, by adopting cloud technologies and architecting their solutions to be focused on limited functionality and achieve rich functionality through a portfolio approach.

doi:10.2760/2119408

RESHAPING THE EARTH-OBSERVATION VALUE CHAIN THROUGH AI-EXPRESS POWERED LOW LATENCY SERVICES FOR SECURITY AND CRISIS RESPONSE

Vito Fortunato
Planetek Italia S.r.l.,
Via Massaua 12, I-70132 Bari, Italy,
fortunato@planetek.it

Leonardo Amoruso
Planetek Italia S.r.l.,
Via Massaua 12, I-70132 Bari, Italy,
amoruso@planetek.it

Cristoforo Abbattista
Planetek Italia S.r.l.,
Via Massaua 12, I-70132 Bari, Italy,
abbattista@planetek.it

Marco Mucci Beltrami
Planetek Italia S.r.l.,
Via Massaua 12, I-70132 Bari, Italy,
muccibeltrami@planetek.it

Alessandro Varriale
Planetek Italia S.r.l.,
Via Massaua 12, I-70132 Bari, Italy,
varriale@planetek.it

Stefano Antonetti
D-Orbit spa,
Viale Risorgimento, 57, 22073 Fino Mornasco CO, Italy,
stefano.antonetti@dorbit.space

Lorenzo Feruglio
AIKO srl,
Corso Castelfidardo, 30/A, 10129 Torino TO, Italy
lorenzo.feruglio@aikospace.com

Gianluca Furano
ESA
ESTEC Keplerlaan 1, 2201 AZ Noordwijk, Holland
gianluca.Furano@esa.int

ABSTRACT

The Earth Observation (EO) sector increasingly demands actionable insights rather than raw data. AI-eXpress (AIX), developed within ESA's InCubed initiative by Planetek Italia, D-Orbit, and AIKO, revolutionizes the satellite-as-a-service paradigm by integrating deep-learning models directly onboard satellites. AIX introduces an open "App Store" ecosystem, facilitating rapid in-orbit demonstrations (IOD) and validations (IOV) of advanced AI applications, such as real-time object detection and anomaly identification. By efficiently managing onboard resources, irrelevant data is automatically discarded, optimizing memory usage and bandwidth.

AIX's low-latency services are crucial for security and disaster responsiveness. Actionable intelligence can trigger autonomous onboard decisions or be transmitted back to Earth as real-time notifications and alerts, ensuring swift responses to critical situations. This capability accelerates the transition from technology "makers" to "enablers."

The AIX mission envelope exemplifies this potential. **AIX-1p ("The Pathfinder")**, launched on 14th January 2025 onboard SpaceX F9 Transporter-12 (T-12), carries a COTS RGB payload (~15 m GSD) assembled in-house by D-Orbit, enabling early demonstrations of onboard AI processing. **AIX-1 ("The Innovator")**, launched on 23rd June 2025 (T-14), follows with an upgraded architecture and expanded service offering. **AIX-1+ ("The Visionary")**, scheduled for the October 2025 Transporter, upgrades the imaging capacity with a dual-head imaging payload: a nadir-looking multispectral camera with ~10 m GSD across Sentinel-2 equivalent bands (from B2 to B8a) plus a panchromatic channel, complemented by a forward-looking wide-swath RGB camera (~150 m GSD). This combination supports both fine-scale tasking and broad-area situational awareness, demonstrating powerful onboard computing, blockchain-enabled services, and near real-time responsiveness.

Index Terms- satellite-as-a-service, ai-augmented onboard processing, low-latency insights delivery, Earth Observation, autonomy, software-defined missions

1. INTRODUCTION

The Earth Observation (EO) sector increasingly demands actionable insights rather than mere raw data. Traditional EO missions face challenges such as limited onboard resources (power and bandwidth), inefficient data downlink (irrelevant or low-quality imagery), and delayed decision-making due to ground-based processing ([1]-[2]-[3]). AI-eXpress (AIX), developed by Planetek Italia, D-Orbit, and AIKO within ESA's InCubed initiative, introduces a paradigm shift towards Satellite-as-a-service. Leveraging powerful onboard computing, artificial intelligence (AI), and blockchain technologies, AIX enables real-time onboard processing, discarding irrelevant data, thus optimizing satellite memory and bandwidth usage. This workflow is referred to as "SpaceStream". In this approach, data is processed where it is most effective, seamlessly bridging space and ground resources. In this context, AIX defines

doi:10.2760/2119408

"Spacedge" services, a novel approach for interfacing EO users with both the satellite and ground segments.

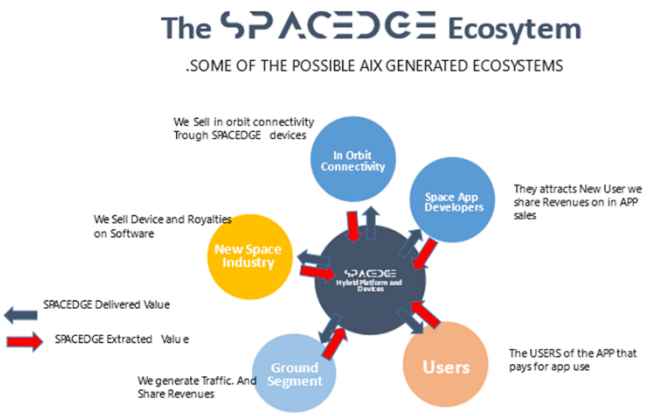


FIGURE 1. ***THE PLANETEK'S SPACEDGE ECOSYSTEM, POWERED BY AIX***
CREDITS: HTTPS://WWW.AIEXPRESS.EU [4]

AIX provides an open 'App Store' ecosystem where users can access pre-built applications (e.g., wildfire detection, vegetation monitoring, maritime anomaly detection) or deploy custom AI algorithms for rapid in-orbit demonstration (IOD) and validation (IOV).

In January 2025, the first AIX satellite (AIX-1p, "The Pathfinder") was successfully launched, demonstrating on-board processing and paving the way for subsequent missions within 2025.

The second successfully launched satellite, AIX-1 ("The Innovator"), incorporates enhanced scalability and near real-time data services, while the final mission, AIX-1+ ("The Visionary"), is scheduled for October 2025, featuring dual-head multi-spectral imaging capabilities and seamless integration with the AIX App Store. All three missions (supported by ESA's InCubed program) are designed to lower barriers for SMEs, accelerate EO innovation, and confirm AIX as a fully flexible, information-driven satellite service platform.

2. BUILDING BLOCKS

The AIX/SpacEdge framework, as shown in Figure 1 consists in several core principles: onboard intelligence, modular software design, and a user-focused ecosystem. Below is a brief overview of how these elements come together:

1. **Space Segment**: Leverages D-Orbit's ION Carrier [5], which comprises two main components: a modular platform and hosted payloads. The ION carrier provides standard spacecraft functions (attitude control, command and telemetry processing, maneuvers) and hosts custom payloads. A dedicated on-board software framework supports the execution of user-defined algorithms, in particular AI-based applications, enabling real-time data filtering, compression and information extraction using on-board accelerators (GPUs, VPUs).

2. **Ground/User Segment**: Provides an app store-like web interface where users can configure satellite resources and tasks, deploy their custom AI applications, and manage data acquisition and processing workflows on-demand. Users interact by selecting sensors, configuring acquisitions, algorithms and pre-processing steps, and then directly retrieve processed results.

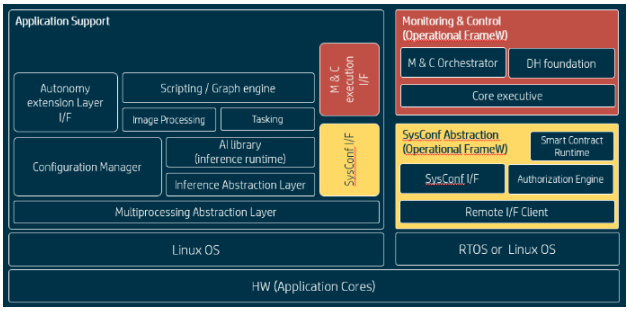


FIGURE 2. AIX FRAMEWORK: HIGH-LEVEL ARCHITECTURAL SCHEME

3. **Software Framework**: As shown in Figure 2, the Spacedge-enabled system comprises layered architecture:
   - *System Controller Abstraction Layer*: Manages secure and abstracted communication between system controllers and applications through remote interfaces, handling authorization and remote procedure calls (RPC).
   - *Monitoring & Control Layer*: Provides basic housekeeping, network, storage and fault detection/recovery services to ensure observability and control reliability.
   - *Application Support Layer*: Hosts runtime services for deterministic image processing and AI tasks, facilitating the deployment of AI algorithms through acceleration devices and parallel processing.
   - *Data Processing Layer:* Utilizes Planetek's proprietary SDK (PkSpacekit), which offers a comprehensive ecosystem for data handling, image processing, multi-platform support, and hardware acceleration through parallelization (e.g., OpenCL, GPUs), augmented by AIKO's services for applications and autonomous operations ([9]);
   - *AI Augmentation Layer*: Enables integration and execution of neural network models using widely-adopted formats (ONNX [6], Apache TVM [7], CNNs), allowing efficient inference and quantization for onboard processing. Figure 3 depicts such a mechanism, sketching the basic behavioral model supported, and extensible, via the framework.

doi:10.2760/2119408

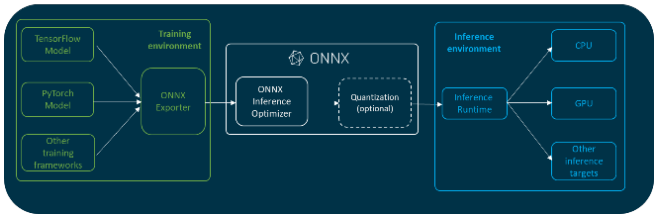


FIGURE 3. AIX SUPPORT FOR NEURAL NETWORKS: THE INFERENCE ENGINE PERSPECTIVE

4. **Blueprint Editor:** As highlighted in Error! Reference source not found., this tool provides a graphical interface for designing custom data-processing and AI workflows. It automatically generates the underlying code, allowing non-expert users to rapidly compose and deploy onboard applications without extensive programming effort.

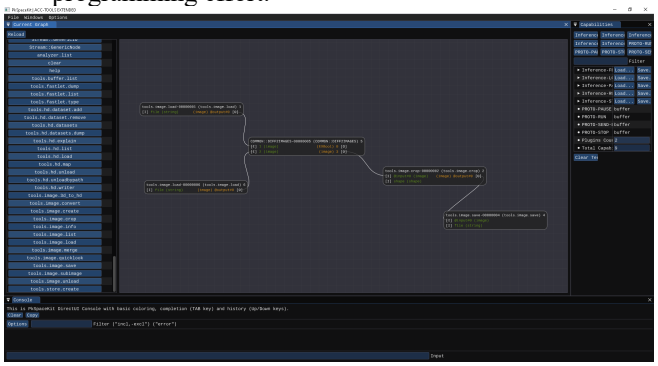


FIGURE 4. PROTOTYPAL DEMO IMPLEMENTED WITH THE IMGUI-NODE-EDITOR OPEN-SOURCE PROJECT [8]

5. **Imaging systems**: The first two AIX missions (AIX-1p and AIX-1) are equipped with a commercial-off-the-shelf (COTS) RGB camera, assembled in-house by D-Orbit, delivering a ground sampling distance (GSD) of ~15 m. These payloads provide rapid prototyping capability for in-orbit AI demonstration and service validation. The third mission, AIX-1+ ('The Visionary'), introduces a dual-head multispectral payload: a nadir-looking imager with ~10 m GSD across Sentinel-2 equivalent bands (B2–B8a) plus a panchromatic channel, complemented by a forward-looking wide-swath camera (~150 m GSD). This dual-head mosaic snapshot camera system (supplied by TSD Space) supports both high-resolution tasking and broad-context monitoring, expanding the range of applications from local asset monitoring to regional-scale anomaly detection.

6. **Operational Workflow**: Users configure services via a web portal, defining sensors, acquisitions time, processing chains, and actionable outputs. Configured tasks are scheduled, uploaded via ground infrastructure, executed onboard, and results returned to users. Resources are automatically released after execution, enabling continuous and optimized resource utilization. Figure 5 shows the entire workflow from the user request to the insights' delivery.

3. CONOPS

By processing data in orbit, AIX fundamentally changes the speed and efficiency of Earth Observation. Instead of collecting raw images and sending them to ground for analysis, the satellite itself evaluates and filters out what's irrelevant - such as cloudy scenes or areas of no interest - and then quickly downlinks only the important results. This eliminates wasted bandwidth, reduces delays and allows faster response to events on the ground. For instance, a wildfire detection app can highlight hot spots and generate an alert almost instantly, rather than waiting for a full image to be downlinked and reviewed later.

The "app store" approach gives researchers and companies a straightforward way to test and deploy their own algorithms in space. It eliminates the long lead times typically required to integrate new software into a mission, making it possible to run quick experiments or deploy updated versions of existing applications whenever needed. This creates an ecosystem where new ideas can reach orbit faster and promising algorithms can be demonstrated under real-world conditions.

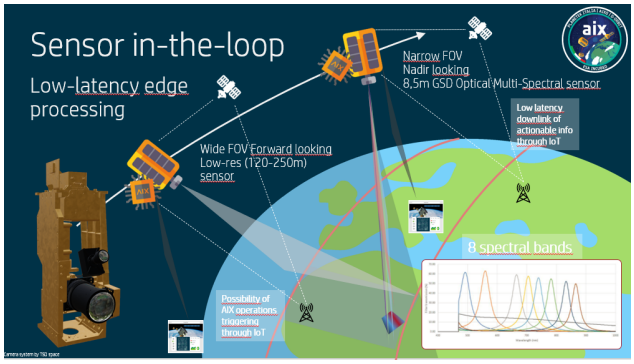


FIGURE 5. AIX OPERATIONAL SERVICES FLOW

AI-eXpress also paves the way for new business models based on on-demand satellite services. Customers pay only for what they need – like a specific type of data product, a short test campaign, or a continuous monitoring service. This makes satellite resources more accessible to smaller companies and research groups, who can now affordably access advanced in-orbit computing capabilities.
In the long term, such flexibility and shared resources could accelerate innovation across the EO sector, shifting the focus from raw data to immediate, actionable insights. By lowering the entry barriers, AIX democratizes access to space technology, enabling a broader range of stakeholders

doi:10.2760/2119408

to participate in and benefit from EO advancements. This democratization fosters a more competitive and dynamic market, where even small-scale enterprises and academic institutions can contribute to and drive technological progress.

Moreover, the on-demand model supports a more sustainable approach to satellite operations. By optimizing the use of satellite resources and reducing unnecessary data transmission, AIX helps minimize the environmental impact of satellite missions. This efficiency not only conserves bandwidth and energy but also aligns with global efforts to promote sustainable practices in space exploration and technology.

AIX 's innovative approach also encourages the development of specialized applications tailored to specific industry needs. For instance, agriculture companies can leverage real-time crop monitoring and predictive analytics to enhance yield and resource management. Disaster response teams can benefit from rapid damage assessment and situational awareness, enabling quicker and more effective interventions. Urban planners and environmental agencies can utilize up-to-date land use and environmental data to make informed decisions and policies.

Furthermore, the integration of blockchain technology within AIX ensures data integrity and security, providing an additional layer of trust and transparency for users. This feature is particularly valuable for applications requiring high levels of data accuracy and reliability, such as financial services, insurance, and regulatory compliance.

Overall, AIX not only transforms the EO sector by providing low-latency, actionable insights but also fosters a collaborative and innovative ecosystem. By making advanced satellite services more accessible and sustainable, AIX is poised to drive significant advancements in various industries, ultimately contributing to a smarter and more connected world.

4. RELEVANCE

AIX is a versatile platform that benefits various industries by providing advanced AI capabilities and low-latency information processing and delivery directly onboard satellites. Few some examples of industries that benefit from such a platform/service, with a special focus on defence, security and sustainability is:

Defence and Security: AIX significantly enhances defence and security operations by providing real-time, actionable intelligence. The platform's ability to process data onboard and deliver low-latency alerts is crucial for timely decision-making in critical situations. For instance, AIX can be used for:

- **Real-time object detection:** Identifying potential threats such as unauthorized vessels in restricted areas.
- **Anomaly detection:** Monitoring and detecting unusual activities or changes in strategic locations, which can be vital for national security.
- **Autonomous decision-making:** Enabling autonomous decision-making based on predefined criteria, such as initiating countermeasures or alerting ground control about potential threats.

Civil protection and emergency response: AIX supports civil protection and emergency response by providing rapid assessments and situational awareness during disasters. This includes:

- **Fire detection:** Identifying and monitoring wildfires in real-time to assist firefighting efforts.
- **Oil spill detection:** Detecting and tracking oil spills to mitigate environmental damage.
- **Disaster management:** Providing real-time data on natural disasters like floods and earthquakes to coordinate rescue and relief operations

Environmental monitoring: AIX aids in environmental monitoring by providing detailed and timely data on various environmental parameters. This includes:

- **Forest health management:** Monitoring forest conditions to detect diseases, pests, and illegal logging activities.
- **Urban planning:** Providing data on land use and environmental changes to support sustainable urban development.

REFERENCES

[1] Prospects for the Small Satellite Market, Euroconsult 2018

[2] Nano/Microsatellite Market Forecast, SpaceWorks 2018

[3] Map of main space private actors organized per technology, Seraphim Capital 2019

[4] AI-express, https://www.aiexpress.eu

[5] D-Orbit ION service, https://www.dorbit.space/inorbitnow

[6] ONNX, Open Neural Network Exchange, https://onnx.ai/

[7] Apache TVM, https://tvm.apache.org/

[8] ImGui node editor, https://github.com/thedmd/imgui-node-editor

[9] AIKO services, https://aikospace.com/products

ACKNOWLEDGMENT

The authors, on behalf of the entire AI-eXpress team, would like to thank the European Space Agency (ESA) for its support in co-founding under the InCubed program the design and implementation of the AI-eXpress initiative.

doi:10.2760/2119408

BRIDGING BIG DATA FROM SPACE DOWN TO EARTH: EXPLORING EO INTEGRATION IN EU LOCAL AUTHORITIES

Elisa Filippi (1), *Antonello Aiello* (2)

(1) elisa.filippi@uniroma1it Sapienza University, Rome, Italy; (2) Planetek, Bari, Italy

ABSTRACT

Earth Observation (EO) satellite data are increasingly recognized for their role in addressing global challenges, including climate change and urban resilience. Despite their potential to enhance monitoring and planning, adoption by Local and Regional Authorities (LRAs) remains limited and underexploited. This paper investigates the perceived drivers and obstacles shaping EO adoption among LRAs, using an integrated framework based on the Technology Affordances and Constraints Theory (TACT) and the Technology-Organization-Environment (TOE) model. A survey was conducted among LRAs engaged in EU-funded EO-related projects. Results from 17 EU countries show a moderate adoption rate (56%), mainly for specific projects. EO data are valued for decision-making support and cost-effectiveness. Still, barriers persist, including technological limitations, lack of skilled personnel, and weak regulatory incentives. Participation in projects and internal organisational needs emerge as key drivers. Despite the limited sample size, these findings offer insights into the end-users' perspectives and enabling conditions for expanding EO data use within LRAs.

Index Terms— Copernicus'end-users' perspectives; Urban Climate adaptation; Space economy.

1. INTRODUCTION

Earth Observation (EO) satellite data are increasingly recognised as strategic assets for monitoring environmental change, contributing to the Sustainable Development Goals (UNOOSA, 2019)[1]). According to the European Commission's Knowledge Centre on Earth Observation (KCEO), EO products and services can contribute to over 28 EU policy areas, including climate adaptation, agriculture, and urban development. The value of EO data is particularly relevant at the local and regional levels, where climate-related risks are most concentrated, and the capacity for monitoring, planning, and resilience-building needs to be supported. EO and other geospatial technologies offer powerful tools for enhancing local decision-making in domains such as climate adaptation, land use, and urban heat mitigation (Dowell M. et al., JRC, 2025) [2]. With nearly 70% of Europe's population residing in urban areas, mainstreaming the use of EO by Local and Regional Authorities (LRAs) is critical for delivering space-enabled policy impact. Nevertheless, despite the growing availability of EO data and platforms, adoption by LRAs across Europe remains limited (EARSC, 2023).[3] In addition, the perspectives of local users—who must translate EO data into policy, planning, or service delivery—are still underexplored in both academic literature and institutional strategies (OECD, 2023 [4] [2]). Building on earlier author's studies in Italy (Filippi et al. 2025) [5], this paper extends the analysis on a European scale, applying an integrated theoretical framework, Technology–Organization–Environment (TOE) [6] and Technology Affordances and Constraints Theory (TACT) [7], to assess factors influencing EO adoption and use. The paper addresses the following research questions: **RQ1:** What are the main technological, organisational, and environmental factors influencing EO adoption by LRAs? **RQ2:** What affordances and constraints are perceived in the use of EO satellite data and services?

2. METHODOLOGY

This study adopts a qualitative, theory-informed approach based on multiple case studies (Yin, 2017) [8] to explore the adoption and use of Earth Observation (EO) satellite data by Local and Regional Authorities (LRAs) across Europe. This design enables both comparative analysis across organisational and territorial contexts and theoretical generalization from prior research conducted in Italy. Given the specialized nature of the target population and limited sample size, the study follows the logic of analytical generalization, aiming to identify transferable patterns and theory-informed insights. **Survey Design and Theoretical Framework.** To address the research questions, a structured survey was designed and administered to a purposive sample of European LRAs. The survey builds on a framework developed in two prior author's studies. It is grounded in two complementary theories: Technology Affordances and Constraints Theory (TACT) (Majchrzak & Markus, 2013) [7], used to assess the perceived benefits and limitations of EO technologies. Technology–Organization–Environment (TOE) (Tornatzky & Fleischer, 1990) [6] was used to identify organisational and environmental conditions influencing adoption. In addition, three more sections have been introduced in the survey: organisation's background, EO usage patterns, Barriers/enablers and peer advice. The 26-question survey employed Likert scales, single/multiple-choice, ranked responses, and one optional open-ended item.

doi:10.2760/2119408

Sampling Strategy and Data Collection. The target population included subnational public authorities (cities, regions, agencies) (potentially) interested in EO satellite data. A purpose-driven sampling strategy was adopted, identifying LRAs engaged in EU-funded projects (i.e. through Horizon Europe, Life+, Interreg, EUI, URBACT) related to EO, climate adaptation/mitigation, or smart cities. A total of 30 LRAs were identified, and the survey was distributed via LimeSurvey (v6.4) in English. Responses from those not using EO were routed to a shorter version (approx. 1 minute); full participation took 10–12 minutes. **Data Analysis** Approach. Given the limited sized of the sample (N=25), the analysis focused on descriptive statistics (frequencies, means, SD) for EO use, affordances, barriers.

3. RESULTS

3.1 Sample overview. Thirty-three individuals representing 55 individuals from 50 LRAs across EU were contacted. A total of 50 responses to the survey have been received, of which 25 are fully complete (completion rate: 50%). Notably, Italian LRAs were deliberately excluded to avoid overlap with the author's prior national-level study and to ensure the European focus of the present analysis. Respondents came from 17 EU countries, with a balanced mix of cities, metropolitan cities, and regional/municipal agencies, while regions are underrepresented. Most were from medium-sized organisations, with over half employing up to 3,000 people.

3.2 EO data usage and adoption patterns. Out of the 25 respondents, N=1 reported not knowing whether their organisation uses Earth Observation (EO) satellite data, and N=1 reported they are not using this technology. N=7 respondents indicated that EO data is used only indirectly, primarily via tools such as Google Earth, without structured access to EO datasets or services. The remaining eight, N=14 respondents, reported direct use of EO satellite data. This group of direct users includes representatives from all organizational types in the sample (cities, metropolitan cities, and regional/municipal agencies), suggesting no clear association between the type of organization and the likelihood of direct EO adoption. As shown in Table 1, the level of EO data adoption is mostly (86%) concentrated at the lowest level (Level 1-2), with residual respondents (14%) reporting a more structured use (Level 3-4).

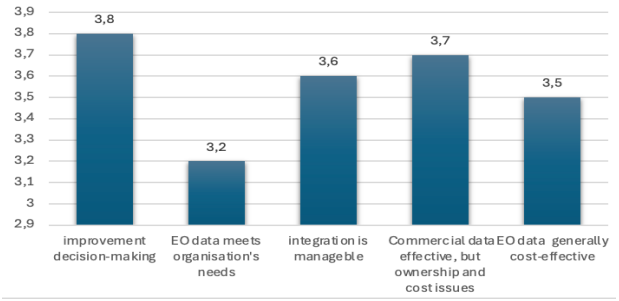
Table 1. Degree of EO satellite data adoption and use among surveyed LRAs (N=14)

Level of EO Usage	Description	Respondents (N)
Level 1	Tested experimentally	5
Level 2	Used occasionally for specific projects or reports	7
Level 3	Regular use in workflows	1
Level 4	Institutionalised use	1

In terms of thematic domains of application, EO satellite data was most frequently used for addressing urban heat islands (N=8), followed by urban planning and urban greening (each N=6). And land use and land cover (N=5). Less frequent domains include air quality (N=3), emergency and Ground motion (N=2) and single mentions for biodiversity, hydro-meteorological monitoring. The primary source of EO satellite data used is the **Copernicus programme**, cited by 71% of users, 21% reported to use primarily national satellite constellations and only 1 (a Dutch org.) a commercial satellite data. In terms of purposes, **monitoring** (N=8) and **planning activities** (N=6) are almost equally represented among the respondents, while no respondents reported using EO data for law-making or forecasting systems.

3.3 Technology Affordances and Constraints. This section was completed only by the 14 respondents who reported direct use of Earth Observation (EO) satellite data within their organisations. Participants were asked to indicate their level of agreement with the following five statements using a 5-point Likert scale.

Figure n.1 Respondents' agreement (5-point Likert scale) on five statements reflecting EO satellite data affordances and constraints. Values represent average values across eight EU LRA respondents directly using EO technologies.

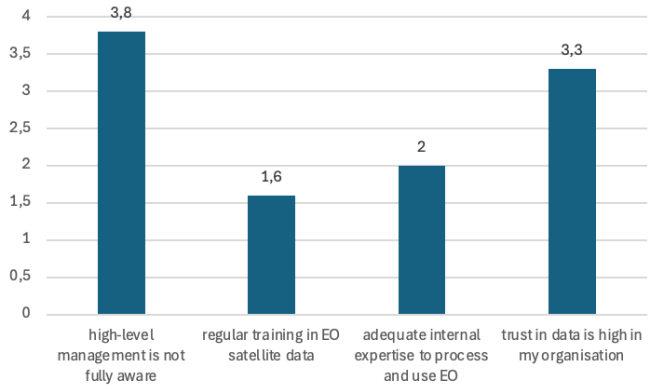


As can be seen in **Figure n.1,** respondents expressed **moderate to moderately high levels of agreement**, suggesting a generally favourable perception of EO technologies, albeit with some reservations. EO satellite data was seen as moderately useful to improve decision-making (mean = 3.8), and they are recongnized as cost-effectives compared to other sources of data (3.5). Also, their integration into an existing platform is perceived as quite manageable (mean = 3,.6). However, concerns about cost and ownership of commercial data are common issues. The resolution adequacy of the EO satellite in meeting the organisations' **operational and knowledge needs** scored lowest (mean= 3.2), indicating perceived limitations. This finding is consistent with the high frequency of "EO satellite data does not fully meet our operational and knowledge needs" being cited as a major barrier in the subsequent section.

doi:10.2760/2119408

3.4 Influential factors. All the respondents, including those using EO only indirectly (total N=21), except for those not using EO satellite data or not aware of it (N=2), participated in the survey.

Figure n.2 **Respondents' agreement (5-point Likert scale) on four statements reflecting organisational factors.** Values represent average scores across respondents N=21



As shown in **Figure 2,** the responses indicate a widely perceived lack of awareness among high-level management regarding EO technologies (mean = 3.8), highlighting a potential leadership-level barrier to adoption. More critically, there is strong evidence of a weak organisational capacity, particularly in terms of **training provision (**mean = 1.6) and **internal technical expertise** (mean = 2). These results suggest that most organisations lack the skills and knowledge required to autonomously process and exploit EO data. On a more positive note, respondents reported **moderate agreement** regarding the existence of a **general data culture and trust in data** (mean = 3.3), which could serve as a partial enabling factor.

3.5 Environmental influential factors. Regarding **regulatory pressures**, among the 21 respondents, only one reported a clear legal or regulatory mandate/encouragement for the adoption of EO satellite data. At the same time, N=13 answered "No" and N=7 selected "Only indirectly". These findings indicate that external institutional pressure is currently minimal. However, the presence of indirect references to support EO in broader policy frameworks suggests a latent or "hidden" potential for regulatory incentives to play a more prominent role in the future.

Regarding **inter-organisational cooperation** at the regional scale, such as collaboration with other cities, provinces, or regions, 10 respondents confirmed active cooperation, while 4 indicated they are not experimenting it.

A particularly strong consensus emerged on the value of **structured exchange and learning opportunities** among public authorities. 21 out of 25 respondents agreed that such opportunities would support EO adoption in their organisation, with only 4 respondents expressing disagreement. This result underscores the importance of peer-to-peer learning, knowledge transfer, and capacity-building networks as enabling conditions for wider and more effective EO integration in public sector contexts. Also, respondents were asked to evaluate if and to what extent **the access to EU funds has been key to start EO data adoption**. The mean value =3.6 indicates a moderately high value of EU funding's perception as an opportunity to support the adoption of EO satellite data.

3.6 Perceived Barriers, Drivers perceived, Peer to Peer advice. All the survey respondents, including those not using EO satellite data (N=25), were asked to identify the main barriers to adopting and using EO data in their organization. They have been invited to choose the three most relevant barriers from 6 different options.

As shown in **Table 2**, respondents highlighted skill shortages (N = 17), budget costraints (N=12) and EO satellite data suitability (N =11), as top barriers.

Table 2. Main perceived barriers to EO data adoption among surveyed LRAs (N = 25)

Barrier	**(N)**
Lack of properly trained staff	17
Limited budget	12
EO data does not fully meet operational/knowledge needs	11
Complexity of the procurement	6
Complexity of the technology	5
Lack of top-management/ legal inputs	5

3.7 Key drivers perceived. Respondents who declared to use directly EO satellite data (N=14) were asked to identify, based on their experience, the key factors enabling EO satellite data adoption, by choosing three among 9 pre-defined options.

Table 3 Main perceived drivers fostering the adoption of EO satellite data among LRAs surveyed using EO data. (N=14)

Driver	**(N)**
Participation in a specific project or initiative	10
Organisation's internal need	9
Collaboration with a university or research institution	6
My own individual awareness or motivation	5
Effective collaboration with an industrial service provider	5
Availability of funding from an EU programme	4
Engagement with ESA	1
Strong commitment from leadership	0
Legal or regulatory requirements	0

As shown in **Table 3**, the most frequent factor (N=10) is the **participation in a specific project or initiative**, followed by

doi:10.2760/2119408

the **organisation's need** for the improvement of monitoring and planning (N=9).

Among secondary factors, **collaboration with universities (N=5) or individual awareness or motivation and collaboration with a company** (N = 5 each) were identified as moderately influential.

3.5 Peer-to-peer recommendations. Drawing on their experience, the eight respondents who directly use EO satellite data were asked what advice they would give to a public authority planning to initiate EO data use by selecting the three most relevant recommendations from a pre-defined list.

Table n.4 Peer-to-peer recommendations among LRAs surveyed (N=14)

Recommendations	(N)
Establish partnerships with research centers/univ.	10
Invest in internal training for staff	7
Collaborate with an industrial service provider to develop tailored solutions	6
Raise awareness and commitment from political/institutional leadership	5
Build direct cooperation with ESA or the National Copernicus Users Forum	4
Allocate funding within ordinary budget	4

As shown in **Table 4,** the most recommended peer-to-peer advice is building academic partnerships (N=10) and (N = 7), followed by service provider collaboration, increased leadership awareness, and engagement with ESA (N=4).

4. CONCLUSIONS

Although preliminary, the results reveal consistent patterns. Despite being distributed among LRAs already involved in potentially EO-relevant projects, the small but institutionally diverse sample (N = 14), comprising cities, metropolitan cities, and regional agencies, highlights that only 57% of respondents directly use EO satellite data, and even then, mainly in experimental or project-based forms rather than through institutionalized adoption.

Organizational barriers are significant. A shortage of skilled personnel, limited training opportunities, and low leadership awareness consistently hinder adoption. Externally, regulatory drivers are weak, with little formal pressure to adopt EO technologies. However, peer-to-peer exchanges are seen as highly valuable, and EU funding plays a moderately supportive role. The potential for regional cooperation remains partially underexploited.

Perceived affordances of EO data are moderately positive. Users value EO's contribution to decision-making and cost-effectiveness despite ongoing concerns about data resolution and cost and ownership issues for commercial data. Not unexpectedly, among key barriers, we find inadequate internal expertise, limited data adequacy for operational needs, and budget constraints, aligning with findings from previous studies (Filippi E. et al., 2025 [8]; JRC 2025 [2]).

Adoption seems to be driven by a combination of internal organisational needs and external opportunities, notably participation in EU-funded projects. This suggests a hybrid model, where bottom-up organisational needs appear to trigger interest in EO technologies, and actual adoption is often activated and supported by externally driven opportunities.

Peer-to-peer recommendations confirm these insights: experienced users emphasise investments in human capital, strategic collaboration with research, industry partners and ESA, and a stronger leadership engagement.

The study presents relevant limitations due to the size of the sample and its composition derived from a purposing sample strategy. However, the survey, grounded in two theoretical models and already tested, could serve as a reference for future research. New investigation is needed with a larger sample to allow for cross-countries and cross-organisation in depth analysis and correlation. Still, the findings highlight that EO satellite data have substantial potential to improve local policy and service delivery but remain only partially underutilized. A comprehensive approach, combining technological upgrades, organizational capacity-building, and targeted policy interventions, is needed to foster broader and more systematic EO integration.

REFERENCES

[1] United Nations Office for Outer Space Affairs. *European Global Navigation Satellite System and Copernicus: Supporting the Sustainable Development Goals—Building Blocks Towards the 2030 Agenda*; United Nations: New York, NY, USA, 2019

[2] EC, Joint Research Centre, Dowell, M., Bernard, S., Kilsedar, C., Gianinetto, M., Speyer, O., Kuffer, M., Grecchi, R., Gliottone, I., Melchiorri, M., *Earth Observation in support of EU policies for urban climate adaptation*, Publications Office of the European Union, Luxembourg, 2025.

[3] EARSC. EARSC Industry Survey 2024. Available online: https://earsc.org/2024/11/18/earsc-industry-survey-2024/ Last access: July 2025

[4] OECD, *The Space Economy in Figures: Responding to Global Challenges*, OECD Publishing, Paris, 2023 https://doi.org/10.1787/fa5494aa-en.

[5] Filippi, E.; Aiello, A. *Barriers, Benefits, and Influential Factors of Adopting Earth Observation Satellite Data at Local and Regional Levels: The Case of the Italian LRAs*. Sustainability **2025**, *17*, 145. https://doi.org/10.3390/su17010145

[6] Tornatzky, L. G., & Fleischer, M., The processes of technological innovation. Lexington, 1, 1990

[7] A. Majchrzak, M.L. Markus, "Technology Affordances and Constraints Theory (of MIS)," in E. Kessler (Ed.), *Encyclopedia of Management Theory*, SAGE Publications, Thousand Oaks, CA, USA, 2013, pp. 832–836.

[8] Yin, R.K. *Case Study Research and Applications: Design and Methods*, 6th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2017

doi:10.2760/2119408

BURNED AREA DETECTION IN GREECE: LEVERAGING DEEP LEARNING AND SENTINEL-2 DATA

Ioannis Kotaridis

Aristotle University of Thessaloniki, Faculty of Engineering, 54124 Thessaloniki, Greece.

ABSTRACT

Precise and regularly updated maps of burned area extents are essential for effective wildfire management. Ground surveys for this purpose are costly and time-intensive, making satellite remote sensing a fundamental and efficient alternative. This study presents a deep learning-based framework for burned area detection in Greece, leveraging Sentinel-2 satellite data. A U-Net model was trained using data from the past two years of fire events cataloged by the Copernicus Emergency Management Service (CEMS). The methodology included preprocessing Sentinel-2 imagery, generating training samples, training the model, and implementing pixel-wise classification with U-Net. To assess performance, the U-Net model was compared to baseline models such as DeepLabV3 and MMSegmentation. The U-Net outperformed the baselines, achieving an accuracy of up to 96% and high F1-scores, while demonstrating computational efficiency. Validation was conducted on two wildfire events (EMSR747 and EMSR767), with overlap analyses showing high agreement (94.5% and 95.7%, respectively) between the U-Net predictions and EMS reference products, indicating that, the U-Net model reliably delineated burned areas across diverse landscapes. This study highlights the potential of deep learning in advancing wildfire monitoring and management, offering scalable, accurate, and efficient solutions for post-fire assessment.

Index Terms— Wildfire, deep learning, classification, remote sensing, U-Net

1. INTRODUCTION

Wildfires are among the most devastating natural hazards, resulting in significant ecological, economic, and social impacts globally. Greece, a country characterized by its Mediterranean climate and diverse ecosystems, has faced increasingly severe fire events over recent years due to a combination of climatic changes and human activities. These events not only cause extensive damage to flora and fauna but also contribute to soil erosion, long-term land degradation and even lead to severe flooding. The importance of effective post-fire management and rapid response strategies necessitates the availability of precise and up-to-date burned area maps [1].

Remote sensing has emerged as a critical tool for wildfire monitoring and assessment, providing a cost-effective and efficient means to gather information across vast and often inaccessible landscapes. Satellite-based remote sensing has long been a cornerstone in the study of wildfires, offering critical insights into pre-fire conditions, active fire detection, and post-fire impacts. For post-fire analysis, it provides essential data for quantifying burn severity and delineating the spatial extent of fire-damaged areas [2], [3]. Sentinel-2, a key mission within the European Copernicus Program, is particularly well-suited for this purpose. With a spatial resolution of 10 meters and frequent revisit cycle (5 days), offers opportunities for detailed and timely burned area mapping [4]. Its multispectral imaging capabilities, combined with its global coverage, enable the detection of subtle spectral changes associated with burned vegetation [5]. These capabilities are particularly relevant in the context of the Greek territory, where fire events are both frequent and spatially extensive.

This study builds upon the existing research and contributes to the growing field of remote sensing and artificial intelligence by developing a deep learning-based framework for burned area detection across the entire Greek territory using Sentinel-2 data. It incorporates the following innovative aspects:

- Two-Year Dataset: Leveraging data from two years events adds temporal depth.
- Model Generalization: Training the U-Net model on a diverse dataset of fire events ensures robustness across varying conditions and regions within Greece. It provides a scale of application (nationwide application) that is less common in existing literature.
- Operational Scalability: The framework is designed to infer burned areas efficiently across the entire Greek territory, enabling rapid deployment in operational settings.

2. MATERIALS AND METHODS

2.1. Study area

Greece, located in Southeastern Europe, is characterized by a diverse topography. This geographical diversity is accompanied by a Mediterranean climate, marked by hot, dry

doi:10.2760/2119408

summers and mild, wet winters. These climatic conditions, combined with human activities, make Greece particularly susceptible to wildfires, especially during the summer months. Over the past decade, the country has experienced numerous devastating fire events (Figure 1), with significant impacts on its ecosystems, economy, and local communities. The study area encompasses the entirety of Greece, including both mainland and island regions. The selection of the study area is critical, given its ecological significance and the increasing frequency of wildfire events.

2.2. Satellite data

This study leverages Sentinel-2 satellite data, which provides multispectral imagery crucial for capturing the spectral changes caused by wildfires. The dataset includes records from fire events across Greece over the past two years, as catalogued by CEMS. This extensive dataset ensures that the U-Net model can be trained to recognize and classify burned areas with high accuracy across the country's diverse landscapes.

In this approach, twenty (20) Sentinel-2 level-2A (Bottom-Of-Atmosphere) images during 2023-2024 were obtained. The criteria for the selection of scenes were the high quality of data and the limited cloud coverage (absence of clouds over the burned areas of interest). A few examples are presented in false color composites (Figure 2).

2.3. Methodological procedure

2.3.1. Tools

For this study, ArcGIS Pro was used as the primary tool for the implementation of the workflow. ArcGIS Pro's Image Analyst extension supports advanced deep learning capabilities, including tools for data preparation, training, and deployment of DL models. Its integration with deep learning frameworks like TensorFlow and PyTorch enables seamless workflows for pixel classification, object detection, and semantic segmentation. All computations were performed on a Windows 10 workstation equipped with an Intel Core i5-8300H CPU, 32 GB of RAM, and an NVIDIA GeForce GTX 1050 GPU.

2.3.2. Initial processing of Sentinel-2 data

Sentinel-2 imagery was preprocessed to ensure consistency and readiness for classification. This included resampling the SWIR-2, NIR, and Red bands [4] using the nearest neighbor algorithm to match a uniform spatial resolution. The bands were then stacked to create a single composite image. This preprocessing step is critical to align spectral data and facilitate efficient training and classification of imagery.

Figure 1. Spatial distribution of the wildfire events in Greece (2023-2024) used as training data for the algorithm, shown as points with their corresponding CEMS codes overlaid on a Sentinel-2 cloudless mosaic.

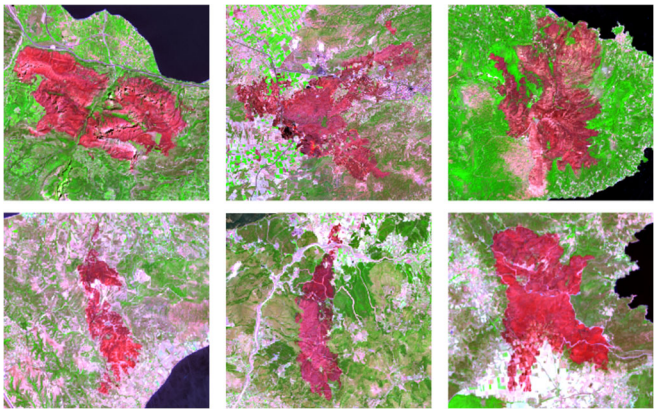


Figure 2. Examples of wildfire events displayed in a false color composite (R: SWIR-2; G: NIR; B: Red).

2.3.3. Classification of imagery

A binary classification scheme was employed to categorize areas into two thematic classes as either burned or not burned. Manual classification was conducted for all images to ensure the accuracy of labels. It has to be mentioned that several representative samples of confirmed land cover were selected from across the scene and not a particular area to capture the spectral variability. This process involved visually inspecting each image and assigning classes based on the observed spectral changes caused by fire events.

doi:10.2760/2119408

2.3.4. Raster to vector

The classified images were converted into vector format, retaining only the polygons representing burned areas. Each polygon was manually reviewed to eliminate false positives, ensuring that the dataset reflected accurate burned area delineation.

2.3.5. Export training data

Once samples have been collected and processed, they were exported to training data in order to be used in the DL model. The training data, consisting of classified tiles in TIFF format (one classified image chip per input image chip), was prepared for pixel classification. ArcGIS Pro's deep learning labeling tools were used to export data in 64*64 tile size (the size of the image chips) and 32*32 stride (the distance in the x direction when creating the next image chips). When stride is equal to tile size, there will be no overlap. When stride is equal to half the tile size, there will be 50 percent overlap, which was the case in this study.

2.3.6. Train deep learning model

The U-Net architecture was selected for training due to its demonstrated efficacy in semantic segmentation tasks, particularly for remote sensing applications. U-Net is a specialized type of fully convolutional neural network (FCN), designed to assign class labels to individual pixels [6].

U-Net builds upon the Fully Convolutional Network (FCN) [7], which replaces the fully connected layers in CNNs with up-convolutional layers, allowing the network to output dense pixel-wise classifications. U-Net comprises a symmetrical U-shaped architecture, consisting of a contracting path and an expansive path [8], [9].

The training process leveraged GPU acceleration for computational efficiency. To enhance feature extraction, several backbone models (e.g., ResNet, VGG) were tested as encoders in the U-Net architecture. These backbone models provide pre-trained convolutional layers, which help to improve training efficiency and accuracy, particularly for datasets with limited labeled samples.

3. RESULTS

3.1. Model performance evaluation

To assess the effectiveness of the proposed methodology, a total of 8,534 training samples were utilized to train and validate the U-Net architecture alongside several baseline models for comparison. The models were evaluated using common performance metrics, including accuracy, precision, recall, and F1-score. Training time was also recorded to assess computational efficiency.

The results indicate that the U-Net model consistently outperformed the baseline models across all metrics. Notably, U-Net paired with ResNet encoders demonstrated exceptional performance, achieving the highest overall accuracy (96%) and balanced scores across precision, recall, and F1 metrics. Additionally, U-Net model exhibited significantly faster training times compared to other architectures, making them not only accurate but also computationally efficient.

A few common error metrics regarding the validation of the models are presented in Table 1.

Table 1. Baseline models.

Model	Encoder	F1	Training time
U-Net	ResNet18	0.95	24' 36''
U-Net	ResNet34	0.95	24' 50''
DeepLabV3	ResNet18	0.90	39' 49''
DeepLabV3	ResNet34	0.89	52' 57''
DeepLabV3	DenseNet-121	0.84	49' 14''
MMSegmentation	HRNet	0.92	1h 58' 26''

3.2. Inference on new data

To validate the trained U-Net model, an inference was performed on two selected regions affected by wildfires, corresponding to CEMS activations EMSR747 and EMSR767. The results were compared with the delineation products that are presented in Figure 3.

Copernicus EMS supports emergency response efforts for various disasters, including wildfires. Upon activation, the EMS provides high-resolution wildfire maps generated from satellite imagery to assess burned areas and severity levels. The delineation product, outlines the extent of burned areas and served as the reference layer for an overlap analysis. This involved comparing the EMS reference layer with the burned area maps generated by the U-Net model (overlay layer). The analysis measured the overlap percentage and the total area of burned regions as detected by each approach. The overlap analysis results are summarized in Table 2.

Table 2. Comparison of burned area delineation and overlap percentages between the U-Net model and the Copernicus EMS reference products for EMSR747 and EMSR767.

Approach	EMSR747		EMSR767	
	Area (ha)	Overlap	Area (ha)	Overlap
U-Net	1088.8		5619.2	
		94.5 %		95.7 %
EMS	1138.4		5829.9	

doi:10.2760/2119408

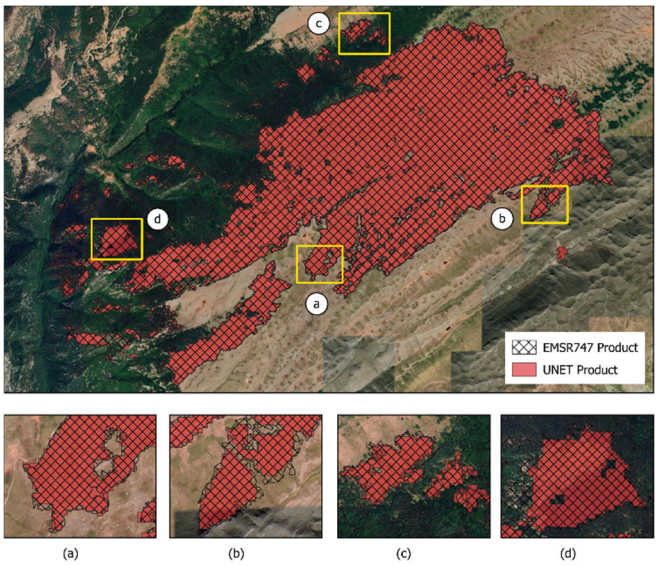


Figure 3. Burned area delineation for EMSR747: Comparison between the Copernicus EMS reference product and the burned area map generated by the U-Net model, including zoom-in views.

4. DISCUSSION

The results of this study underscore the efficacy of the U-Net model for burned area detection. Achieving high overlap percentages with Copernicus EMS delineation products (94.5% for EMSR747 and 95.7% for EMSR767), the U-Net model demonstrates its ability to produce reliable burned area maps that align closely with authoritative reference data.

Comparisons with baseline models further emphasize U-Net's advantages. While alternatives like DeepLabV3 and MMSegmentation provided reasonable performance, they required longer training times and delivered lower accuracies, especially when using encoders like DenseNet-121. U-Net, paired with ResNet encoders, achieved superior precision and computational efficiency, making it a highly suitable tool for operational applications.

Despite these promising results, some discrepancies between the U-Net model and EMS reference data were observed, indicating opportunities for improvement. These minor differences may stem from variations in acquisition times and implemented methodology.

5. CONCLUSION

Monitoring catastrophic events in near real-time is crucial for effective disaster response. The timely detection and accurate delineation of affected regions are essential for providing critical information to public authorities and stakeholders. Such efforts play a pivotal role in minimizing the impact of disasters and supporting recovery initiatives.

The insights gained from this work are not limited to Greece but are also applicable to other Mediterranean regions with similar climatic and topographical challenges that face analogous wildfire risks, thereby demonstrating the broader value of the proposed methodology.

Future research will focus on integrating more diverse datasets, enabling the model to generalize better across varying conditions and regions.

REFERENCES

[1] Chuvieco, Emilio, Inmaculada Aguado, Javier Salas, Mariano García, Marta Yebra, and Patricia Oliva. 2020. “Satellite Remote Sensing Contributions to Wildland Fire Science and Management.” Current Forestry Reports 6 (2): 81–96. doi:10.1007/s40725-020-00116-5.

[2] Chuvieco, Emilio, Florent Mouillot, Guido R. van der Werf, Jesús San Miguel, Mihai Tanase, Nikos Koutsias, Mariano García, et al. 2019. “Historical Background and Current Developments for Mapping Burned Area from Satellite Earth Observation.” Remote Sensing of Environment 225 (May): 45–64. doi:10.1016/j.rse.2019.02.013.

[3] Çömert, Resul, Dilek Küçük Matcı, and Uğur Avdan. 2019. “Object Based Burned Area Mapping with Random Forest Algorithm.” International Journal of Engineering and Geosciences 4 (2). Murat YAKAR: 78–87. doi:10.26833/ijeg.455595.

[4] Filipponi, Federico. 2019. “Exploitation of Sentinel-2 Time Series to Map Burned Areas at the National Level: A Case Study on the 2017 Italy Wildfires.” Remote Sensing 11 (6). Multidisciplinary Digital Publishing Institute: 622. doi:10.3390/rs11060622.

[5] Smiraglia, Daniela, Federico Filipponi, Stefania Mandrone, Antonella Tornato, and Andrea Taramelli. 2020. “Agreement Index for Burned Area Mapping: Integration of Multiple Spectral Indices Using Sentinel-2 Satellite Images.” Remote Sensing 12 (11). Multidisciplinary Digital Publishing Institute: 1862. doi:10.3390/rs12111862.

[6] Kotaridis, I., and M. Lazaridou. 2022. “SEMANTIC SEGMENTATION USING A UNET ARCHITECTURE ON SENTINEL-2 DATA.” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B3-2022 (May): 119–26. doi:10.5194/isprs-archives-XLIII-B3-2022-119-2022.

[7] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 2015. “Fully Convolutional Networks for Semantic Segmentation.” In , 3431–40. IEEE Computer Society. doi:10.1109/CVPR.2015.7298965.

[8] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, 234–41. Lecture Notes in Computer Science. Cham: Springer International Publishing. doi:10.1007/978-3-319-24574-4_28.

[9] Zhang, Pengbin, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. 2018. “Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery.” Sensors 18 (11). Multidisciplinary Digital Publishing Institute: 3717. doi:10.3390/s18113717.

SPACE-DRIVEN GEOSPATIAL ANALYTICS FOR PAN EUROPEAN HUMAN SETTLEMENTS FIXED ASSET VALUATION: ENHANCING EXPOSURE MODELING FOR CLIMATE ADAPTATION POLICY

Michele Melchiorri[a], Andrea Sibilia[b], Christos Bountzouklis[a], Samuel Roeslin[a], Davide Rodomonti[b], Sandro Salari[b], Christina Corbane[a]

a) European Commission, Joint Research Centre, Ispra, 21027, Italy
b) Unisystems Luxembourg S.a.r.l. external service provider of European Commission Joint Research Centre, Via E. Fermi 2749, I - 21027 Ispra (VA), Italy

ABSTRACT

This study addresses the critical gap in harmonized geospatial data on fixed asset values (FAV) across European human settlements by integrating the EU-wide HANZE grid with 100m-resolution built-up surface data from the Global Human Settlement Layer (2020). The methodology enables reliable FAV estimation at grid level, adaptable to diverse spatial scales (e.g., cities, regions) using input data routinely produced by EUROSTAT and Copernicus. Results include Pan-European human settlements FAV metrics (total €, € per capita, € per km²) disaggregated by territorial typology and hazard type for 2000 – 2020 at 5 years interval. Summary statistics are aggregated into NUTS3 units for policy applications. A use case is presented by analysing the exposure of built-up area to river flood hazard using Copernicus Emergency Management Service data. By embedding this dataset into the JRC Risk Data Hub, the research enhances disaster risk management tools and information systems, supporting climate adaptation planning.

***Index Terms*—** Risk Data Hub, Copernicus Exposure Mapping

1. INTRODUCTION

1.1. Addressing the Exposure Information Gap for Human Settlements

Accurate information on human settlements to sustain next generation adaptation and resilience policies must go beyond baseline information. Past efforts have significantly contributed to establishing global information systems like the Global Human Settlement Layer that maps built-up, population and settlements at global level [1]. Yet, information on the characteristics of such settlements is still scarce. The estimation of the monetary value of human settlements exposed to natural hazards is key, but a gap exists for harmonized, high-resolution geospatial data on fixed asset values (FAV) across Europe. This gap limits the ability of policymakers and stakeholders to quantify infrastructure exposed, potential losses, prioritize investments, and design targeted adaptation measures [2]. The need for robust and multi-exposure information is underscored by the increasing frequency and severity of climate-related hazards, and by the requirements of the EU's disaster risk management frameworks, the climate adaptation strategies, the Cities Mission, and the Sendai Framework for Disaster Risk Reduction.

1.2. Data Integration: HANZE Grid and High-Resolution Human Settlement Exposure Baseline Data

To bridge this information gap, our approach integrates the HANZE v2.0 exposure dataset [3] with the Global Human Settlement Layer (GHSL) built-up surface data at 100 m resolution. The resulting asset value grids are embedded within the JRC's Risk Data Hub (RDH), a key platform for disaster risk data and analytics in Europe. Within the RDH, these datasets are combined with hazard layers (e.g., floods, earthquakes, wildfires, subsidence) and territorial typologies, enabling multi-hazard exposure assessments at various spatial scales (from local administrative units to NUTS3 regions). The RDH's analytical tools facilitate the aggregation of FAV metrics (total €, € per capita, € per km²) and their disaggregation by settlement type and hazard, supporting evidence-based policy development and resource allocation. The integration of the FAV dataset in the RDH is planned for 2025. By addressing the exposure information gap through the integration of HANZE and GHSL data within the Risk Data Hub, this work provides a harmonized, scalable, and policy-relevant foundation for pan-European fixed asset valuation.

doi:10.2760/2119408

2. MATERIALS AND METHODS

The core of the implemented methodology is the integration of the HANZE v2.0 exposure grid with the 100m-resolution built-up surface data from the Global Human Settlement Layer (GHSL) via geospatial processing.

2.1. HANZE Grid

The HANZE grid [3] provides a peer-reviewed high-resolution (100m) estimates of fixed asset value (FAV), in various sectors (residential dwellings, services, residential contents, infrastructures, industry, agriculture and forestry) for 42 Pan-European countries over time –from 1870 to 2020 (for this study a shorter time window 2000 – 2020 at 5 years interval is used). The disaggregation from regional socioeconomic statistics (EUROSTAT input data) to grid cells employs a two-step dasymetric approach, combining population density (60%) and land use/soil sealing (40%) as predictors (Copernicus Land data). Fixed assets, expressed in absolute terms at the regional level, are calculated by Paprotny et al. by multiplying the regional GDP by the corresponding wealth-to-GDP ratio for each sector [4].

For the *Total Fixed Asset Value in Human Settlements* the following sectors were considered: Residential and services, industry, residential contents and infrastructures.

2.2. Global Human Settlement Layer

The Global Human Settlement Layer, produced as Exposure Mapping Component by the Copernicus Emergency Management Service delivers global information on human settlements as open geospatial hig-resolution grids. The GHS-BUILT-V and GHS-SMOD products are used in this study to obtain the "settlement mask" to extract HANZE FAV data.

The GHS-BUILT-V [5] product consists in Built-up volume grids (100m), derived from multi-sensor satellite imagery (Landsat, Sentinel-2), available for 1975–2030 at 5-year intervals and provides total and non-residential built-up volume estimates (in m^3).

The GHS-SMOD product [6] classifies each 1km grid cell in a harmonised urban, intermediate and rural classification from the Degree of Urbanisation (urban centre, urban cluster, rural areas), based on population density, implementing EUROSTAT's DEGURBA framework. GHS-SMOD is used to disaggregate zonal statistics to obtain total FAV in urban centres, urban clusters and rural areas.

2.3. Exposure Data:

The exposure from HANZE and GHSL is also combined with the river flood hazard maps for Europe and the Mediterranean Basin region using the 100 years return period [7]. The river flow data used for the maps are generated by JRC using the open-source hydrological model LISFLOOD, while flood inundation simulations are conducted with the hydrodynamic model LISFLOOD-FP. The coverage area includes most of geographical Europe and flood hazard maps are produced for river basins larger than 150 km^2, with each cell value representing water depth in meters. For this study the hazard map has been obtained by thresholding the water depth >20cm.

2.4. Geospatial Processing: Zonal Statistics

The methodology to obtain the *Total FAV in Pan-European Human Settlements* delineated with the GHSL is based on geospatial overlay and zonal statistics. This technique returns a sum in the intersection between HANZE and GHSL grids. The target sum is the FAV by sector for all HANZE grid cells with settlement information (i.e. BUILT-V ≥ 0). The sum is then aggregated to rasterised NUTS3 polygons, and disaggregated by GHS-SMOD grid cell types. The procedure is repeated for all the years (2000-2020:5).

The grid-based approach provides significant spatial flexibility and allows to further integrate this new dataset with other information like hazard data (similarly, the overlay between the *Total FAV in Pan-European Human Settlements* grid and the hazard map is used to quantify the total amount of FAV in exposed areas).

Fig. 1 shows a schematic workflow, input and output data.

With the above method the following indicators are calculated per each year: Total FAV – Monetary Value (EUR): Sum of the monetary value of the assets in each grid cell or aggregated zone (i.e. NUTS3); Per-capita FAV: Ratio of assets to population in each spatial unit (i.e. NUTS3) (EUR per person); FAV per km^2: asset values normalized by the area of the spatial unit (i.e. NUTS3) (EUR per per km^2 of the NUTS3).

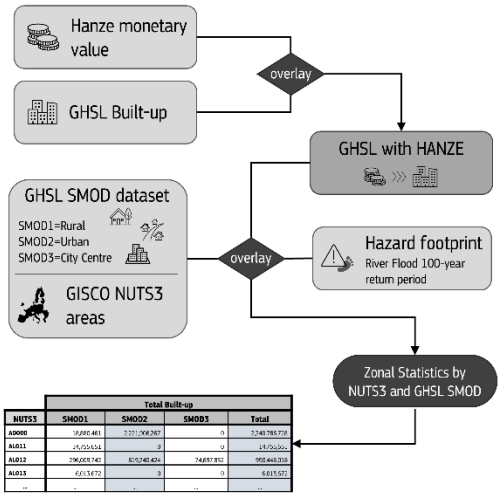


Fig. 1. schematic view of input data, processing chains and outputs.

doi:10.2760/2119408

3. RESULTS

The geospatial data integration returns a Pan-European dataset that quantifies fixed asset values (FAV) in human settlements at multiple geographical scales in grid format (100m and 1000m) and in tabular format from zonal statistics (summary for NUTS3 and country totals) for the year 2000, 2005, 2010, 2015, 2020. The analysis of the obtained dataset reveals significant disparities in FAV:

- The total FAV in Pan-European human settlements in 2020 is in the range of 92 trillion Euro, of which 44% in urban centres, 33% in urban clusters, and 23% in rural areas. Five NUTS3 (Paris, Madrid, Milano, Roma, and Berlin) concentrate more than 1,000 billion in fixed asset values (in 2020 – Fig. 2);

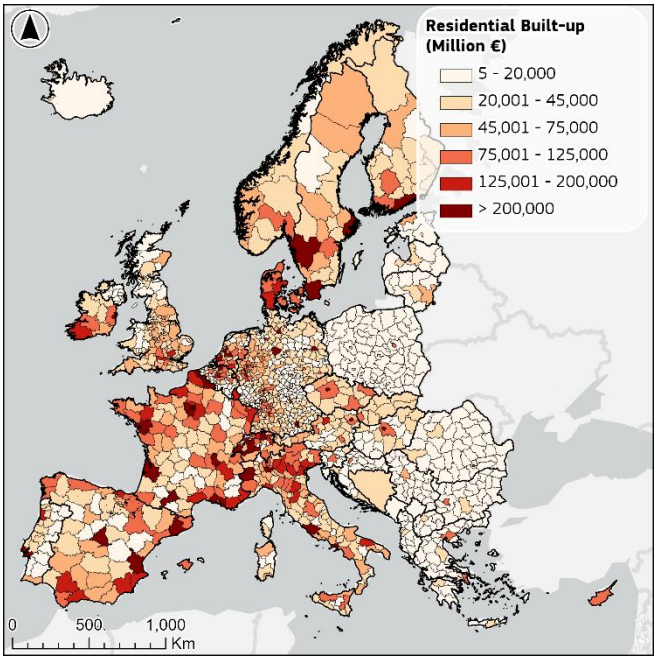


Fig. 2. Absolute Fixed Asset Value in Pan-European NUTS3.

- By territorial typology, 10 NUTS3 exceed 500 billions of residential built-up fixed asset values in urban centres (2020) these include Paris, Berlin, Madrid, Milano, München, Kreisfreie Stadt, Hamburg, Wien, Roma, and Barcelona.
- Between 2000 and 2020 FAV increases by 43% in urban centres and 44% in urban clusters, compared to 35% in rural areas (Fig. 2);
- A typical West/East regional divide clearly emerges (Fig. 3). East Europe Member States have a relatively lower residential built-up absolute FAV;
- The overlay with flood hazard reveals that about 7% of the Pan-European FAV in human settlements is exposed to potential floods with 100 years return period;
- Figure 4 displays the share of FAV exposed to floods across the study area revealing that 16 NUTS3 have more than 35% of their built-up exposed, with an estimated FAV of about 507 billion Euro.

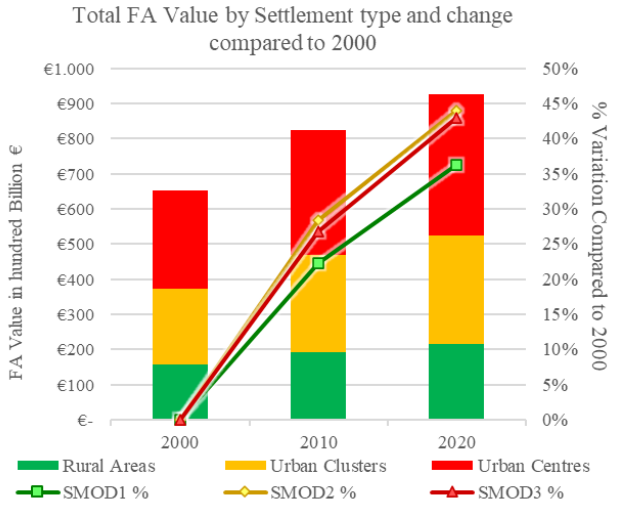


Fig. 3. Absolute Fixed Asset Value in Pan-European NUTS3.

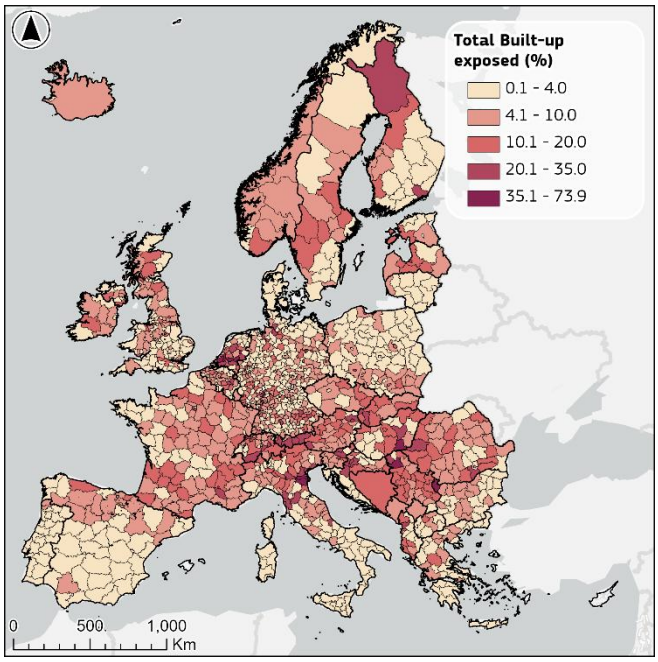


Fig. 4. Relative Fixed Asset Value exposed to 100 years return period riverine flood in Pan-European NUTS3.

4. DISCUSSION

4.1. Improving Exposure Information and Enhancing the JRC Risk Data Hub for Pan-European Risk Assessment

Despite preliminary, the results of this application of remote sensing and other data sources, indicates the feasibility of estimating fixed asset value in Pan-European human

doi:10.2760/2119408

settlements. The integration of high-resolution fixed asset value data for human settlements directly addresses the need for risk-informed urban planning, a cornerstone of modern disaster risk management and climate adaptation strategies. By providing spatially explicit information on the distribution and value of assets exposed to hazards, this dataset empowers planners and policymakers to identify vulnerability hotspots, prioritize risk reduction investments, and implement zoning policies that minimize future exposure. The production of such exposure data is the first step towards improving disaster risk information systems. Embedding the harmonized FAV dataset into the JRC Risk Data Hub (RDH) significantly upgrades the platform's analytical capabilities. With the inclusion of detailed, grid-level asset valuation, the RDH can now support more nuanced, multi-hazard exposure analyses at various spatial scales-from local municipalities to NUTS3 regions. By adding the monetary dimension to human settlement exposure (i.e. built-up as provided by the Copernicus Emergency Management Service Exposure Mapping Component) this study prototypes a new dimension of exposure information. This research contributes to the objectives of the EU Strategy on Adaptation to Climate Change, which outlines a vision for a climate-resilient Europe fully adapted to the impacts of climate change by 2050 aligning with the Strategy's goal of "smarter adaptation" improving climate loss data.

4.2. Limitations and next steps

The choice of the dataset by Paprotny et al. was determined by scalability. In fact, the geospatial covariates used for the HANZE grid are regularly produced by Copernicus (Land Monitoring Service), and the economic data are produced by EUROSTAT with periodic updates.

Despite the availability of FAV by sector in the HANZE grid, we did not include further sectoral disaggregation by sector (besides the built-up residential, the main component of built stock) as the infrastructure assets represented an implausible 4.8% of the total value in the input HANZE data, with this share reducing to 3.2% in the human settlement mask. We therefore computed an overall/gross FAV in human settlements.

Further research might be needed to refine and harmonise proxies for disaggregation (i.e. fully align to GHSL) and to improve the disaggregation by economic sectors. The usage notes and limitations in the HANZE supporting publications are valid for the zonal statistics presented in this article as well [3-4]. Given the continuous availability of input data for later points in time, a routine production of the FAV indicators could be considered.

5. CONCLUSION

This study presented a prototype for a novel, harmonized methodology for estimating fixed asset values (FAV) across European human settlements by integrating the HANZE grid with high-resolution data from the Global Human Settlement Layer. The resulting dataset fills a crucial exposure information gap, enabling spatially explicit and temporally dynamic analyses of asset distribution and hazard exposure at Pan-European scale. A full-scale deployment of the method to Copernicus hazard data, combined with the integration of the resulting exposure metrics into the JRC Risk Data Hub, is expected to substantially enhance the analytical capabilities available for disaster risk management, climate adaptation planning, and policy evaluation at multiple spatial scales.

ACKNOWLEDGMENT

The authors are very grateful for the work of Dr. Dominik Paprotny for the production and validation of the Historical Analysis of Natural Hazards in Europe (HANZE) dataset, and for the interactions during the authors' work with HANZEv2 data.

REFERENCES

[1] Pesaresi, Martino, et al. "Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data." International Journal of Digital Earth 17.1 (2024): 2390454.

[2] McNamara, Karen E., and Guy Jackson. "Loss and damage: A review of the literature and directions for future research." Wiley Interdisciplinary Reviews: Climate Change 10.2 (2019): e564.

[3] Paprotny, Dominik., Morales-Nápoles, O., and Jonkman, S. N.: HANZE: a pan-European database of exposure to natural hazards and damaging historical floods since 1870, Earth Syst. Sci. Data, 10, 565–581

[4] Paprotny, Dominik, and Jakub Śledziowski. "HANZE v2. 1: an improved database of flood impacts in Europe from 1870 to 2020." Earth System Science Data 16.11 (2024): 5145-5170.

[5] Pesaresi, Martino; Politis, Panagiotis (2023): GHS-BUILT-V R2023A - GHS built-up volume grids derived from joint assessment of Sentinel2, Landsat, and global DEM data, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC) [Dataset] doi: 10.2905/AB2F107A-03CD-47A3-85E5-139D8EC63283 PID: http://data.europa.eu/89h/ab2f107a-03cd-47a3-85e5-139d8ec63283

[6] Schiavina, Marcello; Melchiorri, Michele; Pesaresi, Martino (2023): GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA PID: http://data.europa.eu/89h/a0df7a6f-49de-46ea-9bde-563437a6e2ba

[7] Baugh, Calum; Colonese, Juan; D'Angelo, Claudia; Dottori, Francesco; Neal, Jeffrey; Prudhomme, Christel; Salamon, Peter (2024): River flood hazard maps for Europe and the Mediterranean Basin region. European Commission, Joint Research Centre (JRC) [Dataset] doi: 10.2905/1D128B6C-A4EE-4858-9E34-6210707F3C81 PID: http://data.europa.eu/89h/1d128b6c-a4ee-4858-9e34-6210707f3c81

ARTISANAL AND SMALL-SCALE GOLD MINING DETECTION IN THE AMAZON FOREST USING CONTEXTUAL DATA

Selma Dissing, Jan-Christoph Kalo

University of Amsterdam
selma.dissing@student.uva.nl, j.c.kalo@uva.nl

ABSTRACT

Artisanal and small-scale gold mining (ASGM) is a major driver of land cover change in the Amazon, often challenging to detect due to its spectral similarity with other surface features. This study investigates whether incorporating contextual geospatial data from OpenStreetMap (OSM) alongside Sentinel-2 imagery can improve ASGM detection in Venezuela's Bolívar state. Rasterized OSM-derived semantic masks are appended as additional input channels to the satellite imagery and processed through a CNN. This setup enables joint learning of spectral and contextual features, allowing for a more accurate and reliable distinction between ASGM sites and spectrally similar land uses.

1. INTRODUCTION

Artisanal and small-scale gold mining (ASGM) is a largely informal mining practice where individuals or small groups extract gold using rudimentary techniques. It is the primary driver of illegal mining in the Amazon rainforest [1]. Remote sensing has emerged as a key tool for monitoring ASGM [2]. However, current mining detection models rely solely on spectral information, which presents challenges in distinguishing ASGM sites from other land-use changes, such as dried riverbeds or bare soil patches that exhibit similar spectral characteristics [3]. This limitation highlights the importance of integrating complementary data sources that can provide additional contextual information.

One promising way to address these limitations is the integration of contextual geospatial data. A prominent example is OpenStreetMap (OSM) [4], which contains detailed, freely accessible information on infrastructure, transportation networks, and settlements. This study explores whether incorporating such contextual data alongside satellite imagery can improve the detection of ASGM activity by providing additional cues to distinguish it from spectrally similar land-use changes. To explore this issue in detail, this study focuses on the southeastern Venezuelan state of Bolívar, one of the most ecologically significant and mining-affected regions within the Amazon basin. We investigate whether integrating contextual geospatial data from OSM improves the performance of ASGM detection using satellite imagery. Sentinel-2 provides spectral signals related to land disturbance. At the same time, OSM contributes spatial context, such as the presence of rivers, roads, or buildings, which can help distinguish ASGM sites from visually similar land uses. To test this, rasterized OSM-derived semantic masks are appended as additional input channels to the Sentinel-2 imagery and processed together through a convolutional neural network (CNN). This setup enables the model to jointly learn spectral and contextual patterns, allowing for an evaluation of the added value of OSM data for ASGM classification.

2. RELATED WORK

Satellite Imagery for Remote Sensing. Researchers have utilized various geospatial data sources, notably optical satellite imagery, to detect land disturbances associated with ASGM [5]. These approaches build on established remote sensing techniques for LULC change detection, widely applied to monitor deforestation, urban growth, agriculture, and natural disasters [6–9]. Sentinel-2 imagery is particularly effective for identifying ASGM impacts such as vegetation loss, exposed soil, and sediment-laden water, observable via changes in surface reflectance and water clarity [10–12]. However, optical imagery alone has limitations: ASGM sites often resemble agricultural or natural disturbances spectrally, leading to frequent misclassifications [3].

Image Classification with Contextual Data. Incorporating contextual information into image classification improves model performance by enabling reasoning. Applications in product recognition and medical imaging benefit from spatial and semantic context, which enhances the detection of small objects and anomalies [13, 14]. Fusing remote sensing with spatial features such as land types and surroundings provides semantic cues that enhance accuracy [12, 15]. For instance, Gomez et al. [16] used proximity to transport and water networks to detect ASGM activity using SAR imagery and manually derived features. In contrast, our study employs a framework that integrates multispectral Sentinel-2 imagery with geographic OSM data. OSM, a crowd-sourced dataset, offers layers such as roads, rivers, land-use, and buildings. For example, [17] used OSM building footprints to classify

doi:10.2760/2119408

informal settlements via spatial clustering and machine learning. Similarly, [18] aligned OSM features with remote sensing data for LULC classification and infrastructure detection. Despite its utility, OSM poses challenges due to variable quality and coverage, particularly in remote regions such as the Amazon Basin.

3. METHODOLOGY

This study builds on an open-source pipeline for ASGM detection using Sentinel-2 satellite imagery provided by Earth Genome [19]. While the baseline system includes labeled sampling points, a patch extraction process using Google Earth Engine (GEE), and a CNN for binary classification, this work extends it by incorporating contextual geospatial information from OSM, particularly features such as highways, waterways, buildings, land-use, and aeroways. The proposed method processes Sentinel-2 patches with added channels of rasterized OSM-derived semantic masks in the CNN architecture. This setup enables the model to learn spectral and spatial-contextual patterns, with the aim of improving performance.

Sampling Points and Patch Extraction. The study focuses on the Bolívar region of southeastern Venezuela. This region was selected due to its high density of mining operations and the availability of labeled ASGM data.

The baseline process handles Sentinel-2 patch extraction [19], which converts each sampling point into a standardized satellite image patch. Using consistent projection and spatial resolution, a square tile is generated and aligned with Sentinel-2 imagery for each sampling point. Sentinel-2 image data is retrieved from GEE for 2021, corresponding to the period when the sampling points were published on GitHub. The extracted bands include B1–B4, B5–B7, B8, B8A, B9, B11, and B12, covering the visible, red-edge, near-infrared, and shortwave-infrared regions. Band 10 is excluded as it is primarily used for cloud detection and is not relevant for land cover analysis. To reduce the impact of clouds, cloud shadows, and other atmospheric effects, a median composite is generated by averaging all observations across the year. This approach enhances image quality and consistency, particularly in cloud-prone regions such as the Amazon Basin.

The resulting image patches are uniformly sized with the 12 spectral channels. Each patch covers an area, centered around a sampling point labeled as either mining or non-mining. This patch captures the mining activity and the surrounding area, which can provide additional context through OSM data. The dataset includes 156 mining patches and 323 non-mining patches.

OSM Data. Vector-based geospatial features were extracted from OSM. These features were selected based on their relevance to ASGM operations. They comprise: (1) **highway**, including roads, tracks, and footpaths indicating access to mining areas; (2) **waterway**, such as rivers and streams, often used in alluvial mining; (3) **building**, representing nearby informal or residential structures; (4) **landuse**, comprising tagged areas potentially overlapping with ASGM zones; and (5) **aeroway**, which may denote airstrips supporting remote mining operations. See Figure 1 for the spatial distribution of the OSM features in the Bolívar region.

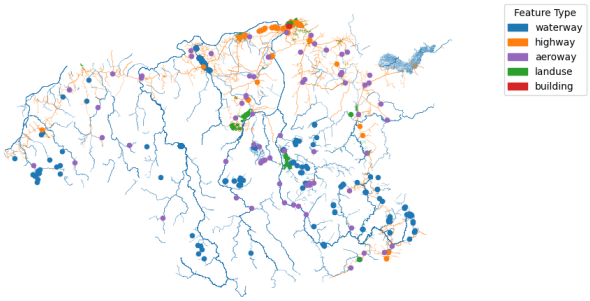


Fig. 1. Spatial Distribution of OSM Features in Bolívar.

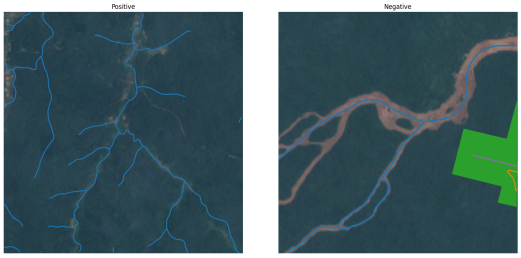


Fig. 2. Visualizations of random patches with OSM overlays.

Rasterization of OSM data. For each image patch, a corresponding subset of OSM features is extracted by clipping the vector data to the same extent. Each feature has its own channel, where a pixel with values for that feature receives a value of 1, and a pixel without that feature receives a value of 0. This ensures that both inputs have identical footprints, allowing the OSM data to be rasterized and stacked with the Sentinel-2 bands before model training. Of the 156 mining patches, 95 contain one OSM feature, representing 60.9% coverage. Among the 323 non-mining patches, 245 include OSM annotations, resulting in 75.85% coverage. This variation reflects the inconsistent availability of OSM data in remote regions, such as Bolívar, where mapping is often sparse.

Model. The baseline model is adapted from Earth Genome's open-source ASGM detection pipeline [19], which employs a lightweight CNN. The architecture consists of nine convolutional layers, each with 32 filters and 3×3 kernels, using 'same' padding and ReLU activation functions. These layers are grouped into three convolutional blocks, each followed by max-pooling layers (2×2 or 3×3) to progressively reduce the spatial resolution. The convolutional feature extraction stage is followed by three dense layers with 64, 64, and 32 neurons. Dropout layers with a rate of 0.3 are included

doi:10.2760/2119408

after each dense layer to prevent overfitting. The final output layer is a sigmoid-activated neuron for binary classification.

4. EVALUATION

4.1. Experimental Setup.

To prevent spatial leakage, the dataset is split using DBSCAN clustering on patch center coordinates, with clusters randomly assigned to train (70%), validation (15%), and test (15%) sets. Patches with excessive masking are filtered out, and Sentinel-2 reflectance values are normalized to the range of [0, 1]. All models use the same CNN architecture trained for up to 160 epochs using the Adam optimizer (3×10^{-4}), binary cross-entropy loss, and a batch size of 16. Data augmentation (random rotations, shearing, zooming, flips) is applied via Keras' `ImageDataGenerator`. Experiments vary by patch size (48px vs. 256px), class imbalance weighting, negative sampling strategy (156 points curated from Earth Genome vs. 780 points randomly sampled), and inclusion of OSM features.

4.2. Results and Discussion.

Ten ASGM detection experiments were conducted under varying patch sizes, negative sampling strategies, class imbalance handling, and the inclusion of OSM features. Performance was evaluated using F1 (with bootstrap confidence intervals), precision, recall, and PR-AUC to capture threshold-dependent and threshold-independent model performance. Table 1 summarizes performance across these settings.

Patch Size. In curated settings, both 256px (Exp 1, F1 = 0.867) and 48px (Exp 3, F1 = 0.711) models achieved strong performance without OSM. Adding OSM improved performance at 256px (Exp 2, F1 = 0.923; –2 FN, –5 FP), but only slightly at 48px (Exp 4, F1 = 0.788; –2 FP). This indicates that OSM features are more effective with larger patches that capture broader spatial context.

Class Weighting. With curated negatives, class weighting (Exp 5) increased FP (+7) and lowered precision (0.594), dropping F1 to 0.728. OSM (Exp 6) further raised FP (+10) and reduced F1 to 0.611, well below the unweighted OSM model (Exp 2). In random-negative settings, class weighting destabilized training: Exp 9 performed poorly (F1 = 0.129; FP = 62), while Exp 10 achieved F1 = 0.898 with perfect recall but lower precision.

Negative Sampling. With random negatives and no class weighting, performance remained strong: Exp 7 (F1 = 0.948) and Exp 8 (F1 = 0.792) both achieved high recall with few FP. Class weighting caused instability, leading one model to collapse (Exp 9) and another to trade precision for recall (Exp 10). Overall, curated negatives yielded more stable results, while random negatives better reflected real-world variability but increased uncertainty.

Confidence intervals highlight variability across experimental setups. Models trained with curated negatives and larger patches showed narrower confidence intervals, while random-negative sampling produced wider intervals, reflecting greater variability and task difficulty.

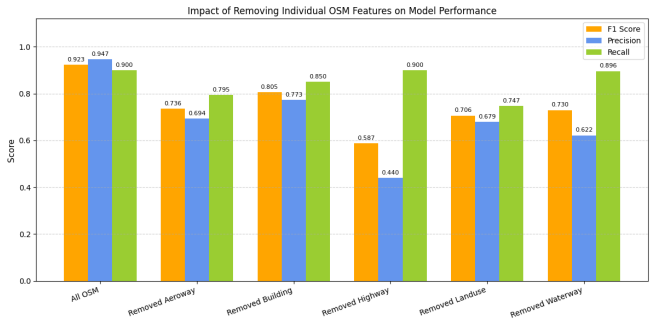


Fig. 3. Impact of removing OSM feature layers on model performance, reported as mean F1, precision, and recall. Results are shown relative to the full OSM model (Exp2).

Ablation Study. Figure 3 shows how the exclusion of different OSM feature layers influences performance. While most features contribute positively, certain layers have a more pronounced effect on the F1 score and precision than others.

Removing highway features resulted in the largest performance decline (F1 = 0.587), underscoring their importance in distinguishing ASGM sites. Land-use removal also reduced performance, while aeroway, waterway, and building layers had moderate effects. Overall, the study shows that multiple OSM layers contribute to detection, with highways providing the most critical contextual cues.

Limitations. This study focused exclusively on the Bolívar region in Venezuela, which limits the generalizability of the findings to other Amazonian settings. Model performance is strongly dependent on OSM coverage, which is uneven in remote areas and may bias results toward data-rich regions. A temporal mismatch between Sentinel-2 imagery from 2021 and OSM features from 2025 could also distort observed associations. In addition, the use of all 12 Sentinel-2 bands, including less informative ones such as B1 and B9, indicates the need for a more critical assessment of feature selection and modeling choices.

5. CONCLUSION

This study demonstrates that integrating contextual geospatial data from OSM with Sentinel-2 imagery can improve ASGM detection, particularly when larger patches and curated negatives are used. The most consistent benefit of OSM was a reduction in false positives, which helped distinguish mining from spectrally similar areas, although this sometimes came at the cost of recall. Notably, in the most challenging setting with random negatives and class weighting, OSM features enabled strong performance, whereas the spectral-only model failed to generalize. Overall, these results show that con-

doi:10.2760/2119408

Exp	OSM	Patch	Neg Type	Class Weight	F1 [95% CI]	PR-AUC	FP	FN	Precision	Recall
1	No	256	Curated	No	0.867 [0.739, 0.960]	0.865	6	0	0.769	1.000
2	Yes	256	Curated	No	0.923 [0.827, 1.000]	0.972	1	2	0.947	0.900
3	No	48	Curated	No	0.711 [0.529, 0.851]	0.850	7	5	0.682	0.750
4	Yes	48	Curated	No	0.788 [0.625, 0.913]	0.782	3	5	0.833	0.750
5	No	256	Curated	Yes	0.728 [0.565, 0.849]	0.866	13	1	0.594	0.950
6	Yes	256	Curated	Yes	0.611 [0.444, 0.744]	0.677	23	1	0.452	0.950
7	No	256	Random	No	0.948 [0.800, 1.000]	0.900	1	0	0.900	1.000
8	Yes	256	Random	No	0.792 [0.533, 0.957]	0.909	3	1	0.727	0.889
9	No	256	Random	Yes	0.129 [0.028, 0.237]	0.069	62	4	0.075	0.556
10	Yes	256	Random	Yes	0.898 [0.727, 1.000]	0.928	2	0	0.818	1.000

Table 1. Performance metrics across ASGM detection experiments with varying use of OSM data, patch sizes, negative sample types, and class weighting settings. Bootstrap confidence intervals are shown for F1-scores.

textual information can substantially enhance detection accuracy, but its effectiveness depends on data quality, temporal alignment, and the training setup. Future work should extend evaluation across regions, develop a tailored model, incorporate historical OSM data to address temporal mismatches, and explore advanced fusion strategies that more effectively align spectral and contextual features.

Acknowledgements

This research is supported by the European Union's Horizon Europe research and innovation programme within the ENEXA project (grant Agreement no. 101070305).

References

[1] S. Wang *et al.*, "Evaluating the Feasibility of Illegal Open-Pit Mining Identification Using Insar Coherence," *Remote Sensing*, vol. 12, no. 3, 2020.

[2] M. A. Alessi, P. G. Chirico, and M. Millones, "Artisanal Mining River Dredge Detection Using SAR: A Method Comparison," *Remote Sensing*, vol. 15, no. 24, p. 5701, 2023.

[3] W. Han *et al.*, "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 87–113, 2023.

[4] OpenStreetMap contributors, *OpenStreetMap: The Free Wiki World Map*, 2025.

[5] P. Kozińska and J. Górniak-Zimroz, "A review of methods in the field of detecting illegal open-pit mining activities," *IOP Conference Series: Earth and Environmental Science*, vol. 942, no. 1, p. 012 027, 2021.

[6] J. Wang *et al.*, "Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects," *Science of The Total Environment*, vol. 822, p. 153 559, 2022.

[7] S. Chen *et al.*, "Review of drivers of forest degradation and deforestation in Southeast Asia," *Remote Sensing Applications: Society and Environment*, vol. 33, p. 101 129, 2024.

[8] Z. Shao *et al.*, "Urban sprawl and its impact on sustainable urban development: A combination of remote sensing and social media data," *Geo-spatial Information Science*, vol. 24, no. 2, pp. 241–255, 2021.

[9] M. Kucharczyk and C. H. Hugenholtz, "Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities," *Remote Sensing of Environment*, vol. 264, p. 112 577, 2021.

[10] S. Camalan *et al.*, "Change Detection of Amazonian Alluvial Gold Mining Using Deep Learning and Sentinel-2 Imagery," *Remote Sensing*, vol. 14, no. 7, p. 1746, 2022.

[11] S. A. Mehta *et al.*, "A Characterization of Land-use Changes in the Proximity of Mining Sites in India," *ACM J. Comput. Sustain. Soc.*, vol. 2, no. 1, 10:1–10:23, 2024.

[12] R. Balaniuk, O. Isupova, and S. Reece, "Mining and Tailings Dam Detection in Satellite Imagery Using Deep Learning," *Sensors*, vol. 20, no. 23, p. 6936, 2020.

[13] J. Leng *et al.*, "Realize your surroundings: Exploiting context information for small object detection," *Neurocomputing*, vol. 433, pp. 287–299, 2021.

[14] M. E. Tschuchnig and M. Gadermayr, "Anomaly Detection in Medical Imaging - A Mini Review," in *Data Science – Analytics and Applications*, 2022, pp. 33–38.

[15] A.-W. Moomen *et al.*, "Assessing the strategic applications of remote sensing for addressing illicit artisanal and small-scale gold mining activities," *GeoJournal*, vol. 89, no. 3, p. 92, 2024.

[16] A. Fonseca Gomez, "Detecting Artisanal Small-Scale Gold mines with LandTrendr multispectral and textural features at the Tapajós river basin, Brazil.," M.S. thesis, University of Twente, 2021.

[17] B. Ayo, "Integrating OpenStreetMap data and Sentinel-2 imagery for classifying and monitoring informal settlements," M.S. thesis, Universidade Nova de Lisboa, 2020.

[18] H. Li *et al.*, "Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection," *Int. Journal of Applied Earth Observation and Geoinformation*, vol. 110, p. 102 804, 2022.

[19] Earth Genome, *Gold Mine Detector*, 2025.

Automated Earth Observation Chain for Wildfire Management in Latin America and Caribbean

Mauro Arcorace[1] , Ruben Ramo[2], Adrian Vicioso[2], Alice Re[1], Fabrizio Pacini[1], Pedro Gonçalves[1], Sofia Teverovsky[1], Caterina Peris[2], and Alberto Lorenzo[2]

[1]Terradue Srl, Rome, Italy, [2]Indra Espacio, Torrejón de Ardoz, Spain

Abstract

Wildfires pose a significant threat to the Latin America and Caribbean (LAC) region, with widespread impacts on ecosystems, public health, and economic stability. The CopernicusLAC Platform delivers Earth Observation (EO) services for comprehensive wildfire management, including fire danger estimation, burned area mapping, and post-fire vegetation recovery. Integrating data from Sentinel-2, Sentinel-3, and VIIRS satellites, the platform offers near-real-time monitoring and analysis tools, supporting preparedness, impact assessment, monitoring and recovery. Demonstrated during the 2023 Chile wildfires, these services are providing timely, high-resolution insights for decision-makers in the LAC region. With growing user engagement, the platform is enhancing regional resilience and aims to scale EO-driven solutions across the region to improve wildfire preparedness and response.

***Index Terms*—** Disaster management, Wildfire, Earth Observation service, NRT monitoring, Copernicus, LAC.

1. Introduction

Wildfires do not impact only forests but have many impacts on human and environmental health, altering the composition of atmosphere, vegetation dynamics, soil erosion, water quality, with many disrupting consequences on socio-economic assets. Wildfires present a serious challenge to the Latin America and Caribbean (LAC) region, as shown in the 2023 fire events that occurred in Argentina, Bolivia, and Chile. Each year, wildfires impact over 3 million square kilometers of total land area burned only in South America (Global Wildfire Information System, 2025) [1]. Monitoring systems based on Earth Observation (EO) data are widely employed to provide timely information across the three key stages of wildfire management: prevention, response, and post-fire recovery. The Copernicus Centre for Latin America and the Caribbean (CopernicusLAC), based in Panama, serves as a regional hub for promoting the use of free and open EO data from the Copernicus missions across the LAC region, building regional capacity in EO applications, and fostering collaboration among local stakeholders and the global EO community, and for strengthening resilience to natural disasters, including wildfires. In this framework, the CopernicusLAC Platform, an exploitation platform developed by Terradue for the CopernicusLAC Centre, provides tailored EO services to address various natural hazards, including floods, wildfires, and landslides. Concerning wildfires, the CopernicusLAC Platform offers customized geospatial services for end-to-end tools for fire danger estimation, early detection, burned area mapping, severity assessment, and post-fire recovery.

2. Material

The EO services of the CopernicusLAC Platform take advantage of a robust ensemble of EO datasets and models, selected for their relevance to fire dynamics, spatiotemporal resolution and environmental context of the LAC region.

2.1. Satellite observations

Radiance Brightness Temperature from daytime and nighttime acquisitions of Sentinel-3 SLSTR data, are used primarily for thermal anomaly detection and hotspot identification over the LAC region. Time series of TOA reflectance from Sentinel-2 MSI data are used for pre- and post-event change detection analysis, focusing on areas surrounding SLSTR-detected hotspot clusters to map burned areas and assess post-fire vegetation recovery.

2.2. Detected fires and meteorological indicator

In addition to Sentinel-3 SLSTR thermal anomalies, VIIRS Suomi-NPP Near Real-Time (NRT) hotspots are also gathered from the NASA FIRMS service [2] to further complement the mapping of wildfire events. The Fire Weather Index (FWI) [3], from the Global Wildfire Information System (GWIS), is also employed in the CopernicusLAC Platform to get both historical (from 2017) and forecast (up to 10 days lead time) fire danger indicator at ~8 km resolution. The FWI is employed because it integrates weather variables including temperature, humidity, wind, and fuel moisture to classify fire danger into five levels, from low to extreme. Also, a Fire Occurrence Probability dataset derived from historical fire occurrence data, using complete time series from MODIS and VIIRS sensors (2000-2024). These dataset help identify areas with higher fire frequency in the past, which is crucial for assessing current and future fire risks.

doi:10.2760/2119408

2.3. Land cover and surface elevation

To support the vegetation-type mapping and fuel modeling, EO services rely on the Copernicus Global Land Cover (GLC) [4], which provides annual land cover maps (2015–2019) at 100 m spatial resolution, categorizing terrestrial surfaces into 23 classes based on the FAO Land Cover Classification System. To have an estimation of vegetation structure, the ETH Global Canopy Height Model (2020) [5] at 10 m resolution, is employed for biomass estimation and fire behavior modeling in forested ecosystems. Elevation data is derived from the FABDEM (Forest and Buildings removed Copernicus DEM) [6], which is a 30 m resolution global digital elevation model that excludes above-ground features, providing an accurate terrain surface for fire spread simulation.

3. Method

The CopernicusLAC platform offers to users a thematic workspace named Wildfire Results Explorer, which is a specialized web application designed for the visualization and analysis of geospatial outputs generated by three wildfire service modules: Fire Danger Mapping, Burned Area Mapping, and Post Fire Vegetation Recovery.

3.1. Fire Danger

The Fire Danger Mapping (FDM) service employs a combination of static variables (e.g., land cover, vegetation height, fire history) and dynamic inputs like the FWI from GWIS to assess localized fire danger. To generate regionally accurate fire danger indices, the service allows users to customize weights and define areas of interest.

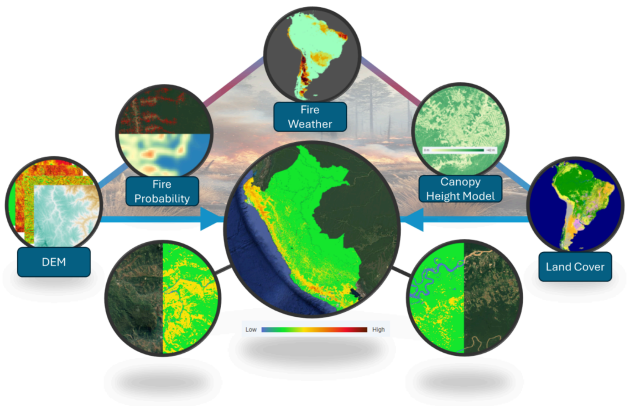


Fig. 1. High level schema describing input data and workflow for the Fire Danger Mapping service.

3.2. Wildfire Events

The Burned Area Mapping (BAM) service operates in two stages: real-time hotspot detection using Sentinel-3 thermal imagery, followed by detailed burned area and severity mapping using the Normalized Burned Ratio 2 (NBR2) (Storey et al. 2016) [7] and the Mid-Infrared Burn Index (MIRBI) (Trigg and Flasse 2001) [8] spectral indexes from Sentinel-2 MSI data and hotspots from VIIRS Suomi-NPP. When a significant density of hotspots is detected by Sentinel-3 SLSTR , the system automatically activates the ingestion of Sentinel-2 MSI L2A data acquired before and after on-going active fires. After the pre-processing of all the needed Sentinel-2 and VIIRS data, the system applies a modified version of the FireCCISFD20 algorithm (Roteta et al, 2021) [9] to map burned areas, in which the pre-fire image is replaced with a multitemporal composite to mitigate the effects of persistent cloud cover.

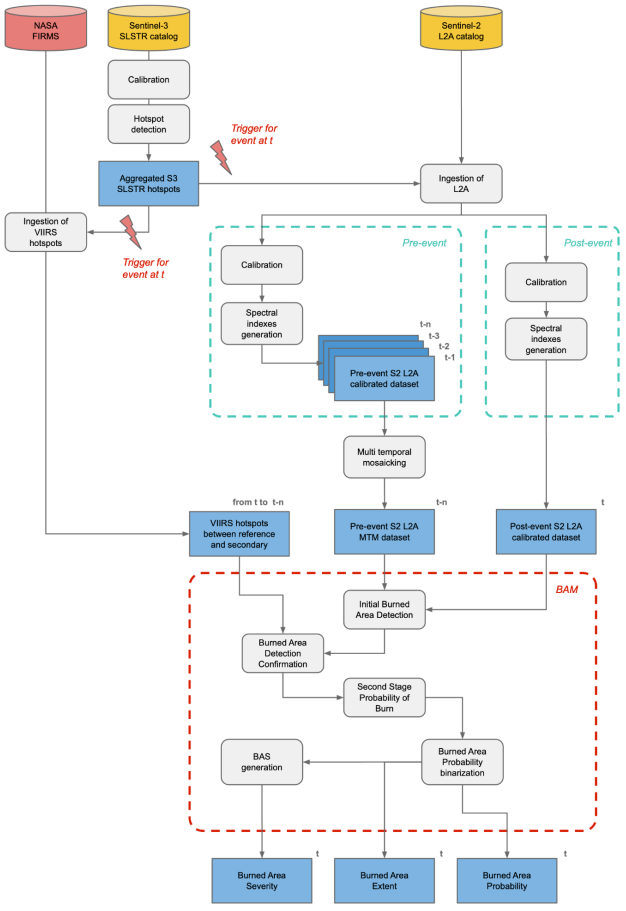


Fig. 2. Workflow of the Burned Area Mapping service.

doi:10.2760/2119408

At each overpass of the Sentinel-2 satellite the BAM algorithm delineates burnt areas using the MIRBI and the NBR2 indexes and the Scene Classification layer derived from Sen2Cor. After that, the service uses VIIRS S-NPP hotspots as ancillary information to confirm the burned areas from the spectral indexes. From this stage a statistical approach based on initially burned statistics is then applied to extract the burned area probability. Later, burned area extent maps are then derived from a thresholding and binarization of this probability. Finally for all the areas mapped as burned the burned area severity is computed, using the Relativized Burn Ratio (RBR) from Parks et al., 2014 [10]. A fully automated chain has been designed to adapt the original algorithm to a near real time execution of the service via a dedicated configuration of the servers. The chain efficiently allocates computing resources only to high-priority zones identified by clusters of aggregated hotspots. The platform offers multiple geospatial products to assess wildfire impacts and guide decision-making. This early detection capability provides a near-instantaneous view of fire activity, allowing for rapid situational awareness and initial response.

3.3. Post-fire Vegetation Recovery

Once fire activity ceases, the Fire Recovery Mapping (FRM) service automatically activates to monitor vegetation recovery. It provides a monthly Normalized Difference Vegetation Index (NDVI) composite and a recovery layer. The percentage of recovery process is derived using NDVI composites and cumulative recovery rasters generated at regular time intervals.

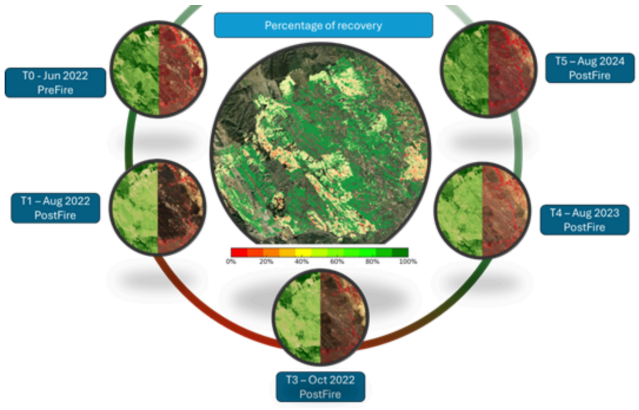


Fig. 3. Schema describing the progression from pre-fire conditions to post-fire recovery with the FRM service.

4. Results

In the CopernicusLAC platform, the FDM, BAM and FRM services are triggered over a region of interest in the LAC and by defining a date from which the wildfire monitoring shall start. Hereinafter are shown examples of results from these services derived for a large wildfire event that took place in central Chile in February 2023. EO services are pre-executed to simulate the whole event in all its stages as if it had been a near-real-time monitoring. The FDM service provides a daily estimation of fire danger from Low to Extreme, providing comprehensive hazard assessments highlighting areas at risk of fire. Fire danger categories are: Low (Minimal fire risk), Moderate (Fire conditions may ignite under certain circumstances), High (Increased likelihood of fire ignition and spread), Very High (Fires can start and spread rapidly), and Extreme (Exceptional fire danger; rapid and intense fire propagation is expected). In figure 4 is shown an example of a Fire Danger Map computed for the day after the beginning of the large wildfire event in the area of Concepción and Los Ángeles, Chile. Areas in orange indicate High fire danger meaning that fires can start and spread rapidly as the area near Los Ángeles.

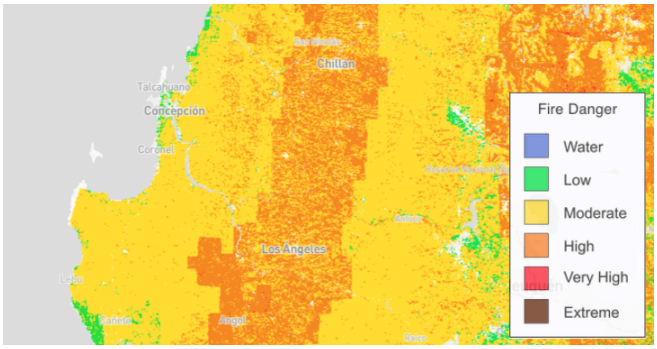


Fig. 4. Visualization of fire danger map over AOI near Los Àngeles, Chile for date 01/02/2023.

Starting from the 3rd of February 2023, the BAS service identified hotspots from Sentinel-3 SLSTS thermal anomalies in the area near Los Ángeles with a peak of active fires registered for the 7th of February 2023 which correspond to the date of the first Sentinel-2 pass over the area after the beginning of the event (see figure 5). Burned area extent and severity products are then derived by the BAM service over all the Sentinel-2 tiles intersecting hotspots, using the acquisitions of the 7th of February as post-event SWIR/NIR reflectances and multi temporal mosaics from previous Sentinel-2 acquisitions as pre-event reference ones. An example of a burned area extent product is shown in Figure 5. In red are shown burned areas, in

doi:10.2760/2119408

green burnable areas, and in grey regions where no valid observations were available for analysis. The mapping of burned areas and its severity is then updated automatically by the system at each following pass of Sentinel-2 over the same tile. The systematic monitoring with the BAM service is maintained up to 30 days after the last hotspots registered.

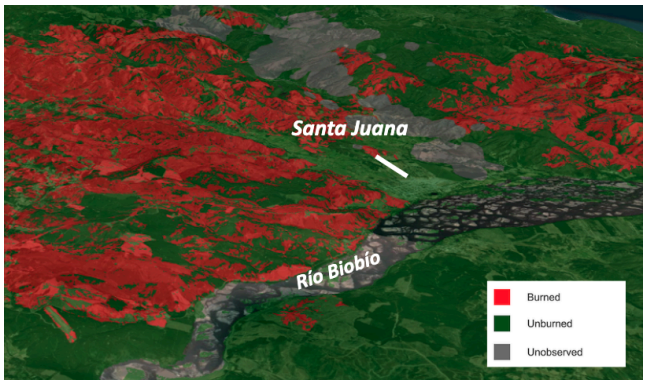


Fig. 5. 3D view of the Burned Area Map derived from Sentinel-2 L2A imagery acquired on 07 Feb 2023 near Santa Juana, Región del Biobío, Chile.

Once the monitoring of on-going wildfires with the BAM service ends, the CopernicusLAC Platform keeps ingesting Sentinel-2 L2A data and automatically triggers the FRM service to assess post-fire vegetation recovery. The FRM service takes as input Sentinel-2 imagery and the last burned areas detected from the BAM service (temporal aggregation of all burned area extents product across the event) and provides post-fire NDVI composite and vegetation recovery rates in percentage at a 15-days frequency. In figure 6 is shown an example of the FRM output map showing the recovery of vegetation in (high values in green) over burned areas after 7 months since the beginning of the event.

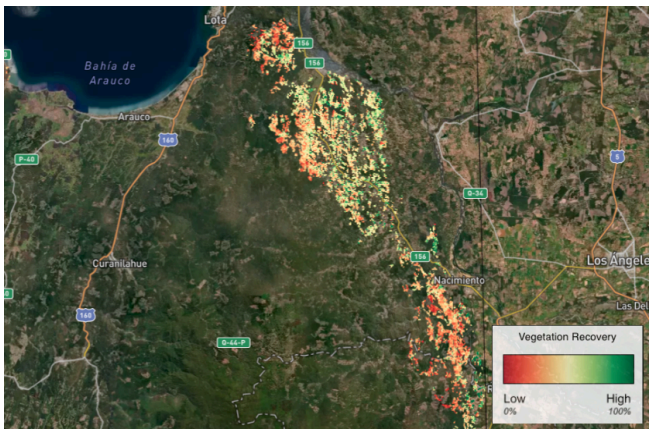


Fig. 6. Visualization of vegetation recovery map over the AOI near Los Àngeles, Chile for 31 Aug 2023.

Vegetation recovery maps measure cumulative regeneration in the area, useful to identify areas needing ecological restoration and guide long-term recovery strategies.

5. Conclusion

CopernicusLAC's integrated approach showcases the potential of EO tools in advancing wildfire management from preparedness to response and recovery. After the engagement with multiple institutions across the LAC region, the wildfire EO services have been tested by users from February to June 2025 in the CopernicusLAC platform for a NRT monitoring of Central America and Colombia. In this demonstration users had the opportunity to engage with the tools, provide feedback, and contribute to the evolution of the service. Registered users are currently 140 from national agencies and ministries of Belize, Colombia, Ecuador, Guatemala, Honduras, Panama, Peru, República Dominicana, and the Caribeean and from UN organizations. Future work will focus on further tailoring EO services to the needs of users from the LAC, strengthening regional resilience with the offering of a scalable model for wildfire management.

REFERENCES

[1] JRC Global Wildfire Information System, GWIS Statistics Portal, Share of the total land area burnt by wildfires each year, South America, 2012 to 2024. Available at: https://gwis.jrc.ec.europa.eu/apps/gwis.statistics/
[2] NASA FIRMS, VIIRS Suomi NPP Fires and Thermal Anomalies Day and Night, 375m. Available at: https://firms.modaps.eosdis.nasa.gov.
[3] JRC Global Wildfire Information System, Fire Weather Index. Available at: https://gwis.jrc.ec.europa.eu.
[4] Copernicus, Land Monitoring Service, Global Dynamic Land Cover. Available at: https://land.copernicus.eu.
[5] Lang, N., Jetz, W., Schindler, K. et al. A high-resolution canopy height model of the Earth. Nat Ecol Evol 7, 1778–1789 (2023). DOI: 10.1038/s41559-023-02206-6.
[6] Hawker, L., Neal, J., Uhe, P. F., Paulo, L., Sosa Moreno, J. E., Savage, J. T. S., and Sampson, C. "A 30m global map of elevation with forests and buildings removed." Environmental Research Letters (2022). DOI: 10.1088/1748-9326/ac4d4f.
[7] Storey E.A., Stow D.A., O'Leary J.F. (2016), "Assessing postfire recovery of chamise chaparral using multi-temporal spectral vegetation index trajectories derived from Landsat imagery". Remote Sens. Environ. 2016, 183, 53–64. DOI: 10.1016/j.rse.2016.05.018.
[8] Trigg, S.; Flasse, S., "An evaluation of different bi-spectral spaces for discriminating burned shrub savanna", Int. J. Remote Sens. 2001, 22, 2641–2647. DOI: 10.1080/01431160110053185.
[9] ESA (2021), "FireCCISFD20 Algorithm Theoretical Basis Document (ATBD) – Sentinel-2 Sub-Saharan Africa", Version 1, Nov. 3, 2021. Available at: https://climate.esa.int.
[10] Parks S. A., Dillon G. K., Miller C. A. (2014) "A New Metric for Quantifying Burn Severity: The Relativized Burn Ratio" Remote Sens. 6, no. 3: 1827-1844. DOI: 10.3390/rs6031827.

RHETICUS® SAFELAND: NEW FRONTIERS IN MULTI-RISK MANAGEMENT

Anna Sblano[1], Michele Antonicelli[1], Marina Zingaro[1], Vincenzo Laurino[1], Raffaele Borrelli[1], Daniela Drimaco[1], Khalid Tijani[2], Davide Oscar Nitti[2], Alessandro Parisi[2], Raffaele Nutricato[2], Gianvito Brandonisio[3], Vincenzo Massimi[1]

[1]Planetek Italia, 70132 Bari, Italy
[2]GAP srl c/o Department of Physics "M. Merlin", University of Bari, 70125 Bari, Italy
[3]Sfera Informatica & Strumentazione SRL,70026 Modugno,Italy

ABSTRACT

Slope instability and forest fires are growing threats to communities, infrastructure and ecosystems, exacerbated by climate change and human activity . Effective management of these phenomena requires integrated monitoring and risk analysis solutions based on up-to-date, interoperable data that can support timely decision-making. In this framework, Rheticus® Safeland 2.0, developed by Planetek Italia as part of the PNRR and AI-LAND projects,integrates high-performance computing, big data analysis and 3D digital models to process multi-thematic information on landslides and fires. The system identifies critical areas and levels of attention by combining evolutionary scenarios and interferometric data, providing a dynamic framework for defining mitigation and intervention strategies. Thus, Safeland 2.0 is configured as an operational tool for proactive land management, in line with the Sustainable Development Goals of Agenda 2030 .

Index Terms— Landslides, fires, multi-risk, monitoring,digital twin

1. INTRODUCTION

Rheticus® Safeland, developed by Planetek Italia, is a land stability monitoring and warning service that, automatically assigns an attention level to each segment of the territory. This classification is based on the analysis of trends and anomalies in ground surface displacements detected via satellite remote sensing. The service offers a synoptic view of the territory, providing constantly updated information on the attention level, distinguishing between stable areas and areas exhibiting signs of landslides or slow subsidence, and thus complements in situ monitoring activities.

Over the years, the Rheticus® Safeland has involved into a multi-risk analysis platform. While in a first version the service provided the user with a classification of the territory based only on interferometric data, the service now allows the user to obtain a more complete picture of the territory by integrating the aforementioned interferometric data with additional information layers considered essential for the identification of areas at greater hydrogeomorphological and fire risk. In particular, new parameters have been defined to improve the performance of multi-temporal interferometric analyses (MTInSAR). The results of the MTInSAR analyses were integrated with auxiliary data necessary for hydro-geomorphological and fire risk characterization, including fire risk, hydraulic hazard and landslide hazard maps. This integration was performed using the Ordered Weighted Averaging (OWA) methodology. This procedure provides an Inspection Priority Score (IPS) for each hexagonal cell: a continuous value between 0 and 1 that quantifies the level of attention required. The IPS enables the territory to be categorized into one of three risk classes: low, medium, or high. The final result is integrated into an interactive, 3D digital twin environment where users can explore the data inputs used to define the IPS, as well as data relating to changes in land cover derived from a machine learning algorithm applied to Sentinel-2 optical satellite data, thematic maps, geological and infrastructure data, climate information and historical movement data. This supports timely diagnoses and targeted interventions. This evolution of the service represents a quantum leap towards a truly operational, multi-risk system capable of providing a holistic, dynamic view of the territory. Compared to the initial version, the new tool enhances the accuracy, predictive capacity and automation of the process, thereby improving territorial resilience and the effectiveness of mitigation strategies.

2. RHETICUS® SAFELAND 2.0: SERVICE DESCRIPTION

The platform interface, as shown in the figure, provides for territorial segmentation based on a grid of hexagonal cells, each covering approximately 5 hectares. These cells are classified by automatic procedures based on analysis that evaluates movement trends and anomaly patterns. The level of attention is determined by combining surface displacement measurements with other parameters that take into account ar-

doi:10.2760/2119408

eas affected by fires, areas at risk of flooding and landslides, and slopes calculated by the DTM. The status of the cells is represented by traffic light colors: green (stable), yellow (requires monitoring), and red (requires immediate inspection) and summarized in a pie chart at the top right of the platform that shows the number and percentage of cells for each color. This color is derived from an inspection priority score associated with each cell of the AOI, which can vary between 0 and 1, also shown in a curve that evaluates the IPS trend of the cells within the platform and is used to classify the urgency of the inspection (Fig. 1). Each cell can be queried and provides information such as:

- Unique cell identifier
- Geographical coordinates (latitude, longitude)
- Municipality and Province
- Presence of transport infrastructure
- Prevailing land cover
- Any changes in land cover
- Area of spatial anomaly cluster
- Area of spatio-temporal anomaly cluster
- Area of burned area
- Elevation and class of slope
- Attention level class
- Inspection Priority Score
- Mean velocity ascending and descending
- Highest hydraulic and landslide hazard values
- Predominant geological cover
- Climate data
- Types of landslide

The service enables end users to explore an area of interest in detail by integrating various customization information layers that can be activated or deactivated directly on the map. These include cell thematisation based on the Inspection Priority Score (IPS), visualization of spatial and spatio-temporal anomaly clusters, ground movement data (PS/DS), and information derived from OpenStreetMap, such as buildings, roads, and railways. Other available features include geological mapping, burned areas, land cover and land use change maps, ERA5 climate data, hydraulic and landslide hazard maps (PAI) and the IFFI inventory.

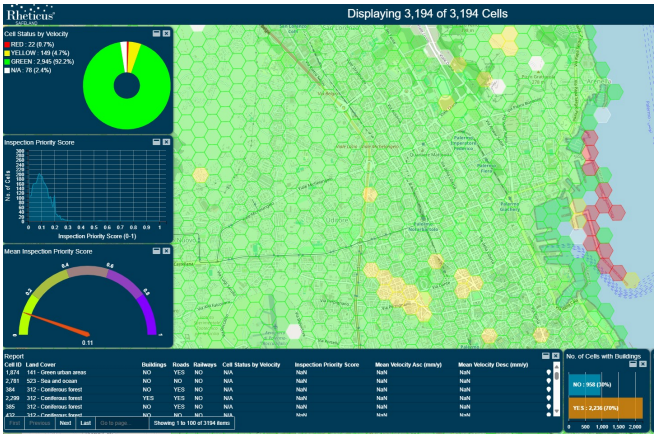


Fig. 1. Rheticus® Safeland Interface

3. CELL RANKING METHODOLOGY TO ESTIMATE THE LEVEL OF ATTENTION (INSPECTION PRIORITY SCORE)

The new version of the Rheticus® Safeland service, which was developed as part of the AI-LAND project, allows users to access a comprehensive and current overview of the state of the territory, classified by level of risk. Unlike the previous version, the system now integrates additional layers of information, as well as interferometric data, to identify areas exposed to greater hydrogeological and fire risk. The territory is divided into hexagonal cells and classified according to the IPS calculation, which assigns each cell a level of attention according to the hazards detected in the area of interest. IPS is determined through a combined analysis of weights derived from different geospatial data sources and amplification factors, enabling areas at higher risk of instability to be identified. The IPS calculation takes into account three main types of data and their respective weights.

- *Spatial anomalies:* The identification of unstable areas is based on the analysis of mean displacement velocity values of Persistent Scatterers (PS) and Distributed Scatterers (DS), measured along the Line of Sight (LOS) and extracted from the EGMS geoportal. For each acquisition geometry, PS and DS exhibiting homogeneous kinematic behavior are grouped into clusters. Each cluster is assigned a unique identifier and is spatially represented by an enclosing polygon;
  The weight assigned to clusters of spatial anomalies in the Inspection Priority Score (IPS) calculation algorithm is determined based on two main parameters: consistency, defined as the ratio between the number of PS/DS points within the cluster (excluding outliers) and the total number of PS/DS in the cluster, and the average speed of movement calculated along the LOS. The weight of the spatial anomalies of each hexagon is calculated as the sum of the weights of the individ-

doi:10.2760/2119408

ual clusters corrected for the fraction of area that each cluster occupies within the hexagon itself.

- *Spatio-temporal anomalies*: Identified by the analysis of displacement time series of PS and DS along the LOS, extracted from the EGMS geoportal. For each acquisition geometry, the first step involves detecting temporal anomalies by identifying breakpoints that appear in the most recent segment of the time series. These breakpoints represent either changes in displacement velocity or abrupt shifts (also referred to as steps), which are temporally well-defined and easily identifiable. Following this, a clustering process is carried out, grouping together PS and DS points that exhibit similar kinematic behavior. Each resulting cluster, characterized by shared motion patterns, is assigned a unique identifier and is spatially represented by an enclosing polygon;
  The contribution of space-time anomaly clusters in the Inspection Priority Score (IPS) calculation algorithm is determined by combining three main parameters: consistency, calculated as the ratio between the number of valid PS/DS points within the cluster, excluding outliers, and the total number of PS/DS points in the cluster, the difference in average speed of movement before and after a trend change, and the displacement step, which represents the significant discontinuity of movement between two consecutive dates. The total weight assigned to spatio-temporal anomalies in each hexagon is calculated as the sum of the weights of the individual clusters, each corrected according to the portion of area actually covered by the cluster within the hexagon itself.

- *Isolated PS/DS, i.e., measurement points that do not fall within the clusters of anomalies*: PS/DS that do not generate spatial/spatio-temporal anomalies. It is planned to use the displacement time series obtained along the LOS and extracted from the EGMS geoportal [1]. The weight of the PS/DS within the algorithm is calculated by taking into account both the average speed of movement and the density of points within the hexagon. This means that areas affected by significant point movements, even if they are not characterised by aggregate anomalies, are adequately considered in the classification process.The weight of the PS/DS within the algorithm is calculated by taking into account both the average speed of movement and the density of points within the hexagon. This means that areas affected by significant point movements, even if they are not characterised by aggregate anomalies, are adequately considered in the classification process.

In addition to ground motion data, the system considers a series of amplification factors that contribute to the prioritization of cell attention levels:

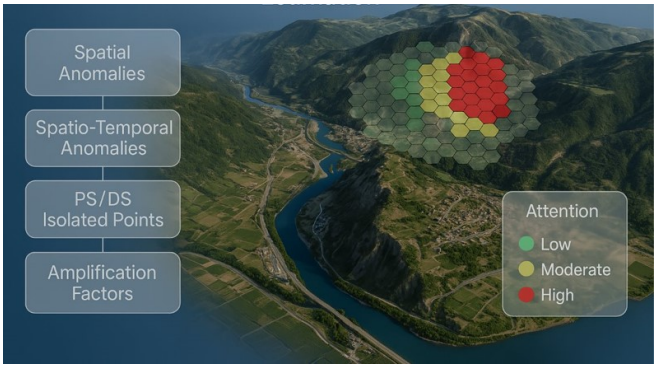


Fig. 2. 3D visualization system

- *Landslide Hazard:* The landslide hazard layer produced by the Istituto Superiore per la Protezione e Ricerca Ambientale (ISPRA) represents a mapping of the areas of the Italian territory exposed to different landslide hazard classes. The classification is based on geomorphological, hydrogeological and historical criteria, which take into account both already occurred landslide events and the susceptibility of the territory to instability phenomena [2].

- *Hydraulic Hazard:* The hydraulic hazard layer produced by the Istituto Superiore per la Protezione e Ricerca Ambientale (ISPRA) represents a mapping of the areas of the Italian territory exposed to different landslide hazard classes [2].

  In the Inspection Priority Score calculation algorithm, each landslide e hydraulic hazard class is assigned an increasing weight based on the level of risk attributed to it. This weight is proportionate to the percentage of the specific class's surface area that falls within each analysis hexagon. If there are multiple hazard classes within a single hexagon, the contributions are added together to obtain the overall value of the amplification factor associated with the landslide and hydraulic risk.

- *Burned Areas:* The burned area layer, produced as part of the European Forest Fire Information System (EFFIS), is a Europe-wide mapping of areas affected by forest fires. This system monitors and records areas affected by fires using high-resolution satellite imagery and automated detection algorithms [3]; Several parameters are considered when calculating the weight associated with forest fires in each hexagon: the percentage of burned area relative to the hexagon area, the average slope of the hexagon, and the time interval since the fire event. The latter parameter is used to modulate the influence of fires based on their date of occurrence. If there are multiple events in the same hexagon, the contributions are combined, taking into

doi:10.2760/2119408

account any spatial overlaps.

The final weight of each cell is calculated by aggregating the normalised contributions of movement data and amplification factors, with a combination of predefined weights applied. This process uses the Ordered Weighted Averaging (OWA) methodology [4], which is a technique that allows multiple factors to be aggregated while taking into account a flexible weighting criterion. OWA assigns weights to different levels of information based on their relevance in determining a cell's attention level. The customer can customise this methodology based on their knowledge of the territory , and an expert user can monitor it. It enables the influence of each contribution to be modulated via a weighting function, thereby ensuring a balanced approach between different data sources. The IPS is therefore an optimal combination of the different contributions. The resulting IPS value is then used to classify cells into one of three attention levels: i)Low (green): stable,no intervention needed; ii)Moderate (yellow): Potential instability, monitoring advised; iii)High (red): High-risk, inspection required.

The final output of the algorithm, the Inspection Priority Score (IPS), was validated by comparing it with available ground truth data, including official landslide inventories, fire records and other documented evidence of instability. Validation was also conducted in collaboration with local authorities, such as the Civil Protection Agency, the Forestry Police and relevant regional structures, to verify consistency between areas classified as highest priority and actual situations on the ground. This iterative calibration process strengthened the model's operational reliability and usefulness for decision support.

In this context, geo-analytical indicators generated by the Rheticus® Safeland 2.0 service, are integrated into an interactive 3D environment.The 3D Digital Twin model, which has been developed, integrates geospatial data, three-dimensional models and high-density point clouds in order to reconstruct accurate and interactive virtual environments. Its architecture enables complex scenarios to be visualised by combining satellite surveys, digital terrain models (DTMs), Light Detection and Ranging (LIDAR) and photogrammetry, and supports dynamic data to simulate evolutionary phenomena. Optimised rendering algorithms, level of detail (LOD) management and progressive streaming techniques ensure high performance, even with large datasets. The platform supports immersive navigation, contextual information overlay and real-time, multi-user interaction, making it an advanced tool for environmental analysis, planning and risk scenario management.(Fig. 2).

4. CONCLUSION

Rheticus® Safeland 2.0 is a significant methodological advance in integrated hydrogeological risk and forest fire analysis. This enhanced system combines multi-temporal interferometric observations with auxiliary data layers, including topographic, climatic and thematic mapping information, to provide a thorough classification of instability conditions. Its hexagonal cell structure enables consistent and scalable spatial analysis, and the Inspection Priority Score (IPS) algorithm facilitates the early identification of critical areas and the transparent and reproducible establishment of intervention priorities. The interconnected nature of these risks means that forest management directly affects erosion control and slope stability, linking fires and landslides in multi-hazard scenarios. The system's output, integrated into a three-dimensional digital twin model, enables current conditions and evolutionary scenarios to be visualised. Continuous monitoring enables previously undetected areas of instability to be identified and tracked over time, thereby strengthening the effectiveness of mitigation strategies and risk reduction planning in line with resilience and sustainable land management objectives.

5. ACKNOWLEDGEMENT

This activity was carried out within the AI-LAND project, under Spoke 5 "Environment Natural Disasters" of the National Centre for HPC, Big Data and Quantum Computing (Project Code CN000013, CUP H93C22000450007), funded by the European Union – NextGenerationEU within Italy's National Recovery and Resilience Plan (PNRR), Mission 4 "Education and Research" – Component 2 "From Research to Enterprise" – Investment 1.4, under D.R. no. 4185 of 15/11/2024 and related annexes, following the Cascade Call issued by D.R. no. 1203 of 29/03/2024.

REFERENCES

[1] Copernicus Land Monitoring Service, "European Ground Motion Service (EU-GMS) - A proposed Copernicus service element", 2021.

[2] A. Trigila, C. Iadanza, B. Lastoria, M. Bussettini, A. Barbano, "Dissesto idrogeologico in Italia: pericolosità e indicatori di rischio – Edizione 2021", 2021.

[3] J. San-Miguel-Ayanz, E. Schulte, G. Schmuck, A. Camia, P. Strobl, G. Liberta, C. Giovando, R. Boca, F. Sedano, P. Kempeneers, D. McInerney, C. Withmore, S. Santos de Oliveira, M. Rodrigues, T. Durrant, P. Corti, F. Oehler, L. Vilar, G. Amatulli, "Comprehensive Monitoring of Wildfires in Europe: The European Forest Fire Information System (EFFIS)", Approaches to Managing Disaster, doi = 10.5772/28441, 2012.

[4] F.V. Ronco,A. Fiore,G. Romani,G.F. Ricci,G. Nolè,F. Gentile, "Multi-criteria decision analysis for monitoring and evaluating soil erosion risk in forest fire-affected areas", Journal of Environmental Management, doi://doi.org/10.1016/j.jenvman.2024.123672, 2025.

GNEO AXIS 3 SAFETY AND SECURITY SERVICE: A GREEK EO-BASED SERVICE SUPPORTING DISASTER RESPONSE AND SECURITY

G. Benekos[1], D. Grigoriadis[1], Th. Valsamidis[1], S. Touloumtzi[2], M. Sdraka[2], I. Papoutsis[2], G. Papadopoulos[3], D.E. Argiropoulos[3], P. Trahanias[3], Th. Giannaros[4], K. Lagouvardos[4], V. Kotroni[4], D. Bliziotis[5], K. Kikaki[5], K. Karantzalos[6]

[1] Planetek Hellas, [2] National Technical University of Athens, [3] Foundation for Research and Technology - Hellas, [4] National Observatory of Athens, [5] Hellenic Space Center [6] Secretary General of Telecommunications & Post, Ministry of Digital Governance

ABSTRACT

The AXIS 3 Safety & Security Service, part of the Greek National Satellite Space Project, is a modular and scalable Earth Observation based solution designed to support near real-time environmental risk monitoring and security operations. This paper presents the system architecture, service components, and integration with the Governmental Hub and EOEPCA processing ecosystem. Through dedicated services for flood risk, wildfire surveillance, and target identification, the AXIS 3 platform delivers timely geospatial insights using SAR, Optical, and Thermal satellite data.

Index Terms— Safety, Security, Earth Observation, Satellite Remote Sensing, Geospatial Products, STAC, EOEPCA

1. INTRODUCTION

The aim of the project is to enhance Greek capabilities in satellite technologies and applications and empower the country to exchange satellite data. AXIS 3 addresses the need for a national Earth Observation (EO)-based infrastructure delivering timely, reliable, and operational geospatial products. It leverages national (AXIS 1.1, 1.2, 2.0) and international assets (e.g., Copernicus, Landsat) to support disaster response, security, and environmental protection. AXIS 1.1 provides thermal infrared imagery for environmental monitoring, AXIS 1.2 offers all-weather SAR imaging for maritime and land surveillance, and AXIS 2.0 delivers high-resolution multispectral and hyperspectral data. The Service designs, develops, validates, integrates, and delivers the Safety and Security Service of AXIS 3, including external data integration. It establishes an end-to-end service chain—from data collection to geospatial product delivery—along with user tools. The goal is to meet national needs for spatial, temporal, and thematic resolution using AXIS satellite data. These needs shape software requirements, resulting in value-added products like thematic maps, insights, and user-friendly services with timely and systematic delivery. Big EO data is handled via batch processing and parallel computing to manage volumes efficiently and accurately.

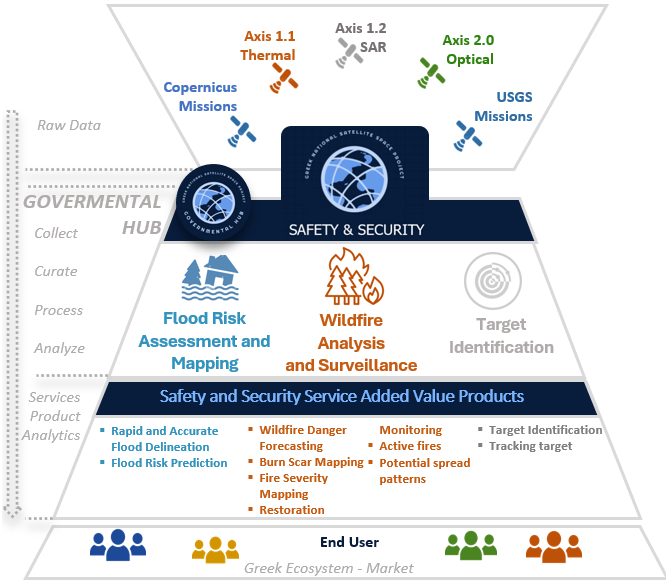


Figure. 1. AXIS 3 Safety & Security Service High-level schematic representation

2. ARCHITECTURE & DESIGN PRINCIPLES

The AXIS 3 Safety & Security Service is built on a modular, scalable architecture, supporting the full EO data lifecycle—from acquisition to analysis and delivery. Each function is encapsulated in a dedicated module, allowing independent development, testing, and scaling without disrupting the rest of the system. The architecture offers key benefits: scalability (modules like ingestion or training can grow as needed); maintainability (updates in one area don't impact others); and reusability (core components adapt across services). It also ensures interoperability for easy integration with national and international EO infrastructures (e.g. AXIS 3 Governmental Hub, HSC Data Hub).

doi:10.2760/2119408

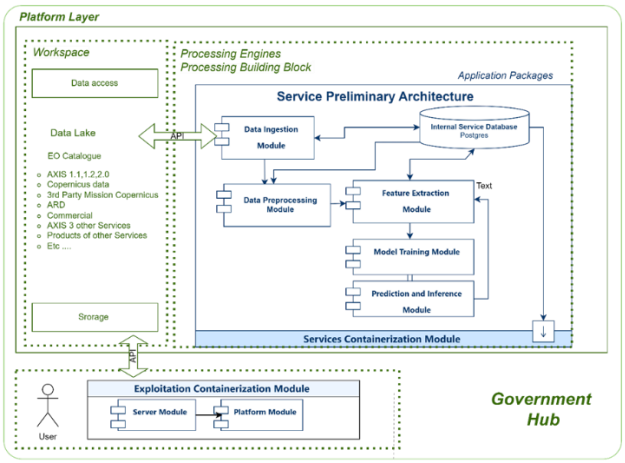


Figure . 2. Service Architecture for Governmental Hub Integration

The Server Services Module forms the processing core and includes: Data Ingestion Module - Collects EO data from AXIS satellites and sources like Copernicus, supporting standard geospatial formats and APIs for compatibility and efficient flow; Data Preprocessing Module - Cleans, filters, and normalizes data, ensuring sensor consistency and filling gaps by aligning third-party datasets to the national mission; Feature Extraction Module - Converts cleaned data into structured features using statistical and geospatial algorithms, preparing outputs for modeling or decision tools. Model Training Module: Trains AI models with EO features, supporting hyperparameter tuning, validation, and transfer learning to optimize models for tasks like flood risk or wildfire forecasting. Prediction and Inference Module: Applies models to new data for batch or real-time insights, supporting services like target tracking and disaster alerts with robust scalability. Each thematic service -Flood Risk Assessment, Wildfire Monitoring, Target Identification- operates as an independent pipeline while sharing a unified backend, following a common flow: ingest → preprocess → extract → train → predict. The use of containers and metadata standards ensure smooth operation with platforms like the Governmental Hub and Earth Observation Exploitation Platform Common Architecture (EOEPCA)[1].

3. THEMATIC EO SAFETY & SECURITY SERVICE IMPLEMENTATIONS

The AXIS 3 Safety & Security Service implements a suite of EO services, built on a modular architecture. These services deliver actionable insights for risk mitigation and security operations across Greece. Each pipeline integrates data ingestion, preprocessing, model-based processing, and product delivery, tailored to its thematic objective.

3.1. Flood Risk Assessment and Mapping

3.1.1. Rapid and Accurate Flood Delineation

Flood extents and water depth are detected using SAR (Sentinel-1, AXIS 1.2) and optical (Sentinel-2, AXIS 2.0) imagery. A Multi-Otsu's thresholding method [2,3] classifies flooded areas, and the FLEXTH tool [4] estimates water depth. Optical data preprocessing includes NDWI and cloud masking [5]. For SAR data, calibration, speckle filtering (Lee filter), terrain correction, and dB scaling [6] are applied.

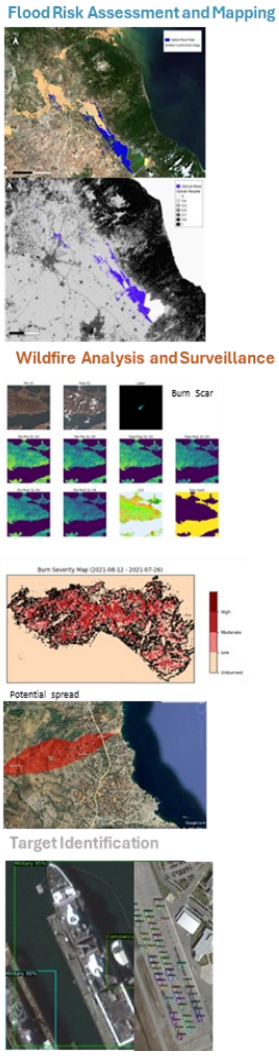


Figure 3 Preliminary results of Safety & Security Services

3.1.2. Flood Risk Prediction

Flood risk is predicted using the LISFLOOD model [7], simulating runoff and hydrological processes across Greece. Inputs include meteorological data (precipitation, temperature, evapotranspiration) and static maps (land use, soil, topography). The model generates daily or weekly risk maps for flood response and mitigation planning.

3.2. Wildfire Analysis and Surveillance

3.2.1. Wildfire Danger Forecasting and Fire Risk

This service predicts fire risk up to three days in advance by integrating satellite data with environmental and human variables. Machine learning models [8] trained on meteorological, fuel, and terrain data produce fire danger maps, anomaly reports, and explainable AI visualizations to support early warnings.

3.2.2. Burn Scar Mapping

Deep Learning models using the FLOGA dataset [9] map burn scars from multispectral and SAR imagery. The system provides accurate post-fire damage assessments, even in inaccessible or cloud-covered areas, aiding recovery, land use, and fire mitigation strategies.

doi:10.2760/2119408

3.2.3. Fire Severity Mapping

This service estimates fire damage intensity, supporting ecological assessments and recovery efforts. The GeoCBI algorithm [10] classifies fire severity using multispectral satellite data, informing reforestation, soil stabilization, and ecosystem resilience planning.

3.2.4. Restoration and Regeneration Monitoring

This service monitors vegetation regrowth across biomes in Greece, using Copernicus and AXIS multispectral data, fire severity, and forest type maps. A remote sensing index [11] classifies recovery status, supporting restoration planning.

3.2.5. Active Fire Monitoring

This module detects active fires in near real-time using thermal-infrared satellite data and EO analytics. It tracks fire fronts and hotspots, helping emergency services optimize suppression strategies and protect critical infrastructure.

3.2.6. Fire Spread and Intensity Forecasting

This service uses the WRF-SFIRE model [12] to predict wildfire spread and intensity, simulating fire-atmosphere interactions. It integrates inputs like fuel moisture, weather forecasts, and terrain to guide suppression tactics and communication.

3.3. Target Identification and Tracking

3.3.1. Target Identification

Machine learning models, primarily Detectron2 [13], detect and classify vessels, aircraft, vehicles, and infrastructure using SAR and optical imagery. Each object category uses dedicated models, with land and sea masks reducing false positives. Infrastructure detection uses only optical data due to SAR complexity.

3.3.2. Tracking Target Identification

Object tracking uses multi-frame spatial correlation and optical flow-based tracking [14]. Multi-SAR (e.g., AXIS 1.2 dwell mode) and multi-optical (e.g., AXIS 2) inputs derive object trajectories. Outputs include LineString paths and velocity metrics for dynamic monitoring.

4. DATA REQUIREMENTS AND INGESTION STRATEGY

The Axis 3 Safety & Security Services ingest a comprehensive and diverse range of satellite and auxiliary datasets to support advanced Earth observation applications across flood risk, wildfire monitoring, and target identification domains. Inputs include multispectral and SAR imagery from Sentinel-1, Sentinel-2, Axis 1.2, Axis 2.0, ICEYE, WorldView, and PlanetScope satellites, as well as terrain models like Copernicus DEM, land cover maps, meteorological time series, and thematic data such as fire hotspots or hydrological parameters.

The Axis 3 Governmental Hub leverages a standardized and scalable data ingestion strategy based on the SpatioTemporal Asset Catalog (STAC) API [15], ensuring that all core datasets ingested into the Hub's internal catalog are STAC-compliant. Conversely, external datasets not registered in the catalogue (e.g., real-time weather feeds, environmental models, hydrological boundaries, and auxiliary layers like CORINE or WorldPop) are accessed through dedicated interfaces or treated as static support layers used exclusively during processing workflows, without direct ingestion into the catalogue.

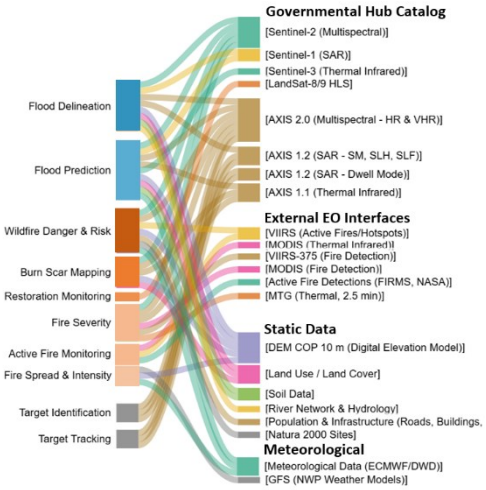


Figure 4 Data Ingestion of various datasets

5. SERVICES OPERATIONAL PROCESS

Figure 5 Operational modes of all services

The operational processes of Axis 3 Safety & Security Services support both systematic monitoring and on-demand response. Core services like Flood Risk Prediction, Wildfire Forecasting, Burn Scar Mapping, and Active Fires operate systematically with predefined cycles (e.g., daily, monthly/ weekly). Services like Rapid Flood Delineation and Target Tracking are triggered on-demand for post-event monitoring or specific AOI operations. Services such as Fire Spread and Intensity Forecasts react to detected events, balancing continuous service with emergency response agility through tasking and dynamic data processing.

doi:10.2760/2119408

6. SERVICES CURATION FOR GREEK GOVERNMENTAL HUB PROCESSING BUILDING BLOCK

Each service—flood mapping, wildfire monitoring, and target identification—is packaged in Docker containers for scalability across Kubernetes clusters. Common Workflow Language (CWL) is used for workflow orchestration, enabling integration and traceability. The application includes a container and a .cwl file defining inputs and execution, ensuring reproducibility across cloud or distributed infrastructures [16].

7. CONCLUSION

The AXIS 3 Safety & Security Service demonstrates the effectiveness of a modular Earth Observation system in meeting national needs for disaster management, environmental monitoring and security. Its architecture enables systematic geospatial service delivery while adapting to time-sensitive and event-driven demands.

Integrating national and international satellite data—such as the national space assets, Copernicus, and other sensors—ensures broad spatial and temporal coverage. The service, based on standardized interfaces and harmonized processing chains, facilitates interoperability across data and system components.

The successful deployment of thematic pipelines for flood, wildfire, and target identification confirms that a well-structured EO system can support these applications. Implemented with containerized modules and orchestrated via CWL, these workflows offer a scalable framework for deployment within the EOEPCA ecosystem.

The service also addresses Big Earth Observation Data challenges by using parallel processing, batch execution, and scalable cloud-native architectures, ensuring efficiency and performance under high-frequency data loads—critical for both systematic and real-time responses.

Overall, the Safety & Security Service demonstrates how a structured EO system supports decision-making and enhances national readiness.

REFERENCES

[1] EOEPCA, "EOEPCA Documentation," [Online]. Available: https://eoepca.readthedocs.io/. [Accessed: 24-Apr-2025].

[2] R. Yadav, "Supervised and Unsupervised Deep Learning Models for Flood Detection", 2023, KTH Royal Institute of Technology, TRITA-ABE-DLT, 2344, 2023

[3] GeoNext, "PyImgProc-1: Waterbodies Extraction using Entropy and Otsu's Threshold," 2024, [Online]. Available: PyImgProc-1: Waterbodies Extraction using Entropy and Otsu's Threshold | by GeoNext | Medium . [Accessed: 24-Apr-2025].

[4] A. Betterle and P. Salamon, "Water depth estimate and flood extent enhancement for satellite-based inundation maps," Feb. 06, 2024, Hydrological Hazards. doi: 10.5194/nhess-2024-22

[5] F. Foroughnia, S. M. Alfieri, M. Menenti, and R. Lindenbergh, "Evaluation of SAR and Optical Data for Flood Delineation Using Supervised and Unsupervised Classification," Remote Sens., vol. 14, no. 15, p. 3718, Aug. 2022, doi: 10.3390/rs14153718.

[6] M. Vassileva, A. Nascetti, F. GiulioTonolo, and P. Boccardo, "Unsupervised flood extent detection from SAR imagery applying shadow filtering from SAR simulated image," in 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy: IEEE, Jul. 2015, pp. 2707–2710. doi: 10.1109/IGARSS.2015.7326372

[7] LISFLOOD, "LISFLOOD hydrological model", Modelling Inventory and Knowledge Management System of the European Commission (MIDAS) [Online] Available : https://web.jrc.ec.europa.eu/policy-model-inventory/explore/models/model-lisflood/ [Accessed 25-Apr-2025]

[8] S. Kondylatos, I. Prapas, M. Ronco, I. Papoutsis, G. Camps-Valls, M. Piles, M.A. Fernandez-Torres, N. Carvalhais., "Wildfire Danger Prediction and Understanding With Deep Learning," Geophys. Res. Lett., vol. 49, no. 17, Art. no. 17, 2022, doi: 10.1029/2022GL099368.

[9] M. Sdraka, A. Dimakos, A. Malounis, Z. Ntasiou, K. Karantzalos, D. Michail, "FLOGA: A Machine-Learning-Ready Dataset, a Benchmark, and a Novel Deep Learning Model for Burnt Area Mapping With Sentinel-2," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 7801-7824, 2024, doi: 10.1109/JSTARS.2024.3381737.

[10] A. De Santis and E. Chuvieco, "GeoCBI: A modified version of the Composite Burn Index for the initial assessment of the short-term burn severity from remotely sensed data", 2009, Remote sensing of Environment, 113(3), 554-562. doi: https://doi.org/10.1016/j.rse.2008.10.011

[11] R. Gibson, L. White, S. Hislop, R. Nolan, and J.Dorrough, "The post-fire stability index; a new approach to monitoring post-fire recovery by satellite imagery, Remote Sensing of Environment, Volume 280, 2022, 113151, ISSN 0034-4257, doi: 10.1016/j.rse.2022.113151

[12] W. C. Skamarock, J.B. Klemp, J. Dudhia, D.O. Gill, Z. Liu, J. Berner, W. Wang, J.G. Powers, M.G. Duda, D.M Barker, X.Y. Huang, "A Description of the Advanced Research WRF Model Version 4," UCAR/NCAR, Mar. 2019. doi: 10.5065/1DFH-6P97

[13] T. Hoeser and C. Kuenzer, "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends," Remote Sens., vol. 12, no. 10, p. 1667, May 2020, doi: 10.3390/rs12101667.

[14] B. Du, S. Cai, and C. Wu, "Object Tracking in Satellite Videos Based on a Multiframe Optical Flow Tracker," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 12, no. 8, pp. 3043–3055, Aug. 2019, doi: 10.1109/JSTARS.2019.2917703

[15] STAC, "SpatioTemporal Asset Catalogs," [Online]. Available: https://stacspec.org/en. [Accessed: 24-Apr-2025].

[16] OGC Best Practice for Earth Observation Application Package, 2021. Available: http://www.opengis.net/doc/BP/eoap/1.0 [Accessed: 24-Apr-2025]

ACKNOWLEDGEMENTS - DISCLAIMERS

"The project is being carried out under an ESA Contract in the frame of the Greek National Satellite Space Project. The Project: Small-Satellites (Measure ID 16855) is implemented by the Hellenic Ministry of Digital Governance with the European Space Agency (ESA) Assistance in the Management and Implementation. The project is part of the National Recovery and Resilience Plan 'Greece 2.0', which is funded by the Recovery and Resilience Facility (RRF), core programme of the European Union-NextGenerationEU"

"Views expressed herein can in no way be taken to reflect the official opinion of the European Union/European Commission/European Space Agency/ Greek Ministry of Digital Governance. Views and opinions expressed are those of the author(s) only and the European Union/European Commission/European Space Agency/ Greek Ministry of Digital Governance, cannot be held responsible for any use which may be made of the information contained therein".

AN INTEGRATED APPROACH FOR ASTEROID IMPACT PREDICTION AND TRAJECTORY VISUALIZATION

Emine Betul Erdogan, Gokhan Bakal

Abdullah Gul University, Türkiye

ABSTRACT

Asteroid collision prediction plays a pivotal role in planetary defence by enabling proactive risk mitigation and informed strategic planning. To address this challenge, we developed a comprehensive framework that integrates historical fireball data and contemporary orbital parameters from NASA's datasets. By deriving shared physical features and applying unsupervised clustering, our system identifies patterns in potential impact scenarios. We further incorporate supervised learning to categorize asteroids based on their threat level. To enhance accessibility and interpretation, the framework includes multi-dimensional visualizations of orbital dynamics and an interactive web application that represents asteroid trajectories in both two and three dimensions. This simulation platform serves both scientific and educational purposes, offering a rich interface for exploring asteroid behaviour. This study demonstrates the potential of combining machine learning, astrophysical modelling, and data visualization to support planetary safety initiatives.

Index Terms— Asteroid prediction, planetary defence, K-means clustering, Random Forest classification, orbit visualization

1. INTRODUCTION

Asteroid monitoring plays a critical role in understanding and mitigating the risks posed by near-Earth objects. As interest in planetary defence continues to grow globally, space agencies and researchers face increasing demands for accurate forecasting systems that can assess potential collision threats in a timely manner. Similarly, Malakouti et al. (2023) highlighted the value of machine learning classification for hazardous asteroid identification [2]. The combination of expanding public datasets and the growing accessibility of computational tools provides new opportunities to model and predict celestial behaviour with greater precision.

In recent years, As Chomette et al. (2024) demonstrated, machine learning significantly improved local asteroid damage prediction and as advancements in data-driven methodologies have enabled researchers to move beyond traditional orbit tracking by incorporating machine learning and simulation techniques into asteroid impact assessment frameworks [1]. Such approaches are essential not only for identifying potentially hazardous asteroids (PHAs) but also for supporting strategic decision-making processes in defence and space governance contexts.

To respond to these emerging requirements, we developed a comprehensive framework that leverages both historical fireball data and real-time orbital records from NASA's databases. The system integrates clustering, classification, and simulation methods to create a unified tool for detecting and visualizing asteroid threats. Section 2 outlines the input data sources, methodology, and machine-learning models employed. Section 3 introduces the simulation platform and technical implementation. Results, insights, and potential applications are presented in the concluding sections.

2. DATA & METHODS

This section presents the data sources, preprocessing strategies, and modelling architecture developed to detect potentially hazardous asteroids and simulate their future trajectories. The workflow combines historical impact records with real-time orbital datasets and integrates unsupervised and supervised machine learning methods for collision risk classification and spatial-temporal modelling.

2.1. Input data

The framework uses two complementary data sources. NASA's CNEOS Fireball and Bolide Data provides an essential dataset for historical impact records [3], capturing prior atmospheric entries with parameters such as velocity, altitude, and total radiated energy. The second source includes orbital elements of currently tracked asteroids in low-Earth and near-Earth space, including semi-major axis, inclination, eccentricity, and close approach data.

Due to limited overlap between these datasets, shared features were derived through transformation techniques. This includes estimated kinetic energy, projected impact probability, and derived velocity vectors. All entries were filtered and standardized to ensure compatibility and relevance to potential Earth-crossing objects. The resulting

doi:10.2760/2119408

unified dataset supports both exploratory clustering and model training.

2.2. Feature engineering and clustering

To reduce data dimensionality and improve model interpretability, a curated feature set was extracted and transformed. Notable variables include vectorized speed components, altitude normalization, miss-distance thresholds, and inferred orbital behaviour based on similarity to past impactors. Recent multimodal anomaly detection methods such as those proposed by Mondal et al. (2025) inspire more advanced asteroid hazard prediction strategies [4].

Initial unsupervised analysis was performed using K-means clustering, segmenting the dataset into distinct risk categories. This step enables a probabilistic view of impact likelihood based on shared physical characteristics. The clusters were later evaluated via visual and statistical methods, including scatter and box plots to validate separability between high and low-risk objects. Figures 1, 2, 3, 4 and 5 shows the results of the clustering mentioned above.

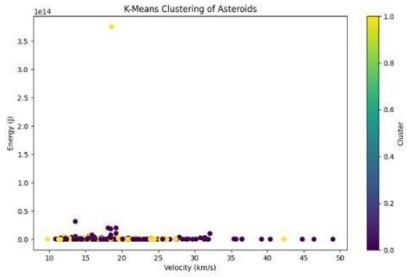


Fig. 1. K-means clustering of historical fireball dataset. Cluster assignments are coloured by velocity and total energy.

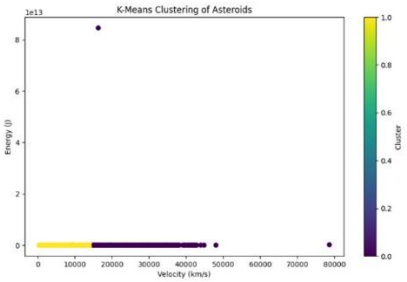


Fig. 2. K-means clustering results on current orbital dataset, visualizing energy vs. velocity with cluster separation.

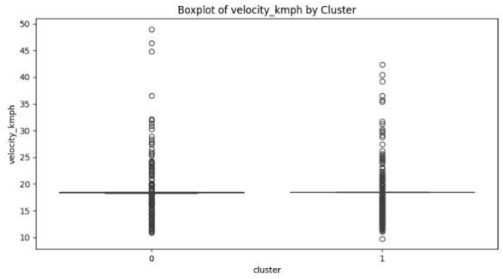


Fig. 3. Boxplot of velocity (km/s) by cluster.

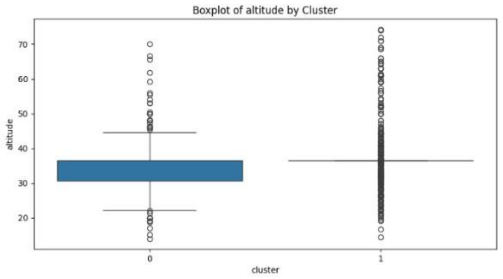


Fig. 4. Boxplot of altitude (km) by cluster.

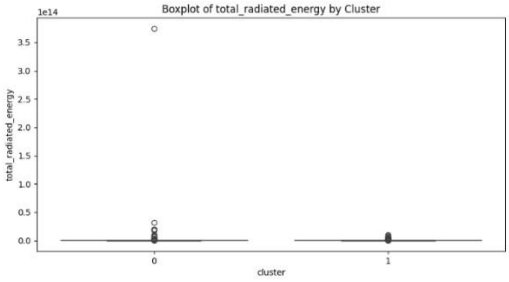


Fig. 5. Boxplot of total radiated energy (J) by cluster.

2.3. Supervised classification

The clustered dataset served as the basis for training a Random Forest classifier to generalize risk prediction to new, unlabelled asteroid entries. Sharma et al. (2024) successfully applied AI-based approaches for detecting hazardous asteroids, showing promise for real-time classification systems [6]. The model leveraged decision-tree ensembles with feature importance tracking, supporting the identification of dominant predictors such as vx, vy, and altitude.

Model training followed a stratified data split and employed standard metrics (precision, recall, F1-score) for evaluation. This classification component was essential for real-time integration, allowing continuous assessment of updated asteroid catalogues.

doi:10.2760/2119408

3. SYSTEM ARCHITECTURE AND DEPLOYEMENT

In addition to the core modelling pipeline, the proposed framework includes a modular, end-to-end architecture for data ingestion, simulation, visualization, and deployment. The system combines machine learning, astrophysics-based simulation, and web technologies to offer a fully integrated platform for asteroid collision risk assessment.

The architecture is composed of three primary layers: a backend data processing layer, a simulation and modelling layer, and a frontend visualization and interaction layer. Each component communicates through lightweight serialized formats (e.g., JSON), ensuring scalability and modular deployment.

3.1.Backend Processing and Modelling

The backend layer, implemented in Python, handles data acquisition from the real-time integration leverages open NASA services such as the Fireball Data API and orbital datasets [5]. It also performs preprocessing steps such as feature extraction, normalization, clustering, and supervised classification. The feature engineering phase is crucial for ensuring consistent scale across heterogeneous datasets.

The simulation engine uses libraries such as Matplotlib, Astropy, and NumPy to render 2D and 3D orbital plots. These plots depict Earth-centric orbits of both historical and predicted high-risk asteroids, enabling static analysis of their trajectories and proximity to Earth as shown in the Figure 6 and 7.

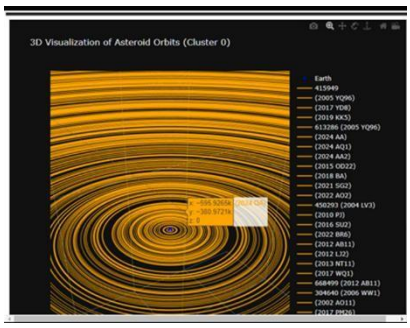


Fig. 6. 3D orbital visualization of near-Earth objects

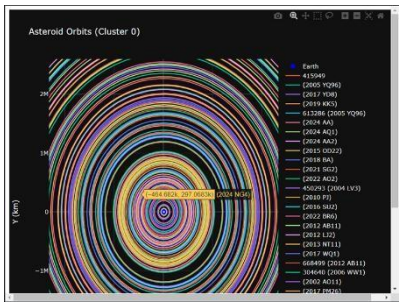


Fig. 7. Example 2D orbital paths of clustered asteroids around Earth.

3.2.Interactive Simulation Frontend

To enhance accessibility and real-time interactivity, a browser-based application was developed using Babylon.js. The simulation recreates a simplified solar system with dynamic asteroid motion, leveraging WebGL for high performance rendering. Each asteroid is plotted in motion based on its orbital parameters, and risk clusters are color-coded for interpretability.

Hover effects reveal detailed metadata (e.g., diameter, velocity, estimated miss distance), enabling users to interact with individual objects as in the Figure 7 and explore their behaviour. The interface is optimized for educational and scientific audiences, promoting transparency and engagement.

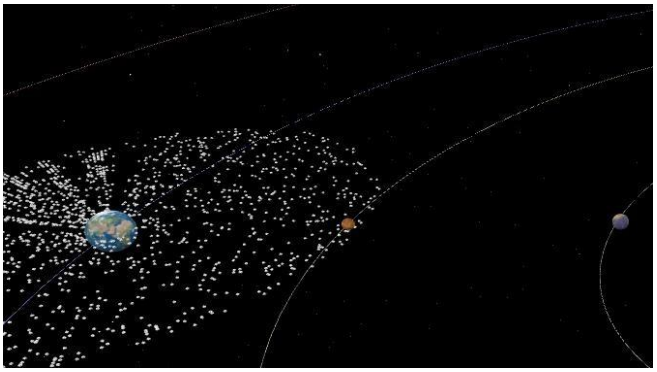


Fig. 7. Interactive web-based asteroid simulation built with Babylon.js. The Earth-cantered view highlights asteroid motion in real-time, with orbital paths and planetary positions rendered in 3D for spatial awareness and educational exploration.

doi:10.2760/2119408

4. RESULTS

This section presents several illustrative outcomes derived from the developed asteroid monitoring framework to showcase its applicability in real-world near-Earth object (NEO) risk analysis. Milani et al. (2005) introduced non-linear impact monitoring techniques, which align with our clustering-based anomaly detection approach [7]. The framework was employed to cluster and analyse close-approaching asteroids based on key physical and orbital parameters such as velocity, total radiated energy, and altitude.

Figures 1 and 2 display the results of K-Means clustering performed on the dataset using velocity and energy as distinguishing features. This clustering helped identify objects with anomalously high energy levels and distinguish them from the bulk population. A second clustering result with a zoomed-in scale is presented in Figure 3, 4 & 5 to better highlight distribution variations among more typical asteroid groups.

To further interpret the characteristics of each cluster, boxplots were generated for velocity (Figure 3), altitude (Figure 4), and total radiated energy (Figure 5). These visualizations provide insights into outlier detection and intra-cluster variability across parameters.

The framework was also applied to a case study involving asteroids with predicted close approaches to Earth between January 1–15, 2024. JSON-formatted data allowed extraction of key features such as object diameter, miss distance, and relative velocity. Among the results, an object named (2024 AR2) was identified with an unusually close predicted miss distance of ~1.26 million kilometres, prompting further risk simulation. The predictions suggest that this methodology can help prioritize observational resources for potentially hazardous objects.

These experiments demonstrate the utility of combining unsupervised clustering, statistical profiling, and feature driven analysis to derive interpretable insights from largescale asteroid datasets, ultimately supporting early warning and planetary defence initiatives.

5. SUMMARY AND CONCLUSION

Farnocchia et al. (2015) and Chodas (2002) emphasized the importance of systematic ranging for impact warning and probability calculation [6], [8]. Considering the significance of the issue we have presented a data-driven approach for assessing and visualizing asteroid collision risk to support planetary defence research and public scientific outreach. By integrating historical impact data with current orbital parameters, the methodology enables intelligent clustering and classification of potentially hazardous asteroids using well-established machine learning models.

Combining statistical feature engineering with visual orbit simulations and interactive web technologies, the system offers an accessible and modular platform for both researchers and the public. The proposed approach supports multi-format visualization, ranging from static plots to fully immersive 3D simulations rendered in web browsers.

The workflow is currently deployed in a simulated environment, enabling consistent monitoring and presentation of close-approaching objects. It is well positioned for integration with real-time data pipelines from external services such as NASA's API, offering a robust, open-source foundation for future planetary defence systems and educational platforms targeting space hazard awareness.

REFERENCES

[1] J. Chomette, A. A. Mathias, and P. Brown, "Machine learning for the prediction of local asteroid damages," *Acta Astronautica*, vol. 204, pp. 15–23, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094576524000572IOSR Journals+2ScienceDirect+2NASA Teknik Raporları Sunucusu+2

[2] A. Malakouti, M. R. Eskandari, and H. R. Pourshahabi, "Machine learning techniques for classifying dangerous asteroids," *Advances in Space Research*, vol. 72, no. 4, pp. 981–990, 2023.
[Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215016123003345ScienceDirect

[3] NASA Center for Near-Earth Object Studies (CNEOS), "Fireball and Bolide Data," 2024. [Online]. Available: https://cneos.jpl.nasa.gov/fireballs/New Space Economy+5cneos.jpl.nasa.gov+5Kaggle+5

[4] A. K. Mondal, N. Aslam, P. Maji, and H. K. Mondal, "A multimodel approach using XAI and anomaly detection to predict asteroid hazards," *arXiv preprint arXiv:2503.15901*, 2025. [Online]. Available: https://arxiv.org/abs/2503.15901 arXiv+1arXiv+1

[5] NASA Center for Near-Earth Object Studies (CNEOS), "Fireball Data API," 2024. [Online]. Available: https://ssdapi.jpl.nasa.gov/doc/fireball.htmlssd-api.jpl.nasa.gov

[6] D. Farnocchia, S. R. Chesley, and P. W. Chodas, "Systematic ranging and late warning asteroid impacts," Icarus, vol. 258, pp. 18–27, 2015.
[Online]. Available: https://doi.org/10.1016/j.icarus.2015.05.023

[7] A. Milani et al., "Nonlinear impact monitoring: line of variation searches for impactors," *Icarus*, vol. 173, no. 2, pp. 362–384, 2005. [Online]. Available: https://doi.org/10.1016/j.icarus.2004.08.005

[8] P. W. Chodas, "Estimating the Impact Probability of Near-Earth Objects Using the Line of Variations," Bulletin of the American Astronomical Society, vol. 34, 2002.

PROCESSING AND DATA ACCESS OF THE GLOBAL FLOOD MONITORING SERVICE

Tobias Stachl, Christoph Reimer, Christian Briese

EODC Earth Observation Data Centre for Water Resource Monitoring GmbH

ABSTRACT

The Global Flood Monitoring (GFM) service, launched in 2021 as an independent component of the Copernicus Emergency Management Service (CEMS), exemplifies the transformative potential of space-based big data in addressing urgent societal challenges, particularly in the context of climate change. Operating continuously, GFM processes all incoming Synthetic Aperture Radar (SAR) imagery from the Copernicus Sentinel-1 satellites using a three-algorithm ensemble approach to generate binary flood masks and flood likelihood layers, along with contextual data such as reference water maps and flood impact indicators. Hosted on cloud infrastructure at the Earth Observation Data Centre (EODC), GFM achieves end-to-end processing times of under 5 hours, with best-case scenarios below 90 minutes. Data access is provided via integration with the Global Flood Awareness System (GloFAS) and the European Flood Awareness System (EFAS), as well as through RESTful APIs, a dedicated web portal, and a STAC-compliant, cloud-optimized catalog. This contribution presents the operational framework, data dissemination infrastructure, and outlines future directions to enhance data access and scalability for climate resilience applications.

Index Terms— Global Flood Monitoring, Sentinel-1, Big Data, Earth Observation, SAR, Datacube, Cloud-optimized

1. INTRODUCTION

Flooding is a pervasive natural hazard that affects millions of people worldwide, necessitating timely and accurate monitoring to mitigate its impacts. Traditional flood mapping methods often rely on manual interpretation of satellite imagery, which can be time-consuming and subject to delays. To overcome these limitations, the Copernicus Emergency Management Service (CEMS), one of the six services of Copernicus, the Earth Observation component of the European Union's space programme, launched the Global Flood Monitoring (GFM) service in 2021. Leveraging Sentinel-1 Synthetic Aperture Radar (SAR) data, GFM operates as a fully automated, 24/7 processing chain. It functions as an independent component of CEMS, supporting both immediate emergency response and long-term disaster risk reduction efforts [11].

All worldwide GFM flood data are freely available in near-real-time (NRT), as well as the historic data from an offline processed archive covering the complete Sentinel-1 observation period (from 2015 to present). As of the time of writing, approximately 1,950,000 Sentinel-1 IW GRDH scenes have been processed, highlighting the extensive coverage and operational scale of the Global Flood Monitoring service.

To support integration into automated workflows and improve accessibility, the growing archive and near-real-time (NRT) outputs of the GFM service have been published as an open-access collection using the SpatioTemporal Asset Catalog (STAC) standard. This enables efficient, programmatic search and filtering by region, time, and metadata such as flood extent. The data is stored in cloud-optimized GeoTIFF (COG) format to support scalable and efficient processing.

2. METHODOLOGY

The GFM service automatically processes all incoming Sentinel-1 SAR images acquired over land in Interferometric Wide (IW) swath mode, Ground Range Detected (GRD) and VV polarization, leveraging a global backscatter datacube that covers approximately 379 billion land surface pixels. After the successful download of the Sentinel-1 IW GRD images, geometrically and radiometrically corrected images of the backscattering coefficient σ° are produced, which are then ingested into a global Sentinel-1 datacube.

The subsequent flood detection is based on three complementary flood mapping algorithms, which outputs are combined using ensemble approaches to produce the main output of the GFM service: a binary flood map and a flood likelihood layer. To enhance interpretation and usability, additional contextual layers are also provided, including a reference water mask (identifying permanent and seasonal water bodies), an exclusion mask, advisory flags, and flood impact layers. The three flood mapping algorithms have been developed by the German Aerospace Centre (DLR), the Luxembourg Institute of Science and Technology (LIST), and the Vienna University of Technology (TU Wien). Further technical details on the flood mapping algorithms, the ensemble methodology, and the contextual layers can be found in Wagner et al. [13] as well as on the Wiki pages of the GFM service (https://extwiki.eodc.eu/en/GFM).

doi:10.2760/2119408

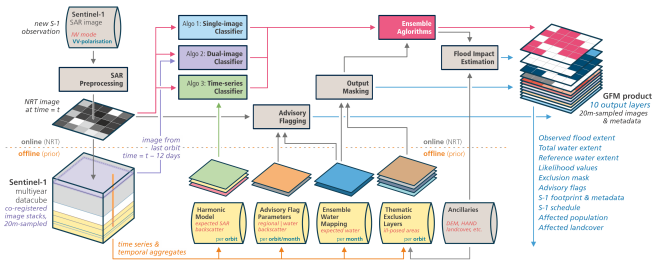


Fig. 1. Overview of the step-by-step GFM NRT production workflow [13].

2.1. Near-real-time workflow

The NRT workflow depicted in figure 1 consists of the acquisition of the latest Sentinel-1 SAR images, its preprocessing to σ° images, production of the flood mapping algorithms including advisory flagging and ensemble approaches and the final preparation and ingestion of the output products to be picked up by the various dissemination systems. Additional inputs that were derived offline by analysing the historic data within the Sentinel-1 datacube along with high-resolution ancillary datasets, are picked up and prepared by the NRT workflow for each incoming Sentinel-1 image.

The entire data production workflow is deployed within a fully independent resource tenant on EODC's cloud infrastructure. This environment is powered by OpenStack, a widely used open-source cloud computing platform that enables flexible, scalable, and automated management of compute, storage, and networking resources. By leveraging OpenStack, we ensure high availability, resource isolation, and efficient orchestration of our processing services, making it ideally suited to the needs of GFM.

To streamline infrastructure management, we use Ansible and Terraform to automate the provisioning, configuration, and maintenance of our environment, ensuring consistency and scalability across all deployments. The current setup includes 480 virtual CPUs (vCPUs) and 1.2 TB of memory of latest processing hardware, distributed across multiple worker units to maintain a service availability of greater than 99%.

Apache Airflow is used for task orchestration, enabling the reliable scheduling, monitoring, and execution of all processing chains including data acquisition, Sentinel-1 backscatter σ° pre-processing and flood mapping workflows within the GFM data production pipeline. Finally, a monitoring and alerting system powered by Grafana provides real-time notifications to operations engineers, ensuring 24/7 service availability and rapid response to any system anomalies.

2.2. Integration with Forecasting Systems

GFM's outputs are integrated in the Global Flood Awareness System (GloFAS) [8] and the European Flood Awareness System (EFAS) [9], enhancing the predictive capabilities of these systems. By combining real-time flood observations with medium- and seasonal-range forecasts, GFM supports proactive flood risk management, enabling authorities to issue timely warnings and implement mitigation measures. The map viewers of GloFAS and EFAS allow users to visualize all GFM product layers.

2.3. Data Access

Alongside the aforementioned integration into GloFAS and EFAS, GFM output products are accessible through a various set of interfaces, following the assumptions of Mostafiz et al. [10], to provide easy access of flood information to maximise its usefulness for both the public and professionals.

A dedicated webportal [4] - which is also integrated in the map viewer of GloFAS and EFAS - allows users to define areas of interests (AOIs), display and download GFM data for the AOIs and configure a notification service for any new available data. Additionally, a set of application programming interfaces (APIs) has been implemented to provide a more flexible way of downloading data. The integration into GloFAS and EFAS as well as the here described webportal and APIs have been developed by GeoVille GmbH [3], member of the GFM consortium.

2.3.1. Cloud-optimized Data Access

Given the continuously expanding volume of GFM output data - including both the complete archive and NRT products - ensuring easy discoverability and programmatic access is essential for integrating GFM data into automated processing workflows and applications [7]. To support this, in addition to the previously mentioned access methods, the GFM data has been published as an open-access collection using the SpatioTemporal Asset Catalog (STAC) specification [5].

STAC is a standardized way to expose and interact with collections of spatial temporal data, which enables users to efficiently search the entire GFM dataset by specifying regions and time periods of interest. It also allows filtering based on metadata attributes specific to the outputs, such as the number of detected flooded pixels. The data is stored in the Cloud-Optimized GeoTIFF (COG) format, which enhances read efficiency and supports scalable processing pipelines.

The GFM STAC catalogue currently comprises approximately 4,8 million STAC items. Each STAC item contains links to all associated output files - specifically, COGs representing the results of the three individual flood mapping algorithms as well as the ensemble product.

Numerous tutorials and example Jupyter notebooks showcasing the benefits of using STAC in simple to complex use cases can be found on EODC's public GitHub repository [2]. The examples range from a simple visualisation of a single GFM output to processing of the maximum flood extent for a

doi:10.2760/2119408

specified AOI and time range using data proximate processing on EODC cloud infrastructure utilizing the Python based parallel computing library Dask [1].

2.4. Timeliness

A core requirement of the GFM service is to deliver its output data as quickly as possible, with a maximum latency of 8 hours following each Sentinel-1 SAR acquisition. The performance of the GFM production workflow is measured and monitored using Key Performance Indicators (KPIs), which are used for quarterly and annual reporting [12]. Figure 2 depicts an overview of the maximum processing durations under regular conditions. In optimal conditions, the GFM system can achieve end-to-end processing times - from sensing to data dissemination - of under 90 minutes. On days when the Sentinel-1 ground segment operates nominally, the total latency typically remains below 5 hours, which can be seen in figure 3.

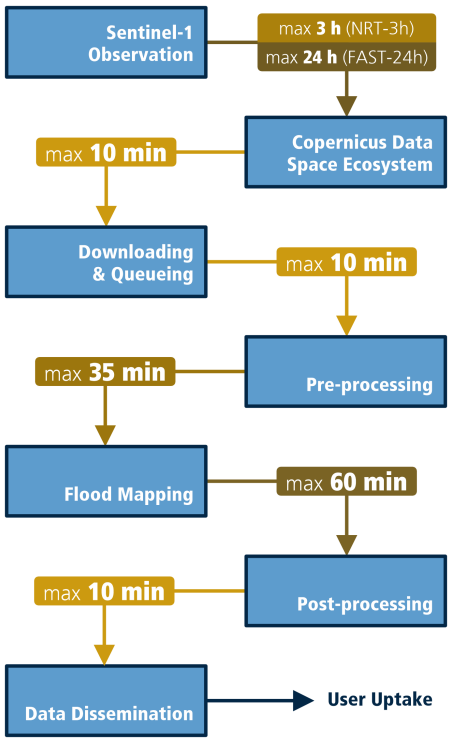


Fig. 2. Timeliness of the GFM service, with maximum durations under regular conditions between acquisition from Copernicus, GFM main processing modules, and product dissemination. NRT-3h and FAST-24h refer to Copernicus' Sentinel-1 timeliness categories [13].

Fig. 3. Grafana dashboard showing the average, minimum and maximum timeliness of a usual day (18. May 2025).

3. OUTLOOK

The Copernicus Sentinel-1 mission was developed as a constellation of two satellites to ensure a repeat cycle of six days. Sentinel-1A and Sentinel-1B were launched in April 2014 and April 2016, respectively. Following the premature loss of Sentinel-1B in December 2021, the third satellite, Sentinel-1C, was successfully launched in December 2024. At the time of writing, integration of Sentinel-1C into the Global Flood Monitoring (GFM) near-real-time (NRT) workflow is ongoing. To maintain the performance and continuity of GFM, the expansion of the Sentinel-1 constellation with the upcoming Sentinel-1D is essential.

Additionally, enhancing and evolving data access methods is critical for enabling seamless integration of the extensive GFM dataset into existing processing workflows. In line with the principles of data-proximate processing, our objective is to improve data access and analytical capabilities, particularly for time series applications, by adopting Zarr [6], a community-driven specification for chunked, compressed, N-dimensional arrays that enables efficient I/O in parallel computing environments. This approach will better support users seeking to exploit the full potential of the nearly decade-long dataset, whether for trend analysis, historical comparison, or large-scale processing.

4. CONCLUSION

The Global Flood Monitoring (GFM) service represents a significant leap forward in operational, near-real-time satellite-based flood detection and monitoring. Leveraging Sentinel-1 SAR data, GFM delivers high-resolution flood maps globally within hours of data acquisition in a fully-automated manner, fulfilling a critical need for rapid and reliable flood information. The service ensures accessibility and performant access to its output data through a robust cloud-based infrastructure and advanced data processing pipeline - featuring independent flood detection algorithms, which outputs are combined using ensemble approaches, integration in the forecasting systems of EFAS and GloFAS, and dissemination via multiple access points including APIs, web portals, and STAC.

As flooding events become more frequent and severe due to climate change, services like GFM are indispensable tools in enhancing global preparedness and resilience. Future developments will focus on further improving algorithm perfor-

doi:10.2760/2119408

mance, expanding data integrations, and refining accessibility to maximize the service's impact across scientific, humanitarian, and operational domains.

REFERENCES

[1] Dask Development Team (2016). Dask: Library for dynamic task scheduling. https://www.dask.org/.

[2] Public GitHub repository of EODC example notebooks. https://github.com/eodcgmbh/eodc-examples/tree/main/demos/GFM.

[3] GeoVille Informationssysteme und Datenverarbeitung GmbH. https://www.geoville.com/.

[4] GFM web portal. https://portal.gfm.eodc.eu/.

[5] STAC Specification. https://stacspec.org/en.

[6] Zarrr - chunked, compressed, N-dimensional arrays. https://zarr.dev/.

[7] Sandro Groth, Marc Wieland, Fabian Henkel, and Sandro Martinis. Global reference water information for flood monitoring: Increasing accessibility with STAC, November 2024. URL https://meetingorganizer.copernicus.org/EGU24/EGU24-3849.html.

[8] Gwyneth Matthews, Calum Baugh, Christopher Barnard, Corentin Carton De Wiart, Juan Colonese, Stefania Grimaldi, Daniel Ham, Eleanor Hansford, Shaun Harrigan, Stine Heiselberg, Helen Hooker, Sazzad Hossain, Cinzia Mazzetti, Leonardo Milano, Francesca Moschini, Karen O'Regan, Florian Pappenberger, Daniel Pfister, Ragindra Man Rajbhandari, Peter Salamon, Arthur Ramos, Kay Shelton, Elisabeth Stephens, Dimitar Tasev, Monica Turner, Marc Van Den Homberg, Julia Wittig, Ervin Zsótér, and Christel Prudhomme. On the operational implementation of the Global Flood Awareness System (GloFAS). In *Flood Forecasting*, pages 299–350. Elsevier, 2025. ISBN 978-0-443-14009-9. doi: 10.1016/B978-0-443-14009-9.00014-6. URL https://linkinghub.elsevier.com/retrieve/pii/B9780443140099000146.

[9] Gwyneth Matthews, Calum Baugh, Christopher Barnard, Corentin Carton De Wiart, Juan Colonese, Damien Decremer, Stefania Grimaldi, Eleanor Hansford, Cinzia Mazzetti, Karen O'Regan, Florian Pappenberger, Arthur Ramos, Peter Salamon, Dimitar Tasev, and Christel Prudhomme. On the operational implementation of the European Flood Awareness System (EFAS). In *Flood Forecasting*, pages 251–298. Elsevier, 2025. ISBN 978-0-443-14009-9. doi: 10.1016/B978-0-443-14009-9.00005-5. URL https://linkinghub.elsevier.com/retrieve/pii/B9780443140099000055.

[10] Rubayet Bin Mostafiz, Robert V Rohli, Carol J Friedland, and Yong-Cheol Lee. Actionable information in flood risk communications and the potential for new web-based tools for long-term planning for individuals and community. *Frontiers in earth science*, 10:840250, 2022.

[11] Peter Salamon, Niall McCormick, Christopher Reimer, Tom Clarke, Bernhard Bauer-Marschallinger, Wolfgang Wagner, Sandro Martinis, Candace Chow, Christian Böhnke, Patrick Matgen, Marco Chini, Renaud Hostache, Luca Molini, Elisabetta Fiori, and Andreas Walli. The new, systematic global flood monitoring product of the copernicus emergency management service. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1053–1056, 2021. doi: 10.1109/IGARSS47720.2021.9554214.

[12] M. Seewald, A. Pasik, C. Gruber, F. Innerbichler, M. Riffler, C. Reimer, T. Stachl, R. Kidd, N. McCormick, and P. Salamon. *Global flood monitoring – Annual product and service quality assessment report 2024*. Publications Office of the European Union, 2025. doi: doi/10.2760/9738940.

[13] Wolfgang Wagner, Bernhard Bauer-Marschallinger, Florian Roth, Tobias Stachl, Christoph Reimer, Niall McCormick, Patrick Matgen, Marco Chini, Yu Li, Sandro Martinis, et al. Preprint: The fully-automatic sentinel-1 global flood monitoring service: Scientific challenges and future directions. 2025. doi: https://dx.doi.org/10.2139/ssrn.5110703.

SURROGATE MODELING AND USER-IN-THE-LOOP EXPERIMENTATION FOR URBAN FLOOD PREDICTION: THE EXTREMEXP APPROACH

Pauline Delporte[1], *Gwendoline Stéphan*[1], *Yasmine Boulfani*[1], *George Papastefanatos*[2], *Vincent Gaudissard*[1]

[1]CS Group, Toulouse, France
[2]Athena Research Center, Athena, Greece

ABSTRACT

Urban flash floods are becoming increasingly frequent due to climate change and land artificialization, posing serious risks to human safety and infrastructure. To address this challenge, the HORIZON-Europe ExtremeXP project [1] provides a user-centric platform integrating explainable AI, visual analytics, and experimentation workflows. This paper presents a use case focusing on flash flood prediction in the city of Nîmes, France. A deep learning surrogate model, based on a UNet architecture enhanced with temporal attention, is trained using data generated by a hydrodynamic model. Input data include topographical information from LiDAR and aerial imagery, as well as high-resolution rainfall data. The ExtremeXP framework enables iterative experimentation, model optimization, and interactive visualization, placing the user at the center of the process. Results show promising predictive performance and highlight the relevance of combining physical simulations and AI within a transparent decision-support system for climate risk mitigation.

***Index Terms*—** machine learning, surrogate model, flash floods, hydrodynamical model

1. INTRODUCTION

In today's data-driven world, machine learning models are increasingly used to solve a wide range of problems. These new methods face new challenges: the "black-box effect" of machine learning models leading to the lack of involvement of user and visualization in the development of their models.

Developing accurate and trustworthy machine learning models is a challenge well known to the scientific community. The ExtremeXP framework (Fig. 1) provides accurate, fit-for-purpose data-driven insights by evaluating different complex analytic variants considering user intents, constraints and feedback. Experimentation is the core concept for generating accurate analytics. AI training pipelines are considered as well as other types of workflows (data analytics, simulation and visualization), including hybrid ones. The user is considered at the center of the process. Its participation takes place at different phases of the experimentation in order to:

- specify intents, constraints and access control policies.
- manage the workflows during the execution of an experiment.
- review the results and provide feedback.

ExtremeXP addresses key societal and industrial challenges through five application-driven use cases (UC): AI-based flash flood forecasting, cybersecurity awareness, predictive maintenance, transport analysis, and disaster response. This article focuses on the first UC: improving flash flood forecasting with artificial intelligence.

Climate change is increasing the occurrence of urban flash flood. Developing prediction models is an interesting tool to reduce human and material damage caused by such events. The data used to drive the models are of two types: topographical and meteorological, including data collected via satellite.

2. EXTREMEXP FRAMEWORK

Fig. 1 presents an overview of the ExtremeXP framework's modular architecture, which orchestrates different subsystems/services. At the core lies the **Experimentation Engine** responsible for designing, scheduling, executing and monitoring an experiment, i.e., the workflow of an AI pipeline that the user wishes to evaluate. The surrounding five subsystems offer additional modular features to the data scientist:

1. The **Analysis-aware Data Integration module** deals with data-processing related challenges and provides novel solutions to automatically select among datasets, and deal with data quality issues such as missing, incomplete, wrong, multilingual and duplicate data points in user-driven reconfigurable workflows.
2. The **User-driven AutoML** offers functionalities for simulation-based data augmentation for ML, constraint-aware ML algorithms, algorithms for model selection based on user preferences and constraints, continual learning of model selection strategies and optimal deployment of ML pipelines in heterogeneous environments.
3. The **Transparent & Interactive Decision Making** offers explanations on the choice and configuration of a ML/data analytics method and interactive visualization

doi:10.2760/2119408

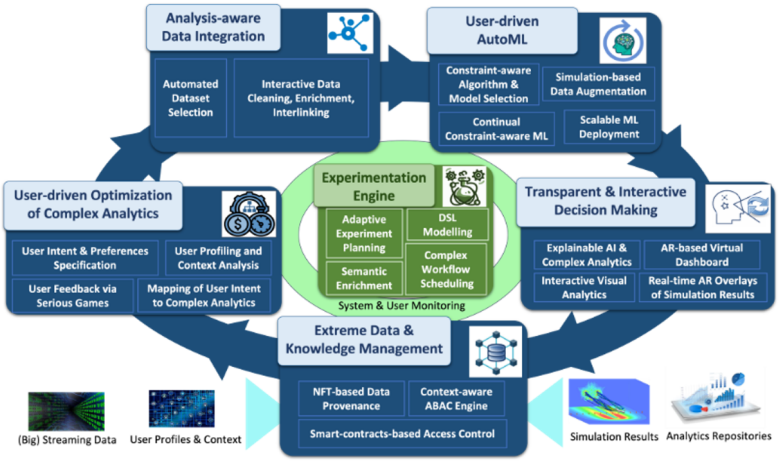


Fig. 1. ExtremeXP framework

and AR technologies to enhance monitoring and decision making.

4. The **User-driven Optimization of Complex Analytics** captures user intents, requirements and constraints as well as user feedback for the optimization of experiments.
5. The **Extreme Data & Knowledge Management** provides capabilities for secure and distributed management of datasets and experimentation-based knowledge assets and learning outcomes.

The framework has been used throughout the entire lifecycle of the flash flooding prediction experiment setup described in this paper, i.e., defining the AI workflow stages, scheduling and evaluating models and hyperparameters that fit the accuracy requirements, inspecting and refining the results through the interactive visualization dashboard. In this UC the user interacts with the frameworks by giving some feedbacks on the results produced by the experiment. This data is then processed by the plateform to automatically improved the experiment setup. With this UC a focus was made on the visualization tools to improve the user experience and to help actor for decision-making.

3. EXTREMEXP UC: FLASH FLOOD PREDICTION

Floodings are recurrent events due to global warming and excessive land artificialization. The damage caused by these phenomena is both human and material, hence the urgent need to develop models because they can be used to immediately evacuate and secure the population, and over time to enable us to better organize the territory.

3.1. Area of study

In this UC, the study focuses on Nîmes, a French city with a special geographical situation, lying between the Mediterranean Sea and the Cévennes mountains. The artificial nature of the soil makes it impermeable, and as water is not absorbed, water runoff is increased. As a result, this area experienced some very serious flooding events in recent years (1988, 2002, 2005, 2014) caused by heavy rainfalls. The last three events are studied here with a focus on the Camplanier catchment (see Fig. 2).

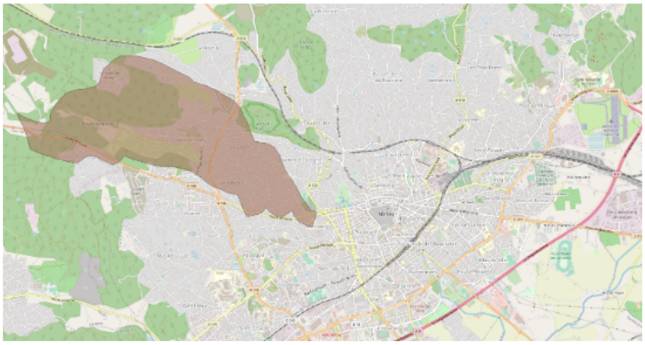


Fig. 2. Camplanier catchment within Nîmes city

3.2. Methodology

The UC is responsible for modeling urban flooding phenomena with machine learning tools. A physical model is developed to generate the data needed to train the deep-learning model. Physical models are generally slow to infer and complex to configure, hence the interest in developing surrogate models based on deep-learning methods. The global approach is based on two main steps:

doi:10.2760/2119408

1. The physical model is configured and run once the input data is set up.
2. The data generated by the HD model is recovered to train the deep-learning model.
3. The surrogate model is validated against the physical one.

3.3. Dataset and data processing

Various types of data are necessary to predict hazard events. In the context of flood prediction, two classes of data are interesting: topographical data providing information about the terrain and its layout, and meteorological data.

3.3.1. Topographical data

The topographical data are open-source data from IGN: the Digital Elevation Model (DEM) RGE ALTI [2] and a buildings mask [3]. In coastal areas, LiDAR is deployed, and, in the mountains, radar is mostly used to acquire the data. For the rest of the territory, aerial images correlation is used. In the model data, the areas with buildings or outside the catchment are marked with a No Data value.

3.3.2. Meteorological data

The city of Nîmes, one of our collaborators, provided rainfall data used as meteorological inputs to characterize the flooding event. They are obtained with the software CALAMAR (CAlcul de LAMes d'eau radAR) [4], a service precipitation at high resolution (0.25 km²) by a hydrometeorological radar. Recovered data has two different types: radar images and pluviometry measurements intended to calibrate radar data.

3.3.3. Data processing

The input data (DEM, rainfall, buildings and catchment geometries), and the water-depth maps calculated by the physical model, are georeferenced in a single tile. In addition to this data the water level situation is added as a 2D map, to inform the model if area is already flooded before the prediction or not. Normalization transformation is applied on the DEM. During training the unique tile is split into small ones and an overlapping can be applied according to the configuration. Sub-tiles with more than 90% of No Data are removed from the training.

3.4. Hydrodynamic model

The physical model, using a hydrodynamical (HD) model, producing water depth maps for the 3 flash flood events in Nîmes works in 3 steps. No benchmark has been done and the in-situ water depth measurements are not used to validate the physical model, yet.

3.4.1. Meshgrid generation

First, we prepare the input data needed by the HD model in a specific format and we produce the calculation meshgrid. It is a Delaunay grid obtained using the software Triangle [5][6]. We also need to process and simplify the buildings geometry to avoid, in the meshgrid, cells with a very small angle which would slow down the HD model tremendously or would prevent the convergence of the results.

3.4.2. HD model

The second step consists of running the HD model, Dassflow-2d [7], producing the water depth images in the form of VTK files. Outflow boundaries are also manually defined, in the south of the catchment, to simulate water escape route and avoid unrealistic accumulation, though no accurate comparisons were performed with in-situ data yet.

3.4.3. Data post processing

Finally, we process those files and the input data, restructuring the Delaunay grid into a cartesian grid, to provide a single netCDF file for the AI model.

3.5. Surrogate Model

3.5.1. Model Architecture

The model's prediction time is a parameter that is generally specific to the city and can vary according to its characteristics. In Nîmes, it is estimated that an event can have a strong impact within 30 minutes. In that way, we are considering 6 rainfall data temporally spaced 5 minutes apparts. The model architecture needs to deal with both spatial and temporal dimension.

The chosen architecture is a UNet associated with a temporal attention layer [8]. It is composed of 3 steps: a spatial encoder, a temporal encoder and a spatial decoder. The chosen model outperforms the original UNet and several RNNs by taking advantage of the attention layer [9]. In the case of flood prediction with spatio-temporal information, attention layer is efficient to recover the important spatio-temporal features from the scene.

3.5.2. Model Training

The model is trained on the ExtremeXP platform through experimentation. The parameter space is browsed during training to find the optimum configuration. L2 loss is used during training.

3.5.3. Results

During training, several metrics are used, as described in the following table (see T. 1). Two types are used classification

doi:10.2760/2119408

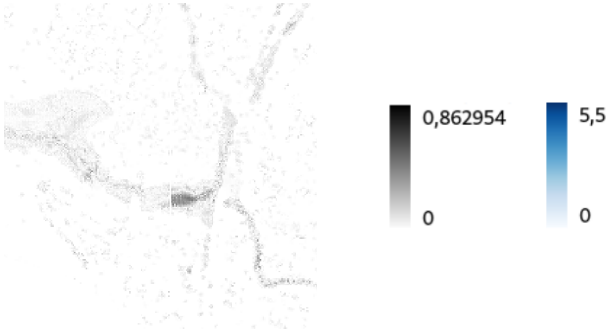
and regression metrics. To calculate some of them, such as recall and precision, it is necessary to return to a classification problem. To do this, we determine a height at which the area is considered flooded. Then, the water depth map is converted into a classification map with two classes : no flood and flood. Here, the threshold is determined at 20 cm. Error maps are generated to analyze the difference between the HD model and AI model predictions. The evaluation of the metrics was made on the the 2005 events. Finally the accuracy is the percentage of well predicted pixels with a certain margin of error, defined at 5 cm.

Table 1. Metrics

Name	Recall	Precision	MSE	Accuracy
N256	0.86	0.91	0.03	0.94
N512	0.91	0.90	0.01	0.97

As the data is georeferenced, it can be projected onto a map (see Fig. 3 (a)). This visualization enables users to identify the possible flooded areas. The water depth representation indicates to the users the level of the flood over the catchment.

(a) Prediction (b) Ground Truth



(c) Error map

Fig. 3. (a) Water depth (in meter) prediction from AI model; (b) Ground truth water depth (in meter) (HD model results); (c) Error maps: the absolute difference between (a) and (b) in meter.

4. CONCLUSION AND PERSPECTIVES

This work presented a flash flood prediction UC developed within the HORIZON-Europe ExtremeXP platform, focusing on the city of Nîmes. By combining hydrodynamic simulations with a deep learning surrogate model, the system predicts water depth 30 minutes ahead, showing encouraging performance in terms of MSE and accuracy.

The ExtremeXP platform has supported the full lifecycle of the experimentation, from data integration to visualization, emphasizing user involvement and explainability.

Future work includes the generalization to other cities with diverse conditions, the addition of incertitude measurement in the model to facilitate decision-making for the user as well as enhanced explainability for better user trust and understanding. ExtremeXP aims to elaborate a proof of concept to demonstrate that a surrogate model is able to replace a physical method once trained. For a better solution in an operational context, it will be important to validate the model against in-situ data. Those axes concerned the AI part of the project, on the hydrological one, it will be interesting to add other infrastructures (pavements, hydrological facilities, etc) in the meshgrid for more realistic simulations.

These results demonstrate the potential of combining AI and physical modeling for effective urban flood forecasting.

REFERENCES

[1] ExtremeXP Horizon-EU: website.

[2] IGN RGE ALTI: Documentation.

[3] IGN BD TOPO: Documentation.

[4] CALAMAR software: website

[5] J.R Shewchuk, "Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator". In: Lin, M.C., Manocha, D. (eds) "Applied Computational Geometry Towards Geometric Engineering". WACG 1996. doi:10.1007/BFb0014497

[6] Jonathan Richard Shewchuk, "Delaunay Refinement Algorithms for Triangular Mesh Generation", "Computational Geometry": Theory and Applications 22(1-3):21-74, 2002.

[7] Dassflow2D: website

[8] Chaudhary, P.; Leitão, J.P.; Schindler, K.; Wegner, J.D. "Flood Water Depth Prediction with Convolutional Temporal Attention Networks". Water 2024, 16, 1286. doi: 10.3390/w16091286.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. "Attention is all you need" in "Advances in Neural Information Processing Systems", 30, 5998–6008. doi:10.48550/arXiv.1706.03762, 2017.

Bringing AI to Geohazard Analysis: The New MLOps Framework in GEP

Simone Vaccari, Alice Re, Parham Membari, Pedro Gonçalves, Herve Caumont

Terradue Srl, Rome, Italy

Abstract

The Geohazards Exploitation Platform (GEP), developed by Terradue, is a cloud-native Earth Observation (EO) platform designed to support the geohazard community. With a user base exceeding 3,500, GEP provides a comprehensive suite of EO processing services that assist researchers and practitioners in monitoring geohazards and responding to disasters. These services include tools for change detection, terrain deformation analysis, and long-term ground motion monitoring. To address increasing demands for automation and scalability, GEP has integrated an end-to-end Machine Learning Operations (MLOps) framework. This enhancement enables users to develop, deploy, and monitor AI models for geospatial applications efficiently. By supporting data ingestion, training, deployment and performance monitoring, GEP facilitates the operationalisation of AI models in areas such as landslide detection, earthquake response, and environmental monitoring. The framework incorporates FAIR principles through the Machine Learning Model (MLM) SpatioTemporal Asset Catalog (STAC) Extension, promoting reproducibility and discoverability. This positions GEP as a robust, scalable solution for advancing geospatial intelligence in science, public safety, and infrastructure resilience.

Index Terms— Geohazards, Thematic Exploitation Platform, GEP, Machine Learning, MLOps

1. Introduction

The Geohazards Exploitation Platform (GEP) is a cloud-based Earth Observation (EO) data processing platform developed and operated by Terradue to support geohazard monitoring, terrain motion analysis, and critical infrastructure assessment.

It serves a diverse user base of over 3,500 researchers, public authorities, and industry professionals, providing access to EO data archives, advanced processing services, and analytical tools.

These services range from systematic data processing workflows, such as generating interferometric deformation maps, to event-triggered processing for rapid response scenarios like earthquake damage assessments. They support a variety of data-driven applications, from data screening and area monitoring to the integration of multi-temporal data for long-term risk assessment.

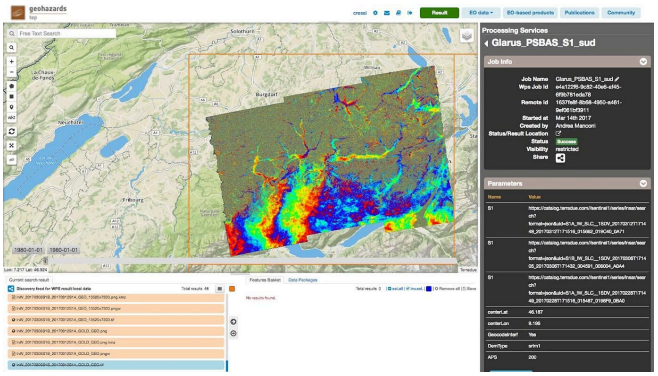


Fig. 1. GEP's community portal.

1.1. A platform purpose-built for geohazards

GEP offers more than 25 specialised EO services, including visualisation services for full-resolution imagery, processing services for event response and change detection as well as advanced SAR and optical data processing dedicated to the analysis of long time series of optical imagery for long-term area monitoring.

The range of processing services dedicated to change detection and event response comprises services for the generation of Digital Surface Models (DSM) and orthoimages from stereo- and tri-stereo very high-resolution satellite imagery [1], as well as services for automatic landslide detection and inventory mapping [2][3]. The portfolio also includes services for the assessment of landslide source volumes and for the flow path assessment of gravitational hazards at the regional scale —including debris flows, rockfalls, rock and snow avalanches and shallow landslides [4]. The change detection service portfolio is further complemented by SAR-based services for generating coherence and intensity composites, and services for detecting changes in SAR amplitude.

Processing services devoted to area monitoring include classic and advanced InSAR for persistent scatterers (PS) and small baseline (SBAS), such as —respectively— the SNAPPING (Surface motioN mAPPING) [5] and the CNR-IREA P-SBAS [6] Sentinel-1 on-demand processing services. Others focus on horizontal ground motion detection from optical time series, such as the Ground

doi:10.2760/2119408

Deformation Monitoring with OPtical image Time series (GDP-OPT) [7] processing services, tailored to landslide monitoring, ice and glacier analysis and earthquake analysis.

The processing services integrated in GEP result from a collaborative effort between service providers, such as research institutions or private companies that develop the algorithm, and Terradue, who provide support for the development, deployment and maintenance of services on a cloud infrastructure. This reflects the collaborative nature of this scientific endeavour, supporting a broad community of practitioners involved in geohazards.

All services operate on cloud infrastructure and have access to over 400 TB of EO data, including from Sentinel-1, Sentinel-2, Sentinel-3, ERS, Envisat, Landsat-8, Spot and Pléiades. Figure 1 illustrates GEP's community portal showcasing the results of a P-SBAS processing job.

1.2 Uptake in the scientific community and typical case studies

GEP caters to a variety of user needs. Service developers can deploy their own algorithms using the EO Application Package [8] model.

Researchers and scientists use GEP for pilot studies, validation campaigns and long-term monitoring. Processing services hosted on GEP have already supported a wide range of use cases and scientific publications, including volcanic activity tracking [9], subsidence analysis [10], earthquake response [11] and landslide detection [12].

The adoption of GEP in scientific research is underpinned by adherence to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The use of Persistent Identifiers (PIDs) for experiments, data and software ensures the reproducibility of scientific experiments. Each application is encapsulated with its execution environment and dependencies, thus promoting consistent outcomes. Adherence to these principles is further facilitated by providing structured documentation and standardised metadata.

Public authorities and civil protection agencies rely on GEP for hazard mapping and disaster response. Rapid response during disasters is supported in a multifold way, specifically through providing access to pre-processed EO datasets, automated services for event-triggered surface deformation mapping and visualisation tools for sharing results with stakeholders in real time. By allowing for systematic data screening and long-term ground motion monitoring, GEP contributes to the operational resilience of critical infrastructure planning.

GEP is in use across multiple continents, including Europe, Southeast Asia, Africa, and the Americas. Some examples of past use cases are summarised in Table 1.

Table 1. Selection of GEP use cases

Title	Region	Use Case Description
Sulawesi Earthquake (2018)	Indonesia	GEP supported rapid mapping of ground deformation and landslides triggered by the earthquake and tsunami. [13]
Central Chile Earthquake (2022)	Chile	GEP provided InSAR analysis to detect surface deformation and assess the earthquake's impact on infrastructure. [14]
Santorini Volcano Unrest Phase (2023)	Greece	GEP services were used to monitor another phase of volcanic unrest at Santorini, one of the most iconic volcanoes in the Aegean Sea. [15]
Turkiye–Syria Earthquakes (2023)	Turkey - Syria	GEP facilitated the generation of interferograms and surface rupture mapping for the M7.8 and M7.7 earthquakes, supporting disaster response. [16]
Morocco Earthquake (2023)	Morocco	Radar interferogram generated using Sentinel-1 acquisitions to analyze the impact of the M6.8 earthquake. [17]

1.3 Expanding the portfolio of GEP with Artificial Intelligence (AI)

To meet the growing demands of its users, GEP's key objective is to expand the portfolio of services that leverage artificial intelligence (AI) and machine learning (ML). The complexity of training, deploying and maintaining ML models at scale, however, poses significant challenges.

These include managing large and diverse EO datasets, ensuring reproducibility, and maintaining model performance over time in dynamic operational environments. Overcoming these obstacles is essential to unlocking the full potential of AI in geospatial applications and opening up GEP to a wider range of data processing services, users, and stakeholders.

The expansion of AI support within GEP was undertaken as part of the AI/ML Enhancement Project. The project's main goal was to integrate an AI/ML processing framework into GEP seamlessly, thereby enhancing its services and enabling service providers to develop and deploy AI/ML models to improve geohazard applications.

doi:10.2760/2119408

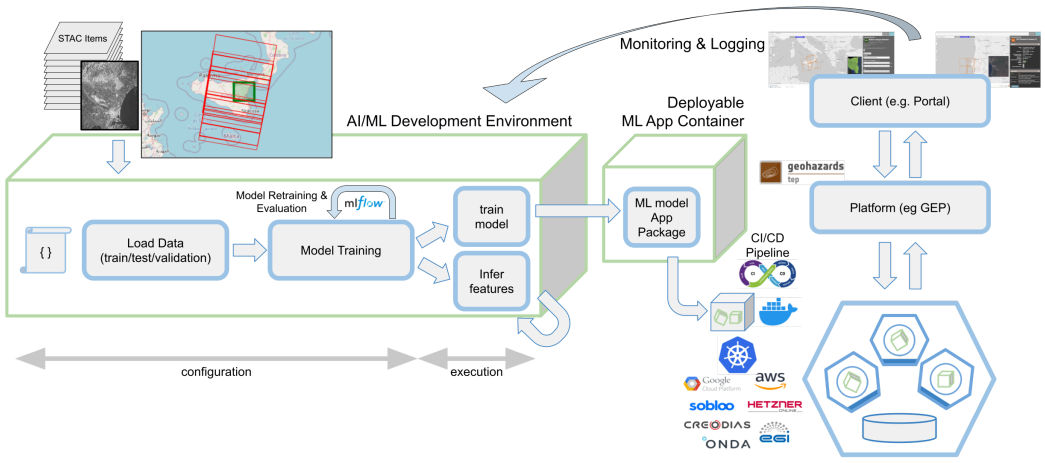


Fig. 2. Schematic representation of the MLOps workflow in GEP.

2. THE NEW MLOPS FRAMEWORK IN GEP

A comprehensive AI/ML processing framework has been integrated in GEP, encompassing the entire machine learning lifecycle. This includes data discovery, training data, model development, deployment, hosting, monitoring and visualisation. A critical aspect has been integrating Machine Learning Operations (MLOps) practices into the platforms' service offerings to ensure the smooth operation of AI-driven applications.

By integrating MLOps directly into its cloud-native architecture, GEP now offers scalable and flexible AI capabilities. This will enable users to deploy advanced geospatial AI applications specifically designed for terrain motion analysis, disaster risk reduction and environmental monitoring.

2.1. The Machine Learning Model (MLM) STAC Extension

The GEP MLOps framework builds upon the work carried out in the context of the Machine Learning Model (MLM) extension [18] —a specification that extends the SpatioTemporal Asset Catalog (STAC) framework in order to catalogue machine learning models, thereby improving their searchability and reproducibility. Originally driven by the need to incorporate FAIR principles, the MLM extension has proven crucial in supporting the identification and documentation of ML models driven by the specific requirements of geospatial applications. The framework supports the development and deployment of ML solutions tailored to applications addressing pressing challenges such as environmental monitoring, agriculture, disaster response and urban planning by supporting effective search and discovery.

2.2. End-to-end ML support

Figure 2 illustrates GEP's integrated MLOps workflow, showcasing data ingestion, model training, feature inference, and scalable deployment on cloud infrastructures.

The newly integrated MLOps framework within GEP provides users with advanced, cloud-native infrastructure, enabling them to manage the complexities of developing, deploying and maintaining AI models for geospatial analysis seamlessly.

Data preparation and management is streamlined through the efficient handling of large EO datasets, with automated splitting for training, testing and validation.

Model development and training are supported by tools and environments that not only facilitate the creation of new ML models but also provide tracking and management functionalities through frameworks such as MLflow. MLflow plays a key role in experiment management by recording key information such as code versions, datasets, and model hyperparameters, ensuring reproducibility and traceability. The MLflow dashboard offers a clear, interactive view of multiple runs side-by-side, enabling straightforward comparison and helping users identify the best-performing model with greater confidence.

GEP supports scalable deployment by enabling trained models to be deployed into operational environments as cloud-native microservices that are compatible with standard geospatial service interfaces, such as Open Geospatial Consortium (OGC) APIs.

Finally, automated pipelines enable continuous monitoring and adaptation: deployed models are monitored for issues such as data drift or performance degradation, triggering re-training processes to maintain optimal performance.

This enables GEP to provide comprehensive support for the creation and operationalisation of AI models. Users can ingest and curate datasets directly within the platform to

doi:10.2760/2119408

facilitate streamlined model development and training workflows. Once trained, AI models can be seamlessly deployed into operational scenarios using automated deployment pipelines. Continuous monitoring ensures that models maintain their performance, accuracy and reliability by detecting issues such as data drift or decreased accuracy and triggering automatic retraining when necessary.

3. CONCLUSION AND PERSPECTIVES

The field of EO research has advanced rapidly, driven by increasing demands for reproducibility, scalability, and adherence to FAIR principles. Yet, significant challenges remain in managing and disseminating EO data, especially within scalable, open science infrastructures.

The integration of ML into EO and geospatial analysis introduces further complexity. It requires effective handling of big data, processing near data sources, and rigorous reproducibility to ensure the reliability and scientific validity of results. Addressing these intertwined challenges, the GEP has expanded its capabilities through a comprehensive MLOps framework.

This framework supports the full ML lifecycle—from data ingestion to deployment and monitoring—tailored to the unique needs of users working in geohazard analysis. GEP's MLOps infrastructure is designed to simplify access to ML models, provide scalable computing resources, and enable the development and operationalisation of advanced, AI-driven processing services. By doing so, GEP empowers researchers, practitioners, and public authorities to develop robust, scalable, and reproducible solutions for terrain motion monitoring, disaster response, and long-term environmental analysis.

REFERENCES

[1] E. Rupnik et al., "Refined Satellite Image Orientation in the Free Open-Source Photogrammetric Tools Apero/MicMac", ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., III-1, 83–90, doi:10.5194/isprs-annals-III-1-83-2016, 2016.

[2] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests", Remote Sensing of Environment, vol. 115, issue 10, doi: 10.1016/j.rse.2011.05.013, 2011.

[3] A. Stumpf et al. "Active Learning in the Spatial Domain for Remote Sensing Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 5, pp. 2492-2507, doi: 10.1109/TGRS.2013.2262052, 2014.

[4] P. Horton et al., "Flow-R, a model for susceptibility mapping of debris flows and other gravitational hazards at a regional scale", Natural Hazards and Earth System Sciences, vol. 13, issue 4, doi: 10.5194/nhess-13-869-2013, 2013.

[5] M. Foumelis et al., "SNAPPING Services on the Geohazards Exploitation Platform for Copernicus Sentinel-1 Surface Motion Mapping". Remote Sensing, 14, 6075. doi: 10.3390/rs14236075, 2022.

[6] F. Casu et al., "SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 8, pp. 3285-3296, doi: 10.1109/JSTARS.2014.2322671, 2014.

[7] F. Provost et al., "Terrain deformation measurements from optical satellite imagery: The MPIC-OPT processing services for geohazards monitoring", Remote Sensing of Environment, vol. 274, doi: 10.1016/j.rse.2022.112949, 2022.

[8] Open Geospatial Consortium (OGC), "OGC Best Practice for Earth Observation Application Package", 2021 [Online, accessed May 23, 2025].

[9] A. B. G. Baumann, "Volcano Trial Case on GEP: Systematically processing EO data", Zenodo. doi:10.5281/zenodo.823533, 2017.

[10] F. Orellana et al. "Measuring Coastal Subsidence after Recent Earthquakes in Chile Central Using SAR Interferometry and GNSS Data.", Remote Sensing, 14, 1611, doi: 10.3390/rs14071611, 2022.

[11] F. Provost et al. "High-resolution co-seismic fault offsets of the 2023 Türkiye earthquake ruptures using satellite imagery.", Scientific Reports 14, 6834, doi: 10.1038/s41598-024-55009-5, 2024.

[12] G. Jaramillo et al., "Multi-technique approaches to diagnose unstable ground: Origin of the structural damages in Ciudad Victoria (Loja, Ecuador) case-study", Journal of South American Earth Sciences, vol. 146, doi: 10.1016/j.jsames.2024.105106, 2024.

[13] F. Pacini in Terradue-Discuss, "GEP for Capacity Building in Palu, Sulawesi - Indonesia" [Online, accessed May 23, 2025].

[14] F. Orellana in Terradue-Discuss, "Measuring Coastal Subsidence after Recent Earthquakes in Chile Central Using SAR Interferometry and GNSS Data", [Online, accessed May 23, 2025].

[15] E. Papageorgiou in Terradue-Discuss, "Santorini Volcano Experiencing Another Unrest Phase", [Online, accessed May 23, 2025].

[16] Terradue-Discuss, "Turkiye-Syria Earthquakes: Displacement maps of the GDM-OPT-ETQ service publicly available", [Online, accessed May 23, 2025].

[17] E. Papageorgiou in Terradue-Discuss, "Radar interferogram over Morocco using the Copernicus Sentinel-1 acquisitions of 30 August 2023 & 11 September 2023", [Online, accessed May 23, 2025].

[18] F. Charette-Migneault et al., "Machine Learning Model Specification for Cataloging Spatio-Temporal Models (Demo Paper)", GeoSearch '24: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data, pp. 36-39, doi:10.1145/3681769.3698586, 2024.

ENHANCING WATER QUALITY MONITORING AND GOVERNANCE THROUGH HELOISA: AN EO APPROACH TO AQUATIC SYSTEM MANAGEMENT

Konstantinos Vlachos[1], Konstantinos Karystinakis[1], Anastasia Moumtzidou[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], Ioannis Kompatsiaris[1], George Keradinidis[2], Katerina Kikaki[3], Dimitris Bliziotis[3], Ariane Mueting[4], Rosario Quirino Iannone[4]

[1]CDXi Solutions P.C., [2]TotalView, [3]Hellenic Space Center, [4]European Space Agency

ABSTRACT

Hellenic Operational Integrated Service for Aquatic systems (HELOISA) is one of the projects of the Greek National SmallSat Programme which implements the Water Monitoring Service. The project builds upon three pillars; Water Quantity, Water Quality and Maritime Surveillance. This paper focuses on the Water Quality module and specifically on the core products and technical approach that encompasses. It utilizes Copernicus data, as well as sensors of the Greek SmallSat constellation that provide optical and thermal data. The module covers various water body types, offering maps of water quality proxy variables such as water temperature, chlorophyll-a and water pollutants. Validation and evaluation activities include exploitation of existing historical and newly acquired data ensuring generated product quality. Overall, the HELOISA system is scheduled to be operational in mid to end of 2026, consistently providing water quality products in the Greek territory supporting authorities in informed decision-making and policy implementation.

Index Terms— *Earth Observation, Small Satellite, Artificial Intelligence, Water Quality, Copernicus, Environmental Monitoring*

1. INTRODUCTION

Water resources are of vital importance to ecosystems, human health, and economic prosperity. In Greece, a country characterized by a complex network of inland, coastal and marine water bodies, effective water monitoring systems are essential for sustainable resource management and environmental protection. Inland waters are especially significant as they provide drinking water to large populations including Athens and Thessaloniki, Greece's largest cities, home to over half of the country's population, as well as they support irrigation in agriculture, hydropower generation, flood regulation and biodiversity conservation. In addition, coastal and marine waters in Greece are vital to the country's economy, environment and security, supporting tourism activity, sustaining rich marine biodiversity, and enabling key sectors like fisheries, aquaculture and maritime, among others. Recognizing those needs and also understanding that space is a key enabler for digital transformation, the Hellenic Ministry of Digital Governance and the Hellenic Space Center (HSC), with the assistance of the European Space Agency (ESA), have initiated the Greek National Satellite Space Project. This project is an important step for the materialisation of the strategy of Greece for the utilisation of space technologies and applications and their uptake in the National economy. It includes the development and launch of a small satellite constellation that will cater applications for inland, coastal and marine water monitoring, precision agriculture, land and forest monitoring, as well as border security. The project consists of three Axes. Axis 1 (1.1, 1.2) and Axis 2 comprise the space components responsible for the development and launch of the smallsats. Axis 1.1 will provide thermal data with two spectral bands in about 200m spatial resolution. Axis 1.2 will provide SLC and GRD Synthetic Aperture Radar (SAR) data in various imaging modes (e.g., Scan, Strip, Spot etc.) and spatial resolutions ranging from 0.25m to 15m. Axis 2 is dedicated to multispectral and hyperspectral data in high and very high resolutions reaching about 0.9m (panchromatic band), 18m (VNIR hyperspectral bands), and 3.3m and 19.3m for VNIR and SWIR bands for the multispectral sensor, respectively. On the other hand, Axis 3 is dedicated to the ground components that include the hardware and software facilities, as well as the end-user thematic applications of the different Earth Observation (EO) services that are expected to be delivered alongside the satellite constellations by mid 2026. The thematic applications are categorized in Land, Water, Forest, Agriculture and Security.

2. APPROACH

HELOISA is the project responsible for the delivery of the Water Monitoring Service which aims to develop an advanced monitoring system tailored to the specific needs and requirements of the Greek territory. The Water Monitoring Service builds upon the foundation laid by previous Earth observation initiatives such as the Copernicus program and leverages cutting-edge technology to enhance spatial, temporal, and thematic resolution. By integrating satellite imagery, advanced sensors, well-established established existing traditional and artificial intelligence algorithms, the system aims to provide comprehensive monitoring of water quantity, quality, and maritime surveillance. In this paper, we focus on the Water Quality module of the HELOISA project. The water quality module will be delivering Level-3 satellite products utilizing both Copernicus and the Greek SmallSat data. The areas of interest that will be covered include the majority of the Greek lakes and some lagoons (around 50), as well as all coastal and marine waters, that go beyond the 12 nautical miles. Some of the products that are associated with sudden natural or human-induced environmental changes will be delivered on a daily basis which is equivalent to about 250GB, while the rest will be delivered with higher latency on a weekly basis which is equivalent to about 220GB, covering a large part of the Greek territory depending on the satellite constellation swath. Technologies such as Docker, Xarray, Dask, GDAL multi-threading, scatter/gather workflow pattern implemented in the Common Workflow Language, and Cloud-Optimized GeoTIFFs are utilized. The expected processing time of the applications is less

doi:10.2760/2119408

than 3 hours. Aligned with the objectives outlined by the Hellenic Ministry of Digital Governance and the HSC, HELOISA encompasses a multi-phase approach, encompassing system definition, technical specifications, component design, platform integration, testing, and operational implementation. Through close collaboration with stakeholders and adherence to stringent quality standards, the project seeks to address critical water management challenges while ensuring the sustainability and resilience of water resources in Greece both long- and short-term.

3. APPLICATIONS

3.1. Ocean Color and Temperature

The first water quality application of HELOISA is the Ocean Colour and Temperature delivering products at 10-60m and about 200m nominal spatial resolution, respectively. In particular, chlorophyll-a and Sea Surface Temperature (SST) are retrieved for the coastal and marine waters of the Greek territory (Fig. 1). Chlorophyll-a retrieval is realized utilizing the Copernicus Sentinel-2 multispectral data corresponding to about 26 to 42 scenes per day (or 180 per week). They undergo an atmospheric correction that specifically treats the ubiquitous sunglint effects such as Polymer [1] that is adapted to Sentinel-2. Since Sentinel-2 is not specifically designed for marine water applications, several types of noise and various effects exist, such as striping/fixed-pattern effect, high frequency noise due to waves, and sporadic artifacts due to ships/ship wakes, among others. The cleaning of all these effects/artifacts is a non-trivial task, therefore special treatment is needed for those that can be cleaned, which is developed on a data-driven basis utilizing detector footprints, image statistics, and other chl-a data when possible (e.g., in situ and Copernicus). The products are foreseen to be delivered daily/weekly depending on the spatial resolution, including all open waters. On the other hand, the retrieval of the SST is realized utilizing the thermal sensor of the Greek SmallSat data from the Axis 1.1. The constellation is originally aimed for forest fire applications, however HELOISA takes advantage of the provided thermal channels (3.8μm, 11.45μm) to different extent each, and develops single and dual-channel [2] SST retrieval methods for day and nighttime. The top-of-atmosphere observations are translated from spectral radiances to brightness temperatures and, as a consequence, to SST. This is made possible by generating multilinear regression coefficients through atmospheric correction using the Libradtran Radiative Transfer Model (RTM) library [3] and atmospheric profiles and SSTs from CAMS and CMEMS, respectively. In addition, a dedicated algorithm will be offering cloud masks of different certainty levels.

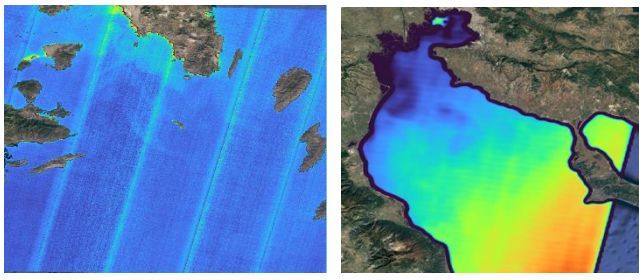


Fig. 1. Marine water quality preliminary products for the Greek territory with chlorophyll-a from Sentinel-2 (left) and preliminary SST from Forest-2 mission (right).

3.2. Muddy Water and Industrial Waste

The second water quality application of HELOISA is the Muddy water and Industrial waste mapping delivering products at a 10m and less than 5m nominal spatial resolution, for Sentinel-2 and Axis 2 data, respectively. The covered territory is foreseen to be almost all of the Greek lakes, as well as coastal waters less than 12 nautical miles. Concerning muddy waters, there are a number of studies attempting to monitor turbid and sediment-laden waters based on satellite remote sensing. Traditionally, the focus has been put on parameter retrieval of turbidity and total suspended matter, but they are not associated with the potential source origin of the polluting sediment. The source could be natural or human-induced, such as industrial waste. The Muddy water and Industrial waste application of HELOISA aims to give semantic information to the sediment-laden waters. The application builds upon Sentinel-2 data that are annotated based on an ensemble methodology as presented in the MUDDAT dataset [4]. An extension of it is implemented, which adds a list of regions presenting coloured waste waters due to industrial activities around the globe (Fig. 2). A custom deep learning framework based on U-Net is trained after performing data preparation such as augmentation and other techniques to adjust for the inherent class imbalance. During inference at least 10496 GPU CUDA cores are utilized. The products undergo post-processing steps such as land-sea masking using the Copernicus 10m Digital Elevation Model (DEM), and filtering to account for systematic and occasional noise effects. Finally, transfer learning is applied so as to generate products using Axis 2 as input data. This is made possible by exploiting the satellite specification similarities but also adjusting for the differences such as different number of spectral bands and pixel size.

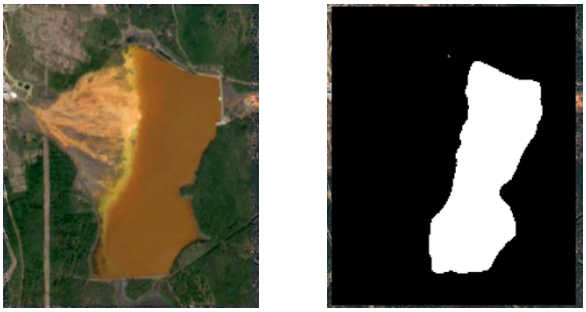


Fig. 2. Industrial waste mapping with a custom U-Net model with the True Color Composite (left) and binary mask (right).

3.3. Oil spills and Surface formations

The third water quality application of HELOISA is the Oil spills and Surface formations mapping, delivering products at a 10m nominal spatial resolution utilizing the Copernicus Sentinel-2 data. The covered territory consists of a list of the Greek lakes. In particular, oil spills are largely identified through radar data, which however present limitations when it comes to inland waters, since the latter demonstrate significant look-alikes due to lake morphology and topography, and low surface roughness due to inconsistent wind conditions. The application fills this gap by offering mapping of oil spill and other suspicious formations at the surface of inland water bodies utilizing multispectral data (Fig. 3). The technological foundation of the approach builds upon the only public multispectral dataset, i.e. Marine Debris and Oil Spill (MADOS) [5] that includes

doi:10.2760/2119408

oil spills, as well as an in-house oil spill Sentinel-2 dataset. A state-of-the-art model (i.e., MariNeXt) which shows high performance is combined with a unique Hydro Foundation Model, which gives higher generalizing power. To make this happen, special modifications of the two models are necessary to adjust for the different number of bands and preprocessing specifications [6]. During inference at least 10496 GPU CUDA cores are utilized. Finally, the product comes with relevant quality flags and masks including sensor viewing and sun geometries to assist the user interpretation.

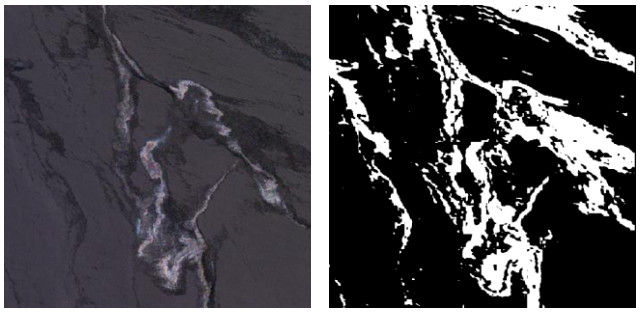


Fig. 3. Oil spill mapping utilizing a custom hybrid deep learning framework with the True Color Composite (left) and the respective binary mask (right).

3.4. Water Quality Features

The fourth, and last, water quality application of HELOISA is the Water Quality Features delivering products at 10m nominal spatial resolution. In particular, the application generates essential water quality variables for inland waters comprising the majority of Greek lakes. This not only includes artificial and natural reservoirs, but also some lagoons. The focus has been put on variables that offer complementary information to the Muddy water and Industrial waste application, which are chlorophyll-a and turbidity (Fig. 4). The estimation of chlorophyll-a and turbidity concentrations in water bodies serves as a major indicator of algal blooms, agricultural practices and pollution. Their retrieval is realized utilizing the Copernicus Sentinel-2 multispectral data after employing water-specific atmospheric correction such as the C2RCC [7] that has been proved to perform well in inland water applications, considering complex Case 2 waters, and also treats adjacency effects, among others. This is possible by using auxiliary information, such as land elevation (e.g., SRTM 30m DEM), water temperature and air pressure (e.g., ERA5), total ozone column, and water salinity. The chl-a retrieval is based on the incorporation of red-edge and near-infrared spectral regions, as well as for turbidity [8]. In addition, auxiliary data will be delivered such as Trophic State Index [9], which indicates the eutrophic state of the waters. Finally, products undergo necessary post-processing and offer relevant quality flag layers.

4. VALIDATION AND EVALUATION

In order to ensure the high quality of the generated Level-3 products, HELOISA and the Water Quality module, in particular, at its core, adopts relevant validation and evaluation practices. To this end, the algorithms of the output products are first verified utilizing proxy/simulated data that are provided in the context of the Greek SmallSat Program, before the advent of the actual Axes data. For instance, in the case of SST, existing Forest-2 mission data ensure the validity of the proposed retrieval algorithm, while in the case of muddy waters, existing Very High Resolution multispectral data are utilized for transfer learning. Furthermore, concerning the marine and inland water quality variables such as chl-a, turbidity and SST, existing in situ historical data are being exploited (e.g., from ARGO [10], project's partner EYATH S.A. and other available data), and new fieldwork campaigns are being conducted for inland, coastal and open waters. Additionally, a match-up analysis protocol has been determined and followed that ensures transparency and quality of outcomes. Finally, manual photointerpretation and quality control of products is conducted by remote sensing experts, and comparison with established existing open datasets from Copernicus and others.

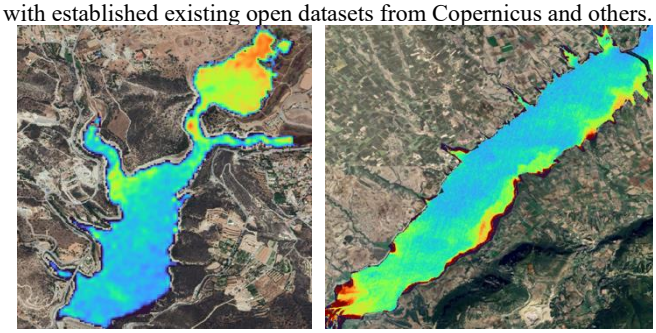


Fig. 4. Inland water quality products for the Greek territory with chlorophyll-a (left) and turbidity (right).

5. IMPACT

The use of EO technology for monitoring water quality brings a wave of positive change across societal, scientific-technical, and economic dimensions. HELOISA leverages national infrastructure and provides satellite-driven insights into inland, coastal and open waters thus addressing pressing environmental challenges while unlocking new opportunities for innovation and growth.

5.1. Societal Impact

Protecting public health stands at the forefront of societal benefits of this EO-driven approach on water quality, as early detection of harmful algal blooms, bacterial contamination, or chemical pollutants allows authorities to issue timely warnings and mitigate risks to drinking water supplies and recreational users. Beyond health, by leveraging national satellite infrastructure, it strengthens the country's autonomy in environmental monitoring, ensuring that critical data for decision-making is generated domestically, thus enhancing national capacity and resilience. Moreover, access to high-quality EO data empowers water authorities to enforce regulations more effectively, supporting compliance with major frameworks such as the Water Framework Directive and the Marine Strategy Framework Directive and shifting governance from reactive crisis management toward proactive and evidence-based management. Furthermore, by integrating EO-derived insights into existing monitoring systems, the project improves the transparency and accountability of water management practices.

5.2. Technical and Scientific Impact

The integration of EO data into water quality monitoring offers critical technical and scientific advantages for both Regional Water Utility operators and National governance bodies. By providing

doi:10.2760/2119408

continuous, wide-scale, and standardized observations, the project enhances the ability of water utilities - who are directly responsible for distributing safe drinking water - to monitor the quality of their source waters more efficiently and with greater spatial and temporal coverage than traditional sampling alone. For Regional Water Utility operators, early identification of emerging threats such as algal blooms, turbidity spikes, or chemical pollutants enables faster, targeted responses that safeguard drinking water treatment processes and distribution networks. Instead of relying solely on periodic field sampling, Water utility operators gain access to near-real-time products, allowing for more proactive management of risks and better protection of public health. At the national level, the project strengthens the technical capacity of Regional Environmental Departments, the Ministry of Environment, and other governmental agencies. With systematic EO data streams, authorities can implement broader surveillance of inland and coastal waters, ensuring regulatory compliance, detecting environmental trends, and evaluating the effectiveness of water protection measures. It provides the scientific backbone for more strategic policymaking, supporting long-term planning for water security, climate adaptation, and biodiversity conservation. The use of national satellite infrastructure also ensures that data sovereignty is maintained, with critical environmental information produced and controlled within the country. This promotes scientific independence and builds national expertise in remote sensing and environmental monitoring technologies. Importantly, the project fosters an integrated approach where EO data is not a replacement for in-situ monitoring but a powerful complement, bridging gaps and optimizing resource allocation. This hybrid monitoring model raises the scientific standard of water quality assessments and offers a replicable framework for future environmental applications, ensuring that both operational needs and strategic priorities are met in a coordinated, technologically advanced manner.

5.3. Economic Impact

By utilizing EO data, the cost of continuous water monitoring is drastically reduced compared to traditional field-based methods, which are resource-demanding and geographically limited. For regional water utilities, this means that broader and more frequent assessments of source water bodies can be achieved without proportional increases in operational expenses. Early detection of potential risks - such as contamination events or seasonal degradation in water quality - allows them to plan interventions more efficiently, mitigating costly and disruptive emergency responses activities. This contributes to a more stable and predictable operational environment, protecting critical infrastructure and minimizing financial risks associated with treatment failures or public health incidents. On the national level, the availability of standardized, large-scale water quality data supports smarter investment planning. Environmental agencies and ministries can prioritize actions based on evidence-based assessments, ensuring that resources are allocated to the most critical areas. Moreover, the integration of national satellite capabilities into operational services can stimulate the growth of value-added industries such as geospatial analytics, environmental consulting, and digital platform development and strengthens the country's positioning in the rapidly growing global market for EO applications. By enhancing water resource governance, the project supports sectors such as tourism, fisheries, and agriculture, ensuring their long-term economic viability, acting as a catalyst for economic modernization, resilience, and sustainable growth.

6. ACKNOWLEDGMENTS

The Project: Small-Satellites (Measure ID 16855) is implemented by the Hellenic Ministry of Digital Governance with the European Space Agency (ESA) Assistance in the Management and Implementation. The project is part of the National Recovery and Resilience Plan 'Greece 2.0', which is funded by the Recovery and Resilience Facility (RRF), core programme of the European Union-NextGenerationEU. The HELOISA consortium comprises TotalView as coordinator, Institute of Communication and Computer Systems (ICCS) of the Technical University of Athens, CDXi Solutions P.C., Center of Security Studies (KEMEA), EYATH S.A.

REFERENCES

[1] Steinmetz, F. and Ramon, D., 2018, October. Sentinel-2 MSI and Sentinel-3 OLCI consistent ocean colour products using POLYMER. In *Remote sensing of the open and coastal ocean and inland waters* (Vol. 10778, pp. 46-55). SPIE.

[2] Merchant, C. J., Harris, A. R., Maturi, E., Embury, O., MacCallum, S. N., Mittaz, J., and Old, C. P. 2009. Sea surface temperature estimation from the Geostationary Operational Environmental Satellite-12 (GOES-12). *Journal of Atmospheric and Oceanic Technology*, *26*(3), 570-581.

[3] Emde, C., Buras-Schnell, R., Kylling, A., Mayer, B., Gasteiger, J., Hamann, U., Kylling, J., Richter, B., Pause, C., Dowling, T. and Bugliaro, L., 2016. The libRadtran software package for radiative transfer calculations (version 2.0. 1). *Geoscientific Model Development*, *9*(5), pp.1647-1672.

[4] Psychalas, C., Vlachos, K., Moumtzidou, A., Gialampoukidis, I., Vrochidis, S. and Kompatsiaris, I., 2024, July. MUDDAT: A Sentinel-2 Image-Based Muddy Water Benchmark Dataset for Environmental Monitoring. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2902-2907). IEEE.

[5] Kikaki, K., Kakogeorgiou, I., Hoteit, I. and Karantzalos, K., 2024. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *210*, pp.39-54.

[6] Anagnostopoulos, C.G.E., Vlachos, K., Moumtzidou, A., Gialampoukidis, I., Vrochidis, S. and Kompatsiaris, I., 2025, June. Hybrid Deep Learning for Oil Spill Mapping Leveraging Sentinel-2 and Foundation Models. In *ESA Living Planet Symposium 2025*. https://zenodo.org/records/15806255

[7] Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S. and Ruescas, A., 2016, August. Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In *Living Planet Symposium* (Vol. 740, p. 54).

[8] Nechad, B., Dogliotti, A., Ruddick, K. and Doxaran, D., 2016, May. Particulate backscattering and suspended matter concentration retrieval from remote-sensed turbidity in various coastal and riverine turbid waters. In *Living Planet Symposium, Proceedings of the conference held* (pp. 9-13).

[9] Carlson, R.E., 1977. A trophic state index for lakes 1. *Limnology and oceanography*, *22*(2), pp.361-369.

[10] Johnson, G.C., Hosoda, S., Jayne, S.R., Oke, P.R., Riser, S.C., Roemmich, D., Suga, T., Thierry, V., Wijffels, S.E. and Xu, J., 2022. Argo—Two decades: Global oceanography, revolutionized. *Annual review of marine science*, *14*(1), pp.379-403.

doi:10.2760/2119408

LEVERAGING SELF-SUPERVISED LEARNING FOR CROP PHOTOGRAPH CLASSIFICATION FROM LIMITED DATA: INSIGHTS FROM THE LUCAS DATASET

Anastasiia Safonova [a], *Stefan Stiller* [a,b], *Momchil Yordanov* [c], *Masahiro Ryo* [a,b]
[a] Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany
[b] Environment and Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany
[c] SEIDOR Consulting S.L., 08500 Barcelona, Spain

ABSTRACT

The dependency on large labeled datasets significantly limits the practical applicability of supervised learning (SL) in remote sensing (RS) applications, particularly in agriculture, where obtaining labeled data is costly. Self-supervised learning (SSL) approaches, which require fewer labeled data, have emerged as viable alternatives. This study evaluates the performance of the VICReg SSL framework against traditional SL models using the Land Use/Cover Area Frame Survey (LUCAS) dataset for crop classification tasks. Our experiments reveal that SSL, particularly using a ResNeXt-50 backbone, achieves superior classification accuracy and robustness, especially under limited data scenarios (5% labeled data), outperforming standard SL methods. We discuss the implications for RS and suggest directions for further research.

Index Terms— Self-supervised learning, VICReg, crop classification, limited data, LUCAS dataset

1. INTRODUCTION

Artificial Intelligence (AI) methods, particularly Supervised Learning (SL), have advanced remote sensing (RS) tasks such as image classification, object detection, and segmentation, enabling precise satellite and aerial imagery analysis for applications like crop monitoring and forest inventory management. Convolutional Neural Networks (CNNs) are extensively employed due to their effectiveness in recognizing spatial patterns [1], [2], [3]. However, SL heavily depends on large labeled datasets, which are costly and difficult to obtain in agriculture, especially in regions facing data scarcity and high annotation costs. Models trained on extensive but region-specific datasets also struggle to generalize to new environments. Developing models that perform well with limited labeled data is crucial for enabling scalable and cost-effective agricultural monitoring, especially in regions where data collection is challenging. Such approaches support timely decision-making for sustainable crop management and food security. Consequently, alternative methods such as Self-Supervised Learning (SSL), Transfer Learning (TL), and semi-supervised learning have gained attention, aiming to leverage unlabeled data or transfer knowledge from related domains [4]. SSL, in particular, reduces labeling dependence by pre-training models on unlabeled data, making it highly suitable for scenarios with limited annotated samples, common in agricultural applications. Despite its potential, SSL remains under-explored in crop classification tasks under small-sample conditions.

This study compares SL and SSL performance for agricultural crop classification under limited data availability. We utilize the Land Use/Cover Area Frame Survey (LUCAS) dataset, consisting of 1,000 images per crop for 10 crop types (common wheat, barley, oats, maize, potatoes, sugar beet, sunflower, rape, soya, and grassland) [5]. Selected SL models (e.g., VGG16 [1], Inception [6], ResNet-18/50 [2], SqueezeNet [7], ResNeXt-50 [8], MobileNet-V2 [9], ShuffleNet [10], EfficientNet-V2 [3], ConvNeXt Tiny [11]) were trained using standard supervised training with cross-entropy loss. For SSL, the same architectures were pretrained using Variance-Invariance-Covariance Regularization (VICReg) [12] and subsequently fine-tuned. Cross-validation (CV), TL, fine-tuning (FT), data augmentation (DA), and varying training ratios (TR) were employed to ensure robust comparisons.

2. LUCAS DATASET

The LUCAS dataset harmonizes land use and land cover information across all 28 EU countries [5]. It comprises approximately 1.35 million observations from 651,780 locations and includes 5.4 million photos collected from 2006 to 2022 (Fig. 1) [13], [14]. Each observation includes photographs taken from four cardinal directions (north, south, west, east), a point photo, and a cover photo, enabling clear crop identification. For this study, we selected ten crops with 1,000 images each, to avoid class imbalance issues present in less represented crops [13]: common wheat (B11), barley (B13), oats (B15), maize (B16), potatoes (B21), sugar beet (B22), sunflower (B31), rape (B32), soya (B33), and grassland (B55) (Fig. 1).

However, the dataset has inherent limitations. Variability arises from manual photo collection using different devices across varied times, dates, and crop stages, causing potential errors in crop labeling and image quality issues (blur, noise,

doi:10.2760/2119408

overexposure, rotations). Additionally, crops can be challenging to distinguish visually, particularly at early growth stages or post-harvest. Nevertheless, the dataset's complexity and diverse perspectives often eliminate the need for data augmentation.

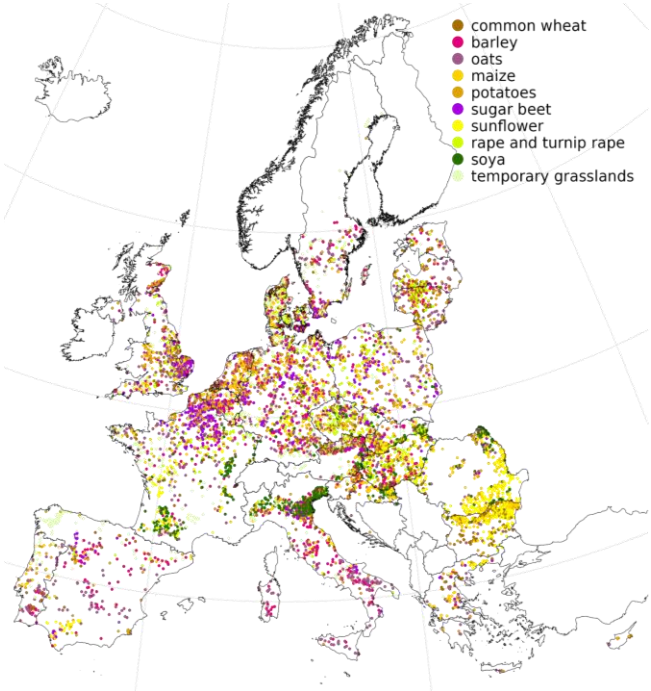


Fig.1. Geographical distribution of 10 crops of the LUCAS dataset across the EU's territory.

3. METHODOLOGY

3.1. Supervised learning

We selected a diverse set of widely recognized neural network architectures, covering both high-capacity models (VGG16 [1], Inception [6], ResNet-18/50 [2], SqueezeNet [7], ResNeXt-50 [8], MobileNet-V2 [9], ShuffleNet [10], EfficientNet-V2 [3], ConvNeXt Tiny [11]).

3.2. Self-supervised learning with VICReg

SSL methods leverage unlabeled data for pre-training, valuable for scenarios with limited labeled data such as crop classification. We used VICReg, a non-contrastive SSL method, to pre-train visual representations from unlabeled LUCAS images [12]. VICReg consists of: (i) a CNN backbone that extracts morphological features; (ii) a projector network that maps features into an embedding space; and (iii) a classification head that is utilized exclusively during supervised fine-tuning. The VICReg loss combines three terms as:

$$\mathcal{L}_{VICReg} = \lambda \cdot s(Z, Z') + \mu \cdot [v(Z) + v(Z')] + \nu \cdot [c(Z) + c(Z')],$$

where the invariance loss $s(Z, Z')$ encourages similarity between embeddings of augmented views, the variance loss $v(Z)$ ensures diversity in features, and Covariance loss $c(Z)$ penalizes redundancy across feature dimensions.

During fine-tuning, the CNN backbone was frozen to preserve pre-trained embeddings, and a supervised classification head was trained. This head consists of a fully connected layer and a Softmax activation, optimized via cross-entropy loss, using the same augmentations as SSL.

We separately optimized SSL and SL hyperparameters using Bayesian optimization and Optuna multi-objective tuning frameworks. Key VICReg hyperparameters included weight decay, cosine annealing schedules, and embedding dimensions. SL fine-tuning hyperparameters (learning rate, weight decay) targeted stable convergence and robust performance.

4. EXPERIMENTAL RESULTS

4.1. Supervised learning

We trained several widely-adopted SL models (see Section 3.1) with and without data augmentation (DA) such as random horizontal flipping, cropping, adjusting brightness and contrast, and randomly rotating. Table 1 summarizes their performance.

Table 1. Comparing performance of selected SL models on the LUCAS dataset with and without DA.

Model	Without DA			With DA		
	Loss	Train Acc	Test Acc	Loss	Train Acc	Test Acc
VGG16	0.49	0.86	0.84	0.58	0.8	0.78
Inception	0.8	0.78	0.76	0.69	0.78	0.76
ResNet-18	0.05	0.98	0.76	0.36	0.88	0.86
ResNet-50	0.43	0.88	0.63	0.38	0.84	0.84
SqueeseNet	0.66	0.85	0.73	0.74	0.81	0.7
ResNeXt-50	0.73	0.99	0.86	0.5	0.8	0.8
MobileNet-V2	0.76	0.74	0.72	0.99	0.66	0.59
ShuffleNet	0.99	0.99	0.82	0.97	0.86	0.73
EfficientNet-V2	0.46	0.96	0.68	0.5	0.92	0.66
ConvNeXt Tiny	0.95	0.76	0.66	0.97	0.76	0.65

To further justify the choice of DA strategies, we conducted additional experiments quantifying its impact on model performance. These experiments confirmed that DA notably improves accuracy for deeper architectures but can be neutral or detrimental for lightweight models due to limited representational capacity. We found that (i) DA significantly improved test accuracy for deeper architectures such as ResNet-18 (from 76% to 86%) and ResNet-50 (from 63% to 85%); (ii) Lightweight models (MobileNet-V2, ShuffleNet) performed worse with DA due to limited representational capacity; (iii) Models like ConvNeXt Tiny struggled regardless of DA, highlighting sensitivity to dataset size. Overall, ResNet models showed robust performance,

doi:10.2760/2119408

balancing accuracy and training efficiency. The strong accuracy was for dominant classes (maize, potatoes, sunflower, rape) but revealed confusion among similar classes (oats, barley, grassland).

4.2. VICReg self-supervised learning

VICReg SSL pretraining utilized unlabeled data (1000 epochs). Embedding dimensions were set at 512 (ResNet-18) and 2048 (ResNet-50, ResNeXt-50). The training loss decreased steadily, indicating effective convergence. Embedding visualizations by UMAP (Fig. 2) demonstrated distinct, semantically meaningful clusters, particularly with ResNeXt-50.

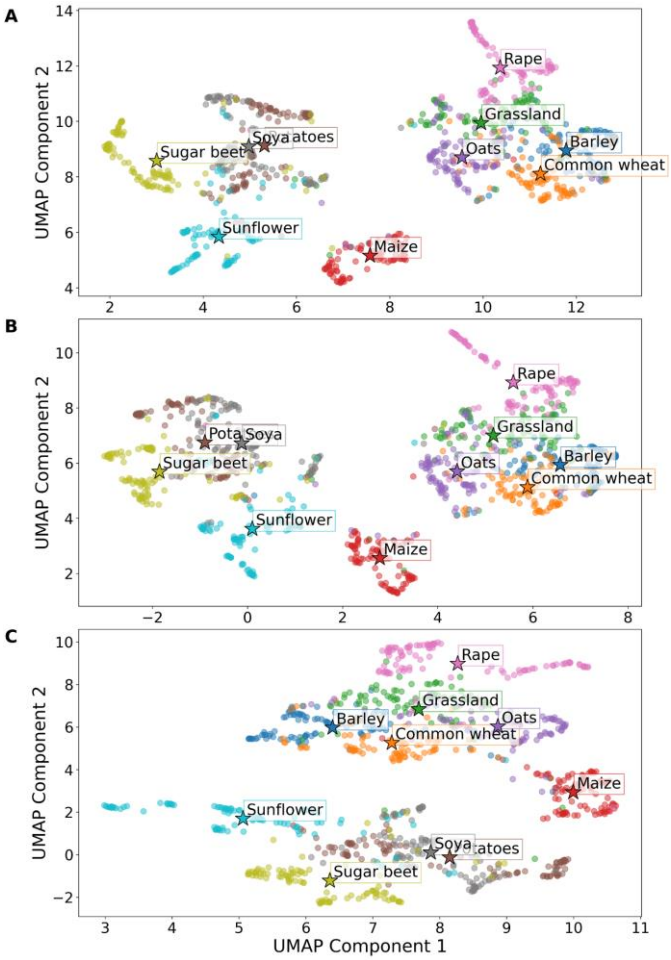


Fig.2. UMAP visualization of learned embeddings on the LUCAS test datasets after VICReg pre-training with ResNet-18 (A), ResNet-50 (B), and ResNeXt-50 (C) backbones.

After fine-tuning (100 epochs, frozen backbone) ResNeXt-50 outperformed other models, achieving 91% accuracy and a macro F1-score of 0.91 (Table 2).

Table 2. A comparison of the training and test performance of VICReg models on the LUCAS dataset without DA.

Backbone	Loss	Acc		F1-score	
		Train	Test	Train	Test
ResNet-18	0.46	0.89	0.87	0.89	0.87
ResNet-50	0.3	0.9	0.89	0.9	0.89
ResNeXt-50	0.17	0.95	0.91	0.95	0.91

4.3. Impact of Training Data Availability

We evaluated SSL robustness using subsets of labeled data (1%-100%). Results (Fig. 3) demonstrated that VICReg SSL achieves competitive accuracy even at minimal supervision levels, surpassing fully supervised baselines at just 5% labeled data. This underscores SSL's effectiveness in limited-data scenarios and suggests saturation of SSL benefits beyond 5-10% training data.

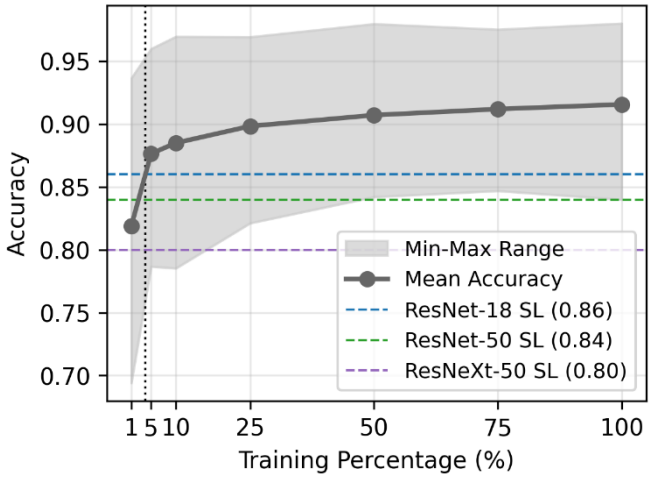


Fig.3. Performance of VICReg with ResNeXt-50 varying proportions of the LUCAS labeled training data compared to the best SL models accuracy.

5. CONCLUSION AND OUTLOOK

Our comparative study between SL and VICReg-based SSL for crop classification using the LUCAS dataset provided several key insights. SL models such as ResNet-18 ResNet-50 demonstrated high accuracy trained on LUCAS dataset with DA. They benefited notably from data augmentation and TL strategies. Conversely, DA had a neutral or even negative effect on lightweight, efficiency-focused models. MobileNet-V2 and ShuffleNet, designed for low-power environments, experienced drops in test accuracy with DA.

In contrast, VICReg-based SSL models exhibited notable robustness under limited labeled data scenarios. Pre-trained SSL models effectively captured transferable features from unlabeled data, significantly enhancing performance even with minimal supervision. Particularly, the ResNeXt-50 backbone consistently outperformed other architectures,

doi:10.2760/2119408

approaching SL model performance using only 5% of labeled data. Embedding visualizations via UMAP further corroborated these quantitative results, showcasing clear and structured crop-type clusters.

However, this study has limitations: (i) the LUCAS dataset's limited size, class imbalance, and geographic restriction to the European Union; (ii) presence of textual labels in images potentially biasing model performance; (iii) exclusive evaluation of VICReg among SSL methods; (iv) the use of frozen backbones during fine-tuning possibly limiting adaptation; and (v) the empirical rather than standardized approach to hyperparameter tuning, model selection, and early stopping.

Future work should address these limitations by evaluating additional SSL frameworks, implementing adaptive fine-tuning strategies, expanding geographic coverage, mitigating textual bias within datasets, and incorporating uncertainty quantification methods to enhance robustness and interpretability. Overall, our findings affirm the significant potential of SSL, particularly VICReg, in efficiently addressing crop classification tasks, especially under conditions of limited annotated data.

ACKNOWLEDGEMENT

This study was supported by the Federal Ministry for Research, Technology and Space (BMFTR – Bundesministerium für Forschung, Technologie und Raumfahrt) project "Multi-modale Datenintegration, domänenspezifische Methoden und KI zur Stärkung der Datenkompetenz in der Agrarforschung (KIKomAG)" (16DKWN089) and the Federal Ministry of Food and Agriculture (BMLEH – Bundesministerium für Landwirtschaft, Ernährung und Heimat) project "Diversifizierte Agrarlandschaftsgestaltung mit vertrauenswürdigem KI-gestütztem Management (DivAg-AIM)" (2823BIJP02).

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[3] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Jun. 23, 2021, *arXiv*: arXiv:2104.00298. doi: 10.48550/arXiv.2104.00298.

[4] A. Safonova, G. Ghazaryan, S. Stiller, M. Main-Knorn, C. Nendel, and M. Ryo, "Ten deep learning techniques to address small data problems with remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 125, p. 103569, 2023.

[5] R. d'Andrimont *et al.*, "Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union," *Sci Data*, vol. 7, no. 1, p. 352, Oct. 2020, doi: 10.1038/s41597-020-00675-z.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 11, 2015, *arXiv*: arXiv:1512.00567. doi: 10.48550/arXiv.1512.00567.

[7] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Nov. 04, 2016, *arXiv*: arXiv:1602.07360. doi: 10.48550/arXiv.1602.07360.

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," Apr. 11, 2017, *arXiv*: arXiv:1611.05431. doi: 10.48550/arXiv.1611.05431.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 21, 2019, *arXiv*: arXiv:1801.04381. doi: 10.48550/arXiv.1801.04381.

[10] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," Dec. 07, 2017, *arXiv*: arXiv:1707.01083. doi: 10.48550/arXiv.1707.01083.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986. Accessed: Aug. 21, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

[12] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," Jan. 28, 2022, *arXiv*: arXiv:2105.04906. doi: 10.48550/arXiv.2105.04906.

[13] M. Yordanov, R. d'Andrimont, L. Martinez-Sanchez, G. Lemoine, D. Fasbender, and M. van der Velde, "Crop Identification Using Deep Learning on LUCAS Crop Cover Photos," *Sensors*, vol. 23, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/s23146298.

[14] M. Ballin, G. Barcaroli, and M. Masselli, "New LUCAS 2022 sample and subsamples design — Criticalities and solutions — 2022 edition." European Union, Jun. 2022. Accessed: Nov. 08, 2024. [Online]. Available: https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-22-005.

ACCURATE MAPPING OF MEKONG DELTA'S MANGROVE DISTRIBUTIONS THROUGH LARGE-SCALE EXTENT LABEL VALIDATION

Duong Cao Phan[1], Anh Vu Vo[2], Quoc Tuan Vo[3], and Quan Le[1]*

[1] Ireland's Centre for AI, School of Computer Science, University of Dublin, Ireland.

[2] School of Computer Science, University of Dublin, Ireland.

[3]Department of Land Resources, Can Tho University, Vietnam.

ABSTRACT

In this study, we present a two-stage, expert-validated labelling pipeline applied over 6,500 km² of southern Vietnam, including the Mekong Delta. We combine Sentinel-2 multispectral time-series, high-resolution Planet NICFI mosaics and ancillary Google basemaps with local knowledge to generate an initial reference map (Label V1), which was systematically refined through targeted expert review to yield Label V2. To quantify the effect of label refinement, we trained and evaluated two LightGBM models on multispectral features, observing a +4.6 % gain in F1-accuracy on the test set (from 87.3 % to 91.9 %). Our best map achieves 99.5 % overall accuracy and 99.7 % F1-score on an independent set of reference points provided by Global Mangrove Watch (GMW), outperforming the GMW baseline (~98.5 % F1), especially in challenging transition zones. In addition, we analyse spatial uncertainty and highlight areas for future SAR integration to support all-weather mapping..

Index Terms—Mangrove ecosystems, Expert-validated labelling, Machine learning, Sentinel-2 MSI, Mekong Delta.

1. INTRODUCTION

Mangroves provide coastal protection, blue-carbon storage, and fisheries habitat, but have declined by ~30 % globally since 1980 [1]. The Mekong Delta is among the hardest-hit regions, where intensive land-use change and accelerated sea-level rise have driven extensive mangrove loss and degradation [2]. Existing global products, e.g., Global Mangrove Watch (GMW) v4.0, report ~95.3 % overall accuracy but misclassify narrow fringing forests and creek edges, perhaps owing to their coarse resolution, causing commission/omission errors in complex transition zones [3], [4]. National-level inventories, while often more detailed at a given point in time, use varying definitions of "mangrove extent" (for example, including all forestland managed by a single agency, regardless of actual canopy cover) and are updated irregularly (some areas every year, others only once every five years), resulting in spatial and temporal inconsistencies [5]. Recent advances in remote sensing and machine learning offer an opportunity to produce maps that are both more accurate and more frequently updated. High-resolution Sentinel-2 imagery (10 m pixels, 5-day revisit) provides rich multispectral data ideal for distinguishing mangrove vegetation [6], while the Planet NICFI basemaps (~5 m) enable fine-scale canopy delineation even in often cloudy tropical regions [7]. However, automated ML approaches remain vulnerable to label noise and boundary uncertainty where training data lack rigorous, expert-validated labels [8].

To address these gaps, we propose an iterative, large-scale, local-knowledge-based annotation protocol coupled with applying LightGBM classifiers to map mangrove extent. Our pipeline (1) generates initial labels by integrating Sentinel-2 multispectral time-series, high-resolution Planet NICFI mosaics and ancillary Google basemaps with local field and stakeholder inputs; (2) refines those labels through multi-expert review to resolve ambiguous edges; and (3) trains a gradient-boosted model on multispectral features to produce high-accuracy, updatable mangrove maps.

2. STUDY AREA

Our study covers approximately 6,500 km² spanning the coastal provinces of southern Vietnam, including the Vietnamese Mekong Delta, characterised by intertidal flats, tidal channels, and mixed aquaculture (Fig. 1). These areas experience high tidal amplitude and frequent cloud cover, which challenge optical-only satellite imagery mapping [5]. The regions include Can Gio (10.4758° N, 106.8650° E), Soc Trang (9.6025° N, 105.9739° E), and Ca Mau (9.1527° N, 105.1961° E). The selection process considered various criteria, with a focus on ecological significance, spatial diversity, accessibility, local knowledge, stakeholder participation, and collaborative efforts between experts in mangrove ecology and artificial intelligence techniques.

Can Gio, designated as a Mangrove Biosphere Reserve, serves as a noteworthy model for the harmonious coexistence of conservation efforts and community livelihoods [9].

[1]* Corresponding author: quan.le@ucd.ie

doi:10.2760/2119408

Notably, Can Gio is distinguished as its strategic location near the largest urban area of Vietnam (Ho Chi Minh City), estuary and coastal zone, coupled with unique biodiversity. In Soc Trang, mangroves play a crucial role along the coastlines, safeguarding them against waves, erosion, and storms [10]. Conversely, mangrove forests in semi-inland Ca Mau exhibit distinct spatial characteristics, distributed along shorelines in both accretion and erosion areas, within a national park, and in production areas marked by intricate patterns and dynamic changes [1], [11].

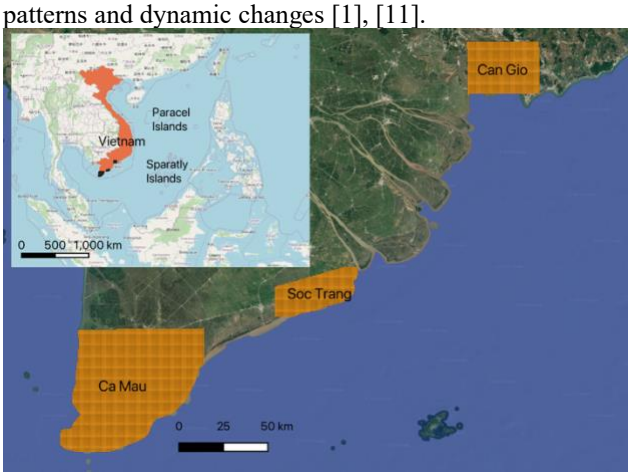


Fig. 1. Study-area map showing the locations of Can Gio (10.4758° N, 106.8650° E), Soc Trang (9.6025° N, 105.9739° E) and Ca Mau (9.1527° N, 105.1961° E) in southern Vietnam (coordinates in decimal degrees, WGS84).

3. DATA & METHODS

3.1. Input Data

Sentinel-2 MSI (L2A): We developed an end-to-end pipeline that automatically ingests all available Sentinel-2 MSI scenes from Google Earth Engine (GEE) for each study tile, applied cloud and cloud-shadow masking using QA60 and scene-classification bands. Then, we computed two key vegetative indices, Normalized Difference Vegetation Index (NDVI) from the 10 m B4 (red) and B8 (NIR) bands, and a Mangrove Vegetation Index (MVI) by incorporating the 20 m B11 (SWIR-1) band alongside B4 and B8.

Planet NICFI Basemaps: Planet NICFI basemaps consist of monthly 4.77 m resolution tropical mosaics, which we used to enhance fine-scale detection of fringing mangroves during our labelling period (2020).

Google Basemaps: High-resolution RGB and street imagery for contextual reference via QGIS QuickMapServices were used to locate each mangrove extent polygon.

3.2. Two-stage Labelling Pipeline

Label Version 1 (V1): Twenty-four trained local annotators delineate mangrove polygons in QGIS using a standardized protocol integrating true-colour composites, NDVI, MVI, mangrove extent maps created by the local government, and local expert notes.

We identified multiple deficiencies in Label V1 that necessitated systematic refinement. Specifically, V1 exhibited omission and commission errors—namely, missing mangrove patches and misclassification of mangrove as adjacent land-cover types—as well as spatial inaccuracies manifested as displaced or imprecise mangrove-extent boundaries; these shortcomings motivated the development of the corrected Label V2.

Label Version 2 (V2): We recruited a team of eight domain experts to produce mangrove annotations across the study area. An independent four-member review panel then assessed the annotated mangrove-extent polygons, reconciled discrepancies, and corrected polygon geometry where required. Through this expert adjudication and polygon-correction workflow we produced the final mangrove extent label (Label V2).

3.3. Feature Engineering & LightGBM Modelling

Feature Set: We implemented an end-to-end, fully automated preprocessing pipeline that ingests Sentinel-2 MSI Level-2A (surface-reflectance) products and extracts ten spectral bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12). The pipeline applies standard preprocessing operations (e.g., cloud and cloud-shadow masking), produces per-pixel median composites over the analysis period, and exports spatially tiled patches as model-ready inputs for downstream modelling, ensuring reproducibility and a complete audit trail of preprocessing steps

Training Splits: For each label set (~62 million pixels of 10x10 m^2), we randomly divided its set of tiles into training/validation/test splits with the ratios70 % / 15 % / 15 % [12].

Classifier: LightGBM with 1,000 trees, max depth = 16, learning rate = 0.05.

Model Variants: Model V1: Trained on Label V1 and Model V2: Trained on Label V2.

3.4. Evaluation Protocol

Pixel-Level Metrics: We evaluated model performance using standard classification metrics, including the macro-average F1-score, precision, and recall.

Point-Based Validation: An independent set of 19,688 reference points provided by the Global Mangrove Watch (GMW) were used to compute the overall accuracy and F1-score of the trained models.

4. RESULTS

4.1. Label Consistency & Model Performance

4.2. Through a structured expert-review process, we enhanced the label quality, improving the F1-score on the test set from 87.3% for Label V1 to 91.9% for Label V2. The other improvement metrics are reported in Table 1. To be

doi:10.2760/2119408

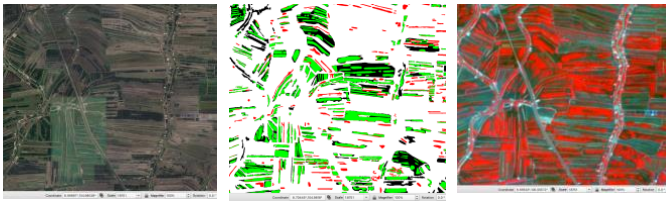
specific, during the review process, we identified and corrected inconsistent mangrove polygons, particularly around complex creek boundaries. Figure 2 illustrates an example of an initial mislabel that was subsequently corrected, while Figure 3 shows a case of disagreement between trained local annotators and senior map experts. These discrepancies were systematically reviewed and resolved through multiple iterative rounds of expert validation. This two-stage refinement yielded our high-confidence V2 dataset.

Table 1. Performance on the test splits of the LightGBM model trained on Label V1 and Label V2.

Metric (%)	Model V1	Model V2	Δ (pp)
Test F1-score	87.3	91.9	+ 4.6
Recall	86.1	90.8	+ 4.7
Precision	88.6	93.2	+ 4.6

(a) (b) (c)

Fig. 2. This is an example of commission labels by the trained local annotators during the labelling process at this location (8.74090, 104.99630): (a) Google high-resolution RGB basemaps, (b) Annotated label, and (c) NICFI basemaps used to detect mangrove extent during the labelling period (2020).



(a) (b) (c)

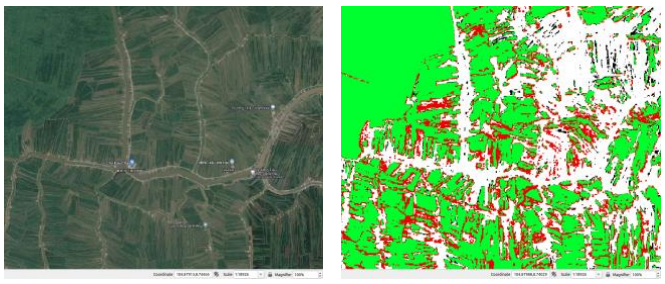
Fig. 3. This is an example of inconsistencies between trained local annotators and experts at this location (8.69533, 105.00572): (a) Google high-resolution RGB basemaps, (b) Labels (Green ~ mangrove labelled by both the trained local annotators and experts; Red ~ experts labelled mangrove while the trained local annotators missed it; and Black ~ experts labelled non-mangrove while the trained local annotators labelled mangrove), and (c) NICFI basemaps used to detect mangrove at the labelling period (2020).

4.3. Validation on the independent reference set

Table 2 shows the accuracy of our map (Model V2) and the GMW 4.0 map validated by a reference set of 19,688 GMW reference points. Model V2 achieved an overall accuracy of 99.5% (F1-score = 99.7%) and GMW's published 95.3 % global accuracy (F1-score of 98.5% for the reference set) [4], [7]. Based on local expert knowledge, our map outperforms the GMW map in narrow-fringe, changeable (plantation or productivity mangrove forests), or mixed-substrate regions (Fig. 4).

Table 2. Comparison of our map and the GMW 4.0 map validated by the GMW independent reference points (19,688 points [4], [7]).

Metric (%)	GMW 4.0	Our map
Accuracy	98.6	99.5
F1-score	98.5	99.7



(a) (b)

Fig. 4. An example to compare our map and the GMW v4.0 map in 2020 at (8.74465, 104.87913): (a) Google high-resolution RGB basemaps and (b) Overlap of our map and the GMW v4.0 map (Green ~ mangrove agreed by both our map and the GMW v4.0 map; Red ~ non-mangrove in our map but mangrove in the GMW v4.0 map, and Black ~ mangrove in our map but non-mangrove in the GMW v4.0 map), and (c) NICFI basemaps used to detect mangrove at the labelling period.

5. DISCUSSION

Applying a mutual, iterative labelling workflow to generate our mangrove training dataset yielded a 4.6 pp gain in F1-score, underscoring the efficacy of repeated local-expert refinement in reducing ambiguous labels [13]. In our approach, initial labels were systematically reviewed and corrected by trained local annotators, particularly along habitat edges and in mixed aquaculture zones, thereby eliminating small but systematic misclassifications before retraining. These successive expert feedback loops translated directly into more accurate segmentation outputs, demonstrating that targeted correction of difficult examples is critical for maximizing accuracy in remote-sensing habitat mapping.

When compared with the Global Mangrove Watch (GMW) v4.0 product, our map exhibits both lower omission and commission errors across heterogeneous landscapes. GMW v4.0 is an improved version of v3.0, which is documented to have global commission and omission rates of ~10–15 % and ~14–16 %, respectively, in its 1996–2020 change assessment [4], and regional error rates can exceed 10 % commission and 20 % omission in fragmented zones [7]. We do not find such detailed documented information in the GMW v4.0. Its overall accuracy reaches 95.3 % compared with 93.1 % of the v3.0 [3], [4]. Interestingly, our model reliably captures small, isolated mangrove patches in aquaculture mangrove mosaics, features that GMW v4.0 often omits and avoids the slight over-prediction in dense cores observed in the baseline product (Fig. 3b) [3].

Accurately defining mangrove patch perimeters remains difficult for several primary geospatial reasons. First, Sentinel-2 L2A products have documented geolocation

doi:10.2760/2119408

accuracies of up to 12 m (95 % conf.) [14] and NICFI basemaps up to 10 m (90 % conf.) [15], leading to co-registration shifts of several pixels along shorelines. Second, even when higher-accuracy co-registration methods are applied, residual misalignments of 2–10 m between Sentinel-1 SAR and Sentinel-2 optical imagery persist, particularly in complex estuarine zones [16]. These small offsets accumulate along convoluted boundaries, inflating mislabelled areas and depressing the mapping accuracy.

Although our map attains 99.5 % overall accuracy against 19,688 GMW reference points, the is still a gap (Fig. 3b) in tackling pervasive, small-scale boundary misalignments. The GMW point samples might tend to cluster in easily recognized core areas and under-sample transitional pixels at complex edges, yielding overly optimistic accuracy estimates [17]. Moreover, misregistration in the GMW v4.0 dataset contributes random errors that inflate confidence intervals for change estimate errors in both omission and commission [4], but these point-based metrics fail to capture systematic boundary drift in fragmented stands.

To further refine boundary precision, we recommend integrating high-resolution, co-registered datasets such as UAV-based data, which has been shown to improve segmentation accuracy of individual mangrove trees beyond canopy-height models [18]. Combined optical–LiDAR approaches, for example, using UAV LiDAR with very-high-resolution WorldView-2 imagery, reduce geolocation uncertainty and enhance delineation of small patches and narrow fringing zones [19]. Additionally, targeted collection of dense boundary-point samples, or the adoption of segmentation masks derived from UAV imagery, will yield more representative validation of edge performance. Above all, maintaining an iterative expert-labelling framework will remain helpful for capturing subtle, site-specific nuances in complex and anthropogenically altered mangrove habitats.

6. CONCLUSIONS & FUTURE WORK

We demonstrate that large-scale, expert-validated labelling combined with LightGBM yields highly accurate mangrove extent maps for the Mekong Delta, surpassing global baselines, especially in critical zones. Future work will (i) scale to the entire Vietnam, (ii) fuse Sentinel-1 SAR for all-weather mapping & test the label datasets with advanced deep learning models, and (iii) operationalize an active-learning loop for continual label refinement.

Acknowledgement: This work was funded by Taighde Éireann - Research Ireland grant 22/FIP/SDG/10899 "AI Solutions for Mangrove Blue Carbon in Vietnam"
under the SFI Future Innovator Prize 2022.

REFERENCES

[1] B. Choudhary, V. Dhar, and A. S. Pawase, "Blue carbon and the role of mangroves in carbon sequestration: Its mechanisms, estimation, human impacts and conservation strategies for economic incentives," *J Sea Res*, vol. 199, p. 102504, 2024, doi: https://doi.org/10.1016/j.seares.2024.102504.

[2] H. Nguyen *et al.*, "Transformation of rural landscapes in the Vietnamese Mekong Delta from 1990 to 2019 : A spatio-temporal analysis," *Geocarto Int*, 2022.

[3] M. Leal and M. D. Spalding, Eds., "The State of the World's Mangroves 2024," Jul. 2024. doi: 10.5479/10088/119867.

[4] P. Bunting *et al.*, "Global Mangrove Extent Change 1996–2020: Global Mangrove Watch Version 3.0," *Remote Sens (Basel)*, vol. 14, no. 15, Aug. 2022, doi: 10.3390/rs14153657.

[5] D. C. Phan *et al.*, "First comprehensive quantification of annual land use/cover from 1990 to 2020 across mainland Vietnam," *Sci Rep*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-89034-5.

[6] M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sens Environ*, vol. 120, pp. 25–36, 2012, doi: https://doi.org/10.1016/j.rse.2011.11.026.

[7] P. Bunting *et al.*, "Global Mangrove Watch: Monthly Alerts of Mangrove Loss for Africa," *Remote Sens (Basel)*, vol. 15, no. 8, Apr. 2023, doi: 10.3390/rs15082050.

[8] C. M. Roelfsema *et al.*, "Workflow for the Generation of Expert-Derived Training and Validation Data: A View to Global Scale Habitat Mapping," *Front Mar Sci*, vol. 8, Mar. 2021, doi: 10.3389/fmars.2021.643381.

[9] H. T. Le *et al.*, "Characterizing spatiotemporal patterns of mangrove forests in can gio biosphere reserve using sentinel-2 imagery," *Applied Sciences (Switzerland)*, vol. 10, no. 12, Jun. 2020, doi: 10.3390/APP10124058.

[10] K. Schmitt, T. Albers, T. T. Pham, and S. C. Dinh, "Site-specific and integrated adaptation to climate change in the coastal mangrove zone of Soc Trang Province, Viet Nam," *J Coast Conserv*, vol. 17, no. 3, pp. 545–558, 2013, doi: 10.1007/s11852-013-0253-4.

[11] L. T. M. Nguyen, H. T. Hoang, H. Van Ta, and P. S. Park, "Comparison of mangrove stand development on accretion and erosion sites in Ca Mau, Vietnam," *Forests*, vol. 11, no. 6, Jun. 2020, doi: 10.3390/F11060615.

[12] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, doi: 10.1002/sam.11583.

[13] Z. Wang *et al.*, "A Deep Learning Based Platform for Remote Sensing Images Change Detection Integrating Crowdsourcing and Active Learning," *Sensors*, vol. 24, no. 5, Mar. 2024, doi: 10.3390/s24051509.

[14] R. Rengarajan, M. Choate, M. N. Hasan, and A. Denevan, "Co-registration accuracy between Landsat-8 and Sentinel-2 orthorectified products," *Remote Sens Environ*, vol. 301, p. 113947, 2024, doi: https://doi.org/10.1016/j.rse.2023.113947.

[15] H.-E. Wei, M. Grafton, M. Bretherton, M. Irwin, and E. Sandoval, "Evaluation of the use of two-stage calibrated PlanetScope images and environmental variables for the development of the grapevine water status prediction model," *Technology in Agronomy*, vol. 3, no. 1, pp. 0–0, 2023, doi: 10.48130/tia-2023-0006.

[16] Y. Ye, C. Yang, B. Zhu, L. Zhou, Y. He, and H. Jia, "Improving Co-Registration for Sentinel-1 SAR and Sentinel-2 Optical Images," *Remote Sens (Basel)*, vol. 13, no. 5, 2021, doi: 10.3390/rs13050928.

[17] J. Tang, J. Cheng, D. Xiang, and C. Hu, "Large-Difference-Scale Target Detection Using a Revised Bhattacharyya Distance in SAR Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3161931.

[18] H. You, Y. Liu, P. Lei, Z. Qin, and Q. You, "Segmentation of individual mangrove trees using UAV-based LiDAR data," *Ecol Inform*, vol. 77, p. 102200, 2023, doi: https://doi.org/10.1016/j.ecoinf.2023.102200.

[19] Y. Zhu, S. W. Myint, K. Liu, L. Liu, and J. Cao, "Integration of UAV LiDAR and WorldView-2 images for modeling mangrove aboveground biomass with GA-ANN wrapper," *Ecol Process*, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13717-024-00566-w.

OPTIMISATION OF SAMPLING DESIGN FOR MULTIVARIATE SOIL MAPPING WITH MACHINE LEARNING

Jeonghwan Choi[1], Alexander Kmoch[1], Evelyn Uuemaa[1]

[1]Department of Geography, University of Tartu, Tartu, Estonia

ABSTRACT

The quality of data is important for building reliable Machine Learning (ML) models in Digital Soil Mapping (DSM). Insufficient or unrepresentative training data often limits the ML model's accuracy. This study introduces a dissimilarity-driven sampling framework, which integrates Area of Applicability (AOA) and Iterative Dissimilarity-Driven Sample Selection (IDDSS) methods to enhance the reliability of predictions for key soil properties, such as Soil Organic Carbon (SOC). The framework identifies targeted sampling locations that are environmentally distinct from the current training data and underrepresented in the covariate space. Using Estonia as a case study, we determined that approximately 25% of the country lies outside AOA, which indicates where predictions are unreliable. A total of 41,930 targeted sampling locations were identified within these areas. This proposed framework provides a robust data-driven strategy for optimising future fieldwork, which ensures that new samples most effectively enhance the reliability of ML models in DSM.

Index Terms— Machine Learning, Soil Organic Carbon, Digital Soil Mapping, Area of Applicability, Dissimilarity Index.

1. INTRODUCTION

The representativeness and volume of training data influence the performance of Machine Learning (ML) models in Digital Soil Mapping (DSM) [1]. However, the soil legacy data used for training are often sparse and contain spatial biases due to conventional sampling limitations and accessibility constraints [2, 3]. ML models trained on these biased datasets may produce inaccurate predictions when applied to areas with unrepresented environmental conditions [4].

To address current sampling limitations, we proposed a novel dissimilarity-driven sampling framework to optimise the soil fieldwork. Our proposed framework integrates two methods: the Area of Applicability (AOA) method, which identifies areas where ML-based predictions are reliable [4], and the Iterative Dissimilarity-Driven Sample Selection (IDDSS) method, which iteratively selects the most dissimilar locations from underrepresented areas while ensuring they are not too similar to each other in the multivariate feature space. This approach provides a systematic and data-driven method for identifying and addressing the critical gaps in current training datasets. By optimising the selection of new soil samples in this way, our study presents a replicable sampling strategy designed to enhance the reliability of ML predictions.

2. DATA AND METHODS

2.1. Data and Preprocessing

In the study, we used 924 Soil Organic Carbon (SOC) observations from across Estonia (Figure 1), which were derived from 3 data sources [5, 6, 7]. These data, along with 11 environmental covariates (Table 1), served as the basis for predictive Random Forest (RF) modelling.

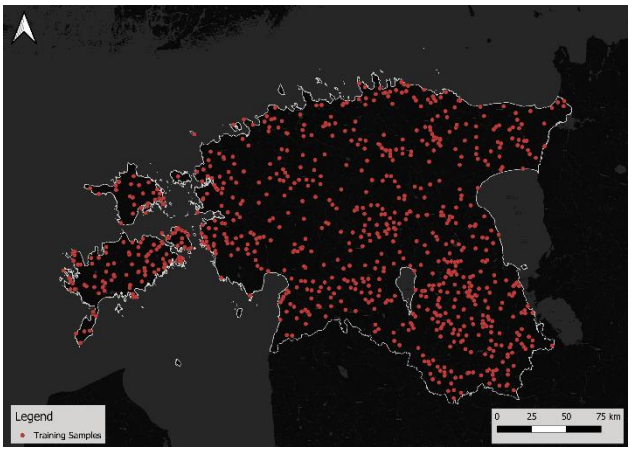


Fig. 1. The distribution of training samples across Estonia.

Our framework identifies targeted sampling locations at the scale of individual soil mapping units from the EstSoil-EH dataset [5]. To prepare the data, we first harmonised all environmental covariates into a single unified dataset to ensure consistency. Using the zonal statistics method from the rasterstats Python package [8], we aggregated raster-formatted covariates (e.g., vegetation dynamics, topographic attributes, land use) to these vector soil units. This process calculated the mean values for continuous covariates (vegetation and

doi:10.2760/2119408

topography) and the majority class for the categorical land use covariate. Subsequently, a spatial join function in GeoPandas [9] combined the harmonised covariates with SOC observations based on geographic coordinates, which created the final modelling dataset for the study.

Table 1. Environmental Covariates used in the study.

Category	Covariates	Type
Vegetation Dynamics [10]	▪ Normalised Difference Vegetation Index (NDVI)	Raster
Soil Texture [5]	▪ Clay ▪ Silt ▪ Sand ▪ Rock	Vector
Topographic Attributes [11]	▪ Slope ▪ LS-factor ▪ Terrain Roughness Index (TRI) ▪ Topographic Wetness Index (TWI)	Raster
Hydrological Features [12]	▪ Drainage ditches	Vector
Land Use [13]	▪ Land use types	Raster

2.2. Framework Overview

Our methodological framework is implemented through the two-phase workflow summarised in Figure 2. The first phase identifies regions where our baseline RF model is likely to be unreliable through an AOA analysis. The second phase uses that information to determine targeted sampling locations within those regions, which can guide future soil fieldwork to enhance the representativeness of training data.

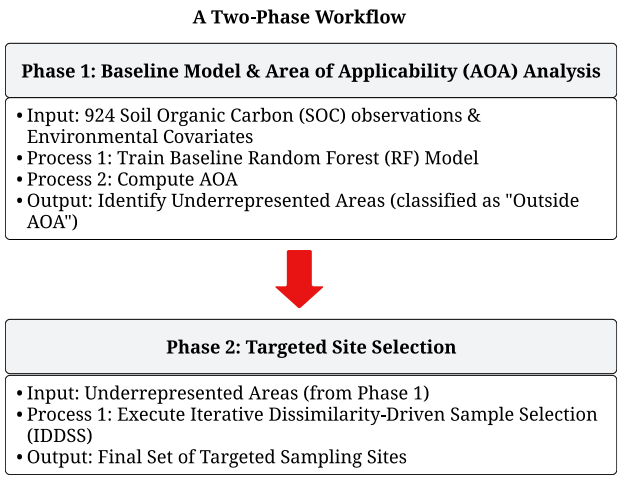


Fig. 2. Methodological framework for the targeted soil sampling design.

2.2.1. Baseline model

The workflow starts with building a baseline RF model using Python's scikit-learn package [14]. This initial model was trained on the current 924 SOC observations and optimised using a 5-fold cross-validation grid search. The determined hyperparameters are shown in Table 2. Based on this model, the SHapley Additive exPlanations (SHAP) values were calculated and subsequently utilised for the AOA computation [15].

Table 2. Hyperparameters of RF.

Hyperparameter	Value
bootstrap	False
criterion	squared_error
max_depth	10
max_features	sqrt
min_samples_leaf	2
min_samples_split	5
n_estimators	100
random_state	25

2.2.2. AOA analysis

Next, we identified underrepresented regions using the AOA methodology. The AOA is defined by a Dissimilarity Index (DI), which measures the degree of difference between a new location and the training data used for ML modelling in terms of its environmental covariates. A DI threshold is then derived through a cross-validation, which determines the boundary of reliable ML model application. A location with a DI value above this threshold is considered outside the AOA (unreliable predictions), making it inappropriate to use the trained RF model in that location.

2.2.3. Targeted site selection

Following the identification of underrepresented regions, the IDDSS method was applied to select targeted sampling locations (Figure 3). The selection process is based on the AOA results, which selects the candidate with the highest DI value at the start of each iteration. Then, a filtering step uses this newly selected site to trim the candidate pool; its multivariate distance to all other remaining locations is calculated, and any found to be too similar are permanently discarded. This iterative process ensures that targeted sampling locations are both highly informative relative to the current training data and distinct from one another.

doi:10.2760/2119408

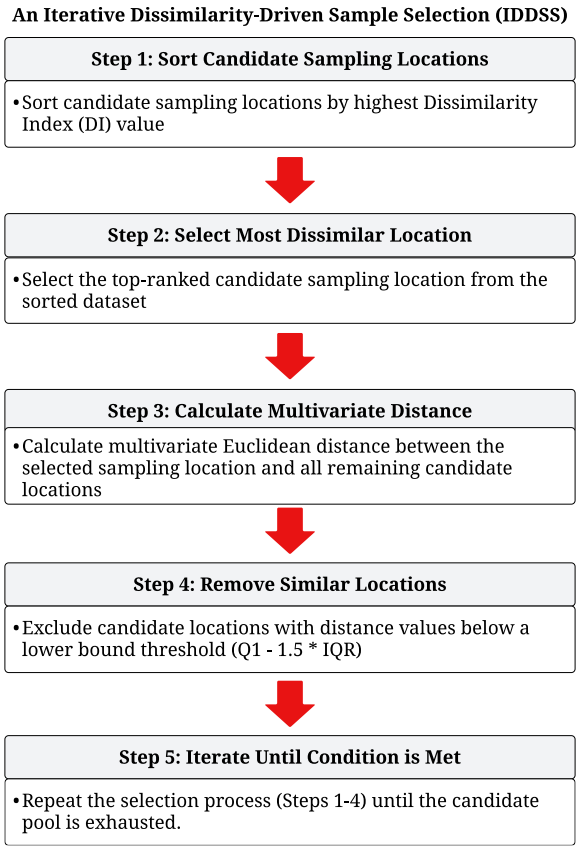


Fig. 3. A workflow for targeted sample selection.

3. RESULTS

Figure 4 shows the spatial distribution of DI values across Estonia. Colours closer to red indicate higher dissimilarity relative to the current training dataset. We found that approximately 25% of Estonia lies beyond the AOA threshold, with a DI greater than 0.11. This result suggests that a substantial part of the country would be unreliable using the RF model trained on the current training data (Figure 5).

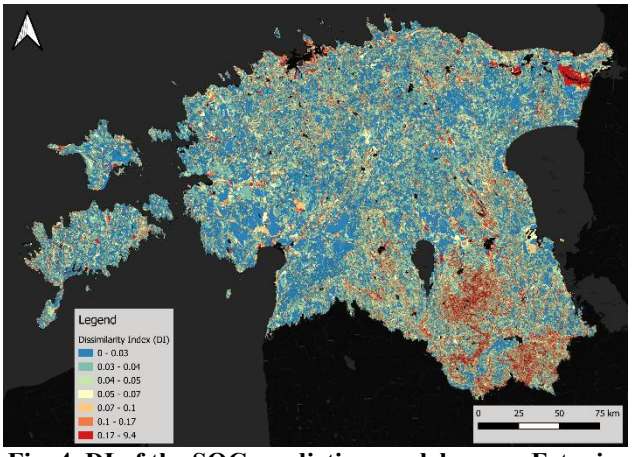


Fig. 4. DI of the SOC prediction model across Estonia.

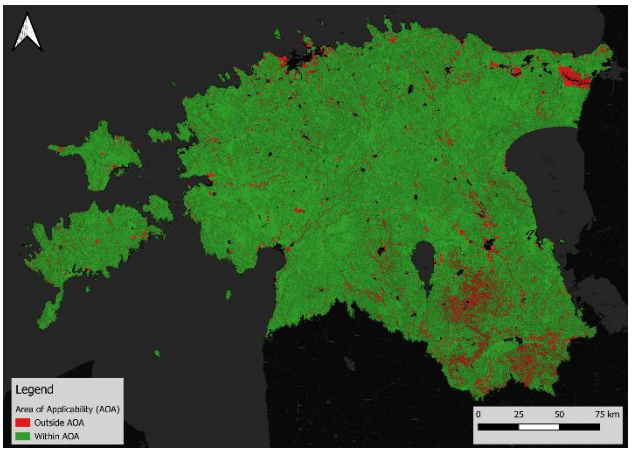


Fig. 5. AOA of the SOC prediction model across Estonia.

Based on the AOA result, we applied the IDDSS method to identify 41,930 targeted sampling locations situated exclusively within these high-uncertainty regions (Figure 6). This analysis highlights the usefulness of the AOA and IDDS frameworks in guiding strategic data acquisition, as they help target underrepresented and environmentally distinct regions not well captured in the current training data.

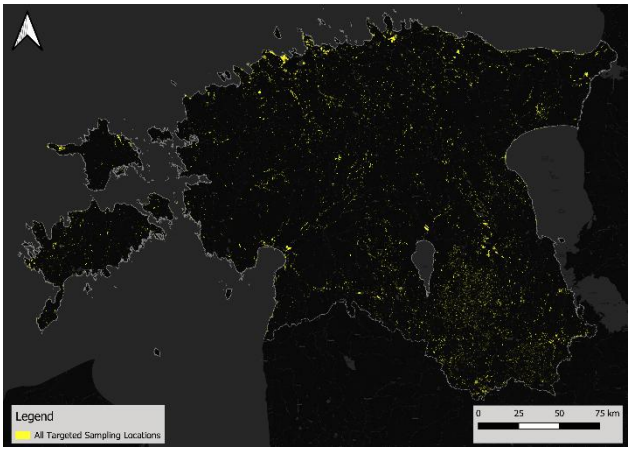


Fig. 6. All Targeted Sampling Locations for SOC across Estonia.

4. DISCUSSION AND CONCLUSION

In this study, we introduced a novel dissimilarity-driven sampling framework to optimise soil sampling design. The proposed framework successfully integrates AOA analysis with an IDDSS method to create a targeted data-driven sampling plan. The application in Estonia shows the usefulness of this approach, which identified that a significant portion of the country (~25%) lies in regions where the current SOC prediction model could be unreliable. These areas are primarily located in southern Estonia alongside the northern coast.

doi:10.2760/2119408

Although conventional sampling strategies like grid sampling or random sampling ensure spatial coverage of the study area, they may inefficiently place new samples in the multivariate feature space that are already well-represented by the current training data. Our proposed framework is designed to be more efficient by directly targeting areas of high ML model uncertainty, thereby improving the predictive reliability of the model. The integration of AOA and IDDSS refines the process of sampling location selection by ensuring that the selected locations are both informative relative to the current training data and distinct enough from one another. We determined a total of 41,930 targeted sampling locations, which represent the full candidate pool for enhancing the quality of the current training data. From this candidate pool, researchers could narrow down a smaller number of sampling locations based on their practical constraints, such as budget, accessibility, and the needed sample size for their research.

The main contribution of this study is the demonstration of this complete methodological framework for soil sampling from the initial RF model training to the final selection of targeted sampling locations. Although a comprehensive validation of the generated sampling plan through a dedicated fieldwork is beyond the scope of this study, the work presented here provides a robust and replicable strategy to guide such an effort. To support reproducibility in research, the Python codes used for the proposed sampling selection approach will be made publicly available after the publication of the related work.

REFERENCES

[1] C. A. Ramezan, T. A. Warner, A. E. Maxwell, and B. S. Price, 'Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data', *Remote Sens.*, vol. 13, no. 3, Art. no. 3, Jan. 2021, doi: 10.3390/rs13030368.

[2] J. Koch, S. Stisen, J. C. Refsgaard, V. Ernstsen, P. R. Jakobsen, and A. L. Højberg, 'Modeling Depth of the Redox Interface at High Resolution at National Scale Using Random Forest and Residual Gaussian Simulation', *Water Resour. Res.*, vol. 55, no. 2, pp. 1451–1469, 2019, doi: 10.1029/2018WR023939.

[3] B. Minasny *et al.*, 'Soil Science-Informed Machine Learning', *Geoderma*, vol. 452, p. 117094, Dec. 2024, doi: 10.1016/j.geoderma.2024.117094.

[4] H. Meyer and E. Pebesma, 'Predicting into unknown space? Estimating the area of applicability of spatial prediction models', *Methods Ecol. Evol.*, vol. 12, no. 9, pp. 1620–1633, 2021, doi: 10.1111/2041-210X.13650.

[5] A. Kmoch *et al.*, 'EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia', *Earth Syst. Sci. Data*, vol. 13, no. 1, pp. 83–97, Jan. 2021, doi: 10.5194/essd-13-83-2021.

[6] O. Fernández-Ugalde *et al.*, *LUCAS 2018 soil module: presentation of dataset and results*. Publications Office of the European Union, 2022. Accessed: Aug. 04, 2025. [Online]. Available: https://data.europa.eu/doi/10.2760/215013

[7] A. Helm, 'Role of Grasslands in Mitigating Climate Change, 2022-2023'. EIC Environmental Program Project. University of Tartu, 2023.

[8] M. Perry, *rasterstats: Summarize geospatial raster datasets based on vector geometries*. (2023). Python. Accessed: Oct. 31, 2024. [OS Independent]. Available: https://github.com/perrygeo/python-raster-stats

[9] K. Jordahl *et al.*, *geopandas/geopandas: v0.8.1*. (Jul. 15, 2020). Zenodo. doi: 10.5281/zenodo.3946761.

[10] European Space Agency, 'Sentinel-2 – Documentation'. Accessed: August 30, 2025. [Online]. Available: https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel2.html

[11] Estonian Land Board, 'Elevation Data'. 2022. [Online]. Accessed: August 30, 2025. Available: https://geoportaal.maaamet.ee/eng/Maps-and-Data/Elevation-data/Download-Elevation-Data-p664.html

[12] Estonian Land Board, 'Estonian Topographic Database'. 2023. Accessed: August 30, 2025. [Online]. Available: https://geoportaal.maaamet.ee/eng/spatial-data/estonian-topographic-database-p305.html

[13] Estonian Ministry of the Environment, 'ELME2 project final report', 2023. [Online]. Available: https://loodusveeb.ee/sites/default/files/inline-files/ELME2_LOPPARUANNE_fin_151123.pdf

[14] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[15] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions', *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Jul. 10, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

THE FOREST IN A FUNCTION: DEMOCRATIZING DEEP LEARNING FOR FLEXIBLE AND SCALABLE EO ANALYSIS

Loïc Dutrieux[1], *Keith Araño*[*2], *Pieter Kempeneers*[1]

[1]European Commission, Joint Research Centre, Italy
[2]Arcadia Sistemi Informativi Territoriali SRL, Italy

ABSTRACT

The application of Deep Learning (DL) to Earth Observation (EO) has yielded remarkable advancements across diverse tasks, including land cover classification and biophysical variable regression. However, a significant gap persists between the development of sophisticated DL models and their routine deployment in operational settings. This disconnect largely stems from the technical expertise needed to handle complex DL workflows, which differs from standard geospatial practices. To address this challenge, we introduce `xinfereo`, a Python package designed to bridge the divide between xarray Datasets, a common data structure for geospatial data handling, and spatio-temporal DL models. We showcase the package's potential through a tree cover percentage mapping application using Sentinel-2 data, emphasizing its capacity for flexible analysis and scalable processing. Our results underscore the promise of `xinfereo` in democratizing DL for EO, paving the way for wider adoption and operational integration of these powerful techniques.

***Index Terms*—** deep learning, Earth Observation, xarray, Sentinel-2, scalability, remote sensing

1. INTRODUCTION

The past decade has witnessed extraordinary progress in the application of Deep Learning (DL) methodologies to a broad spectrum of EO tasks [2, 9]. From precise mapping down to tree level to retrieval of drivers of deforestation, DL models have demonstrated a capacity to extract valuable information from complex remote sensing data [1, 5]. Despite these successes, a critical bottleneck impedes the transition of these models from experimental tools to operational assets. The technical expertise demanded by DL, particularly in managing specific data structures and intricate pre-processing steps, presents a substantial barrier. DL models often require data to be structured in specific formats (e.g., torch or tensorflow tensors with strict dimensions and various kind of preparation steps such as positional encoding, padding or normalization), and the absence of user-friendly tools to automate these transformations limits wider adoption. Consequently, DL for EO remains, to a large extent, the domain of specialists.

However, this challenge is not insurmountable. Innovative technical solutions can effectively bridge the gap between cutting-edge DL techniques and the practical needs of geospatial scientists. In this paper, we present `xinfereo`, a Python package developed to facilitate the integration of xarray Datasets, a fundamental data structure in geospatial analysis, with spatio-temporal DL models. Xarray Datasets provide a robust framework for representing multidimensional data as data cubes with clearly defined dimensions and coordinates. Furthermore, the xarray ecosystem offers seamless interoperability with other essential components of the geospatial toolkit, such as Spatio Temporal Asset Catalog (STAC) via odc-stac, the Geospatial Data Abstraction Library (GDAL) via rioxarray, and dask.

The `xinfereo` package, in its current prototype form, simplifies the application of DL models to EO data. Users can generate model outputs by executing a single function on a Sentinel-2 data cube encapsulated within an xarray Dataset. The package handles the complexities of model execution, including retrieving model parameters (which can be stored on an online platform such as Zenodo), verifying data compatibility, and performing necessary transformations such as normalization, padding and positional encoding. The model's prediction is then returned to the user as a NumPy array.

To illustrate the capabilities of `xinfereo`, we focus on the task of mapping tree cover percentage. While our primary objective is to demonstrate the package's functionality and ease of use, we provide an overview of the model architecture and training strategy for context. The core emphasis remains on enabling accessible and scalable model deployment. We highlight the package's flexibility in accommodating various data scenarios and its ability to scale to large-area processing.

Section 2 describes the model, data, and scalability experiment, section 3 presents and discusses results, and section 4 summarizes findings and future work.

*Consultant for the Joint Research Center

doi:10.2760/2119408

2. MATERIAL AND METHODS

2.1. `xinfereo` package

The `xinfereo` package is designed to streamline the integration of DL models into typical geospatial workflows. It provides a user-friendly interface that simplifies the application of pre-trained DL models to xarray Datasets. The core functionality of `xinfereo` is encapsulated in a single function. This function accepts an xarray Dataset as input and produces the model's prediction as a NumPy array. Importantly, this function abstracts away much of the complexity associated with preparing data for DL models. It handles the following critical steps:

- Data conformance: The function begins by validating that the input Dataset conforms to the model's requirements, checking for necessary variables and dimensions.
- Normalization: Any required normalization or scaling of the input data is performed automatically.
- Temporal encoding: If the model requires it, temporal information (e.g., positional encoding) is generated and incorporated into the input data.
- Padding: Necessary padding of time or channel dimensions is applied to ensure compatibility with the model's input shape requirements.
- Model execution: Finally, the pre-trained DL model is loaded and executed on the prepared input data.

To ensure portability and ease of use, pre-trained DL models and their associated parameters are stored in the Open Neural Network Exchange (ONNX) standardized format. `xinfereo` handles the retrieval of these models and parameters from their online location, relieving the user of the burden of manual management. The core inference function supports xarray Datasets backed by either NumPy arrays for in-memory computation or Dask arrays, thereby offering both flexibility in execution and inherent scalability for large data volumes.

Model-specific requirements, such as expected input variables and dimensions, are documented using a JSON-based metadata schema. This metadata model is currently an evolving aspect of the package and is designed to align closely with the STAC Machine Learning Model (mlm) extension. However, STAC mlm had limitations (e.g., temporal restriction or optional bands), so our implementation, while inspired by it, retains flexibility and doesn't strictly adhere yet.

2.2. Tree Cover Percentage Mapping Model

To demonstrate the practical application of `xinfereo`, we developed a DL model designed to map tree cover percentage from Sentinel-2 time series data. This model aims to predict the proportion of tree cover within each pixel, generating valuable information for forest monitoring and management activities.

The training dataset for our model was derived from the Copernicus 2018 Tree Cover Density (TCD) layer [4]. To create this dataset, we randomly selected 4800 bounding box locations, each 1280 m x 1280 m in size, across the European continent. For each of these locations, we extracted the corresponding TCD layer and resampled it to a 20 m resolution to align with the spatial resolution of the Sentinel-2 data. The resulting dataset was then partitioned into training (3200 patches), validation (800 patches), and testing (800 patches) subsets.

It is crucial to acknowledge the inherent limitations of this training data generation approach. The Copernicus TCD layer is already the result of a model and represents an estimate of tree cover with potential innacuracies. However, we reiterate that the primary focus of this study lies in demonstrating the `xinfereo` package's ability to simplify the application of DL models in a streamlined and scalable manner, rather than producing a definitive tree cover percentage product. In real-world scenarios, users frequently need to generate custom forest masks or tree cover estimates tailored to specific regions, timeframes, or input data characteristics. For instance, Near Real-Time monitoring systems often rely on up-to-date forest masks, which may not always be readily available [3]. While alternative methods, such as rule-based forest mask generation, exist, they may not be optimal in all situations [10].

In addition to the TCD data, we extracted one year (from January 1st to December 31st 2018) of Sentinel-2 data at a 20m resolution for each of our selected bounding box locations. We retained the following spectral bands, relevant for land surface analysis: B02 (blue), B03 (green), B04 (red), B05 (red edge 1), B06 (red edge 2), B07 (red edge 3), B08A (Near InfraRed), B11 (Short-Wave InfraRed (SWIR) 1), and B12 (SWIR 2), as well as the Scene Classification Layer (SCL) to allow the masking of clouds and shadows during training and testing.

Our model architecture for tree cover percentage retrieval from Sentinel-2 time series is based on a simple 1D Convolutional Neural Network (CNN), complemented by a temporal attention mechanism [6, 7], and is remarkably compact (42KB in ONNX format). The 1D CNN processes each pixel's time series independently, extracting spectral-temporal features directly from raw input sequences. This approach effectively handles irregular observation patterns and heterogeneous data (e.g., varying observation counts or residual clouds) without requiring prior temporal binning or interpolation. The temporal attention mechanism complements the CNN feature extraction by weighing the contribution of each time step, enabling the model to focus on the most informative segments for accurate retrieval. This combination offers significant flexibility and robustness to varying data conditions.

To enhance the model's versatility and robustness, we implemented a range of data augmentation techniques during the training process. These techniques simulate potential data

doi:10.2760/2119408

availability scenarios that users may encounter in practice, including:

- Cloud and shadow contamination: Randomly leaving portions of the input data unmasked to mimic the presence of residual clouds and shadows.
- Missing spectral bands: Randomly omitting spectral bands to train the model to handle situations where certain bands are unavailable.
- Partial temporal extent: Training the model on random subsets of the one-year time series, enabling it to work effectively with incomplete time series data.

This data augmentation strategy aims to improve the model's flexibility and adaptability to diverse data conditions. While we anticipate the model to perform optimally when provided with the full complement of Sentinel-2 bands, properly masked for clouds and shadows, and the complete one-year time series, it is also designed to provide reasonable predictions even when faced with degraded data availability.

2.3. Scalability Experiment

To assess `xinfereo`'s scalability, we applied the tree cover density (TCD) mapping model to the entirety of the 31TFK MGRS tile, utilizing Sentinel-2 data from the year 2024. The input dataset consisted of 7 months of observations, encompassing 44 Sentinel-2 acquisitions and amounting to approximately 16 GB of uncompressed data. This task was executed using a Dask-based workflow on a 10 workers local cluster within `xinfereo`, which processed the data in spatial chunks of 244x244 pixels. Identical workflow can easily be deployed on High Throughput Computing clusters, such as the one available on the Big Data Analytic Platform (BDAP) of the Joint Research Center [8], anticipating near-linear scaling.

3. RESULTS AND DISCUSSION

3.1. Model Performance

The model's predictive accuracy was assessed using the 800 patches designated for the test split, with the Root Mean Square Error (RMSE) calculated across a matrix of varying input data scenarios. As anticipated, the model performed optimally when provided with a full year of Sentinel-2 data, encompassing all specified spectral bands and with clouds and shadows masked, achieving an RMSE of 11.14 % (Figure 1). Notably, many alternative scenarios involving degraded input data such as omission of certain spectral bands, reduced temporal coverage, or the absence of cloud masking, resulted in only marginal reductions in performance. This observation underscores the model's versatility and robustness to variations in data availability and quality. For instance, these results suggest that in practical applications, users may not need to strictly adhere to the maximum data requirements;

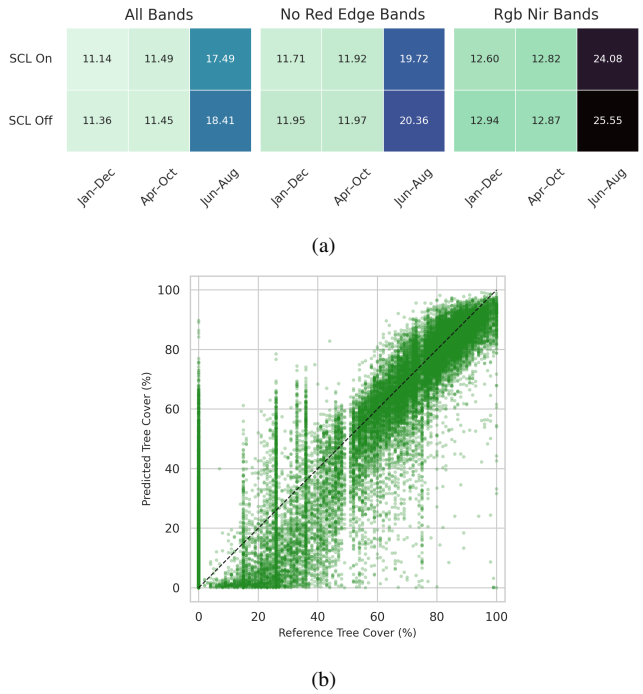


Fig. 1: Model performance on the dataset's test split: (a) Root Mean Square Error (RMSE) of predicted tree cover percentage for various input data modalities. (b) Scatter plot of predicted versus reference tree cover percentages for the best-performing input modality.

an acceptable tree cover density layer could potentially be generated for a given year using as few as six spectral bands and seven months of data. The most significant performance degradation was observed when the input data was limited to only three months, even if these were the summer months. This finding suggests that observations from outside the peak growing season are beneficial for accurately assessing tree cover and unambiguously discriminating trees from other land cover types.

3.2. Scalability

Processing the 16 GB 31TFK MGRS tile on the 10-worker local Dask cluster took approximately 5 minutes (including approximately 2 minutes for data loading from the EOS distributed file system). While tile-specific timing may vary, this initial benchmark indicates the approach's inherent scalability. Consequently, mapping an EEA38-sized area (5.8 million km²) is projected in hours on an HTC infrastructure, confirming `xinfereo`'s suitability for large-scale operational tasks.

3.3. Discussion

The findings of this study underscore the potential of the `xinfereo` package to democratize DL for EO analysis. By offering a user-friendly interface that seamlessly connects xarray Datasets with diverse DL models, including emerging

doi:10.2760/2119408

foundation models, `xinfereo` lowers the barrier to entry for geoscientists who may not possess specialized expertise in DL. This can foster wider adoption of these powerful techniques across a range of EO applications.

Furthermore, the modular design of `xinfereo` facilitates its integration into various geospatial processing workflows. This adaptability allows users to deploy `xinfereo`-based inference across diverse execution environments, from local Dask clusters to large-scale cloud platforms. Emerging infrastructures like the Copernicus Data Space Ecosystem (CDSE), which provide extensive access to Copernicus data and associated processing capabilities, represent one such type of environment where `xinfereo` could be effectively utilized by users for their specific analysis needs.

The demonstrated scalability highlights `xinfereo`'s potential for generating EO products over extensive areas in a timely and efficient manner, a capability relevant for various large-scale monitoring applications. While the tree cover map produced in this study serves as an illustration of the package's technical capabilities, it is important to distinguish it from officially validated data products. Operational services, such as the Copernicus Land Monitoring Service, rely on products that undergo rigorous accuracy assessment and validation processes, typically overseen by entities like the European Environment Agency (EEA), to ensure their fitness for designated applications. The approach demonstrated here has not been subjected to such extensive validation. Nevertheless, `xinfereo` can empower users to generate custom/interim products when up-to-date outputs are needed, official versions are unavailable, or existing products don't meet specific needs.

Importantly, `xinfereo` challenges the notion that ease of use and scalability must be mutually exclusive. Our package demonstrates that it is indeed possible to provide a user-friendly experience without compromising performance or the ability to handle large datasets. This approach has the potential to transform the way high-level EO products are generated and delivered, perhaps even leading to the development of new Copernicus services.

4. CONCLUSION

In this paper, we introduced `xinfereo`, a Python package designed to democratize the use of DL in EO analysis by effectively bridging the gap between xarray Datasets and spatio-temporal DL models. We showcased the package's utility through a tree cover percentage mapping application using Sentinel-2 data, highlighting its flexibility and scalability. Our results indicate that `xinfereo` can significantly simplify the integration of DL models into standard geospatial workflows, thereby promoting broader adoption and operational implementation of these powerful techniques.

Future research will focus on expanding the capabilities of `xinfereo` by developing a wider range of modular building blocks for data preparation and pre-processing, enabling it to accommodate a greater variety of models and EO applications. We also plan to explore tighter integration with the STAC mlm extension to facilitate more comprehensive model documentation and metadata management.

REFERENCES

[1] M. Brandt, C. J. Tucker, A. Kariryaa, K. Rasmussen, C. Abel, J. Small, J. Chave, L. V. Rasmussen, P. Hiernaux, A. A. Diouf, et al. An unexpectedly large count of trees in the west african sahara and sahel. *Nature*, 587(7832):78–82, 2020.

[2] M. Brandt, J. Chave, S. Li, R. Fensholt, P. Ciais, J.-P. Wigneron, F. Gieseke, S. Saatchi, C. Tucker, and C. Igel. High-resolution sensors and deep learning models for tree resource monitoring. *Nature Reviews Electrical Engineering*, 2(1):13–26, 2025.

[3] L. Dutrieux and J. Viehweger. nrt: operational monitoring of satellite image time-series in python. *Journal of Open Source Software*, 9(100):6815, 2024.

[4] European Environment Agency and Copernicus Land Monitoring Service. High resolution layer: Tree cover density 2018 (raster 20m), europe, 3-yearly, 2020. Date of publication: 2020-09-07.

[5] R. N. Masolele, D. Marcos, V. De Sy, I.-O. Abu, J. Verbesselt, J. Reiche, and M. Herold. Mapping the diversity of land uses following deforestation across africa. *Scientific Reports*, 14(1): 1681, 2024.

[6] C. Pelletier, G. I. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019.

[7] M. Rußwurm and M. Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 169:421–435, 2020.

[8] P. Soille, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, and V. Vasilev. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems*, 81:30–40, 2018.

[9] D. Tuia, K. Schindler, B. Demir, X. X. Zhu, M. Kochupillai, S. Džeroski, J. N. van Rijn, H. H. Hoos, F. Del Frate, M. Datcu, et al. Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward. *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[10] Z. Zhu, C. E. Woodcock, and P. Olofsson. Continuous monitoring of forest disturbance using all available landsat imagery. *Remote sensing of environment*, 122:75–91, 2012.

Code Availability. The source code for the `xinfereo` Python package is publicly available at: https://code.europa.eu/jrc-forest/xinfereo. The code, models, and necessary resources to reproduce the experiments presented in this paper are available at: https://code.europa.eu/jrc-forest/reproducibility/bids25.

doi:10.2760/2119408

GNEO AXIS 3 GOVERNMENTAL HUB: A GREEK CLOUD-NATIVE EO PLATFORM FOR SCALABLE, FEDERATED EARTH OBSERVATION SERVICES

Maria Ieronymaki[1], *George Benekos*[1], *Stratos Gerakakis*[1], *Francesco Bruni*[1], *Prof. K. Karantzalos*[2], *Dimitris Bliziotis*[3]

[1]Planetek Hellas, [2]Secretary General of Telecommunications & Post, Ministry of Digital Governance, [3]*Hellenic Space Center*

ABSTRACT

The AXIS 3 Governmental Hub is part of Greece's National Satellite Space Project, which aims to enhance the country's capabilities in satellite technologies and applications. Specifically, AXIS 3 focuses on developing geospatial services tailored to meet the needs of the Greek public sector. The AXIS 3 Governmental Hub is a modular, cloud-native Earth Observation (EO) infrastructure supporting the delivery of five thematic services: Land, Forest, Water, Agriculture, and Safety & Security. Developed under the Greek National Earth Observation (GNEO) mission, the Hub hosts and integrates data from GNEO AXIS 1.1, 1.2, and 2 satellite missions, along with Copernicus and Landsat products. Its architecture combines EO Exploitation Platform Common Architecture (EOEPCA) components and Kubernetes orchestration to enable scalable data ingestion, metadata harmonization (via a SpatioTemporal Asset Catalog (STAC) -compliant catalogue), user-triggered tasking, and automated processing. The Hub's microservices architecture supports federated access, metadata-driven workflows, and asynchronous orchestration of EO services. This paper presents the Hub's infrastructure, highlighting innovations in service orchestration, interoperability, and modular processing pipelines.

Index Terms— GNEO, AXIS 3 Governmental HUB, Earth Observation, Cloud-native architecture, STAC, EOEPCA, Kubernetes, ARD

1. INTRODUCTION

The Greek National Satellite Space Project strengthens national EO capabilities, equipping the public sector with advanced tools for monitoring, analysis, and decision support. At its core, it delivers geospatial services in five domains, Land, Forest, Water, Agriculture, and Safety & Security, through standardized processes ensuring reliability and operational impact (Figure 1).

Recent EO platforms increasingly rely on cloud infrastructures, data cubes, standardized APIs, and AI-driven methods. Representative examples include Google Earth Engine, Sentinel Hub, Open Data Cube, SEPAL, openEO, JEODPP, pipsCloud, EarthDataMiner, GeoCube, and the DestinE Platform [1–10]. These solutions demonstrate advances in scalability, interoperability, and real-time access, but most focus either on global analytics or single domains, with limited integration of heterogeneous data and end-to-end orchestration.

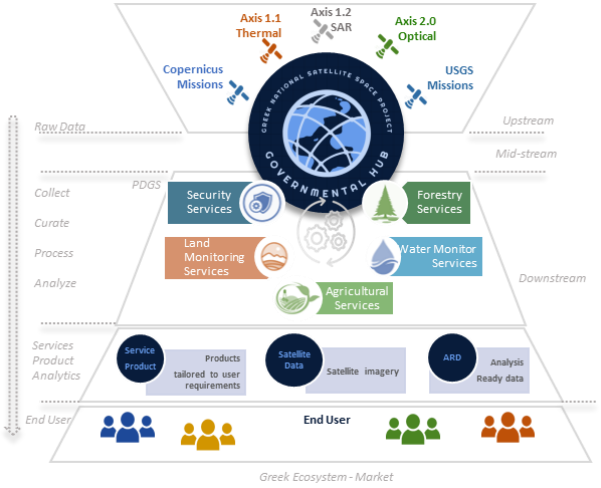


Figure 1 High-level overview of the AXIS 3 Hub

The management and analysis of big Earth observation (EO) data increasingly rely on cloud infrastructures, data cubes, standardized APIs, and AI-driven methods. Key platforms include Google Earth Engine, Sentinel Hub, Open Data Cube, SEPAL, openEO, JEODPP, pipsCloud, EarthDataMiner, GeoCube, and the DestinE Platform.

Google Earth Engine (GEE) enables petabyte-scale cloud access, APIs, and machine learning for large-scale analysis [1]–[4]. Sentinel Hub offers real-time streaming and GIS integration [1], while the Open Data Cube (ODC) provides an open-source spatio-temporal model for national ARD systems [1], [5], [6]. SEPAL supports land and forest monitoring [1], and openEO ensures interoperability through standardized APIs [1], [7]. JEODPP and pipsCloud deliver high-performance EO processing [1], [6]. Emerging platforms such as EarthDataMiner apply AI analytics [8], GeoCube enables raster–vector fusion [5], and the ESA DestinE Platform provides access to Earth-system digital twins for climate and environmental monitoring [9], [10].

The Governmental Hub draws upon these advances by combining EOEPCA-based modules with Kubernetes orchestration, custom developed components, STAC-compliant cataloguing, and dedicated hardware infrastructure, enabling interoperable and operational services tailored to governmental needs.

doi:10.2760/2119408

2. GOVERNMENTAL HUB SYSTEM ARCHITECTURE & DESIGN PRINCIPLES

The Hub's architecture is cloud-native, using EOEPCA [11] components and Kubernetes orchestration for scalable data ingestion, metadata harmonization through a STAC [12]-compliant catalogue, user-triggered tasking, and automated processing. This enables efficient handling of large data volumes with high performance and reliability. The AXIS 3 Hub adopts a microservices architecture supporting federated access, metadata-driven workflows, and asynchronous orchestration of EO services. Its modular design ensures flexibility to tailor pipelines for each thematic domain, while providing analysis-ready data (ARD) for efficient and accurate EO analysis.

2.1. Cloud-Native & Microservices Architecture

The Hub uses a containerized, microservices-based architecture, allowing independent deployment and scaling of components (Figure 2). Each component (e.g., ingestion, processing, cataloguing) communicates through APIs and runs in Kubernetes-managed containers. This approach ensures high availability and dynamic scaling based on processing demands.

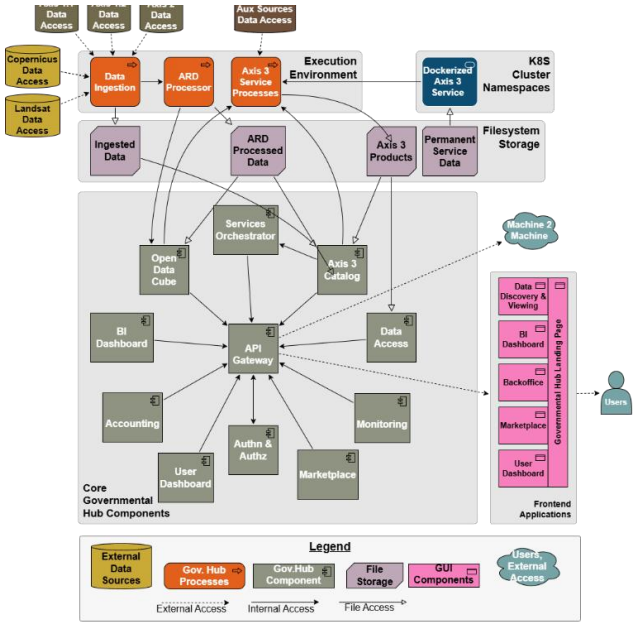


Figure 2 High-Level System Architecture of the AXIS 3 Governmental Hub

2.2. AXIS 3 Hub Core Components

The AXIS 3 Hub leverages the open-source EOEPCA, which provides a reference framework for EO data management and exploitation services. The core EOEPCA modules integrated into the Hub include:

- Resource Catalogue (STAC-compliant) for product discovery and metadata indexing;
- Processing Component for Common Workflow Language (CWL)-based scalable job execution;
- Data Access providing standardized interfaces for accessing geospatial data assets stored within the platform, supporting both human users and machine users.

Beyond the EOEPCA integration, the AXIS 3 Hub introduces a set of core components that enable the operational delivery of governmental thematic services. These include:

- Data Ingestion and ARD Processor, responsible for harmonizing external and mission-specific data streams (e.g. Copernicus, Landsat, AXIS missions) into Analysis Ready Data (ARD) formats.
- Services Orchestrator, which manages the sequencing and execution of workflows across multiple processing modules.
- Tasking Orchestrator, which extends Hub functionality to manage satellite tasking requests with external Data Hubs.
- API Gateway, providing unified access to Hub functionalities, ensuring secure interaction between internal services and external applications.
- Monitoring, Accounting, and Dashboards, which support operational control, usage tracking, and visualization for both administrators and end-users.
- Axis 3 Service Processes, that include the thematic service algorithms (land, water, forest, agriculture, and safety & security services).

2.3. Hardware Infrastructure and High-Performance Resources

The AXIS 3 Governmental Hub is powered by a high-performance compute and storage backbone tailored to the demands of large-scale Earth Observation data management and processing. The compute infrastructure combines general-purpose servers with GPU-accelerated nodes, enabling both standard processing and advanced AI/ML workflows. The storage layer is based on an all-flash enterprise system complemented by high-capacity disks, providing more than 2 petabytes of raw storage and over 500 TB of NVMe flash for rapid data access and intensive analytics. Cybersecurity is ensured through firewalls and centralized monitoring, while the networking layer delivers high-bandwidth switching with 100/400G uplinks and 10/25G access, ensuring low-latency, resilient interconnectivity across the system. Together, these resources provide the foundation required for the Hub's services.

3. DATA INTEGRATION & HARMONIZATION

3.1. Data Integration and Management

The AXIS 3 Governmental Hub manages EO data from both national and international missions to support operational workflows. Key sources include open datasets from the Copernicus Data Space Ecosystem (CDSE) and Landsat, accessed via ESA and USGS services, with open data retained under a rolling one-year policy. National data includes AXIS 1.1 (thermal, nominal and tasking), AXIS 1.2 (SAR, tasking-based), and AXIS 2.0 (optical, high-acquisition rate across HR, VHR, hyperspectral, IoT, and AIS sensors). Each mission provides multiple processing levels (L0, L1C, L2A) (Figure 3).

doi:10.2760/2119408

Data is ingested via STAC-compliant APIs for seamless integration and cataloguing.

While open data follows a limited retention policy, AXIS mission data is archived for the full operational lifecycle. This supports scalable, timely access for downstream processing and thematic services.

The Hub uses a dual-mode strategy to manage both systematic and on-demand data. Automated pipelines ingest continuous streams from open sources and nominal GNEO missions. Tasking workflows for GNEO missions allow users to request data by area, time, or priority.

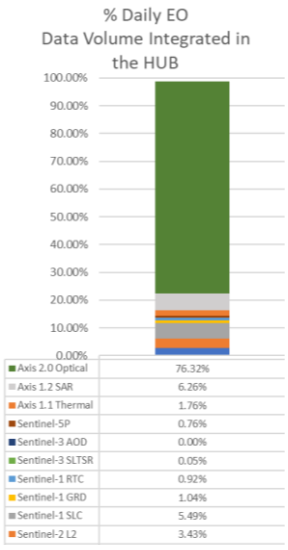


Figure 3 Daily Percentage Distribution of EO Data Volume Integrated into the AXIS 3 Governmental Hub

Ingested data are routed to mission-specific storage, and metadata is automatically generated with spatial/temporal tags, sensor info, and asset links. Once validated, products are indexed and published in the Governmental Hub Catalogue (Section 3.4) for immediate discovery and use across all thematic domains.

3.2. Analysis-Ready Data (ARD)

The AXIS 3 Governmental Hub adopts a standardized ARD strategy to ensure all ingested datasets are preprocessed, harmonized, and immediately usable across its thematic services. ARD products serve as the primary input for downstream processing.

The ARD generation workflows, inspired by CEOS-aligned specifications include radiometric calibration, geometric and terrain corrections, and atmospheric corrections where applicable.

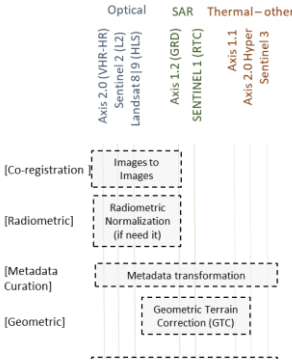


Figure 4 ARD Harmonization Workflow

To support integration with national datasets, all ARD products are transformed into the Greek Grid (EPSG:2100). The output formats follow open standards such as Cloud-Optimized GeoTIFF (COG) to enable efficient access and cloud-native storage. Priority is given to open Copernicus data, including Sentinel-1, Sentinel-2, and Sentinel-3, to maintain continuity with AXIS 1 and AXIS 2 products. Across all EO sources—optical, SAR, thermal, and hyperspectral—the ARD workflow ensures that data inputs are consistently transformed into standardized, analysis-ready outputs that are both operationally reliable and scientifically robust (Figure 4).

The AXIS 3 Governmental Hub generates and manages ARD products across a range of sensor types to ensure standardized processing and interoperability. For optical imagery, the Hub processes Sentinel-2 Level 2A products, AXIS 2.0 high-resolution and very-high-resolution multispectral data, and Harmonized Landsat Sentinel (HLS) products from Landsat. For Synthetic Aperture Radar (SAR), it includes Sentinel-1 GRD products and various imaging modes from AXIS 1.2, such as Scan, Strip, Spot High-Resolution, Spot Fine, and Spot Extended. In the domain of thermal and hyperspectral data, ARD is produced from Sentinel-3 SLSTR, AXIS 1.1 thermal infrared and AXIS 2.0 hyperspectral.

3.3. STAC Strategy

To ensure consistent and interoperable metadata management, the AXIS 3 Governmental Hub adopts a STAC (SpatioTemporal Asset Catalog) compliant strategy across all EO products. As data from GNEO missions, Copernicus, and Landsat is ingested, metadata records are automatically generated with standardized spatial, temporal, and sensor attributes, and references to data assets and processing levels. All EO data—from raw satellite inputs to products generated by AXIS 3 Thematic Services—are mapped to STAC Collections, Items, and Assets (Figure 5). This approach provides uniform cataloguing and exposure of datasets to internal and external users, ensuring discoverability, accessibility, and compliance with emerging EO standards. STAC Items and Collections are validated before indexing into the Hub Catalogue, enabling federated access and scalable, cloud-native discovery and processing across all services.

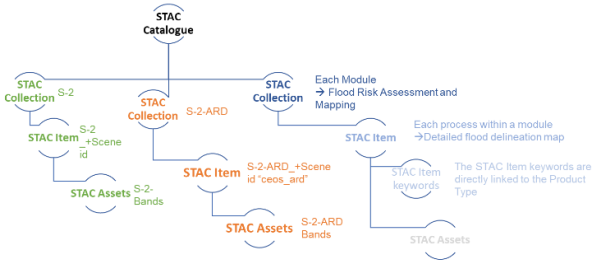


Figure 5 STAC Catalogue Structure Strategy

3.4. GOVERNMENTAL HUB Catalogue

The Governmental HUB Catalogue will serve as the central metadata repository for all datasets within the AXIS 3 Hub, supporting comprehensive discovery and access. It will cover both internally generated and externally sourced EO datasets. Internally, this includes raw and processed data from national missions—AXIS 1.1 thermal, AXIS 1.2 SAR, and AXIS 2.0 optical and hyperspectral imagery—along with value-added products from thematic services and standardized ARD. Externally, it will index datasets from Copernicus, USGS (Landsat), and other open-access providers (Figure 6).

doi:10.2760/2119408

Designed as a centralized interface, the catalogue provides metadata-driven access to multi-source datasets with federated querying across thematic domains via STAC-based APIs. All metadata will follow STAC specifications, ensuring machine-readability, integration with external platforms, and interoperability across the wider EO ecosystem.

4. THEMATIC SERVICES AND ADDED VALUE PRODUCTS

The AXIS 3 Governmental Hub operates five thematic services—Land, Forest, Water, Agriculture, and Safety & Security—each built on standardized EO pipelines aligned with public-sector priorities. Outputs are published through the Hub Catalogue with STAC-compliant metadata, ensuring traceability, discoverability, and integration with national geospatial infrastructures.

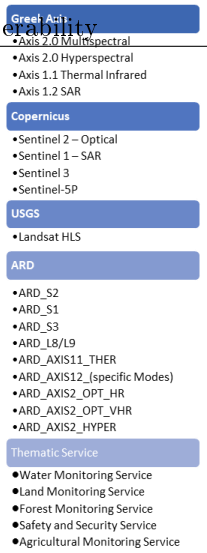


Figure 6 Governmental Hub Catalogue

The Land Service supports land cover classification, change detection, InSAR deformation (PSI, SBAS), and urban analytics including heat island effects, air quality, and health indicators. The Forest Service covers national forests and NATURA 2000 areas through forest type and fuel mapping, health monitoring, biodiversity analysis, and threat detection, delivering outputs such as fuel maps, tree cover density, and biodiversity hotspot maps.

The Water Service includes Water Quantity, Water Quality, and Maritime Surveillance modules, producing inland and coastal water maps, indicators like chlorophyll-a, SST, turbidity, oil spill detection, and maritime object identification. The Agriculture Service provides crop type prediction, vegetation indices (NDVI, SAVI, PSRI), radar and texture metrics, growth estimates (LAI, Ccc, Cdmc), and soil moisture analysis through optical–SAR fusion.

The Safety & Security Service addresses floods and wildfires, generating high-resolution flood maps, depth and risk assessments, wildfire danger forecasts, burn scar mapping, fire spread predictions, and object tracking with bounding boxes and trajectories.

5. INNOVATIONS AND CHALLENGES

The AXIS 3 Governmental Hub adopts a forward-looking strategy based on open standards, modular architecture, and cloud-native design. By integrating EOEPCA components and leveraging Kubernetes, it ensures scalability, resilience, and interoperability across EO data ecosystems. Metadata-driven workflows, federated access, and microservices-based automation enable seamless end-to-end service delivery—from ingestion to analysis and visualization. Strategically, the Hub acts as a national anchor for secure EO operations while interoperating with European and global platforms, supporting cross-border collaboration, rapid thematic service deployment, and long-term adaptability to new missions, sensors, and user needs.

6. CONCLUSION

The AXIS 3 Governmental Hub establishes a scalable, secure, and interoperable EO service delivery model, demonstrating how cloud-native systems can support strategic decision-making, environmental resilience, and digital transformation. Supporting both national priorities and European cooperation, it provides future-ready infrastructure that turns EO data into actionable insights. With its modular, standards-based design, the Hub enables rapid deployment of thematic services while fostering continuous innovation in data exploitation.

REFERENCES

[1] V. Gomes, G. Queiroz, and K. Ferreira, "An overview of platforms for big Earth observation data management and analysis," *Remote Sensing*, vol. 12, no. 8, p. 1253, 2020, doi: 10.3390/rs12081253.

[2] H. Tamiminia, B. Salehi, M. Mahdianpari, L. Quackenbush, S. Adeli, and B. Brisco, "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review," *ISPRS J. Photogramm. Remote Sens.*, vol. 164, pp. 152–170, Jun. 2020, doi: 10.1016/j.isprsjprs.2020.04.001.

[3] A. Seyed *et al.*, "Google Earth Engine cloud computing platform for remote sensing big data applications: A comprehensive review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5326–5350, Sep. 2020, doi: 10.1109/JSTARS.2020.3021052.

[4] Q. Zhao, L. Yu, X. Li, D. Peng, Y. Zhang, and P. Gong, "Progress and trends in the application of Google Earth and Google Earth Engine," *Remote Sensing*, vol. 13, no. 18, p. 3778, 2021, doi: 10.3390/rs13183778.

[5] F. Gao *et al.*, "A multi-source spatio-temporal data cube for large-scale geospatial analysis," *Int. J. Geogr. Inf. Sci.*, vol. 36, no. 9, pp. 1853–1884, 2022, doi: 10.1080/13658816.2022.2087222.

[6] X. Yao *et al.*, "Enabling the big Earth observation data via cloud computing and DGGS: Opportunities and challenges," *Remote Sensing*, vol. 12, no. 1, p. 62, 2019, doi: 10.3390/rs12010062.

[7] B. Pondi, M. Appel, and E. Pebesma, "OpenEOcubes: An open-source and lightweight R-based RESTful web service for analyzing Earth observation data cubes," *Earth Sci. Inform.*, vol. 17, pp. 1809–1818, 2024, doi: 10.1007/s12145-024-01249-y.

[8] J. Liu and H. Zhong, "EarthDataMiner: A cloud-based big Earth data intelligence analysis platform," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 509, no. 1, p. 012032, 2020, doi: 10.1088/1755-1315/509/1/012032 .

[9] European Space Agency, "What is Destination Earth?," *ESA – Destination Earth*, 2025. [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Destination_Earth

[10] Copernicus Programme, "Destination Earth: Building a highly accurate Digital Twin of the Earth," Copernicus Sentinel Success Stories, Jul. 24, 2025. [Online]. Available: https://sentinels.copernicus.eu/web/success-stories/-/destination-earth-building-a-highly-accurate-digital-twin-of-the-earth

[11] EOEPCA, "EOEPCA Documentation," [Online]. Available: https://eoepca.readthedocs.io/ . [Accessed: 14-Apr-2025].

[12] STAC, "SpatioTemporal Asset Catalogs," [Online]. Available: https://stacspec.org/en . [Accessed: 14-Apr-2025].

[13] Copernicus Data Space Ecosystem, "Data Collections," [Online]. Available: https://dataspace.copernicus.eu/explore-data/data-collections . [Accessed: 14-Apr-2025].

[14] U.S. Geological Survey, "APIs," [Online]. Available: https://www.usgs.gov/products/web-tools/apis. [Accessed: 14-Apr-2025].

[15] CEOS, "Committee on Earth Observation Satellites," [Online]. Available: https://ceos.org/. [Accessed: 14-Apr-2025].

[16] Open Geospatial Consortium, "OGC Cloud Optimized GeoTIFF Standard," [Online]. Available: https://docs.ogc.org/is/21-026/21-026.html . [Accessed: 14-Apr-2025].

[17] CEOS, "CEOS Analysis Ready Data," [Online]. Available: https://ceos.org/ard/ . [Accessed: 14-Apr-2025].

[18] GitHub, "CEOS-ARD Extension Specification," [Online]. Available: https://github.com/stac-extensions/ceos-ard . [Accessed: 14-Apr-2025].

DISCLAIMERS

The project is being carried out under an ESA Contract in the frame of the Greek National Satellite Space Project.

The Project: Small-Satellites (Measure ID 16855) is implemented by the Hellenic Ministry of Digital Governance with the European Space Agency (ESA) Assistance in the Management and Implementation. The project is part of the National Recovery and Resilience Plan 'Greece 2.0', which is funded by the Recovery and Resilience Facility (RRF), core programme of the European Union-NextGenerationEU

doi:10.2760/2119408

XCUBE AND THE PYTHON DATA SCIENCE ECOSYSTEM: FEDERATED ACCESS, INTEROPERABILITY, AND DISSEMINATION FOR BIG EARTH OBSERVATION DATA

Yogesh Kumar Baljeet Singh, Gunnar Brandt, Pontus Lurcock, Tejas Morbagal Harish, Konstantin Ntokas, Norman Fomferra

Brockmann Consult GmbH, Hamburg, Germany

ABSTRACT

The rapid growth of data from Earth Observation (EO) and models presents significant challenges and opportunities for the scientific community, emphasizing the need for scalable, interoperable, and flexible analytical infrastructures. ESA's ongoing support for cloud-native Data Processing-as-a-Service (DPaaS) platforms and standardization efforts (such as EO Exploitation Platform Common Architecture project (*EOEPCA*), *openEO*, and Application Propagation Environments (*APEx*)) has substantially streamlined access to these datasets. At the same time, Python's data science ecosystem, particularly the *Pangeo* stack (based around *xarray*, *Dask*, and *Zarr*), has gained popularity due to its flexibility, transparency, accessibility, and extensive analytic capabilities. The *xcube* toolkit integrates seamlessly into both ecosystems, offering powerful tools for uniform data access and harmonization, and analysis-ready EO data cubes and other gridded datasets. The *xcube* datastore framework has recently been enhanced with plugins for the new Earth Observation Processing Framework (*EOPF*) *Zarr* format, integration of STAC-compliant collections such as those from the Copernicus Data Space Ecosystem (*CDSE*), and dataset harmonization and combination via *xcube-multistore*. These additions, together with advanced visualization features in the *xcube viewer* and its new *chartlets* extension, significantly enrich the framework. Further advances include the xcengine processing tool and integration into the *EarthCODE* project, particularly the *deep-code* initiative, emphasizing reproducibility and computational efficiency. This paper describes the features, datastore architecture, recent developments and integration capabilities of *xcube*, and highlights its role in enhancing interoperability, flexibility, and analytical effectiveness within the EO community.

Index Terms— EO Data Cubes, Cloud Computing, Interoperability, Python Data Science, Gridded data, xcube

1. INTRODUCTION

The exponential growth of EO data volumes driven by programs such as Copernicus requires sophisticated and scalable solutions for effective data processing and analysis. ESA has fostered various initiatives supporting a *DPaaS* approach, such as *EOEPCA* [1], *openEO* [2], and *APEx* [1], to facilitate setting-up and operation of a processing service as well as to improve standardization and interoperability of platforms. While these services substantially simplify EO and gridded data processing and provide convenient access to scalable cloud processing capabilities, many users, particularly those with sufficient programming background, appreciate the benefits of the long-established, flexible, transparent, and reproducible programming environment offered by Python's data science ecosystem. Tools like *xarray, Dask*, and *Zarr*, and the large collection of associated packages that make up the *Pangeo* stack, are particularly favoured for their capacity to handle large EO and gridded datasets effectively across a wide range of computing environments, from individual workstations to large-scale clusters, all with the same code base. Moreover, the wide adoption of the *Pangeo* stack in many user communities has turned it into a quasi-standard with minimum risk for lock-in with a specific service provider or computational environment.

The *xcube* framework is fully integrated with the *Pangeo ecosystem*[2], offering powerful functionalities along the entire workflow chain from data access to publication and dissemination of data and workflows. Particularly, several *xcube* tools effectively bridge the Python ecosystem and ESA's growing world of cloud services and applications.

2. PYTHON DATA SCIENCE ECOSYSTEM FOR EO AND GRIDDED DATA

Python has emerged as a leading language for EO and gridded data analysis, driven by libraries like *xarray* for handling multidimensional arrays and *Zarr* for efficient, cloud-friendly storage of large datasets. This ecosystem offers considerable advantages in flexibility, reproducibility, and transparency, making it very attractive for scientific researchers, data scientists, and developers of EO-related software. A key success factor has been the *Dask* package for distributed computing, which allows for working with data sets that are larger than available memory and effectively abstracts away from users the burden of scaling

[1] https://apex.esa.int/
[2] https://pangeo.io/#ecosystem

doi:10.2760/2119408

workflows to multiple parallel compute nodes. It enables users to seamlessly process very large, cloud-based datasets from any machine running a Python environment.

Despite these strengths, users often encounter significant challenges when attempting to integrate and harmonize diverse EO and gridded datasets from various sources, formats, and access methods, make their workflows reproducible and shareable, disseminate their results through standard services, or visualize them. While technically feasible, each of the above steps may entail substantial effort, leading to inefficiencies and raising barriers to rapid, iterative, and collaborative progress in EO research.

3. THE XCUBE FRAMEWORK

The *xcube*[3] framework addresses these typical integration challenges directly, providing structured methods for creating harmonized, analysis-ready data cubes from multiple sources through its robust datastore framework and rich functionalities, which offers seamless integration of multiple services and platforms offering EO and gridded data and by this means facilitates their integration into Python data science workflows.

xcube is a Python toolkit designed for generating, analyzing, and publishing EO data cubes. It leverages the *xarray* library for handling multi-dimensional data arrays, *dask* for parallel computing, and *zarr* for efficient storage. Foundational libraries like *rioxarray* extend *xarray* to handle geospatial raster data (utilizing *rasterio* for reading and writing *GeoTIFFs* and managing coordinate reference systems), while *Pangeo* provides a cloud-native ecosystem for scalable data analysis with *xarray* and *dask*. Building upon these tools, *xcube* offers higher-level abstractions for data processing, including spatial rectification, resampling, rechunking, subsetting, optimizing data cubes etc. all while maintaining compatibility with existing *xarray* workflows. A core strength of *xcube* is its robust datastore framework, which integrates multiple EO and gridded data sources and services seamlessly into Python workflows, allowing users to abstract data retrieval and harmonization through simple, reusable components. For dissemination and publication, *xcube* includes fully integrated *xcube server* and *xcube viewer* capabilities, supporting standard APIs for interoperability and offering powerful visualization and interactive analysis tools directly within Python environments as well as Jupyter Notebooks. This comprehensive integration enables users to transition smoothly from satellite products to analysis-ready data cubes, facilitating efficient spatio-temporal analysis and visualization.

[3] https://xcube.readthedocs.io/en/latest/

To make Python user workflows reproducible and integrate them with ESA's *EarthCODE*[4] initiative, two recently developed packages, *deepcode*[5] and *xcengine*[6], enable users to add their Python workflows to ESA's Open Science Catalogue, either as plain Jupyter Notebooks or after turning them into an EO Application Package [3]. The *xcengine* component can then run such packages and provide them as a service.

4. ADVANCED DATASTORE FRAMEWORK AND INTEGRATION CAPABILITIES

xcube's datastore framework facilitates integration through a continually growing set of specialized datastore plugins. Any new datastore can be created as a plugin following the design pattern shown in Fig. 1.

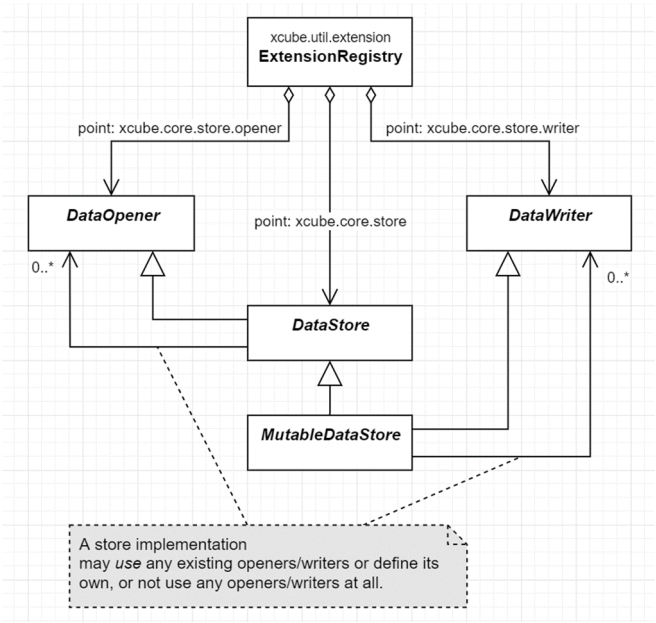


Fig 1: xcube datastore design.

Recent developments include the following:

The *STAC*[7] Store plugin[8] allows dynamic discovery and retrieval of datasets from *STAC*-compliant catalogues like that of the CDSE [4]. It streamlines the creation of analysis-ready data cubes by implementing a standardized and efficient method for querying available data and metadata and includes stitching and mosaicking of individual products as shown in Fig. 2.

[4] https://earthcode.esa.int
[5] https://github.com/deepesdl/deep-code
[6] https://github.com/xcube-dev/xcengine
[7] https://stacspec.org/en
[8] https://github.com/xcube-dev/xcube-stac

doi:10.2760/2119408

```
[3]: %%time
store = new_data_store("stac-cdse-ardc", **credentials)

CPU times: user 743 ms, sys: 90.7 ms, total: 834 ms
Wall time: 986 ms

[4]: %%time
ds = store.open_data(
    data_id="sentinel-2-l2a",
    bbox=[9.1, 53.1, 10.7, 54],
    time_range=["2020-07-15", "2020-08-01"],
    spatial_res=10 / 111320,  # meter in degree
    crs="EPSG:4326",
    asset_names=["B02", "B03", "B04", "SCL"],
    add_angles=True,
    tile_size=2048,
)
ds

CPU times: user 6min 6s, sys: 1.76 s, total: 6min 8s
Wall time: 7min 12s

[4]: xarray.Dataset

► Dimensions:  (time: 11, lon: 17813, lat: 10020, angle_lon: 37, angle_lat: 22, angle: 2, band: 3)

▼ Coordinates:
  time          (time)          datetime64[ns]  2020-07-15T10:15...
  spatial_ref   ()              int64           0
  lon           (lon)           float64         9.1 9.1 9.1 9.1 ... 1...
  lat           (lat)           float64         54.0 54.0 54.0 ... 5...
  angle_lon     (angle_lon)     float64         9.1 9.146 9.191 ... ....
  angle_lat     (angle_lat)     float64         54.04 54.0 53.95 .....
  angle         (angle)         object          'zenith' 'azimuth'
  band          (band)          <U3             'B02' 'B03' 'B04'

▼ Data variables:
  B02           (time, lat, lon)                float32  dask.array<chunk...
  B03           (time, lat, lon)                float32  dask.array<chunk...
  B04           (time, lat, lon)                float32  dask.array<chunk...
  SCL           (time, lat, lon)                float32  dask.array<chunk...
  solar_angle   (angle, time, angle_lat, angle_lon)  float32  dask.array<chunk...
  viewing_angle (angle, band, time, angle_lat, angle_lon)  float32  dask.array<chunk...

► Indexes:  (7)

► Attributes:  (3)
```

Fig 2: Sentinel-2 L2A analysis-ready datacube from CDSE [5] SAFE-Format.

The *EOPF* Store plugin[9] supports the new *EOPF*[10,11] format, optimized for cloud environments. Leveraging the *Zarr* storage format, this plugin provides rapid, efficient access to large Sentinel data collections directly within Python, significantly simplifying pre-processing and cube generation. The datastore builds upon the recently released *EOPF* engine for *xarray*[12].

The newly developed *Multi-Source Data Store*[13] enables users to federate datasets from multiple heterogeneous sources into a unified source. This capability eliminates intermediate steps typically required in manual harmonization, enabling cross-source analyses with minimal effort.

The new datastores complement the long list of existing stores e.g., for *Copernicus Marine Service (CMEMS)*[14], *Copernicus Land Monitoring Service (CLMS)*[15], *Climate Change Initiative (CCI) Open Data Portal*[16], *Zenodo*[17], and others as listed in the documentation[18].

[9] https://github.com/EOPF-Sample-Service/xcube-eopf/
[10] https://zarr.eopf.copernicus.eu/
[11] https://eopf.copernicus.eu/eopf/
[12] https://eopf-sample-service.github.io/xarray-eopf/
[13] https://xcube-dev.github.io/xcube-multistore/
[14] https://github.com/xcube-dev/xcube-cmems
[15] https://github.com/xcube-dev/xcube-clms
[16] https://github.com/esa-cci/xcube-cci/tree/main

5. VISUALIZATION AND INTERACTIVE ANALYSIS

Beyond data integration and processing, visualization plays a key role in analyzing EO and gridded data and in communicating results to others. *xcube* addresses this demand through its interactive *Viewer*[19] component as shown in Fig. 3 — a powerful and intuitive visualization tool connected directly to the backend *xcube Server*. It may be run as a stand-alone webpage but is also available as part of the Python package for interactive development within Jupyter Notebooks. With only three lines of code, users can start a new *Server* instance, add an *xarray* dataset to it, and launch a new *Viewer* instance, which can then be explored interactively inline or in a new tab, with the full set of functionalities of the app.

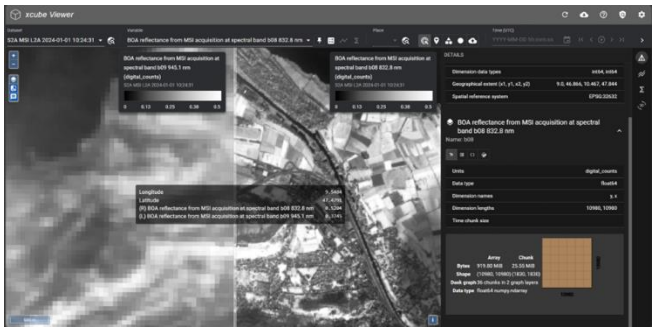


Fig 3: xcube Viewer.

The *Viewer* recently introduced the *chartlets*[20] extension, enhancing analytical functionality. *Chartlets* are interactive, customizable widgets allowing users to create specialized visual analyses — such as histograms, scatter plots, and time-series — directly within the *Viewer*. This significantly enhances user experience, providing immediate insights without additional external analysis and visualization tools.

Beyond its role as the *Viewer's* back end, *xcube Server* is a robust tool with an extensible architecture that enables low-cost development of plugins for additional interfaces. It currently supports the following API endpoints shown in Table 1.

Table 1: xcube Server APIs

Endpoint	Purpose
/viewer	Providing the viewer application and configurations.
/meta	Server information and maintenance operations
/places	Places API providing vector and feature data
/datasets	Datasets API for accessing data cubes
/tiles	XYZ tiles API for retrieving tiles as PNG

[17] https://github.com/xcube-dev/xcube-zenodo
[18] https://earthsystemdatalab.net/data/#available-data-sources
[19] https://xcube-dev.github.io/xcube-viewer
[20] https://github.com/bcdev/chartlets

doi:10.2760/2119408

	from image pyramids
/timeseries	Timeseries API for getting time-series data from data cubes
/stats	Statistics API for computing statistics from given coverages
/volumes	Volumes API for getting a 3D volume from a data cube
/ows/stac	OGC STAC API for the configured datasets
/ows/coverages	OGC Coverages API for the configured datasets
/ows/wmts	OGC WMTS API (an OGC wrapper of the more flexible XYZ /tiles endpoint)
/s3	AWS S3 compliant API for directly accessing the configured datasets as lazily loaded Zarr datasets

6. COMPUTATIONAL BACKEND: XCENGINE

xcengine is a recent addition to the *xcube* ecosystem, providing tools to convert Python Jupyter Notebooks to containerized, standalone '*compute engines*'. These engines can be run both in interactive mode (providing an *xcube Server* instance supporting a wide range of standard and specialized APIs and an interactive *xcube viewer* instance) and in batch mode as a standardized EO Application Package, suitable for integration into larger cloud-based processing workflows. *xcengine* thus helps to bridge the local, interactive and cloud-based, workflow-controlled processing models with minimal additional effort required from the user.

7. EARTHCODE INTEGRATION

The *EarthCODE* initiative seeks to ensure that EO workflows created by researchers on ESA platforms are persistent, findable, and reproducible as shown in Fig. 4. A key element for Python workflows developed in Jupyter Notebooks is the *deep-code* package, part of the Deep Earth System Data Laboratory (*DeepESDL*)[21] project, which leverages *xcengine* and *xcube* to facilitate efficient, reproducible processing workflows on EO datasets. *deep-code* enables detailed versioning, tracking, and publication of workflows encoded in Jupyter notebooks as well as EO datasets in the ESA Open Science catalogue. The software also supports converting a Jupyter Notebook into an interoperable, and reproducible EO Application Package, which can then be executed on other compatible platforms.

[21] https://earthsystemdatalab.net/

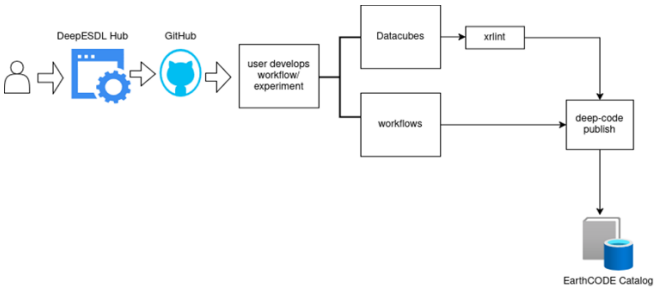


Fig 4: Simplified EarthCODE publication workflow.

8. CONCLUSION

The *xcube* framework significantly advances interoperable, federated gridded data processing and dissemination tasks, directly addressing diverse scientific community needs not covered by the other standard packages of Python's data science ecosystem. It particularly facilitates working with Earth Observation data in ESA's growing ecosystem of platforms and services and enables Python users to comply with the emerging and challenging requirements for reproducibility and interoperability.

REFERENCES

[1] Conway, R., Hinton, J., Taposeea, C., Iacopino, C., Pinto, S., and Hunter, S.: EOEPCA+: a method for an open-sourced EO Exploitation Platform Common Architecture, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-10583, https://doi.org/10.5194/egusphere-egu25-10583, 2025.

[2] Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., & Reiche, J. (2021). The openEO API – Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. Remote Sensing, 13(6), 1125. https://doi.org/10.3390/rs13061125, 2021

[3] Gonçalves, P., Brito, F., Landry, T., Charette Migneault, F., Conway, R., Luna Cobos, A., Barrilero, O., Vretanos, P., Lopes, C., Romeo, A., Sacramento, P., Lavender, S. and Neagul, M. (2021). OGC Best Practice for Earth Observation Application Package. https://doi.org/10.13140/RG.2.2.29220.69763, 2021

[4] Niemyjski, M. and Musiał, J.: Building the Copernicus Data Space Ecosystem STAC Catalog: Methodologies, Optimizations, and Community Impact, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-17171, https://doi.org/10.5194/egusphere-egu25-17171, 2025.

[5] Milcinski, G., Musial, J., Leszczenski, J., Clarijs, D., and de la Mar, J.: Copernicus Data Space Ecosystem - Platform That Enables Federated Earth Observation Services and Applications, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-4942, https://doi.org/10.5194/egusphere-egu24-4942, 2024.

COLOR33 – A CLOUD-BASED SERVICE FOR AUTOMATED SEMANTIC ENRICHMENT OF OPTICAL SATELLITE IMAGES

Martin Sudmanns[1/2], Matthias Laher[2], Steffen Reichel[2], Markus Kerschbaumer[2], Andrea Baraldi[2], Dirk Tiede[1]

[1]University of Salzburg, Department of Geoinformatics, Salzburg, Austria, [2]Spatial Services GmbH, Salzburg, Austria

ABSTRACT

color33 is a cloud-based service that enables semantic enrichment of optical Earth observation (EO) images such as from Sentinel-2. It does not require training samples, is worldwide applicable, explainable, and transferrable. The service uses the SIAM software to categorizing reflectance values into colour names (spectral categories). Thus, color33 supports general-purpose and sensor-agnostic downstream applications. The service is based on a modular, containerized architecture and facilitates automated EO workflows through standards-based APIs using OGC API – Processes and STAC. As a foundational component and building block for workflows, color33 enables downstream applications and offers a fast, flexible alternative to end-to-end classification approaches.

Index Terms— Sentinel-2, Copernicus, semantic enrichment, OGC-API Processes, STAC, automatic workflows

1. INTRODUCTION

The Copernicus Sentinel-2 mission is now operational for ten years with a continuous data delivery. To date, there are more than 105 Mio. Images with 51 PB of data in volume. The main goal has been since then to create information and, in turn, maps that can be used to make decisions.

In the efforts to create workflows that suit different end-user needs a multitude of different approaches have been proposed. Today, many of them use deep learing (DL) artificial intelligence (AI), requiring training samples and -time. Several limitations and challenges exist and are documented [1]. Notable limitations and challenges that we want to address are the following four: (1) fixed legends, (2) end-to-end workflows, (3) limited transferability, (4) explainability.

(1) Due to relying on training, the legend, i.e., the output classes, is dependent on the available training samples. Users requiring different or extended sets of target classes face the tedious work of creating new training samples.

(2) AI models are typically trained directly on the reflectance values and do not provide reusable, intermediate layers that can be used as building blocks for ad-hoc combinations and analysis. All start with the original reflectance values from scratch, although the work on analysis-ready-data (ARD) removes some of the burden of repeated processing steps on the user side [2].

(3) Training samples typically cover local areas. AI models for worldwide use are limited by specific domain, e.g. clouds [3] or a huge undertaking with very few examples, such as the Dynamic World (DW) dataset [4].

(4) Operational AI models are not explainable. To overcome this, explainable artificial intelligence (XAI) is an active research field. However, end users requiring explainability today are still facing challenges.

We present a cloud-based service, called color33, that can be used without Earth observation (EO) knowledge and parameters to perform a semantic enrichment of optical satellite images worldwide. This service is called color33 because the main output is a scene classification map (SCM) with a pixel-based categorization of reflectance values into 33 color names (spectral categories). It is a general-purpose shared legend, that can be generated from other images such as Landsat, Sentinel-2, or Sentinel-3.

2. BACKGROUND

Semantic enrichment (SE) is known from the Semantic Web with the aim of meaningfully annotating documents to create a larger framework to link them. Thus, a knowledge graph is created, that can be queried and information that is only implicitly available can be produced. In the context of EO imagery we refer with SE to interpreted content of EO imagery (i.e., mapping data to symbols that represent stable concepts). In contrast to SE providing textual descriptions of EO image content, we refer to dense SE on a per-pixel level.

It is noteworthy that SE should provide general levels of semantic granularity to allow interoperability and transferability. For example, users requiring specific legends, e.g., a specific vegetation type, can start with general vegetation categories. This approach is also known from land cover classification frameworks such as the FAO LCCS [5]. Here, a dichotomous phase precedes a modular-hierarchical phase, stepwise approaching more specific target classes from general ones. It supports reusable components and significantly reduces processing time and workflow

doi:10.2760/2119408

complexity as well as transferability to larger areas or at even the entire Earth, given that terms such as "water"/"vegetation" are sufficiently generalized.

In color33 we use the satellite image automated mapper (SIAM) software [6], which generates low-level, generic, data-derived SE that is application-independent. SIAM encodes a per-pixel-, spectral-, and physical-model-based decision tree (DT) that can be applied to any EO imagery that is calibrated to at least top-of-atmosphere (TOA) reflectance. Thus, no training samples or user-defined parameter are required. Instead, SIAM uses a priori knowledge encoded into the DT, mapping each calibrated observation to one stable, sensor-agnostic multi-spectral colour name (i.e., category) based on its location in a multi-spectral reflectance hypercube. The result is a discrete and finite vocabulary for observations. This vocabulary (i.e., colour names/categories) is mutually exclusive (i.e., every observation belongs to one single partition) and totally exhaustive (i.e., the entire multi-spectral reflectance hypercube is partitioned).

SIAM can produce different granularities (i.e. number of colour names) from coarse (i.e., 18 colour names) to fine (i.e., 96 colour names), as well as additional data-derived information layers (e.g., multi-spectral greenness index, brightness). The 33 colour names are shared across different sensors and, therefore, transferrable.

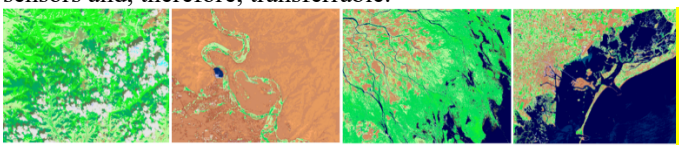


Fig. 1. Examples of SE for Sentinel-2 images worldwide: Cloud & cloud shadow in Austria, Oasis in the Sahara Desert, Volga river delta, Venice lagoon.

Consequently, like the SE for the Semantic Web, the semantic enrichment of EO images does not aim to create specific land cover classes, but is a first necessary, not sufficient step within a larger modular processing chain that is in contrast to the limited end-to-end classification.

3. ARCHITECTURE

This vision of a modular, hierarchical EO analytics workflow can be translated into a scalable architecture based on several independent cloud services of which color33 is the first of its kind. This architecture is modern, exchangeable, and scalable. Fig. 2. illustrates how color33 can be used in a broader workflow. It is noteworthy that the direct usage of color33 outputs themselves is limited if the analysis goal is beyond very simple categorisation, e.g., a vegetation mask. Thus, users can run both Jupyter notebooks on-premise after obtaining the semantically enriched images or connect to external cloud-based processes, although they are currently limited.

color33 employs a cloud-native, scalable, container-based architecture. While users can interact with a graphical user interface (GUI) or a command-line-interface (CLI) for automated usage and batch processing as a frontend, the backend consists of several interconnected steps that require internal synchronization between them (Fig. 3).

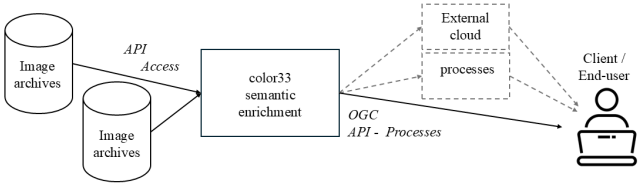


Fig. 2. Conceptual view on the usage of color33. It is a building block of a service-based architecture.

The service is offered through a standardized OGC API – Processes endpoint and users only specify the area-of-interest (AOI) and the time interval as well as optional settings, e.g., the coordinate reference system (CRS). The API is implemented in a Python-based Django backend and uses a PostgreSQL database system with the PostGIS spatial extension for persistent storage.

color33 automatically obtains the accessible links to the images that matches the search parameter. Several archive connectors allow flexible use of different archives, specified through administrative settings during the deployment. However, for big data processing, the use of an archive in close proximity to the processing deployment is preferred. In European environments examples are the Earth Observation Data Centre (EODC), CREODIAS, or the Open Telekom Cloud (OTC).

The processing includes pre-processing (PP), the SE and an output phase. The PP includes format conversions, potentially re-projection, and stacking the bands to a six-band composite, which constitutes the spectral signature as input to SIAM. SIAM is scaled through OpenMP in a multi-CPU setting. The parameter of the number of CPUs is tuned by the administrator based on the available CPUs and the number of images that are supposed to be processed in parallel. Using elastic cloud virtual machines (VMs) this parameter can be adjusted during re-sizing. The final steps include the generation of cloud-optimized geotiffs (COG) and creating a STAC catalogue.

The entire process is managed using Prefect as a workflow orchestration tool. We chose Prefect because it is developed in Python like the color33 service (except SIAM) and uses annotations as an easy approach to elevate Python scripts into managed workflows without extensive reprogramming. This is ideal for projects that start small but require a managed workflow at a later stage. In color33, Prefect manages the described tasks in a job with subflows that can be individually controlled. In situations with more requests than available resources, the jobs are queued and sequentially processed. Priority queues allow bypassing in case of important jobs. Like the SIAM task, the number of concurrent subflows is a tuning parameter that scales with the available resources. Experiences did not reveal problems scaling to common cloud flavours ranging from xlarge to

doi:10.2760/2119408

8xlarge and did not require anything else than scaling the VM and updating the tuning parameters.

The output is made available to the users as COGs within a spatio-temporal asset catalogue (STAC). Each user has their own STAC and the images of one job (typically a complete AOI and a time interval) are added to a new collection, while the job name presented to the user through the GUI or CLI is the collection name. Thus, users' STACs are growing over time allowing a very flexible downstream use using standard tools, e.g. STAC clients in different programming languages or GIS such as ArcGIS or QGIS. The users' STAC are protected through token-based access, which they can obtain using their username and password, for a better security in programming-based usage.

For stable operation of the color33 service, Keycloak is used for identity management providing standard OAuth authentication services. Authorization is accomplished using the Django backend. System and performance metrics are collected via Grafana. Several test jobs a day ensure consistency in the operation and detecting errors rapidly.

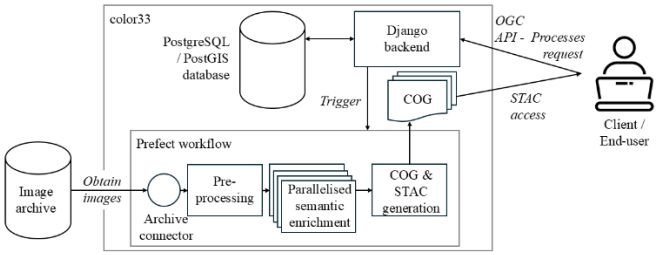


Fig 2. Architecture overview of the color33 service.

4. USE-CASE EXAMPLES

Application for SE-based analysis are as varied as the questions they are meant to address and range from ad-hoc analysis to dedicated processing pipelines. They include application scenarios such as producing tailored land cover classifications, automated asset monitoring, multi-temporal change-detection and emergency response activities. As a use-case example, analysing the spread of forest fires based on color33 outputs allows for both continuous monitoring as well as historical analysis and can be easily automated by simple computation of SE categories (Fig 3). A major benefit is the full automation including cloud-free image selection and worldwide transferability without changes.

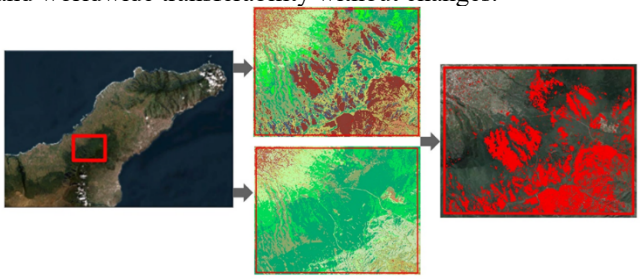


Fig 3. Detail view of the change map of the forest fire in Teneriffe, Spain, on 15.08.2023 obtained using the example workflow and color33's semantic enrichment.

Monitoring damages of vegetation and crops due to droughts is another use-case where fully automated categorical time series analysis can be beneficial. Questions such as how long, how large can be answered with countable observations. The automation and transferability of color33 is again a major advantage and facilitates production of masks and layers that can serve as the basis for such efforts (Fig 4) In contrast to approaches based on continuous variables, e.g NDVI, the vegetation duration can be analysed providing useful complementary information (Fig 5).

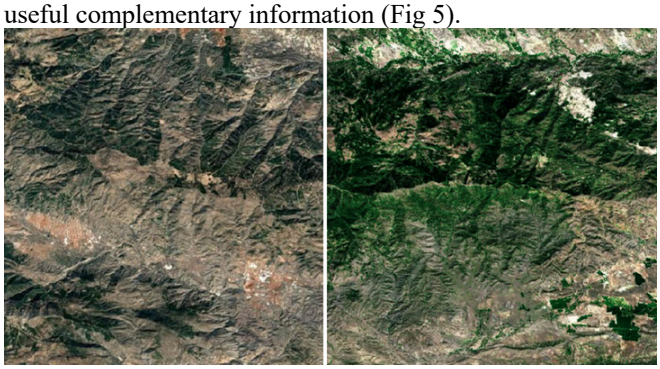


Fig 4. Comparison of a Sentinel-2 scene and overlayed with a mask produced from vegetation categories.

Fig 5: 2021 vegetation mask for Almeria (Spain) produced via color33. Note: Values represent the number of vegetation observations aggregated over the whole year.

5. DISCUSSION

Cloud-based approaches are a now common and typical, but typically offer locked-in processing chains, e.g. Google Earth Engine, or specific applications, e.g. Sen4CAP.

color33 provides a building block for a multi-service-based architecture that can consume cloud services, e.g. image archives, but also be consumed by cloud services, e.g., for downstream applications. This is accomplished by standards-based interfaces. color33 is operationally available at https://color33.io and operated by the company Spatial Services GmbH. As color33 is a new service, only few downstream applications were developed so far but it has been commercialized in selected use cases with very promising results.

doi:10.2760/2119408

The decentralized system goes beyond technical considerations of a federated service-based architecture: The SE provided by color33 is not the final output. Instead, it is a building block of a larger environment, avoiding the typical end-to-end classification, similar to the FAO LCCS. Primitives, i.e. spectral categories, are intermediate products for easy reuse to generate final products, e.g., a specific land cover map with a tailored legend. The DW is meant to be used similarly [4]. A main difference is that the DW lacks a generic approach using spectral categories, which is a colour naming and not land cover classes. Indicated by the number of classed in the DW (nine, without cloud) in comparison to at least 33 spectral categories (including cloud) of color33, the need for a dynamic worldwide classification is evident.

In current, ongoing discussions of energy consumption, lack of explainability, transferability, and reusability of DL approaches, more approaches that overcome these limitations will be developed. Although color33 does not yet provide land cover classes, as they can be generated downstream, it is a step to overcome these limitations. The algorithm is explainable, does not require training samples or energy-consuming training, is worldwide applicable, and can be used for different EO, e.g. Landsat, Sentinel-2, Sentinel-3, or very-high-resolution (VHR) [6].

The highest impact and usability of color33 is achieved with dense, long time series of EO images. A major advantage is that a categorical sequence is very informative about the Earth's surface dynamics and leverages the fully automated approach, shown by the two use-cases. Other use-cases for agricultural monitoring [7] or vegetation dynamics [8] demonstrated potential.

color33 was developed with scalability in mind, uses OpenMP and Prefect for parallelization at different levels. Scaling up and down works seamlessly. The algorithm itself has been verified [9], although validation of downstream applications needs to be done in case-by-case on user side.

There are also limitations, e.g., operationalization of downstream applications needs to be tackled to leverage full potential. Categorical variables, although very small (e.g. approx. 30 Mb / Sentinel-2 image) are limited in their re-usability in different CRS. Reprojection is limited (e.g. nearest neighbour) as it may have a strong effect on spatial accuracy. Reprojection is required prior to SE and needs to be repeated for different CRS. However, only few CRSs are usually required, e.g. UTM or a national grid.

6. CONCLUSION

color33 is a cloud-based service for SE of optical EO images, in the first development for Sentinel-2. In contrast to approaches based on DL, SE provided through color33 is on a lower semantic granularity supporting reusability and transferability. Users can include the results in cloud-based or local workflows to refine the semantics towards their target classes using spatial and temporal neighbourhoods or series, or by additional geodata such as a digital elevation model and derivatives such as height, slope, or aspect. This approach is similar to the FAO LCCS concept and is different from approaches that create target classes from reflectance values. Still, if color33 is considered as a building block, it can be included in DL-based workflows.

We showed the usability and potential in use-cases focusing on automation, stepwise refining the target legend, transferability, and explainability. Since color33 is applicable worldwide, any application can be instantiated by users on their own.

Future work will include an extension to additional EO satellite images (e.g. Landsat or Sentinel-3) and providing additional analysis capabilities.

REFERENCES

[1] D. Tuia, K. Schindler, B. Demir, XX. Zhu, M. Kochupillai, S. Džeroski, … and V. Markl, "Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward" "IEEE Geoscience and Remote Sensing Magazine", doi: 10.1109/MGRS.2024.3425961, 2024.

[2] G. Giuliani, B. Chatenoux, A. De Bono, D. Rodila, JP. Richard, K. Allenbach, … and P. Peduzzi, "Building an earth observations data cube: lessons learned from the swiss data cube (sdc) on generating analysis ready data (ard)" "Big Earth Data, 1(1-2)", doi: 10.1080/20964471.2017.139890, 2017.

[3] C. Aybar, L. Ysuhuaylas, J. Loja, K. Gonzales, F. Herrera, L. Bautista, ... and L. Gómez-Chova, "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2" "Scientific data, 9(1)", doi: 10.1016/j.dib.2024.110852 , 2022.

[4] CF. Brown, SP. Brumby, B. Guzder-Williams, T. Birch, SB. Hyde, J. Mazzariello, ... and AM. Tait, "Dynamic World, Near real-time global 10 m land use land cover mapping" "Scientific Data, 9(1)", doi: 10.1038/s41597-022-01307-4, 2022.

[5] A. Di Gregorio, "Land cover classification system: classification concepts and user manual: LCCS (Vol. 2)" "Food & Agriculture Org.", https://www.fao.org/4/y7220e/y7220e00.htm, 2005.

[6] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery—Part I: System Design and Implementation", "IEEE Trans. Geosci. Remote Sensing, vol. 48, no. 3", doi: 10.1109/TGRS.2009.2032457, 2010.

[7] A. Hartmann, M. Sudmanns, H. Augustin, A. Baraldi, and D. Tiede, "Estimating the temporal heterogeneity of mowing events on grassland for haymilk-production using Sentinel-2 and greenness-index" "Smart Agricultural Technology, 4" doi: 10.1016/j.atech.2022.100157, 2023.

[8] M. Sudmanns, H. Augustin, A. Baraldi, T. Strasser and D. Tiede, „How green is Austria? An Austrian-wide information layer based on semantic querying of EO data". Proc. of the 2023 conference on Big Data from Space (BiDS'23). Publications Office of the European Union, doi: 10.2760/46796, 2023.

[9] A. Baraldi, ML. Humber, D. Tiede and S. Lang, „GEO-CEOS stage 4 validation of the Satellite Image Automatic Mapper lightweight computer program for ESA Earth observation level 2 product generation–Part 2: Validation" "Cogent Geoscience, 4(1)", doi: 10.1080/23312041.2018.1467254, 2018.

ENGAGING SENTINEL DATA USERS TO ADOPT THE ZARR DATA FORMAT: THE EOPF TOOLKIT

Julia Wagemann[1], Sabrina H. Szeto[1], Gisela Romero Candanedo[1], Emmanuel Mathot[2], James Banting[3]

[1] thriveGEO GmbH, Ringeisenstr. 15, 86381 Krumbach, Germany

[2] Development Seed, Travessa da Pereira 16A, Armazém 12A, 1170-313 Lisbon, Portugal

[3] Sparkgeo UK Limited, 6 St. Colme Street, EH3 6AD Edinburgh, Scotland

ABSTRACT

The Earth Observation Processor Framework (EOPF) Toolkit is a community-driven set of resources that facilitates the adoption of the Zarr data format for Copernicus Sentinel data users, specifically targeting users who are new to cloud computing. The Sentinels EOPF Toolkit is developed by Development Seed, thriveGEO and Sparkgeo, with a group of champion users. The main resource that has been developed is EOPF 101, a user-friendly online book consisting of documentation, Jupyter Notebooks and open-source libraries and plug-ins that showcase the use of Zarr format Sentinel data for applications across multiple domains. In addition, community engagement activities such as a notebook competition helpsSentinel users to explore the new data format while getting community support.

Index Terms— cloud-native data formats, Zarr, EOPF, Sentinels, user engagement, earth observation

1. INTRODUCTION

Over the last few years, cloud-native tools like STAC for data discovery and access and cloud-optimized data formats such as Zarr and COGs have emerged and are currently revolutionising how scientific communities work with large-scale geospatial data. The European Space Agency (ESA), through the Earth Observation Processor Framework (EOPF), is currently reprocessing Sentinel-1, -2, and 3 data into cloud-optimized data formats. Through the EOPF Sentinel Zarr Samples Service, Copernicus Sentinel data users get early access to "live" sample data from the Copernicus Sentinel missions -1, -2 and -3 in the new Zarr data format.

The EOPF Toolkit [1,2] engages the Sentinels user community to facilitate the adoption of the new Zarr data format, starting with the data published by the EOPF Sentinel Zarr Samples Service [3]. The Toolkit team acts as test users of the Samples Service data. In addition, the resources being developed complement existing user engagement efforts by specifically targeting users who are new to cloud computing.

2. ABOUT EOPF

The Earth Observation Processor Framework (EOPF) is an initiative led by the European Space Agency (ESA) designed to modernise and harmonise data from the Copernicus Sentinel Missions. With the upcoming Copernicus Expansion missions in 2028, the amount of data produced daily will significantly increase. EOPF is ESA's solution to organise Sentinel data in a way that works seamlessly with modern cloud technology. This will make it easier to find, access, and process the information needed. The new approach provides user-friendly access, simplifies maintenance, and helps keep costs down, guaranteeing reliable access to Sentinel data in the long run.

Under the EOPF Sentinel Zarr Sample service, ESA provides access to re-engineered EOPF Zarr products. The Sentinel-1, Sentinel-2, and Sentinel-3 missions are the first to be re-processed and have been made available under the EOPF Sentinel Zarr Samples STAC catalog [6].

3. ABOUT ZARR

Zarr is an open-source, cloud-native protocol for storing multi-dimensional arrays. It is specifically designed to work well with cloud storage and larger-scale computing systems and can be seen as a cloud-native alternative to older formats like HDF5 or NetCDF. A key advantage to traditional formats is that the Zarr specification stores large multi-dimensional arrays in chunks, which are smaller pieces of the larger array. Chunks can be accessed individually, or multiple chunks can be read and written in parallel, making data access highly efficient.

Zarr works across different storage systems, including local file systems, cloud object storage, as well as distributed file systems, offering a greater flexibility compared to traditional file formats.

doi:10.2760/2119408

In addition, Zarr embeds metadata directly alongside the data. This makes Zarr self-descriptive, as each data array contains descriptive information about itself, such as data type, dimensions or additional attributes.

4. EOPF RESOURCES

The EOPF Toolkit contains several useful resources to help Sentinel data users get started with using the Zarr data format. These resources include: an online book, open source integrations and plug-ins, case studies that showcase the use of Zarr format Sentinel data for applications across multiple domains and a notebook competition. All materials are open source and freely accessible online on Github [4].

The resources that have been developed under the EOPF toolkit are addressing common questions that users are asking around data access and use resulting in a comprehensive resource about accessing and processing Sentinel Zarr data. The toolkit highlights example workflows by champion users and gives the user community the chance to try it out themselves via a notebook competition.

4.1. EOPF 101

The EOPF 101 online book [5] is a community resource where data users learn how to discover, access and process data from the EOPF Sentinel Zarr Samples Service by ESA.

The book is divided into five chapters: (i) Introduction to EOPF, (ii) Introduction to Zarr, (iii) Introduction to EOPF STAC, (iv) Tools to work with Zarr and (v) EOPF in Action.

Chapter 1 - 'Introduction to EOPF' provides a high-level easy-to-understand overview of the EOPF project by ESA. Chapter 2 - 'Introduction to Zarr' provides a practical introduction to the cloud-optimized Zarr data format. It shows the advantages of the format, gives an overview of the data structure and includes performance comparisons with other formats, too. Chapter 3 - 'Introduction to EOPF STAC' gives easy-to-understand practical examples on how to discover and access data with the EOPF STAC catalog. Chapter 4 - 'Tools to work with Zarr'provides a collection of practical examples of libraries and plug-ins that support users in working with data from the EOPF Samples Service. Chapter 5 - 'EOPF in Action' is a collection of hands-on practical end-to-end workflows featuring the use of Sentinel Zarr data in different application areas. Part of these case studies are co-developed with a set of early adopters, also referred to as champion users.

4.2. Open source integrations and plugins

As part of toolkit a series of open-source integrations and plugins for using the EOPF Sentinel Zarr Samples Service (Table 1). These libraries and plugins will make it easier for users to use the Samples Service data together with their favourite programming languages and software. Table 1 provides an overview of integrations and libraries the EOPF toolkit team is currently working on.

Table 1. Name and description of plugins and integrations

Plugin / Library	Description
STAC + Zarr	EOPF Zarr Access from STAC
GDAL Zarr Driver	Enhanced GDAL driver for EOPF Zarr
QGIS Plugin	Native QGIS integration for EOPF Zarr
R Integration	R libraries for EOPF Zarr access
Julia Integration	Julia packages for EOPF Zarr
TiTiler Multidim	Multidimensional data support for TiTiler
Stackstac Optimizations	Enhanced Stackstac for EOPF

4.3. Case studies

A set of applied case studies along with Jupyter Notebooks will be developed by a set of champion users. These champion users bring specific domain expertise and will provide example workflows for a variety of applications. Each Champion User will also share their experience working with Zarr as well as the advantages and disadvantages they see using the new data format for relevant workflows. The Jupyter Notebooks will serve as easily modifiable templates for Sentinel data users to develop their own workflows.

The case studies can be categorised as thematic and technical case studies. Technical case studies demonstrate workflows such as Zarr with QGIS, Zarr with R, multi-scale Zarr and dataset screening with rio-tiler and lonboard. On the other hand, thematic case studies demonstrate workflows from different application areas such as wildfire assessment with Sentinel-3 and Sentinel-2 data or monitoring coastal dynamics in cloud-prone regions using Sentinel-1, or flooding analysis in Valencia with Sentinel-1.

4.4. Notebook competition

To engage with Sentinel data users interested in working with data from the EOPF Samples Service, a notebook competition will take place between October 2025 and January 2026. During this competition, the Sentinel data user community is invited to try out the new Zarr data format for themselves by creating workflows in Jupyter Notebooks. A set of judges will decide at the end on the top ten notebooks based on published evaluation criteria. The top ten notebooks will be published on Github and the winning notebook authors will be announced in an online webinar.

doi:10.2760/2119408

5. CONCLUSION

The resources from the EOPF Toolkit help Sentinel data users who are new to cloud computing become familiar with working with the Zarr data format in general and reprocessed Sentinel Zarr data from the EOPF Zarr Samples Service in particular. The additional development of open-source integrations and plugins enable users to explore the reprocessed sample data in their preferred programming languages and tools.

REFERENCES

[1] Wagemann, J., Szeto, S. H., Mathot, E., and Banting, J.: The Sentinels EOPF Toolkit: Community Notebooks and Plug-ins for using Copernicus Sentinel Data in Zarr format, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-17137, https://doi.org/10.5194/egusphere-egu25-17137, 2025.

[2] Szeto, S. H., Wagemann, J., Mathot, E., and Banting, J.: The Sentinels EOPF Toolkit: Driving Community Adoption of the Zarr data format for Copernicus Sentinel Data, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-15864, https://doi.org/10.5194/egusphere-egu25-15864, 2025.

[3] Briese, C., Reimer, C., Briese, C., Reck, C., Papadakis, D., Claus, M., Brandt, G., Fouilloux, A., and Odaka, T.: From SAFE to Zarr: The EOPF Sample Service Initiative, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-21202, https://doi.org/10.5194/egusphere-egu25-21202, 2025.

[4] EOPF Toolkit Github repository, https://github.com/eopf-toolkit

[5] ESA EOPF 101, https://eopf-toolkit.github.io/eopf-101/

[6] EOPF Sentinel Zarr Samples Service STAC API, https://stac.browser.user.eopf.eodc.eu/?.language=en

doi:10.2760/2119408

doi:10.2760/2119408

EOPF SENTINEL ZARR SAMPLES – A REVIEW ON THE NEW DATA FORMAT

Christoph Reimer[1], Stefan Reimond[1],Sean Hoyal[1], Christoph Reck[2], Mario Winkler[2], Esther Millet[3], Michele Claus[4], Konstantin Ntokas[5], Tina Odaka[6], Anne Fouilloux[7]

[1] EODC Earth Observation Data Centre for Water Resources Monitoring GmbH, Austria

[2] DLR German Aerospace Center, Earth Observation Center (EOC), Germany

[3] Evenflow, Belgium

[4] Eurac Research, Institute for Earth Observations, Italy

[5] Brockmann Consult GmbH, Germany

[6] Ifremer, UMR LOPS, France

[7] Simula Research Laboratory, Norway

ABSTRACT

A new data format will be introduced for all Copernicus Sentinel mission data. This marks a significant change in the distribution and access of Sentinel data for users and data providers. The EOPF Sentinel Zarr Samples service provides early access to the new Zarr data format with the objective to ensure user adoption. Jupyter notebooks and open-source libraries are developed to support users on their journey with the new format. Benchmarks of the EOPF Zarr data format in comparison to the legacy SAFE format are presented. Various benchmarking scenarios are defined to give an objective comparison and to highlight potential challenges and opportunities inherent to this data format change for users and data providers.

Index Terms— Sentinel, Zarr, Cloud-native, Data formats

1. INTRODUCTION

Over the past decade, the operational data processors for Copernicus Sentinel missions have produced vast amounts of Earth observation data, supporting a wide range of scientific research and commercial applications. However, the current data format used by ESA, known as SAFE (Standard Archive Format for Europe), has become increasingly outdated. In response, ESA initiated the transition to a modern, cloud-native format based on Zarr. This fundamental change is supported by the EOPF Sentinel Zarr Samples service, offering early access to the new data format and relevant documentation to guide users through this transition.

This evolution is key to enabling a cloud-friendly, interoperable data ecosystem that enhances accessibility and integration with today's data processing platforms. The overarching goal is to harmonize data formats across all Copernicus Sentinel missions, facilitate scalable cloud-based processing, and ensure seamless compatibility with modern data science tools. The initiative is designed to minimize disruption while maintaining continuity for existing users, applications, and services.

2. SERVICE COMPONENTS

The EOPF Sample Service consists of several core components. The EOPF Core Platform transforms incoming SAFE-format data into the new cloud-optimized EOPF Zarr format and offers access through STAC and S3 APIs. The EOPF Analytical Hub provides additional services for users, including JupyterHub, Dask, and a STAC Browser. This toolset was selected with the objective to lower the barrier to cloud-based applications and to enhance data discovery capabilities. Furthermore, a strong focus of the service is on user support and engagement. High quality Jupyter notebooks are provided complementing and demonstrating the use of open-source software libraries exploiting the new data format. User engagement is carried out through regular user interactions via webinars, newsletters and moderation of a community support forum.

doi:10.2760/2119408

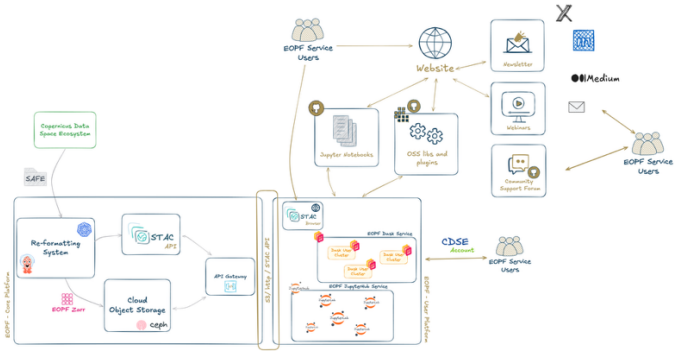


Fig. 1: EOPF Sentinel Zarr Samples Service components

2.1. EOPF Core Platform

The EOPF Core Platform is deployed on a robust and scalable infrastructure hosted and operated by EODC. The re-formatting system is part of the EOPF Core Platform, responsible for converting existing data from the Copernicus Data Space Ecosystem (CDSE) in the current SAFE format into the new Zarr-based EOPF format. At the core of this system is the workflow engine, powered by Argo Workflows and related projects (Argo CD, Argo Events). Argo Workflows is an open-source container-native workflow engine for orchestrating parallel jobs on Kubernetes. The Object Storage Service is realized via CEPH, an open-source storage platform, exposing a S3 compatible interface (S3 API). The STAC API, based on eoAPI, enables interactive exploration of the EOPF Zarr data offering. The STAC and S3 API is behind an API Gateway to enforce security policies as required.

2.2. EOPF Analytical Hub

The EOPF Analytical Hub is a central component of the service offering with the objective to facilitate user adoption activities by providing a robust framework for data analysis and computation. It includes the services JupyterHub, Dask (Dask Gateway) and STAC Browser, all orchestrated on Kubernetes to guarantee scalability. The EOPF Analytical Hub is a fully managed service providing pre-configured cloud environments to interact with the new EOPF Zarr data.

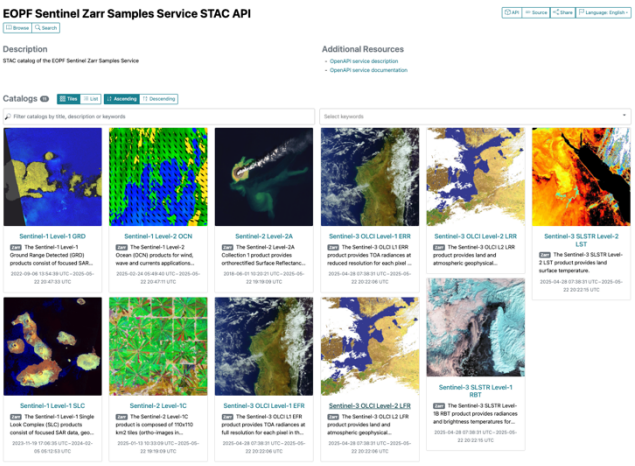


Fig. 2. STAC Browser

2.3. User Support and Engagement

A suite of Jupyter Notebooks has been developed by consortium experts, including members of the Pangeo community to support user adoption. These notebooks demonstrate the practical use of the new format and are regularly refined based on user feedback. Notebooks are complemented by a set of open-source libraries and plug-ins such as the xarray-eopf backend[1]. Development contributions to standard libraries, such as GDAL, are carried out to further enhance the user adoption in other communities. Community engagement will be done via various channels. The centerpiece of the activity is the website, holding and linking all relevant resources for users. Users can subscribe to the EOPF Zarr Newsletter to get all the latest information. In total, 9 webinars are planned about the EOPF Zarr format and to showcase dedicated use cases demonstrations. Furthermore, a community support forum is actively maintained and managed for direct interaction with users and to collect feedback.

3. QUALITY ASSURANCE

Quality assurance procedures are implemented to ensure the service operates according to functional and non-functional requirements. GitHub and GitHub Actions are utilized to implement these quality assurance procedures as depicted in Fig. 3. Test cases are executed regularly via GitHub Actions against the various user facing components of the EOPF Sentinel Zarr Samples service.

[1] https://github.com/EOPF-Sample-Service/xarray-eopf

doi:10.2760/2119408

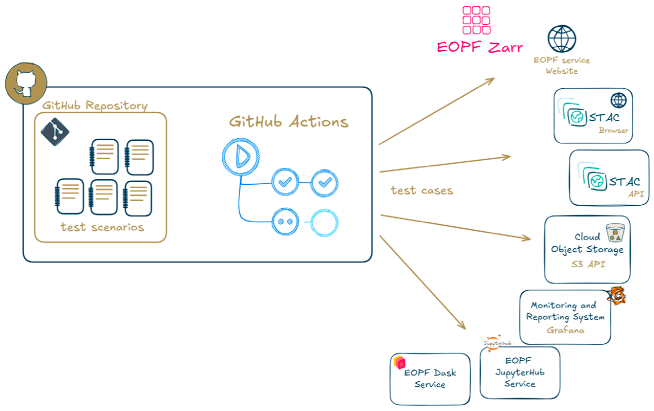


Fig. 3. Quality assurance approach

One important aspect of the quality assurance is to guarantee the EOPF Zarr standard and its specifications. The conformance of the products is verified in respect to its metadata properties, the structure of the Zarr product and the STAC related entities. The underlying testing framework used will be presented in the final paper. Moreover, benchmarking tests of the EOPF Zarr products are conducted.

4. PRODUCT BENCHMARKS

The EOPF Zarr product benchmarks will focus on data access performance in comparison to the legacy format SAFE and potentially other data formats. Objective benchmark tests will be developed considering the fundamental differences in data format and access between the formats. Furthermore, the performance of the EOPF Zarr data format on different cloud object storage solutions will be presented. A set of scenarios will be developed for all benchmarks, considering the various options provided by Zarr to organize data. These options will represent typical access patterns and use case scenarios resulting in different data chunking and compression approaches of the data. The performed benchmarks will analyze the resulting storage requirements and access speeds compared to the SAFE format and others.

5. CONCLUSION

Introducing a new data format for all Copernicus Sentinel data products represents a significant change in how data get consumed. The present paper outlines the challenges and opportunities inherent to this data format change for data providers and users.

doi:10.2760/2119408

doi:10.2760/2119408

Synthetic hyperspectral PRISMA Data Generation From Existing Landsat 8 and 9 Multispectral Data

Ari Gjerazi[1], Valeria La Pegna[2], Fabio Del Frate[2]

[1] Computer Engineering Department, Epoka University, Tirana, Albania

[2] Dipartimento Ingegneria Civile ed Ingegneria Informatica, Università degli Studi di Roma "Tor Vergata", via del Politecnico 1, 00133 Rome, Italy

Abstract

Remote sensing applications focused on hyperspectral imaging have proven to be able to produce detailed and comprehensive results in Earth Observation applications compared to multispectral data. All recent implementations of hyperspectral data, particularly those integrating machine learning tools, suffer from relatively small datasets. PRISMA mission, launched in 2019 by Italian Space Agency (ASI), in specific, provides significant data through its high spectral resolution, but it faces low revisit time. In light of this, the production of synthetic hyperspectral PRISMA images starting from multispectral data, becomes all the more important. This work has explored different Neural Network (NN) models, such as Multi Layer Perceptron (MLP) and Generative Adversarial Networks (GANs) to generate synthetic PRISMA data and further proposes a number of new technologies or combinations thereof, in the endeavor to optimally generate synthetic hyperspectral data from multispectral ones, acceptable for the purposes of training learning models.

Index Terms— PRISMA, synthetic data, hyperspectral data generation, neural network, MLP, GAN

1. Introduction

Demand for high volumes of hyperspectral data is continuously growing due to their usefulness in a variety of fields ranging from fields like agricultural studies, climatology, environmental science, land classification, mineral recognition and mapping, lithologic mapping, mineral resource prospecting, mining environment monitoring, and leakage monitoring of oil and gas [1, 2]. Remote sensing technologies have seen remarkable advancements over the past few decades, offering unprecedented capabilities in environmental monitoring, resource management, and scientific research [3]. Among these technologies, hyperspectral imaging has emerged as a powerful tool for capturing detailed spectral information by having very narrow bands available across the entire VNIR range of wavelengths, providing insights that go far beyond the capabilities of traditional multispectral imaging. However, despite the promise of hyperspectral imaging, several challenges remain, particularly when it comes to the collection of the acquisition and the utilization of robust data for these applications.

A key challenge in remote sensing, especially with hyperspectral imagery, lies in the availability of large, high-quality datasets. The application of machine learning methods often requires vast amounts of annotated data to build effective predictive models. Unfortunately, hyperspectral datasets, especially those derived from specific platforms such as the PRISMA (PRecursore IperSpettrale della Missione Applicativa) satellite by Italian Space Agency (ASI), are typically limited in both size and accessibility. While the PRISMA satellite provides hyperspectral data with high spectral resolution, its data acquisition is both costly and challenging. This makes it difficult to rely solely on real-world data for training machine learning algorithms, which depend on large volumes of labeled examples.

In light of these limitations, the generation of synthetic hyperspectral images has become a crucial area of research [4, 5]. Synthetic data can offer a solution to the scarcity of real-world data by providing an accessible and controlled environment in which to generate large, diverse datasets. Specifically, the focus of this work is on developing synthetic PRISMA-like images that can be used for training machine learning models. While the concept of synthetic data generation has been explored in other areas of remote sensing, the production of high-quality synthetic PRISMA images remains a relatively unexplored domain. Currently, there are little to no existing metrics or benchmarks for entirely synthetic PRISMA data due to such generation not having been previously attempted in full.

This paper investigates various technologies and approaches for generating synthetic hyperspectral images that closely mimic the characteristics of PRISMA data, produced from multispectral Landsat 8/9 data. These two missions share the same ground spatial resolution with pixel size of 30 m, which is convenient to avoid and limit discrepancies in the production of synthetic data. The acquisitions of the two satellite sensors were chosen in the same area and on the same day, ensuring matching pairs.

doi:10.2760/2119408

The primary objective is to explore the usage of Feed Forward Networks (FFNs) to generate synthetic imagery with particular focus on Multi Layer Perceptrons (MLPs) and Generative Adversarial Networks (GANs) that can produce synthetic data that is not only computationally efficient to generate but also sufficiently representative of the original data.[11] The remainder of this paper is structured as follows: Chapter 2 provides a review of the different methods that can be used to synthetically generate image data, focused in particular on deep-learning based approaches to synthetic generation. Chapter 3 discusses the experimental results as a proof of concept of PRISMA-data generation. Chapter 4 presents a few proposals regarding other techniques that are explored and compared with what has already been done.

2. Materials and Methods

PRISMA data were chosen, specifically examining the VNIR range [1]. Attempts at synthesizing hyperspectral data have already been made for various applications: of particular note is pansharpening [6]. However, wholesale generation of hyperspectral PRISMA data out of a multispectral source is still a largely unexplored direction.

Several approaches can be taken to the generation of hyperspectral data, for example methods that employ band interpolation between the multispectral and hyperspectral bands [6, 7]. However these standard approaches fail when met with more complex band relationships that do not directly correlate. This necessitates the usage of more powerful modes of simulation: in particular, deep learning techniques [8].

The first technique used in this study is that of a deep feed-forward network (FNN), specifically a Multi Layer Perceptron (MLP), which is used for simulating the relationships between multispectral and hyperspectral bands. In this study we also discuss two potential alternatives to feed-forward networks.

The second technique used in this study is a Generative Adversarial Network (GAN) which supports the necessary structures for learning complex relations between input and output features [8, 9]. GANs have a history of being used in the field of remote sensing for similar applications [10-12].

However, few attempts have been made to generate PRISMA data specifically and the production of synthetic PRISMA images remains a mostly unexplored field, with room for further experimentation [8].

This study relies on two primary data sources: hyperspectral PRISMA and multispectral Landsat 8/9 imagery. The PRISMA data consists of 63 bands spectrum in the VNIR with reflectance values expressed in the form of digital numbers (DN) scaled to a [0,1] range. The Landsat 8/9 data brings a total of 5 bands in the VNIR range of the spectrum.

Part of preliminary preprocessing operations includes the corregistration of the acquisitions from the two sensors, to ensure precision in the estimation of the synthetic spectra. Extraction of reflectance values for each channel present in the VNIR range is then conducted for both satellite datasets. A deep learning approach was employed to predict the synthetic hyperspectral PRISMA data based on multispectral Landsat 8/9 input. The model architecture is a FFN with parameters optimized over multiple trials. It consists of a traditional multi-layer fully connected neural network, implemented using the Keras deep learning framework with TensorFlow backend. An optimization library (Optuna) is used and repeated trials are performed to check for the optimization of layer count, neuron count, epoch and batch size and especially activation function choices, selected from a pool of the following: ReLU, Sigmoid, ELU, SeLu, Softplus, Tanh, Leaky ReLU, SiLU, ReLU6, GELU and lastly, Linear.

2.1. Training Procedure

The training procedure involved making use of a dataset of 4 pairs of images, of 3548 data points, augmented with an additional 1787 simulated points and then splitting the dataset into training and validation subsets. The training subset contained 70% of the available Landsat 8/9 and PRISMA image pairs, while the validation subset contained the remaining 30%. Data augmentation techniques such as random rotation, scaling, and flipping were applied to the training set to improve model generalization.

The architecture for the MLP consists of two layers of 52 and 22 neurons respectively, as well as GELU and ReLU6 activation functions, followed by a dropout layer (float = 0.1163) and a Linear activation function.

The training architecture for the GAN consists of a generator with three processing branches. The first branch is a dense 64-neuron layer, followed by batch normalization, while the second branch is a skip connection branch based on the concept initially presented in ResNet architecture [13]. This is followed by a secondary dense layer of 32 neurons. Lastly, the Noise Processing branch which processes noise input through a minimal 16-neuron dense layer with ReLU activation and batch normalization. These three branches are concatenated to form a coherent pipeline that feeds into a progressive expansion architecture with three layers. Based on 30 total experiments (each containing 200 trials) with parameter variation and the optimal values for the network parameters were derived. .

Figure 1 and Figure 2 show the behaviours of the Model Loss for MLP and GAN models during the training phase according to the number of epochs.

doi:10.2760/2119408

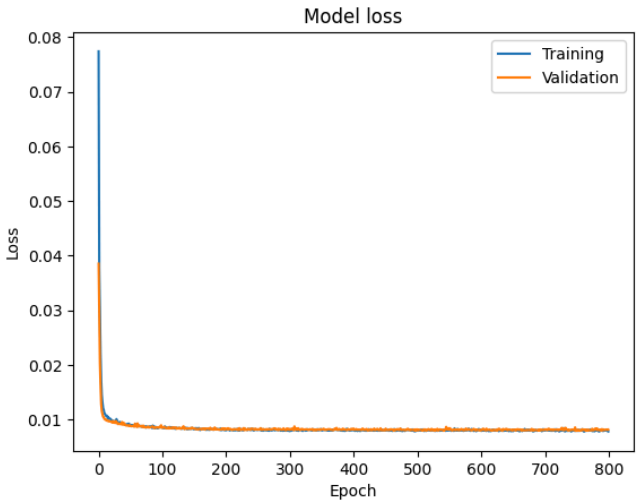


Fig. 1. Plot of model loss for the MLP model during training phases

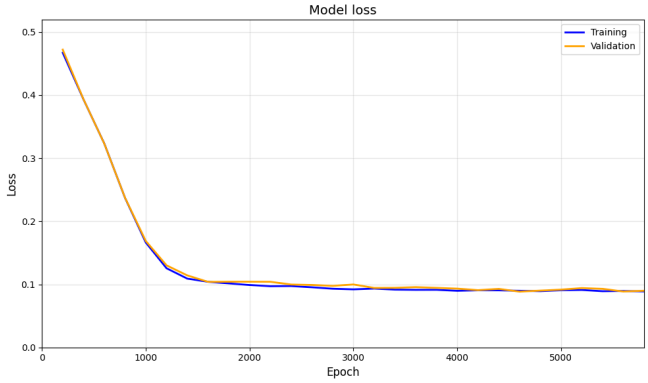


Fig. 2. Plot of model loss for the GAN model during training phases

3. Results

Following the above training procedure, from both models 63 synthetic hyperspectral bands were simulated starting from 5 multispectral Landsat 8/9 acquisition.

Model performances were assessed by evaluating the synthetic PRISMA reflectances obtained against the original ones from the validation set. Metrics were compared across the whole spectra. These include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson Coefficient and Coefficient of Determination (R^2) and were used to quantify the accuracy of the model's predictions.

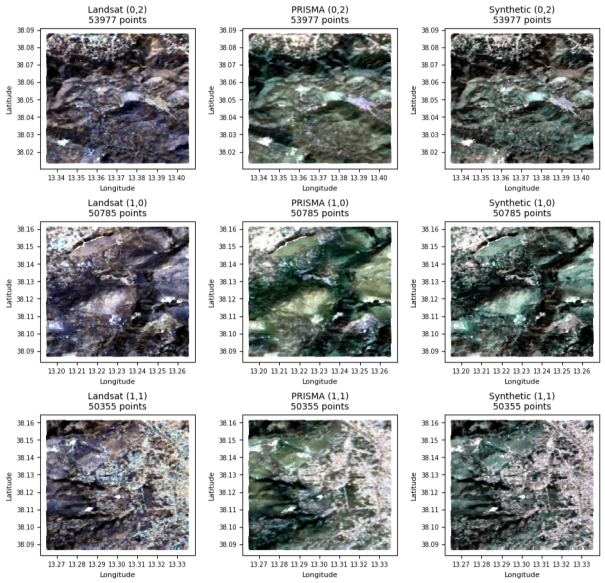
Table 1. Results of quality metrics of the MLP NN training from the original and synthetic PRISMA images.

Set	R^2	RMSE	MAE	Pearson
Train	0.479	0.016	0.01	0.692
Test	0.447	0.017	0.011	0.668
Validation	0.441	0.017	0.011	0.664

Table 2. Results of quality metrics of the GAN NN training from the original and synthetic PRISMA images.

Set.	R^2	RMSE	MAE	Pearson
Train	0.843	0.097	0.069	0.91
Test	0.844	0.097	0.068	0.918
Validation	0.842	0.098	0.069	0.918

While the RMSE and MAE for the GAN approach is worse than the MLP network, keeping in mind that the generator has to compete against a robust discriminator architecture, the R^2 score is significantly better, indicating a much stronger correlation. Synthetic MLP samples and synthetic GAN samples in RGB representation are provided in Figure 2 and Figure 3.



doi:10.2760/2119408

Fig.3. Visual comparison in RGB composite of the original multispectral Landsat 8/9, original hyperspectral PRISMA and synthetic PRISMA generated with the GAN model.

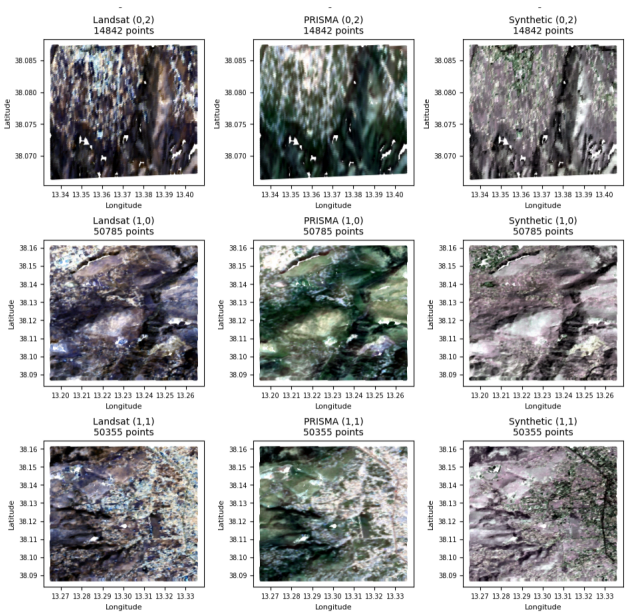


Fig.4. Visual comparison in RGB composite of the original multispectral Landsat 8/9, original hyperspectral PRISMA and synthetic PRISMA generated with the MLP model.

4. Conclusions and Future Research

The results obtained from the two NN models are promising for the application of simulating hyperspectral bands starting from multispectral ones. The MLP approach certainly yields promising results, but the GAN approach has proven to be significantly better at synthesizing images. The GAN network presented in this study currently simulates images at a higher similarity than the known benchmarks, reaching values of R^2 of 0.84 [8].

Future alternatives for research include exploration of alternative network architectures within the MLP or GAN umbrella. The implementation of CNNs, Variational Autoencoders (VAEs) and diffusion techniques could also prove promising outcomes.

This study was carried out in the framework of AIOBSERVER project (https://ai-observer.eu/) titled ”Enhancing Earth Observation capabilities of the Eratosthenes Centre of Excellence on Disaster Risk Reduction through Artificial Intelligence”, that has received funding from the European Union’s Horizon Europe Framework Programme HORIZON- WIDERA- 2021-ACCESS-03 (Twinning) under the Grant Agreement No. 101079468.

5. REFERENCES

[1] Gianinetto, Marco & Lechi, Giovanmaria. (2004). The development of Superspectral approaches for the improvement of land cover classification. Geoscience and Remote Sensing, IEEE Transactions on. 2670 - 2679. 10.1109/TGRS.2004.835347.

[2] Li Zhizhong et al., "A review on the geological applications of hyperspectral remote sensing technology," 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Shanghai, 2012, pp. 1-4, doi: 10.1109/WHISPERS.2012.6874235.

[3] Schaepman, M.E. & Green, Robert & Ungar, Stephen & Curtiss, Brian & Boardman, Joe & Plaza, Antonio & Gao, Bo-Cai & Ustin, Susan & Kokaly, Raymond & Miller, John & Jacquemoud, Stéphane & Ben-Dor, Eyal & Clark, Roger & Davis, Curtiss & Dozier, Jeff & Goodenough, David & Roberts, Dar & Swayze, Gregg & Milton, E.J. & Goetz, Alexander. (2006). The Future of Imaging Spectroscopy Prospective Technologies and Applications. International Geoscience and Remote Sensing Symposium (IGARSS). 2005 - 2009. 10.1109/IGARSS.2006.519.

[4] D. A. Abuhani, I. Zualkernan, R. Aldamani and M. Alshafai, "Generative Artificial Intelligence for Hyperspectral Sensor Data: A Review," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 6422-6439, 2025, doi: 10.1109/JSTARS.2025.3538759.

[5] C. Giacomo et al., "The Hyperspectral Prisma Mission in Operations," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020, pp. 3282-3285, doi: 10.1109/IGARSS39084.2020.9323301.

[6] Paul, S., & Kumar, D. N. (2020). Transformation of Multispectral Data to Quasi- Hyperspectral Data Using Convolutional Neural Network Regression. IEEE Transactions on Geoscience and Remote Sensing, 1–17.

[7] Hoang, Nguyen & Koike, Katsuaki. (2017). Transformation of Landsat imagery into pseudo-hyperspectral imagery by a multiple regression-based model with application to metal deposit-related minerals mapping. ISPRS Journal of Photogrammetry and Remote Sensing.

[8] M. Monaco, G. A. Licciardi, M. L. Battagliere, R. Guarini, M. G. C. A. Cimino and L. Candela, "A Machine-Learning Approach for Generating Synthetic Prisma Hyperspectral Images from Multispectral Data," IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 2024, pp. 3659-3662, doi: 10.1109/IGARSS53475.2024.10641756.

[9] Kurach, K., Lučić, M., Zhai, X., Michalski, M., & Gelly, S. (2019, May). A large-scale study on regularization and normalization in GANs. In International conference on machine learning (pp. 3581-3590). PMLR.

[10] Hennessy, A., Clarke, K., & Lewis, M. (2021). Generative adversarial network synthesis of hyperspectral vegetation data. Remote Sensing, 13(12), 2243.

[11] Nie, Y., Li, L., Gan, Z., Wang, S., Zhu, C., Zeng, M., ... & Wang, L. (2021). MLP architectures for vision-and-language modeling: An empirical study. *arXiv preprint arXiv:2112.04453*.

[12] Audebert, N., Le Saux, B., & Lefèvre, S. (2018, July). Generative adversarial networks for realistic synthesis of hyperspectral samples. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 4359-4362). IEEE.

A FRAMEWORK FOR GLOBAL HIGHWAY NETWORK CHANGE DETECTION APPLIED TO LANDSAT DATA

[1]Johannes H. Uhl, [2]Luca Maffenini, [3]Panagiotis Politis, [4]Katarzyna Krasnodębska, [1]Martino Pesaresi, [5]Taylor Jaworski, [6]Carl T. Kitchens

[1]Joint Research Centre (JRC), European Commission, Ispra (VA), Italy
[2]UniSystems Luxembourg Sàrl, Bertrange, Luxembourg
[3]European Dynamics Belgium S.A., Brussels, Belgium
[4]Institute of Geography and Spatial Organization, Polish Academy of Sciences, Warsaw, Poland
[5]Department of Economics, University of Colorado Boulder, Boulder (CO), USA
[6]Department of Economics, Florida State University, Tallahassee (FL), USA

ABSTRACT

Multi-temporal geospatial data measuring the evolution of transportation networks is scarce, impeding our quantitative knowledge on the dynamics of highway and other transportation networks at global scale. To tackle this issue, we developed a framework that integrates contemporary road network data with road presence probabilities extracted from historical, multispectral Landsat data (1990-2024), enabling the measurement of highway network growth from 1990 onwards. The framework also supports earlier Landsat data, other geohistorical data such as historical maps or pre-Landsat aerial imagery. First experiments conducted for a study area in the United States yield promising results, achieving Area-under-the-Curve values of up to 0.88.

Index Terms— Road network, symbolic machine learning, historical transportation networks, accessibility.

1. INTRODUCTION

Transportation networks are the backbone of economic activity, enabling the shipping of goods, work-related commuting, and transportation in general. The evolution of such networks reflects patterns of economic development, technological changes, or effects of transportation-related policies, and is linked to issues of accessibility, often reflecting social and economic inequalities [1].

While data on present-date transportation networks such as roads and railroads are abundant, due to volunteered geographic information (e.g., OpenStreetMap - OSM), industry-fueled data harmonization efforts (e.g., Overture, TomTom Traffic Stats, Microsoft MLRoads), or cartographic products of national mapping agencies, digital geospatial data on past states of transportation networks are scarce, and typically involve considerable manual labor (e.g., [2]). The evolution of local roads *within* settlements (i.e., cities, towns, and villages) can often be inferred from construction year information of buildings and properties, or neighborhood-level development information [3]. In contrast, the evolution of roads *connecting* settlements (e.g., highways, federal and regional roads) cannot be easily inferred by measurable proxy variables.

Hence, researchers have manually digitized such road network geometries from road atlases and similar resources [2], or developed automated methods to extract road network data from historical maps [4]. However, these efforts are typically regionally constraint. Thus, to our knowledge, no harmonized, multitemporal dataset on highway networks at global scale is available.

Herein, we describe a framework that aims to account for this shortcoming. Specifically, we integrate present-day highway network vector data from OpenStreetMap with multispectral Earth observation data from the Landsat archive (1990-2024). We first convert the historical Landsat data into probabilistic estimates of road presence. Then, we measure differences in road presence probability over time, and attribute these change signals to vector representations of the highway network. This way, contemporary highway network data can be enriched with an age estimate. The proposed framework is modular and extendible to other data sources, e.g., using different supervised or unsupervised statistical methods, incorporate other remote sensing data, or scanned and georeferenced historical maps.

The resulting, enriched road network data will enable the measurement of highway network growth, and aim to provide a data source for modelling accessibility and the interactions of connectivity and economic indicators over extended time periods. This includes the multi-temporal modelling of travel-time based commuting zones to better represent the dynamics of functional urban and rural areas [5,6], supported by the Global Human Settlement Layer (GHSL) project [7].

We test our method using manually digitized, as well as authoritative historical highway network data for Interstate highways covering the conterminous United States. While the experiment presented herein focuses on a relatively small area and uses Landsat data, the method is scalable to global level, and can be used with heterogeneous input data.

doi:10.2760/2119408

2. DATA

The choice of study area is mainly driven by the availability of reference data and the occurrence of change in the highway network. Herein, we focus on an area of approximately 720 km × 380 km covering the region around the Atlanta metropolitan area (United States), see Figure 1.

We use contemporary highway network data from OSM, reflecting the state in 2024. We filtered the OSM database to get only major roads such as interstates, national and state highways[1] using the Overpass API. Moreover, we acquired cloud-free, top-of-atmosphere reflectance from multispectral data from the Landsat 8 sensor (OLI) for 2024 ("contemporary" data), and from Landsat 4 and 5 (MSS, TM) for the year 1992 ("historical" data), using the USGS Landsat Collection 2, Level 1 data available via Amazon Web Services (s3://usgs-landsat/collection02/level-1).

We use two sources of reference data: (a) manually digitized highway network data [8]: These data have been curated from contemporary road network data overlaid with historical road atlases. Road segments not existing in a given year were manually deleted from the set of contemporary roads to produce historical depictions of the interstate, US highway, and state highway networks. (b) We use authoritative, multitemporal highway network data from the TIGER/Line dataset, for the earliest available epoch 1992 and for 2020 [9] (Figure 1).

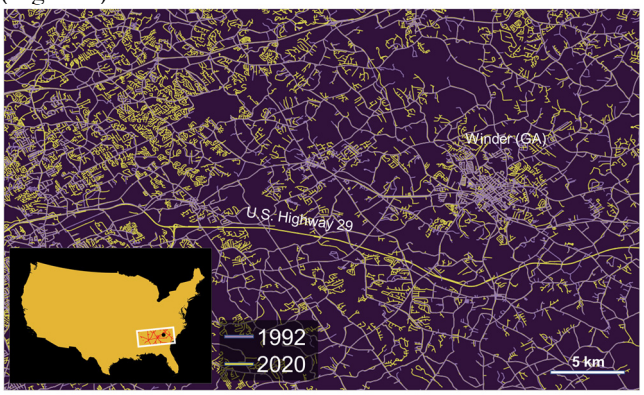


Fig. 1. Example of multi-temporal road network reference data, including changes in the highway network, shown for a subset of the Greater Atlanta study area (USA) near Winder, Georgia: Authoritative TIGER/Line roads in (a) 1992 and (b) 2020. The full study area (white box in inset map) covers 720 km × 380 km.

3. METHODS

3.1. Definition of the "highway" domain

The proposed framework assumes that the highway network grows, and does not shrink over time. Thus, we constrain our analysis to the contemporary highway network as obtained from OSM, buffered by a given distance (e.g., 200m) to account for positional uncertainty in the data involved in the analysis and potential resulting offsets. This domain contains samples of ***stationarity*** (i.e., road presence in contemporary and historical data) as well as samples of ***change*** (i.e., road presence in contemporary, but not in historical data). This makes the framework computationally efficient, as only Landsat data within this highway domain need to be considered.

3.2. Symbolic machine learning (SML)

For the areas within the defined "highway domain", we first create training labels, by rasterizing the contemporary OSM highway vector data in a spatial grid. This grid can be user-defined, or can be dictated by the Landsat scenes to be analyzed. Grid cells intersecting with the OSM highway lines (buffered by a distance reflecting approximate road width) are considered "positive" labels, i.e., containing relevant road signals. Grid cells within the highway domain, but not intersecting the buffered OSM highway lines, are considered "negative" labels, containing signals of non-highway objects (e.g., vegetation, buildings, other). Next, "contemporary" Landsat data (i.e., temporally close to the date of the OSM highway data) are used to learn the relationship between highway presence / absence and multispectral Landsat data. In our first tests, we use R,G,B, and Near Infrared bands only, but the flexible framework allows for derivation of further features (e.g., band ratios, directional or morphological features, or more complex representations).

For the experiment described herein, we use the Symbolic Machine Learning method (SML; [10]), a weakly supervised, frequentist classification method. SML has been proven successful for global extraction of information related to human settlements from multi-temporal stacks of multispectral remote sensing data, in the context of the GHSL project [7]. The SML method yields class-specific probability surfaces, and these probabilities are denoted as Φ_{POS} for road presence, and Φ_{NEG} for road absence. As the SML method has been proven to be highly invariant to label noise, the method can also be applied to *historical* Landsat scenes, despite potential discrepancies between contemporary labels and historical multispectral information due to occurred changes. Thus, for a given pair of contemporary and historical Landsat scenes, the SML method yields probability surfaces of road presence and absence in the contemporary and historical epochs, denoted as $\Phi_{POS\_CONTEMP}$, $\Phi_{NEG\_CONTEMP}$, Φ_{POS_HIST}, and Φ_{NEG_HIST}.

[1] We used the following query: tag = 'highway' and values IN ('motorway', 'motorway_link', 'primary', 'primary_link', 'secondary', 'secondary_link', 'trunk', 'trunk_link')

doi:10.2760/2119408

3.3. Vector-raster data integration

In order to link the raster-based road presence and absence probability surfaces with the highway network vector data, we developed an efficient vector-raster integration approach (Figure 2). This method is based on the contemporary OSM highway network data, which is typically very detailed, including individual lanes, ramps, frontage roads, etc. (Figure 2a). We first buffer these polylines, using a defined buffer distance, and create the union of all buffered surfaces (Figure 2b). We then create the center line of the resulting polygons using a skeletonization method, representing a generalized depiction of the highway network (Figure 3c). The skeleton line is then segmented into chunks of equal length (as defined by the user; Figure 3d), and these segments are then buffered by a given distance (Figure 3e). The resulting squares are called buffered road segments (BRS) and represent the analytical unit for any subsequent analyses. We then calculate zonal statistics such as mean and median road presence probability for each BRS based on the Landsat-derived probability surfaces for contemporary and historical epochs (Figure 2f,g). For evaluation purposes, the reference labels are also attached to the BRS based on intersection with reference road network data (Figure 2h).

3.4. Calculation of change indicators and evaluation

We calculate and test highway network change indicators, including the absolute difference of probabilities Δ_Φ between historical and contemporary epoch (Φ_H and Φ_C, respectively), and the normalized difference index $NDI = (\Phi_C - \Phi_H) / (\Phi_C + \Phi_H)$ [11]. Moreover, differences can be calculated pixel-wise *before* calculating zonal

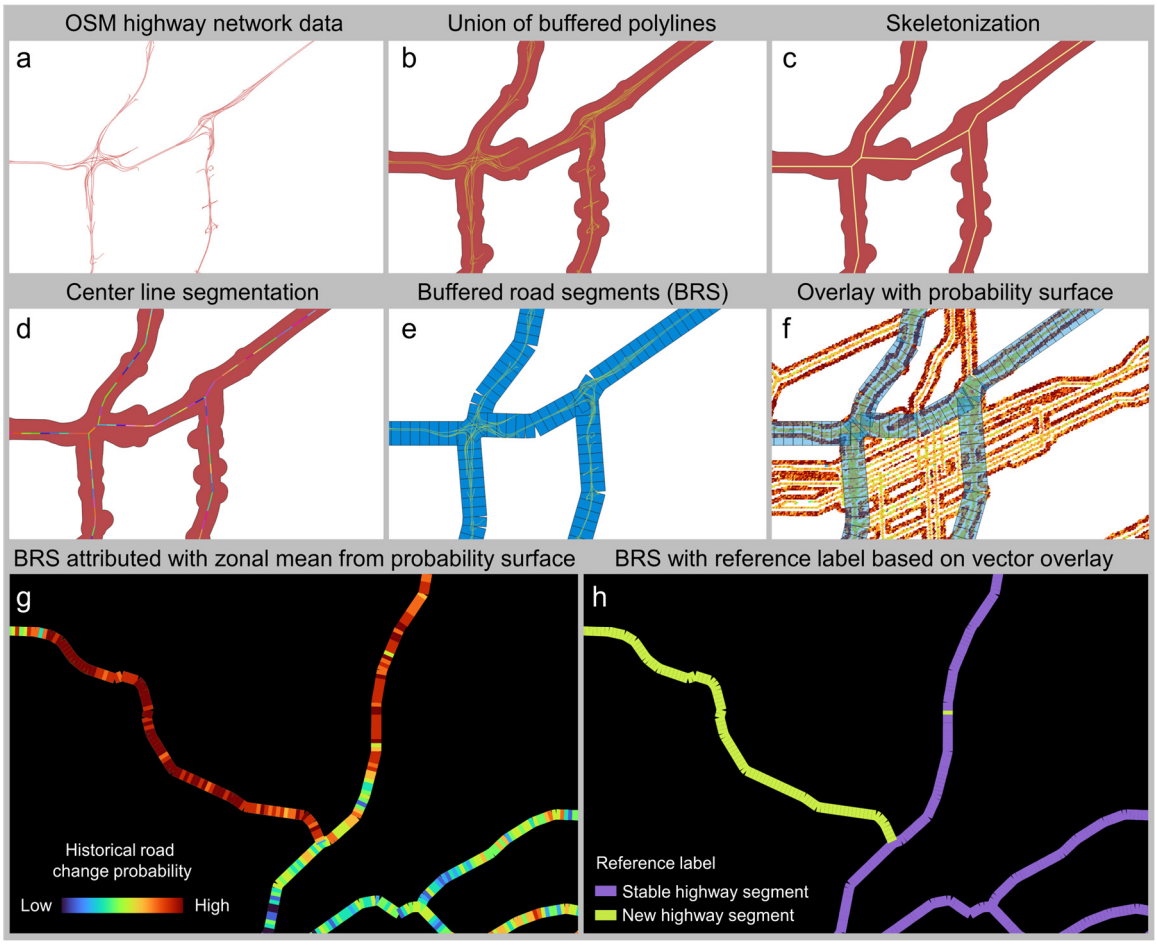


Fig. 2. Illustrating the vector-raster data integration concept for highway network change detection. (a) Contemporary OSM highway data, (b) union of buffer areas (distance to (a) = 100m), (c) skeleton line (yellow) of the union of buffered lines, (d) segmented skeleton line (i.e., separated into equal-length chunks of 100m, (e) buffered road segments (BRS) based on the skeleton line segments (buffer distance 150m), overlaid with (a); (f) BRS overlaid with a road absence probability raster surface derived from Landsat data in proximity of the contemporary major road network. Panel (g) shows the historical road change probability after conducting zonal statistics (i.e., zonal mean) for each BRS. The BRS constitute an abstract representation of the contemporary highway network. Panel (h) shows the reference label for each BRS obtained from spatial overlay of the BRS with the reference data.

doi:10.2760/2119408

statistics, or *after*, by subtracting the BRS-level zonal aggregates. In the pilot experiment presented herein, we use Receiver Operating Characteristic (ROC) analysis (i.e., the area-under-the-curve - AUC metric) to test the responsiveness of these change indicators to the changes observed in the reference data. The data processing pipeline and experiments are implemented in Python, using scikit-learn, numpy, rasterio, geopandas, rasterstats, pygeoops, and shapely python packages.

4. RESULTS

Based on the (continuous) change indicators and the binary change / no change reference labels at the BRS level (e.g., Figure 2g,h) we observe considerably high AUC values, for all tested change indicators, indicating that Landsat data in combination with the SML-based probability surfaces and derived change metrics are responsive to the change reported in the reference data (Table 1). Notably, there are differences between the two reference datasets used. For the manually digitized reference data AUC values are all >0.85, while for the TIGER/Line comparison, AUC values range between 0.65 and 0.77. This is possibly due to discrepancies in the definition of highways between TIGER/Line and OSM data, or potentially due to effects of road reclassification not captured by the Landsat signals. Moreover, higher levels of positional uncertainty in the 1992 TIGER/Line data could contribute to the observed lower levels of AUC values.

5. CONCLUSIONS AND OUTLOOK

Herein, we described a framework to detect changes in highway networks over extended time periods (1990 to 2024). First results show promising performance. Importantly, these results are specific to the analysed study area, and may not be generalizable to other regions where land cover and spectral signatures of roads are different.

Thus, in future work, we will test the method at national and global scale. The geographic non-stationarity of road presence/absence signatures in multi-temporal, multispectral imagery will likely require geographically adaptive models. We will also test more complex, Landsat-derived spectral and morphological features. We are currently experimenting with cloud-based remote sensing data and processing infrastructure to scale the method up. The use of spatio-temporal asset catalogues (STAC) will facilitate the use of annual composites, to reduce the effects of cloud cover, and cloud-based processing will increase the efficiency of the method. We will incorporate refined change indicators, as well as earlier Landsat data (1975-1990) and historical maps or historical aerial imagery to further extend the temporal coverage of highway network changes measured within this framework.

REFERENCES

[1] T. Jaworski, and C.T. Kitchens, "National policy for regional development: Historical evidence from Appalachian highways", Review of Economics and Statistics, 101(5), 777-790, 2019.

[2] A.M. Condeço Melhorado, P. Christidis, L. Dijkstra, "Replication Data for: Travel speed changes along the European core road network for the period 1960–2030: an application of octilinear cartograms". European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-tem-10005, 2018.

[3] K. Burghardt, J.H. Uhl, K. Lerman, and S. Leyk, "Road network evolution in the urban and rural United States since 1900", Computers, Environment and Urban Systems, 95, 101803, 2022.

[4] C. Avcı, E. Sertel, and M.E. Kabadayı, "Deep learning-based road extraction from historical maps", IEEE Geoscience and Remote Sensing Letters, 19, 1-5, 2022.

[5] A.I. Moreno-Monroy, M. Schiavina, and P. Veneri, "Metropolitan areas in the world. Delineation and population trends", Journal of Urban Economics, 125, 103242, 2021.

[6] L. Dijkstra, and C. Jacobs-Crisioni, "Defining Functional Rural Areas", Publications Office of the European Union, Luxembourg, JRC138561, doi:10.2760/702937, 2024.

[7] M. Pesaresi, M. Schiavina, P. Politis, S. Freire, K. Krasnodębska, K., J.H. Uhl, et al., "Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data", International Journal of Digital Earth, 17(1), 2024.

[8] T. Jaworski, C.T. Kitchens, and S. Nigai, "Highways and globalization", International Economic Review, 64(4), 1615-1648, 2023.

[9] U.S. Census Bureau "TIGER/Line Shapefiles", available online: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. Last Access: 02 May 2025.

[10] M. Pesaresi, V. Syrris, and A. Julea, "A new method for earth observation data analytics based on symbolic machine learning", Remote Sensing, 8(5), 399, 2016.

[11] L. Zou, and B. Wei "GND-RI: A Normalized Difference Form More Suitable for Remote Sensing Applications", IEEE Transactions on Geoscience and Remote Sensing, 61, 1-12, 2023.

Table 1. ROC analysis results (i.e., area-under-the-curve; AUC) of highway network change indicators using manually digitized reference data (1990) and TIGER/Line road network data from 1992.

Zonal statistics of cell-level differences	AUC Digitized ref. data 1990	AUC TIGER/ Line 1992	Differences of zonal statistics	AUC Digitized ref. data 1990	AUC TIGER/ Line 1992	Normalized differences of zonal statistics	AUC Digitized ref. data 1990	AUC TIGER/ Line 1992
Mean($\Delta\Phi_{POS}$)	0.872	0.662	Δ(Mean(Φ_{POS}))	0.872	0.662	NDI(Mean(Φ_{POS}))	0.852	0.651
Median($\Delta\Phi_{POS}$)	0.875	0.672	Δ(Median(Φ_{POS}))	0.873	0.673	NDI(Median(Φ_{POS}))	0.856	0.684
Mean($\Delta\Phi_{NEG}$)	0.880	0.770	Δ(Mean(Φ_{NEG}))	0.880	0.770	NDI(Mean(Φ_{NEG}))	0.874	0.770
Median($\Delta\Phi_{NEG}$)	0.876	0.764	Δ(Median(Φ_{NEG}))	0.879	0.761	NDI(Median(Φ_{NEG}))	0.871	0.762

LIGHTWEIGHT CNNS FOR EMBEDDED SAR SHIP TARGET DETECTION AND CLASSIFICATION

Fabian Kresse[1,3], Georgios Pilikos[1,4], Mario Azcueta[2], Nicolas Floury[1]

[1]Wave Interaction & Propagation, RF Payloads & Tech. Div., Elec. Dept., ESA/ESTEC, NL-2200AG Noordwijk, The Netherlands
[2]Copernicus Sentinel-1 Payload, Copernicus Space Component, ESA/ESTEC, NL-2200AG Noordwijk, The Netherlands
[3]Institute of Science and Technology Austria (ISTA), AT-3400 Klosterneuburg, Austria
[4]Department of Neuroscience, Erasmus MC, NL-3000CA Rotterdam, The Netherlands

ABSTRACT

Synthetic Aperture Radar (SAR) data enables large-scale surveillance of maritime vessels. However, near-real-time monitoring is currently constrained by the need to downlink all raw data, perform image focusing, and subsequently analyze it on the ground. On-board processing to generate higher-level products could reduce the data volume that needs to be downlinked, alleviating bandwidth constraints and minimizing latency. However, traditional image focusing and processing algorithms face challenges due to the satellite's limited memory, processing power, and computational resources. This work proposes and evaluates neural networks designed for real-time inference on unfocused SAR data acquired in Stripmap and Interferometric Wide (IW) modes captured with Sentinel-1. Our results demonstrate the feasibility of using one of our models for on-board processing and deployment on an FPGA. Additionally, by investigating a binary classification task between ships and windmills, we demonstrate that target classification is possible.

Index Terms— Deep Learning, raw echo data, ship detection, synthetic aperture radar (SAR), Field-Programmable-Gate Array (FPGA)

1. INTRODUCTION

Synthetic Aperture Radar (SAR) satellite data enables all-weather maritime monitoring. Traditional, on-ground Constant False Alarm Rate (CFAR) [2] detection on focused SAR images entails a costly downlink, focus, analyze pipeline, incurring latency and limited contact windows. Processing the data directly on the satellite significantly reduces the volume of data that needs to be downlinked by generating higher abstraction level outputs (e.g., pixel coordinates of detected ships) instead of raw data. Deep learning models have shown promising results in this context [11, 12, 13, 4, 17], offering the opportunity to optimize for inference on embedded devices straightforwardly. Yet, obtaining fully focused SAR images on satellite is computationally and memory intensive, prompting research into onboard ship detection using the intermediate raw [3, 10] or range-compressed data products [16, 9, 8]. Overall, an effective onboard SAR ship detection algorithm must process small data segments to accommodate limited on-board memory, maintain a compact model suitable for Field-Programmable-gate Arrays (FPGAs), directly output ship coordinates to reduce downlink and storage requirements, and still achieve high accuracy. While prior work addresses individual elements of this pipeline, none achieve real-time, accurate detection on unfocused or range-compressed data under embedded FPGA constraints.

In this work, we propose a range of lightweight deep-learning models designed for Stripmap and Interferometric Wide (IW) SAR data obtained with Sentinel-1, addressing the requirements for on-board data processing. Our model configurations, processing raw and range-compressed data, are one-stage detectors built on ResNet blocks [5], allowing for flexible model sizing. They employ a single-stage detection layer as the final step, which performs coordinate predictions and target classification following the grid-based approach of the YOLO architecture [14]. Our smallest model operating on Stripmap mode data delivers near-perfect ship detection results in our evaluation scenes. Additionally, we demonstrate that this model can be successfully deployed on a Xilinx Zynq UltraScale+ MPSoC ZCU104 FPGA, achieving sufficient throughput for real-time processing. For IW data, we evaluate the performance of multiple model sizes, achieving competitive results in open water scenes and offering valuable insights for future research. We also show that our model can perform binary target classification, distinguishing between windmills and ships in the IW dataset we employ.

2. PROPOSED DEEP LEARNING ALGORITHM

Fig. 1 shows the processing pipeline for both the traditional method, including range and azimuth compression, and the approach adopted by us. Similar to previous work [3], for Stripmap mode, we operate directly on raw data, while for IW data, we perform range-compression.

Stripmap Preprocessing: We shift the raw data by half

F. Kresse and G. Pilikos performed this work while at ESA and are no longer affiliated with the agency.
Correspondence: Nicolas.Floury@esa.int

doi:10.2760/2119408

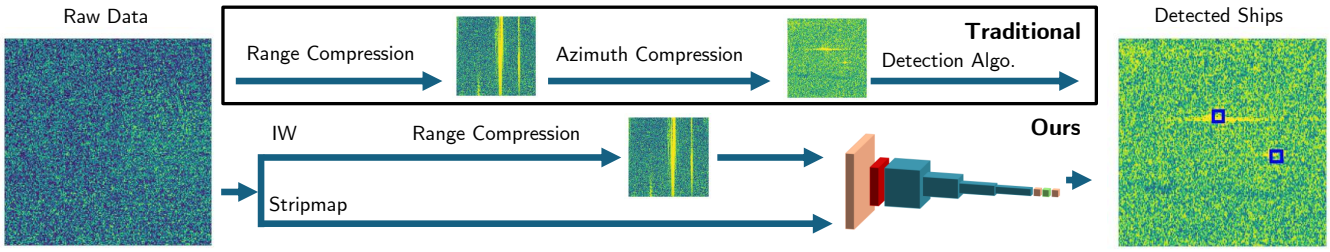


Fig. 1. The traditional pipeline (black box) focuses data before detection. Our approach bypasses this: we detect directly on raw Stripmap echoes [3] and on range-compressed IW data.

the chirp length in the range dimension so that the feature response aligns with the Single-Look Complex (SLC) labels, accounting for mode- and chirp-specific acquisition offsets.

IW Preprocessing: Range-compressed IW features and SLC labels are misaligned, so we apply a locally consistent mapping between their pixel spaces. We compute the offset from the center pixel of each range-compressed crop and use it to shift the corresponding label crop.

Model: We employ a YOLO-style architecture for its efficient single-stage design, enabling fast inference without the overhead of region proposals [14]. The backbone consists of four layers of ResNet blocks, preceded by a 7×7 convolution with 64 kernels. As in previous work, we treat the complex-valued SAR data as two separate input channels [3]. We apply the network to crops of the original SAR image. The network outputs predictions on a YOLO-style grid, where each grid cell predicts the coordinates (x, y) and an objectness score indicating the presence of a target. Since ships in our datasets occupy a narrow range of sizes, we omit multi-scale anchor-based detection typical in YOLO. We also perform binary classification between ships and windmills for IW data, adding two additional outputs. Our output grid has size $S \times S$, with each cell corresponding to a 32×32 pixel region in the input data. For example, an input crop size of 128×128 results in an output grid of size 4×4. The model configurations evaluated in our experiments are listed in Table 1.

Table 1. Each Resnet-block contains two convolutions with kernels of size 3×3, Batchnorm and ReLU activation.

Param. (Size)	Blocks per Layer	Channels
96800 (S)	1, 2, 2, 1	16, 16, 32, 32
1222368 (M)	3, 4, 6, 3	64, 64, 64, 64
11222880 (L)	2, 2, 2, 2	64, 128, 256, 512

Loss Function: We adopt the standard YOLO loss, omitting only the bounding-box size regression term while retaining all other components unchanged.

Prediction of Ship Locations: After obtaining predictions, we assign a fixed 50-pixel bounding box to each detected ship. As in the YOLO pipeline [14], we apply non-maximum suppression (NMS) to remove overlapping boxes, keeping the one with the highest confidence. The acceptance threshold is set as the lower of the Youden J threshold [15] and the minimum distance threshold from validation data, rounded down to two decimals. We compute a distance-based $F1$ score $F1_{30}$ counting a prediction as correct if it lies within 30 pixels of a ground-truth label; unmatched predictions and labels are false positives and negatives, respectively. Given SAR resolutions (5 m × 5 m for Stripmap, 5 m × 20 m for IW), this tolerance equals 150 m in range and up to 600 m in azimuth, accommodating minor localization errors while remaining well below the fixed box size.

3. EXPERIMENTS

We evaluate our proposed model using two different datasets. The first dataset consists of raw SAR Sentinel-1 Stripmap mode (S6) VV polarization data, as previously utilized by [3]. We show that the model trained on this dataset can be deployed on an FPGA, achieving real-time throughput. We then investigate an IW dataset from the Shanghai port. For this dataset, we show good off-shore detection performance and the ability of our model to distinguish between windmills and ships. Both our datasets where originally obtained from the Copernicus Data Space Ecosystem [1]. All experiments with standard deviations given were performed with three seeds.

3.1. Sentinel-1 Stripmap Mode

The dataset from [3], consist of a total of 12 SAR images from the São Paulo port, with one image used for validation (84 ships), two for testing (155 ships), and the remaining for training (726 ships). We investigate the performance of the smallest model outlined in Section 2 with a crop size of 128×128, achieving an $F1_{30}$ score of 0.98 ± 0.00 and 0.97 ± 0.01 on our first and second test image (see Fig. 2 for qualitative results). After manually inspecting the few erroneous predictions, these can be attributed to double detections of ships, the NMS not being aggressive enough, and label ambiguities.

doi:10.2760/2119408

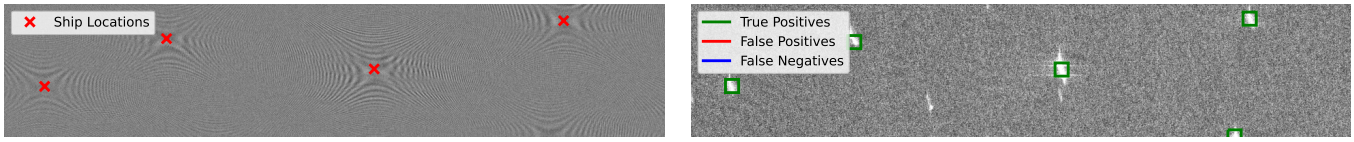


Fig. 2. Left: real part of raw Stripmap SAR data, shifted by half-chirp to align SLC labels; Right: manually labeled SLC image overlaid with the network's post-processed detections (contains modified Copernicus Sentinel Data [1]).

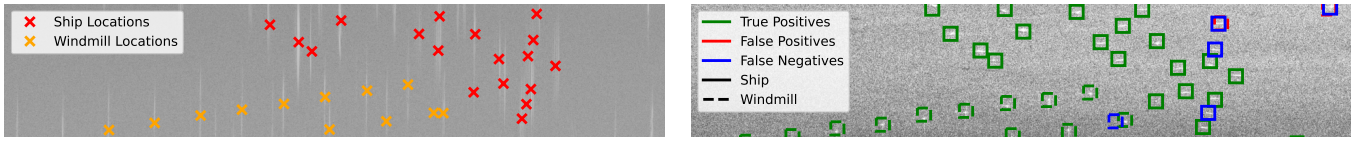


Fig. 3. Left: range-compressed IW SAR crop (input to the network) with center-based offset correction (Sec. 2); Right: post-processed detections overlaid on the labeled SLC image for the (M) model (contains modified Copernicus Sentinel Data [1]).

3.2. Embedded Inference for Sentinel-1 Stripmap Mode

We deployed our 8-bit AdaQuant-quantized [6] model (via Vitis AI 3.0 [7]) on a Zynq UltraScale+ MPSoC ZCU104 FPGA with no accuracy loss. Real-time operation requires $\geq$ 2027 FPS (PRF = 1664 lines/s $\times$ 19950 samples/line); DPU inference (excluding preprocessing & NMS), achieved 3527 ± 23 FPS with four CPU threads, exceeding the target.

Table 2. Comparison of Models: Results on Interferometric Wide data for off-shore ships on our test image (Range Compressed Input Data). Input denotes the crop size.

Size	Input	$F1_{30}$	Ship $F1_{30}$	Wind. $F1_{30}$
L	256	0.87 ± 0.01	0.77 ± 0.01	0.71 ± 0.04
M	256	0.87 ± 0.05	0.78 ± 0.05	0.78 ± 0.02
S	256	0.79 ± 0.01	0.56 ± 0.07	0.54 ± 0.06
S	128	0.72 ± 0.04	0.52 ± 0.02	0.51 ± 0.03

3.3. Sentinel-1 Interferometric Wide Mode

Our hand-labeled IW dataset from the Shanghai port comprises 10 bursts of size 20760 $\times$ 1617 at 5 $\times$ 20 m resolution. We split 8/1/1 bursts for training/validation/test sets. The training set contains 1163 ships and 460 windmills; after excluding near-shore ships, the test set contains 66 ships and 19 windmills. Preliminary experiments with ships located close to shore showed significant performance degradation, possibly due to the complex and ambiguous backscatter in these areas. Therefore, we exclude them from our final evaluation. The IW data presents additional challenges due to the continuous antenna pattern steering in azimuth during the acquisition and the more complex nature of the scenes investigated. Initial attempts using our models with small input sizes on raw IW data did not yield adequate performance. As a result, we utilized range-compressed data.

Table 2 reports detection metrics and class-wise $F1_{30}$ (see Fig. 3 for quantitative results); class scores exclude the other class's labels to isolate per-class performance. Our smallest model already performs well, and increasing crop size and parameter count further boosts $F1_{30}$, though gains plateau—likely due to overfitting. After manual inspection, a large number of remaining errors arise from closely spaced ships, due to NMS limitations, and ambiguous labels.

4. DISCUSSION

We conduct experiments on raw Stripmap and range-compressed IW data. For Stripmap, excellent results are achieved by shifting the raw input by half the chirp size, bypassing the computationally expensive image-focusing step and enabling direct predictions on small crops of raw SAR data. The model remains compact in both parameters and forward-pass complexity, and we deploy it on a Zynq UltraScale+ MPSoC ZCU104 FPGA, demonstrating suitability for real-time embedded processing. However, the Stripmap scenes used here are relatively simple with similar backscatter, so future work should test robustness under more complex sea conditions and diverse environments.

In our evaluation of IW data, we find that raw data alone, in contrast to Stripmap data, does not yield satisfactory results. As a result, we perform range-compression, resulting in improved performance. We attribute this improvement to the target energy being more concentrated and, hence being easier to identify with the small Field-of-View of our neural network. By increasing both the input crop size and the network complexity compared to our Stripmap model, we achieve good performance on IW data. Additionally, we demonstrate successful target classification, performing binary classification between ships and windmills.

doi:10.2760/2119408

A main limitation of our method is that it struggles to detect ships very close to shore, so we exclude these cases from the final evaluation. This low performance is likely due to strong backscatter from surrounding structures and vessels, and may be mitigated with more diverse training data.

Both the limitations observed in Stripmap and IW data can be largely attributed to the availability and quality of the datasets. A key challenge is the lack of sufficiently large and diverse labeled raw SAR datasets, particularly with varying sea conditions. The upcoming Sentinel-1 satellites, equipped with AIS tracking antennas, holds promise for addressing this gap by potentially enabling automated labeling through AIS data correlation. Additionally, since numerous algorithms already exist for processing SLC images, future work could leverage these methods to generate large-scale datasets by aligning SLC-based detections with raw SAR data, further expanding the pool of labeled data for training and evaluation.

5. CONTRIBUTIONS

We demonstrated the feasibility of real-time ship detection using deep-learning models applied to raw Stripmap data. Furthermore, we deploy our Stripmap model on a Zynq UltraScale+ MPSoC ZCU104 FPGA, demonstrating its practical use for real-time, onboard processing in resource-constrained environments. For IW data, we show that using range-compressed data and larger input sizes improves detection and classification performance, enabling binary classification of ships and windmills.

Acknowledgment. The authors thank Sanath Muret, Max Ghiglione and Maris Tali for FPGA support and discussions.

REFERENCES

[1] Copernicus Data Space Ecosystem. Copernicus data space ecosystem, 2024.

[2] David J Crisp. The state-of-the-art in ship detection in synthetic aperture radar imagery. 2004.

[3] Kevin De Sousa, Georgios Pilikos, Mario Azcueta, and Nicolas Floury. Ship detection from raw SAR echoes using convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[4] Max Ghiglione, Ernesto Imbemomba, Salvatore D'Addio, Georgios Pilikos, Nicolas Floury, Marc Zimmermanns, Luis Mansilla Garcia, Nicolas Longépé, Roberto Del Prete, Stefan Graham, Mario Azcueta, and Martin Suess. Technology developments and activities at the European Space Agency for cognitive synthetic aperture radar payloads. In *International Astronautical Congress*, 2024.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv:2006.10518*, 2020.

[7] Xilinx Inc. *Vitis AI: AI Inference Development Platform*, 2023.

[8] Sushil Kumar Joshi, Stefan V Baumgartner, Andre BC da Silva, and Gerhard Krieger. Range-Doppler based CFAR ship detection with automatic training data selection. *Remote Sensing*, 2019.

[9] Xiangguang Leng, Jin Wang, Kefeng Ji, and Gangyao Kuang. Ship detection in range-compressed SAR data. In *IEEE International Geoscience and Remote Sensing Symposium*, 2022.

[10] Xiangguang Leng, Kefeng Ji, and Gangyao Kuang. Ship detection from raw SAR echo data. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[11] Jianwei Li, Jie Chen, Pu Cheng, Zhentao Yu, Lu Yu, and Cheng Chi. A survey on deep-learning-based real-time SAR ship detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[12] Georgios Pilikos, Mario Azcueta, Roberto Camarero, and Nicolas Floury. Raw data compression for synthetic aperture radar using deep learning. In *Int. Workshop On-Board Payload Data Compression*, 2022.

[13] Georgios Pilikos, Mario Azcueta, Roberto Camarero, and Nicolas Floury. Raw SAR data compression with deep learning. In *IEEE International Geoscience and Remote Sensing Symposium*, 2024.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE conference on computer vision and pattern recognition*, 2016.

[15] William J Youden. Index for rating diagnostic tests. *Cancer*, 1950.

[16] HongCheng Zeng, YuTong Song, Wei Yang, Tian Miao, Wei Liu, WeiJie Wang, and Jie Chen. An Incept-TextCNN model for ship target detection in SAR range-compressed domain. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[17] Tianwen Zhang and Xiaoling Zhang. ShipDeNet-20: An only 20 convolution layers and $<$ 1-MB lightweight SAR ship detector. *IEEE Geoscience and Remote Sensing Letters*, 2020.

doi:10.2760/2119408

DEVELOPING A CUBESAT FOR EDUCATIONAL PURPOSES

Viktors Gutakovskis, Aleksandrs Okunevs, Atis Vanags, Daniils Marcenko

Riga Technical University, Riga Technical College

ABSTRACT

Developing CubeSat for educational purposes is an excellent way to engage students in hands-on learning about space, engineering, and science. It offers experience across various fields, including physics, computer science, mechanical engineering, and electronics.

Educational goals define the specific learning outcomes you aim to achieve for students involved in the project. Key educational goals could include Technical and Engineering Skills:

Understanding Satellite Design: Teach students how satellites are designed and how systems integrate into a small form factor like a CubeSat. This includes knowledge of structural design, subsystem integration, and component selection.

Subsystem Design and Function:

Provide experience with designing or working with power, communication, attitude control, and payload subsystems. Students can learn the specifics of each system, such as how solar panels provide power or how the attitude control system stabilizes the satellite.

Programming and Software Development: Involve students in programming the onboard computer (OBC) to perform data acquisition, communication protocols, and mission control functions. This also includes developing ground station software for data retrieval and command sending.

Scientific and Analytical Skills:

Data Analysis and Interpretation: Teach students how to analyze and interpret data from sensors on the CubeSat, including any scientific or Earth observation data. For example, students could analyze temperature variations in orbit or assess radiation levels at different altitudes.

Research Skills: Introduce students to the scientific method by having them define hypotheses, design experiments (for experimental missions), collect data, and analyze results.

System Engineering and Integration: Educate students on how complex systems are integrated into a single mission, covering concepts like interface requirements, modularity, and systems testing. In this publication, the real implementation of the 1U CubeSat is represented. It was developed in 3 months, in joint team from Riga Technical College (4 students) and Riga Technical University (2 students) and 2 mentors. The result was represented on the vocational education competition in the ZRKAC (Jelgava) by MASOC on 11.04.2025.

1. INTRODUCTION

Educational goals define the specific learning outcomes you aim to achieve for students involved in the project. These goals are based on the mission's complexity, the students' skill level, and the depth of knowledge the project can provide. Key educational goals could include:

Technical and Engineering Skills:

Understanding Satellite Design: Teach students how satellites are designed and how systems integrate into a small form factor like a CubeSat. This includes knowledge of structural design, subsystem integration, and component selection.

Subsystem Design and Function: Provide experience with designing or working with power, communication, attitude control, and payload subsystems. Students can learn the specifics of each system, such as how solar panels provide power or how the attitude control system stabilizes the satellite.

Programming and Software Development: Involve students in programming the onboard computer (OBC) to perform data acquisition, communication protocols, and mission control functions. This also includes developing ground station software for data retrieval and command sending.

Scientific and Analytical Skills:

Data Analysis and Interpretation: Teach students how to analyze and interpret data from sensors on the CubeSat, including any scientific or Earth observation data. For example, students could analyze temperature variations in orbit or assess radiation levels at different altitudes.

Research Skills: Introduce students to the scientific method by having them define hypotheses, design experiments (for experimental missions), collect data, and analyze results.

Project Management and Teamwork:

Project Planning: Guide students through the stages of project planning, budgeting, and scheduling. This includes understanding timelines, deliverables, and deadlines, which are essential in real-world engineering projects.

System Engineering and Integration: Educate students on how complex systems are integrated into a single mission, covering concepts like interface requirements, modularity, and systems testing.

Collaboration and Communication: Encourage teamwork and collaboration across different roles (e.g.,

doi:10.2760/2119408

engineering, software development, testing), as well as communication with external stakeholders like sponsors, school officials, or industry advisors.

Hands-On Practical Skills, which were in this work:

Construction and Assembly: Allowed students to participate in assembling and testing the CubeSat, teaching theoretical and practical content about handling sensitive electronic and mechanical assembly.

Testing and Troubleshooting: Gave students experience with testing protocols, including vibration, thermal, and vacuum tests. This teaches them about the importance of testing in space missions and how to troubleshoot issues that arise.

Operation and Maintenance: Teach students how to operate and monitor the CubeSat once deployed in desktop mode, including interpreting telemetry data and responding to potential issues in orbit/ or in practice on the table.

STEM Engagement and Outreach:

Inspire Interest in Space and Engineering: using the CubeSat as a vehicle to engage broader student interest in STEM fields.

Documenting and Sharing Findings: students are encouraged to document the project and share their findings through presentations, reports, or outreach events. This inspired other students and schools to pursue similar educational projects.

Collaborations with Other Schools or Institutions: Partner with other schools or universities are impressed by project's impact and allow students to engage in a collaborative, multi-institutional project.

2. EXAMPLE OF EDUCATIONAL MISSION OBJECTIVES AND GOALS

Mission Objective:

Deploy a 1U CubeSat as a desktop model to collect temperature, humidity, GPS location and images for analysis.

Educational Goals:

1.Teach students the fundamentals of satellite systems, including power, communication, and payload subsystems.

2. Develop skills in programming the onboard computer to collect and transmit data.

3. Engage students in data analysis and presentation of findings, providing hands-on experience with real-world satellite data.

4. Inspire students to pursue careers in aerospace, science, and engineering through public outreach and project showcases.

3. DEVELOPING A 1U CUBESAT USING ARDUINO AND RASPBERY PI

In this work, a 1U CubeSat educational model was created. The frame was created using moder CAD, 3D printed, and CNC milled on the Hardford CNC mill. The internal structure could be seen on the figure 1.

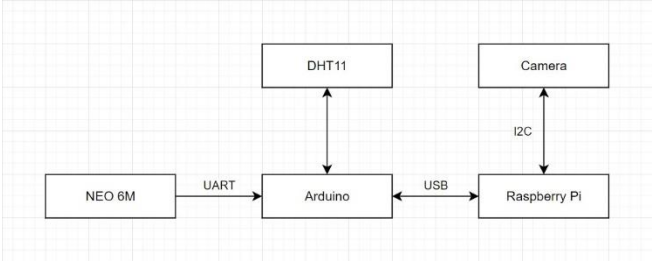


Fig. 1. CubeSat structure design graph

The Arduino, Raspberry Pi and sensors connection process can be seen on figure 2 and 3. Figure 4 represents the result model of the developed CubeSat. Figures 5 and 6 represent the result presentation of the CubeSat educational model in local competitions and conferences.

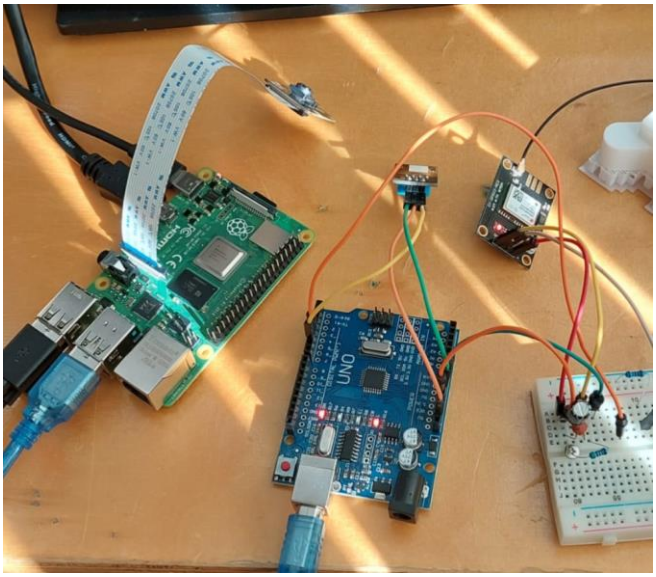


Fig. 2. Basic electronics connections and testing

doi:10.2760/2119408

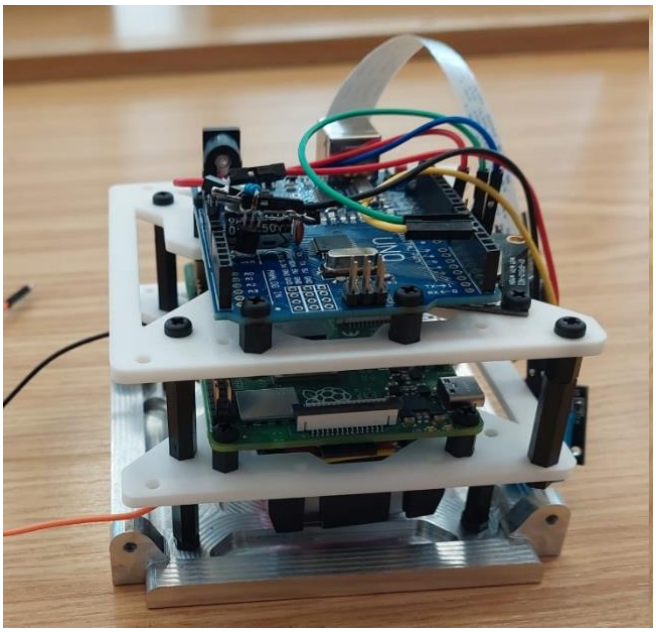


Fig. 3. Assembly process of the electronic connection process.

Fig. 4. The result model of the developed CubeSat.

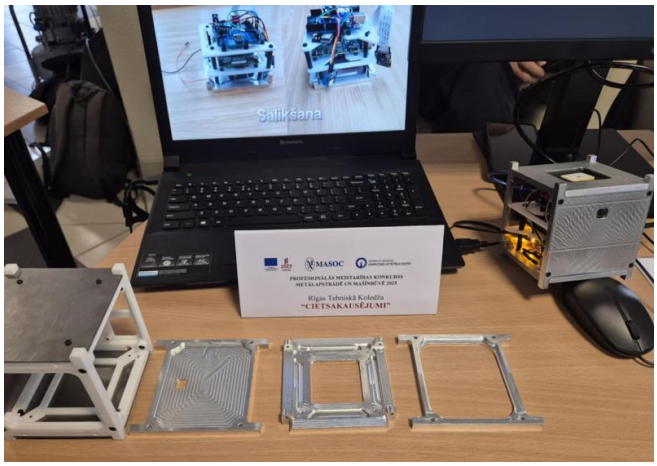


Fig. 5. 3D printed prototype, CNC milled parts and ready-to-use for educational process CubeSat model (represented in the vocational skills competition, ZRKAC, April 2025, Latvia, Jelgava).

Fig. 6. Visual representation of CubeSat possibilities (represented on 15.-16.05.2025., DEEP TECH ATTELIER 2025, Latvia, Riga).

4. CONCLUSIONS

1. Establishing clear mission objectives and educational goals is critical to creating an impactful educational CubeSat project. This gave us an opportunity to design, produce and to program the 1U Cubesat.

doi:10.2760/2119408

2. The mission objectives focus on achievable, educationally rich space activities, while the educational goals align with students' learning needs. When combined, these objectives create a practical, exciting, and informative experience that can have lasting benefits on students' understanding of space technology and inspire their future careers.
3. Creating a CubeSat for educational purposes involves assembling a team with a variety of skills and securing funding to support the project. Below are the steps and considerations for forming an effective team and securing necessary funding.
4. 1U educational CubeSat was created, represented in this work. The process took a lot of time from developing CAD frame files and Mastercam CNC design. Arduino and Raspberry Pi asked for skillful students work. In total the result is very impressive.

REFERENCES

[1] Programming for the Raspberry Pi with Arduino IDE, https://forum.arduino.cc/t/programming-for-the-raspberry-pi-with-arduino-ide/884634

[2] Raspberry Pi Arduino Serial Communication – Everything You Need To Know, https://roboticsbackend.com/raspberry-pi-arduino-serial-communication/

[3] "Internet of Things Projects with Arduino and Raspberry Pi" by Rajesh Singh. 2020.,208.p.

doi:10.2760/2119408

A COMPARISON OF THE ATLAS OF HUMAN SETTLEMENTS AGAINST THE GLOBAL HUMAN SETTLEMENT LAYER

Georgios K. Ouzounis, Andrew A. Vekinis

Atlas AI
AI Research
Palo Alto, California, USA

ABSTRACT

The article presents the findings of a quality evaluation exercise between the most recent releases of two planetary-wide built-up basemaps; the Atlas of Human Settlements or AHS of Atlas AI and the Global Human Settlement Layer or GHSL of the Joint Research Center of the European Commission. The exercise was split in two parts; a qualitative and a quantitative analysis, both powered by data collected over 8 regions across the world. The comparison yields an overwhelming lead of the AHS over the GHSL.

***Index Terms*—** ahs, ghsl, built-up, regression, evaluation

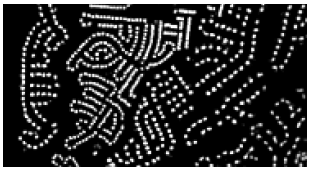
1. INTRODUCTION

The Atlas of Human Settlements or AHS [10] is a built-up basemap of global extent, delivered at a nominal spatial resolution of 10m and updated annually, with historical records dating back to 2016. The primary data layer, referred to as the Built-up Index or **BuI**, reports the percentage of built surface within a spatial unit of 100 square meters in size. The equivalent term in GHSL [3, 8] is called the Built-up Fraction or **BUFRAC** - Fig. 1. The BuI layer is generated using a state-of-the-art deep-learning model implemented on a U-Net-like, multi-scale convolutional attention encoder to transformer-decoder architecture using multi-scale attention. Further to the BuI, the AHS delivers the Built-up Confidence or **BuC**; a raster image that reports the confidence of the regressor in producing the BuI value for each input pixel.

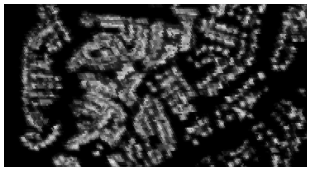
The AHS is generated from two different models, one that reports built-up (BuI) in the developed world, and a second one tailored for built-up in the developing countries, capturing residential buildings and structures in places where housing and living conditions are poor. It uses a set of model weights that shift the focus on smaller, more dense and radiometrically more diverse patterns of built-up. This analysis is focused on the former model due to limited reference data availability.

Reporting the built-up systematically, and in a globally consistent manner allows for accurate change detection, free of parallax-related artifacts, which in turn delivers actionable data on the growth, regression, or stagnation patterns witnessed in various human settlements across the world. This wide area monitoring (WAM) service [1] is geography, scale, climate, living standards and prior data agnostic and powers a wide range of applications in key industry verticals. Fig. 2 shows an example of the state of built-up in the South-East Corridor, Georgia, USA in early 2024 and built-up change detection between the years 2021 and 2023.

(a) Google Maps basemap view of a residential neighborhood



(b) The AHS-BuI layer

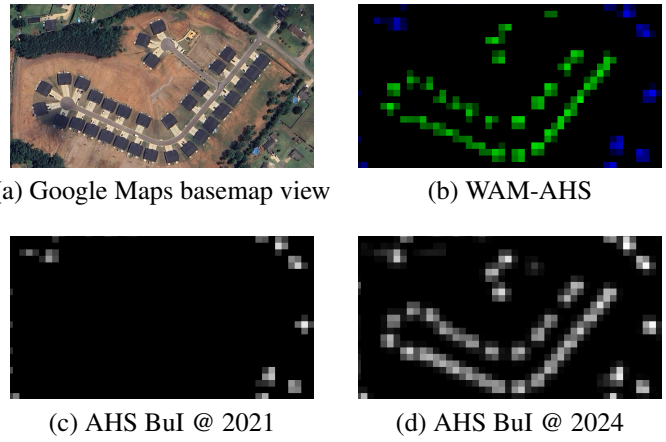


(c) The GHSL-BUFRAC

Fig. 1. AHS and GHSL views of the S-E Corridor, GA, USA.

The only known alternative to the AHS, i.e. being of global coverage and delivering the same semantics, is the Global Human Settlement Layer offered by the JRC of the EC. GHSL-BUFRAC is available at 10m resolution for a single year, 2018. The GHSL underwent several model improvements to enhance BUFRAC quality, each one referred to as a *Release* for the 2018 epoch. All references to the GHSL in this article point to the R2023 release/ 2018 epoch.

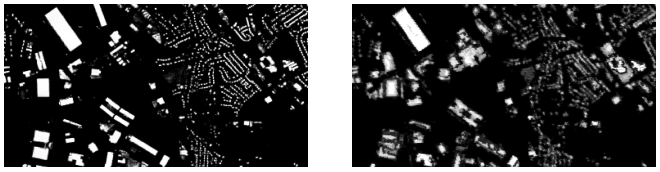
doi:10.2760/2119408



(a) Google Maps basemap view (b) WAM-AHS

(c) AHS BuI @ 2021 (d) AHS BuI @ 2024

Fig. 2. Change detection using WAM-AHS. The building color coding for (b) is green: new, blue: unchanged.



(a) AHS - BuI (b) GHSL - BUFRAC

Fig. 3. Example of big building extraction consistency between the two layers

Having only two layers of the kind available in the market today, this article attempts to identify the strengths and weaknesses of each through a qualitative and a quantitative exercise presented in Sections 2 and 3 respectively. A summary of findings and discussion of results is given in Section 4.

2. QUALITATIVE ANALYSIS

To evaluate qualitatively the two layers, we attempted to capture the data scientist user-experience when confronted with both layers as analysis ready data. We consulted four geospatial-data analysts to identify key features that best describe their engagement with the data layers and recorded their experiences when re-visiting each one separately. The findings are discussed in Section 4. The features are:
- built-up surface completeness: empirical estimate of completeness of the binary built-up surface,
- noise in-between built-up: false BU positives inside the studied settlements,
- noise outside settlements: false BU positives outside the studied settlements,
- visual clarity of built-up: overall appreciation of the visual scene - Fig. 1 (b, c),
- big building segmentation: suitability for segmenting clearly distinguishable big buildings- Fig. 3,
- built-up statistics: suitability for computing statistics,
- change detection: suitability for change detection, Fig. 2.

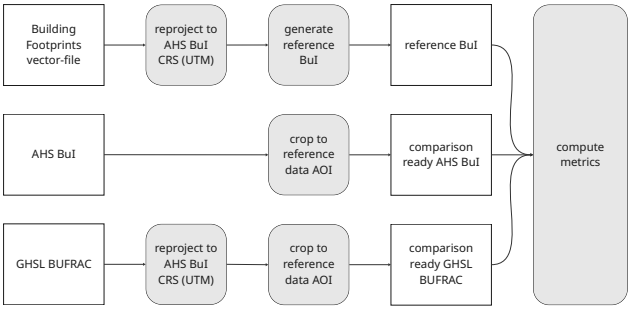


Fig. 4. Evaluation workflow.

3. QUANTITATIVE ANALYSIS

The quantitative analysis of both layers was orchestrated as a comparison against reference data. The latter was assembled from 8 areas of interest (AOIs) in 6 different countries: China - Jinxiang, Japan - Hamamatsu, Japan - Kyoto, Japan - Maebashi, Poland - Warsaw, South Africa - Cape Town, UK - Southampton, USA - Worcester (MA). They are referred to as AOI 1 through 8, [5, 2, 4, 6, 9, 7].

Manually delineated building footprints were collected for each AOI and translated into test BuI surfaces, i.e. the same material consumed by both AHS and GHSL models during training. The term test BuI is referred to as reference data instead of *ground truth*, as the quality of the building footprints is subject to human interpretation and skill. Each set of building footprints differs from the others in two ways, the date of production and in the clarity/resolution of the underlying image used to produce them. The reference data used was selected based on the production date being as close to the end of the year 2018 as possible.

To make this a fair comparison we computed the AHS-BuI of each AOI for the year 2018. The quality figures reported do not necessary reflect the true quality of the AHS (underestimation) but allow for a one-to-one comparison against the GHSL. The results are biased by the fact that we compare findings computed on annual median composites (AHS and GHSL) against the reference data generated for a certain time stamp later in time; $t > 2018$. This propagates the same error to both layers thus does not impact the outcome of the comparison. A further, very minor bias in favor of the AHS comes from the fact that both the reference data and the GHSL layers used, were re-projected to the AHS CRS (UTM) - Fig. 4. The latter was selected over the World-Mollweide of the GHSL for its iso-tropic pixel representation, appreciated in both ML model training and deployment.

doi:10.2760/2119408

To compute the 'closeness' of each layer to the reference data we utilized two sets of metrics; one for segmentation and one for regression. The former set was used to evaluate the completeness of the built-up surface, and the latter for the evaluation of the pixel content accuracy. In the following P, N, TP, TN, FP and FN stand for the number of pixels that are *positive* (BuI>0), *negative* (BuI=0), *true positive*, *true negative*, *false positive*, *false negative* respectively. The segmentation metrics used are as follows:

$$\textbf{Binary Accuracy}: \frac{TP+TN}{P+N}, \tag{1}$$

$$\textbf{Precision}: \frac{TP}{TP+FP}, \ \textbf{Recall}: \frac{TP}{TP+FN} \tag{2}$$

$$\textbf{F1 Score}: \frac{2\times PR\times RE}{RP+RE} = \frac{2\times TP}{2\times TP+FP+FN} \tag{3}$$

Matthew's Correlation Coefficient:

$$\frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

The regression metrics used are as follows:

$$\textbf{Round Mean Square Error}: \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2} \tag{5}$$

$$\textbf{Mean Average Error}: \frac{1}{n}\sum_{i=1}^{n}|\hat{X}_i - X_i| \tag{6}$$

and were computed for each inference image as a whole, and for the BuI>0 and BuI=0 pixel sets separately.

4. DISCUSSION OF RESULTS

Qualitative analysis findings: Responses on the Built-up Surface Completeness suggested that both layers, if treated as binary surfaces, deliver a satisfactory IoU with reference data, i.e. they do not miss and do not over-represent built-up. Some in-between built-up noise appears in both, examples of which are due to highly reflecting road segments, construction materials, aggregation of metal sheets that are not parts of roofs (containers, trucks), parking lots, etc. The GHSL slightly under performs due to increased sensitivity to parking lots/loading bays. Some minor false positives appear outside settlements and are primarily attributed to highly reflecting rocks and in some instances to minor water bodies.

The visual clarity of built-up is the highest discriminator between the two layers. While built-up is mostly captured accurately in both, the GHSL delivers a blurry view making it hard to trace individual buildings. By contrast to the GHSL, the AHS presents clearly distinguishable buildings, even the smallest ones, in cases they are further than 10m apart, i.e. 1 spatial unit separation. An example is shown in Fig. 1.

Big buildings can be extracted from both layers as standalone structures that are useful and in-demand for supply-chain and real-estate management applications. While generally highlighted well in both layers, in the GHSL big buildings appear with textural noise running through their extent that makes it harder for simple/fast computer vision scripts to return a single segment per building, Fig. 3.

Built-up statistics can be computed from both layers, offering insights such as how much of a settlement surface area is built, how dense is the built-up, etc. A limitation that relates to the earlier observation on big buildings is that in the case of the GHSL the built-up cannot be trivially binned to size histograms reporting how many buildings are there for specific size ranges. In case of very dense built-up both layers cannot discriminate between individual buildings and this is primarily due to the spatial resolution of the input data.

Lastly, the GHSL being a single-year release cannot be utilized directly in change detection, where as AHS-WAM detects change robustly and among any two BuI instances from the present date back to 2016. Change detection can be computed by ML models on multispectral image pairs, but at the cost of model retraining for each AOI to prevent drift, and increased sensitivity to local built-up patterns. The unavailability of historical GHSL data weakens its adoption by what appears to be one of the highest priority commercial use-cases.

Quantitative analysis findings: An initial observation from Table 1 is that the GHSL has a major imbalance between precision and recall. Low precision and high recall means that the model is good at finding all the actual positives (BuI>0) but among the instances the model predicts as built-up, many are actually false positives. The lack of model sophistication leads to poor generalization erring on the side of predicting positive BuI and pretty much 'catches everything' but inaccurately. The AHS maintains a far better balance between the two metrics. Looking at the F1 and MCC scores (holistic view of model performance), the AHS leads by 14% and 13% respectively and this is consistent throughout all 8 AOIs.

Table 2 lists the regression metrics' scores related to the BuI values. The AHS layers on average and across all pixels show a relative reduction of about 30% in regards to RMSE and about 35% in regards to MAE when compared to the GHSL layer values. These relative reductions persist in empty (BuI=0) and non-empty (BuI>0) pixel sets too.

5. CONCLUSIONS

In this paper we evaluated the AHS on selected global AOIs, comparing it to the GHSL. Both qualitative and quantitative results show AHS has a competitive edge. As a continually evolving product, further improvements are expected.

doi:10.2760/2119408

Table 1. Segmentation Metrics

metric	layer	AOI1	AOI2	AOI3	AOI4	AOI5	AOI6	AOI7	AOI8	average
Binary Accuracy	GHSL	0.95449	0.95114	0.87118	0.81972	0.78339	0.94021	0.92876	0.85596	0.88810
	AHS	**0.97388**	**0.98036**	**0.92371**	**0.93487**	**0.93890**	**0.96911**	**0.96635**	**0.94827**	**0.95443**
Precision	GHSL	0.46971	0.45751	0.67842	0.43199	0.37874	0.55974	0.45056	0.37651	0.47539
	AHS	**0.63336**	**0.71959**	**0.81203**	**0.72438**	**0.75916**	**0.74474**	**0.68024**	**0.68203**	**0.71944**
Recall	GHSL	**0.96417**	**0.96787**	**0.98799**	**0.98565**	**0.97163**	**0.98625**	**0.92083**	**0.95821**	**0.96782**
	AHS	0.84213	0.86114	0.93088	0.85008	0.79426	0.90091	0.81496	0.77707	0.84642
F1 Score	GHSL	0.63169	0.62132	0.80445	0.60071	0.54503	0.71416	0.60506	0.54060	0.63287
	AHS	**0.72297**	**0.78403**	**0.86740**	**0.78221**	**0.77631**	**0.81541**	**0.74153**	**0.72646**	**0.77704**
MCC Score	GHSL	0.65567	0.64714	0.74130	0.57813	0.52037	0.71788	0.61520	0.54678	0.62780
	AHS	**0.71753**	**0.77725**	**0.81794**	**0.74751**	**0.74121**	**0.80304**	**0.72707**	**0.69984**	**0.75392**

Table 2. Regression Metrics

metric	layer	AOI1	AOI2	AOI3	AOI4	AOI5	AOI6	AOI7	AOI8	average
RMSE	GHSL	0.10310	0.09340	0.21453	0.17004	0.18895	0.12891	0.10707	0.11957	0.14069
	AHS	**0.09198**	**0.06411**	**0.17369**	**0.11263**	**0.12570**	**0.09689**	**0.07332**	**0.09188**	**0.10377**
MAE	GHSL	0.02259	0.02074	0.10674	0.07090	0.08300	0.03612	0.02743	0.04053	0.05100
	AHS	**0.01846**	**0.01217**	**0.08121**	**0.03771**	**0.03909**	**0.02454**	**0.01638**	**0.02450**	**0.03175**
RMSE - non empty	GHSL	0.40256	0.33715	0.33785	0.33721	0.35086	0.38365	0.33130	0.30600	0.34832
	AHS	**0.37423**	**0.27614**	**0.31414**	**0.28217**	**0.30835**	**0.30553**	**0.26902**	**0.26250**	**0.29901**
MAE - non empty	GHSL	0.31843	0.26816	0.27299	0.26933	0.27451	0.30701	0.26004	0.23602	0.27581
	AHS	**0.29383**	**0.21599**	**0.25382**	**0.21822**	**0.23220**	**0.24036**	**0.20739**	**0.19934**	**0.23264**
RMSE - empty	GHSL	0.06654	0.06691	0.14866	0.12764	0.15276	0.08033	0.07613	0.08580	0.10059
	AHS	**0.05542**	**0.03452**	**0.07821**	**0.05219**	**0.06771**	**0.05342**	**0.03892**	**0.05622**	**0.05457**
MAE - empty	GHSL	0.01047	0.01073	0.04835	0.04136	0.05590	0.01504	0.01401	0.02345	0.02741
	AHS	**0.00718**	**0.00393**	**0.02060**	**0.01080**	**0.01180**	**0.00775**	**0.00538**	**0.00924**	**0.00958**

REFERENCES

[1] Atlas AI. Introducing aperture® pulse: A new era of scalable economic change detection. https://www.atlasai.co/blog/introducing-aperture-pulse, 2025.

[2] Source Cooperative. Japanese building footprint data. https://source.coop/repositories/pacificspatial/flateau/description, 2024.

[3] Christina Corbane, Vasileios Syrris, Filip Sabo, Panagiotis Politis, Michele Melchiorri, Martino Pesaresi, Pierre Soille, and Thomas Kemper. Convolutional neural networks for global human settlements mapping from sentinel-2 satellite imagery. *Neural Computing and Applications*, 33(12):6697–6720, 2021. doi: 10.1007/s00521-020-05449-7.

[4] geoportal.gov.pl. Topographic objects database (bdot10k). https://www.geoportal.gov.pl/en/data/topographic-objects-database-bdot10k/, 2023.

[5] Planetek Hellas. Jinxiang building footprints dataset. manual annotation, 2023. URL https://www.planetek.gr/.

[6] City of Cape Town. 2d building footprints. https://odp-cctegis.opendata.arcgis.com/datasets/cctegis::2d-building-footprints/about, 2023.

[7] MassGIS (Bureau of Geographic Information). Massgis data: Building structures (2-d). https://www.mass.gov/info-details/massgis-data-building-structures-2-d#downloads-, 2024.

[8] Martino Pesaresi, Marcello Schiavina, Panagiotis Politis, Sergio Freire, Katarzyna Krasnodebska, Johannes H. Uhl, Alessandra Carioli, Christina Corbane, Lewis Dijkstra, Pietro Florio, Hannah K. Friedrich, Jing Gao, Stefan Leyk, Linlin Lu, Luca Maffenini, Ines Mari-Rivero, Michele Melchiorri, Vasileios Syrris, Jamon Van Den Hoek, and Thomas Kemper and. Advances on the global human settlement layer by joint assessment of earth observation and population survey data. *International Journal of Digital Earth*, 17(1):2390454, 2024. doi: 10.1080/17538947.2024.2390454.

[9] Ordnance Survey. Great britain's national geographic database. https://www.ordnancesurvey.co.uk/, 2024.

[10] Abe Tarapani. Introducing the fourth release of the atlas of human settlements. https://www.atlasai.co/blog/introducing-the-fourth-release-of-the-atlas-of-human-settlements, 2024.

EMPOWERING HUMANITARIANS: INTRODUCING RISE FOR RAPID GEOSPATIAL INTELLIGENCE AND ADVANCED EO ANALYTICS

Alessandra Anaya[1], Paolo Campanella[1], Lucas Falk[1], Laura Giustarini[1], Valentina Leone[1],
Marco Menapace[1], Nikhil Mohan[1], Cristiano Nattero*[1], Alberto Tasso[1], Jihed Ben Zarb[1]
Marco Chini[2], Yu Li[2], Anis Amziane[2], Aolin Jia[2], Kanishka Mallick[2], Patrick Matgen[2]
Ana Carolina Helena[3], Daniel Ledesma Nicrosi[3]
Michelle Joseph[4], Bethany Plant[4]

1: WASDI, Luxembourg
2: Luxembourg Institute of Science and Technology (LIST), Luxembourg
3: Aide internationale de la Croix-Rouge luxembourgeoise (AICRL) - Shelter Research Unit (SRU), Luxembourg
4: World Food Programme (WFP) - Innovation Accelerator, Germany

ABSTRACT

Remote Imaging Support for Emergencies (RISE) is a novel web application democratizing access to actionable geospatial intelligence from Earth Observation (EO) data. Built on the WASDI cloud platform, RISE transforms complex EO workflows into user-friendly products for rapid crisis response. It automates data ingestion, processing, and dissemination, enabling near real-time monitoring and impact assessments. With capabilities from flood mapping to drought monitoring, RISE delivers critical intelligence quickly and cost-effectively. Successful field validation, coupled with positive beta feedback, underscores RISE's potential to revolutionize emergency management.

Index Terms— Humanitarian, Geospatial Intelligence, Emergency, RISE

1. INTRODUCTION

Timely geospatial intelligence from EO data (optical, SAR, thermal) is vital for humanitarian crisis response, damage assessment, and risk reduction. The foundational engine, WASDI [1, 2], is a powerful cloud platform for large-scale EO processing, with a proven record in high-impact events: (i) Pakistan Floods (2022): Rapid preliminary damage assessment for the Asian Development Bank; (ii) Cyclone Amphan, Bangladesh (2020): Flood assessment support for WFP; (iii) Typhoon Kammuri, Philippines (2019): Damage monitoring for WFP; (iv) Ukraine Wildfires (2020): Monitoring support for ESA.

While WASDI demonstrated significant capabilities, its technical complexity limited direct use by many humanitarians. This firsthand experience highlighted the need for a more accessible solution. Data latency and the "human-in-the-loop" are key bottlenecks; lengthy map production times limit utility to post-event analysis or preparedness, whereas rapid intelligence is crucial for in-emergency intervention.

RISE addresses this by leveraging WASDI's power through an intuitive interface. Developed by WASDI with LIST and AICRL-SRU, and funded by WFP's Humanitarian Innovation Accelerator, RISE automates EO workflows. Its core mission is to provide humanitarians with direct, timely, and cost-effective geospatial intelligence, specifically to guide interventions during acute emergency phases. Beta feedback and the results from the final validation exercise are highly positive.

2. RISE: CAPABILITIES ENABLED BY WASDI

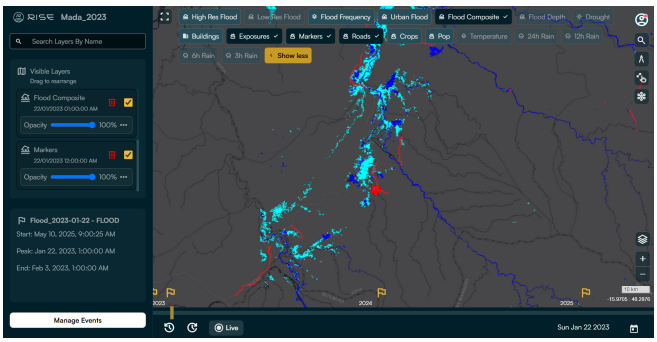


Fig. 1. The UI of RISE.

*cristiano.nattero@wasdi.cloud

Thanks to the WFP's Humanitarian Innovation Accelerator Program for having funded the development of RISE.

doi:10.2760/2119408

RISE offers a diverse suite of geospatial products, operationalized on WASDI's cloud platform. While initial wide-area screening for events like floods is supported by data from providers like NOAA (e.g., VIIRS), feedback from our humanitarian partners during the validation exercise confirmed that these lower-resolution products were insufficient for their operational needs. This finding validated our strategic decision to build the core of RISE's high-resolution analytical capabilities upon the Copernicus programme. Our primary data source for detailed, actionable flood mapping, urban analysis, and impact assessment is Sentinel-1 SAR, complemented by Sentinel-2 optical and Sentinel-3 thermal data. This focus on the Copernicus suite ensures operational continuity, aligns with European data infrastructure, and most importantly, delivers the high-quality intelligence our users require, thereby mitigating dependencies on other data sources for our most critical products.

A cornerstone of RISE is daily automated monitoring and event identification, orchestrated by WASDI, with options to upload ancillary data for contextualization.

Figure 1 shows the UI of RISE, with a flood and impact analysis.

2.1. Flood Management and Monitoring:

Low-Resolution Daily Flood Maps: Derived from NOAA's VIIRS (Visible Infrared Imaging Radiometer Suite) data (typically 375m resolution for flood products), providing consistent, wide-area screening using thresholding on specific spectral bands and indices. Processing and dissemination are managed via WASDI.

High-Resolution Flood Maps: Generated on WASDI using LIST's HASARD algorithm suite [3, 4, 5, 6, 7]. HASARD applies advanced change detection and SAR backscatter analysis techniques to Sentinel-1 data to delineate floodwater extent with high accuracy. The execution of HASARD at scale is a key WASDI capability.

Urban Flood Maps: A specialized HASARD module for urban environments, utilizing multitemporal interferometric SAR coherence analysis from Sentinel-1 to detect subtle changes indicative of flooding amidst complex urban structures. This computationally intensive analysis is performed on WASDI.

Composite Flood Map: An integrated flood product produced on WASDI, potentially employing rule-based integration or weighted fusion of different flood layers (e.g., VIIRS and HASARD) to provide a more comprehensive overview.

Flood Relative Frequency Maps: Constructed on WASDI through the systematic reprocessing of historical Sentinel-1 archives using HASARD, enabling the identification of areas with high recurrent flood probability. This relies on WASDI's capacity for large-volume data processing.

Flood Depth Maps: Estimated on WASDI by integrating SAR-derived flood extent with Digital Elevation Models (DEMs), using hydraulic principles or calibrated empirical relationships.

Historical Flood Archives: RISE can reconstruct historical flood events using the HASARD algorithm, providing valuable long-term data for risk assessment, trend analysis, and planning, all stored and processed within WASDI.

2.2. Rain Observation

IMERG Rain Observation: Accumulated rainfall data from NASA's Integrated Multi-satellitE Retrievals for GPM (IMERG) product, offering precipitation estimates at approximately 0.1° x 0.1° resolution with near real-time latency (latency of a few hours for late run products). Data is ingested and made available through WASDI.

2.3. Impact Assessment

Flood Impact Maps: Detailed assessments of flood impacts on buildings, road networks, agricultural land (crops), and estimated population affected. Produced on WASDI by integrating flood extent/depth maps with exposure data layers (e.g., building footprints from CityWatch, global road datasets, land cover classifications, and population density grids like WorldPop or GHSL) using CIMA Research Foundation's RASOR (Rapid Analysis and SOlutions for Response) methodology. This geospatial analysis is orchestrated by WASDI.

2.4. Drought Monitoring

Land Surface Temperature (LST): Derived from thermal infrared data from Sentinel-3.

Drought Monitoring: RISE incorporates LIST's innovative Radiative Thermal Inertia (RTI) index [8], a physically-based model for monitoring soil moisture and vegetation water stress. The RTI model integrates LST and other EO data (e.g., albedo, solar radiation) and has demonstrated a strong correlation (coefficient of 0.62) with in-situ soil moisture measurements across multiple sites—outperforming conventional drought indicators such as the Keetch-Byram Drought Index (KBDI) [9] and Apparent Thermal Inertia (ATI) [10]. These models are run on the WASDI platform.

2.5. Urban and Settlement Analysis

Settlements and Urban Area Maps (CityWatch): Leveraging LIST's CityWatch algorithm [11], RISE provides regularly updated maps of urban areas and settlements, with processing hosted on WASDI.

CityWatch Baseline: Global mapping of built-up areas at 10-meter resolution using a synergistic approach with Sentinel-1 SAR and Sentinel-2 optical data. This involves automated training data sampling and a label-noise robust cross-fusion neural network.

doi:10.2760/2119408

CityWatch Premium: Higher detail by incorporating commercial Very High-Resolution (VHR) optical imagery. This solution employs advanced AI techniques, potentially involving transfer learning or weakly supervised learning, to classify VHR images using lower-resolution labels from CityWatch Baseline as a training source.

3. TECHNICAL ARCHITECTURE

RISE features an intuitive UI presenting key indicators upon entry: estimated people affected, ongoing events, and alerts. All processing occurs on WASDI, a PaaS optimized for EO, offering (i) Scalable Distributed Computing; (ii) Co-location of Data and Processing; (iii) Interoperability via standards; (iv) Robust Data Management.

RISE's architecture uses containerized plugins deployed on WASDI, interacting via APIs. RISE is an intelligent orchestrator and user-friendly front-end to WASDI's specialized EO processing.

The decision to build RISE on the WASDI cloud platform was a strategic one, driven by efficiency and cost-effectiveness. As the developers of both RISE and the underlying WASDI platform, our team was able to ensure rapid development and a robust, seamlessly integrated user experience. While this provides a strong foundation, we recognize the importance of interoperability for wider adoption. The WASDI platform is inherently designed to offer this "off the shelf," as it supports approximately 90 different standards across file formats, protocols, languages, and architecture. These capabilities facilitate the rapid integration of new functionalities from third-party providers through its compliance with established standards integral to the European EO cloud ecosystem, including OGC WxS, OpenEO, and the Earth Observation Exploitation Platform Common Architecture (EOEPCA). This is demonstrated by our current infrastructure; WASDI already operates several nodes on the CloudFerro cloud, and exploits CREODIAS as one of its data providers. Building on this, our future roadmap includes developing interfaces with other major European platforms and services, including the Copernicus Data Space Ecosystem (CDSE), WEkEO, the Copernicus Emergency Management Service (CEMS), and the Destination Earth initiative, as well as UN services, such as those by UN SPIDER, FAO, and World Bank, to name but a few. This expansion is not only a priority but also a streamlined process, allowing RISE services to be discovered, accessed, and chained within a more open and collaborative ecosystem.

4. MITIGATING LATENCY AND ENHANCING EFFICIENCY FOR EMERGENCY INTERVENTION:

RISE significantly reduces latency to maximize EO data utility for active emergency intervention with (i) **End-to-End Automation**: From image retrieval (e.g., Copernicus Hubs) through WASDI processing to RISE display, and (ii) **Proactive, Request-Free Product Generation**: Core products (e.g., daily flood maps) are generated proactively by scheduled WASDI processes, ensuring readiness.

Computational gains are substantial: LIST's HASARD, traditionally taking days, runs in hours on WASDI. The 2022 Pakistan flood assessment, manually requiring a 4+ person team, is now automatable in hours. The primary speed improvement is removing the human-in-the-loop from critical processing, transforming EO data from a retrospective tool into a vital asset for immediate emergency decision-making – RISE's core purpose.

5. VALIDATION IN REAL-WORLD SCENARIOS

RISE is undergoing rigorous validation. Beta user feedback (WFP, Red Cross) is highly encouraging. The field validation, initially planned for Niger, was redirected to Madagascar (a region with Red Cross presence and relevant flood/drought events) due to security. It simulates response to the Cheneso Cyclone (Jan 2023), focusing on Maroantsetra, selected for (i) Humanitarian Data Availability, (ii) Optimal Sentinel-1 Coverage, (iii) Local Red Cross team presence.

5.1. Validation Methodology

The exercise employed a comparative methodology designed to establish a clear baseline for RISE's impact. It is important to note that this comparison was structured for analytical purposes; in a real-world scenario, RISE is intended to be a complementary tool that enhances, rather than replaces, traditional methods. For this validation, however, the two approaches were deliberately separated to measure the platform's standalone contribution:

Team 1 (Standard), conducted the assessment using traditional methods, including field surveys, secondary data review, and key informant interviews. This required 26 personnel and took 8 hours to complete.

Team 2 (RISE-Equipped), used only RISE's analytical products to guide their strategy. This required just 2 staff members and was completed in 1 hour and 40 minutes.

The results of this baseline comparison were extraordinary and demonstrate significant operational gains:

Quantitative Gains: The data shows that the RISE-Equipped team was able to produce its initial assessment with an 85% reduction in time and an 90% reduction in associated costs (personnel and transport) compared to the team using standard methods.

Accuracy and Effectiveness: The final report from the exercise concluded that RISE provides "much more accuracy of data, when compared to standard methods, especially linked to historical data... and no or hard to access areas." This is critical in humanitarian contexts where access can be limited or dangerous.

doi:10.2760/2119408

User Feedback: Qualitative feedback was highly positive, with the final report concluding that RISE could make needs assessments "more accurate, time, cost-effective and safe." A key advantage noted was the increased safety for humanitarian workers, as the platform allows for rapid assessment without deploying personnel to potentially unsafe areas.

These findings provide tangible evidence of RISE's capacity to significantly enhance the speed, efficiency, and effectiveness of humanitarian response. Comprehensive results will be presented at the BIDS 2025 conference.

6. THE FUTURE OF RISE: EXPANDING CAPABILITIES AND TRANSFORMING EMERGENCY MANAGEMENT

RISE's plugin-based architecture on WASDI allows agile expansion. Roadmap includes (i) Deforestation Monitoring: Sentinel-1/2 and Landsat time-series; (ii) Active Fires and Burned Area Mapping; (iii) Landslide Detection; (iv) Earthquake Impact Assessment. Enhancements will cover impact assessment (socio-economic data, vulnerability models), data fusion, and predictive analytics within the WASDI-RISE ecosystem.

7. BROADER IMPACT AND RECOGNITION

RISE's potential was showcased at the 2024 STI Forum, reflecting demand for automated, cloud EO solutions. Its adaptability is proven by the validation redirection.

The successful validation in Madagascar, coupled with positive feedback from the beta program, provided strong evidence of RISE's operational benefits.

8. CONCLUSIONS

RISE significantly advances the delivery of sophisticated EO data in an accessible, actionable format for humanitarians. By synergizing advanced algorithms, the scalable WASDI cloud platform (its core processing engine), and a user-centric design with immediate key indicators, RISE empowers timely, data-driven emergency decisions. Its architecture addresses latency and efficiency through automation and optimized cloud processing, transforming EO maps from historical records into vital tools for in-emergency intervention.

The conclusive results from the Madagascar validation, alongside positive beta feedback, have provided robust evidence of RISE's significant operational benefits. RISE is set to establish a new standard in operational geospatial intelligence, catalyzing change by effectively leveraging powerful backend platforms like WASDI through user-focused applications for a more agile and effective global humanitarian response.

REFERENCES

[1] G. J. Schumann, P. Campanella, A. Tasso, L. Giustarini, P. Matgen, M. Chini, and L. Hoffmann, "An online platform for fully-automated eo processing workflows for developers and end-users alike," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 8656–8659.

[2] C. Nattero, R. Rudari, M. Chini, P. Campanella, and M. Menapace, "Wasdi, a cloud platform for earth observation and natural hazards," in *EGU General Assembly Conference Abstracts*, 2023, pp. EGU–15 882.

[3] P. Matgen, R. Hostache, G. Schumann, L. Pfister, L. Hoffmann, and H. Savenije, "Towards an automated sar-based flood monitoring system: Lessons learned from two case studies," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 36, no. 7-8, pp. 241–252, 2011.

[4] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A hierarchical split-based approach for parametric thresholding of sar images: Flood inundation as a test case," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 6975–6988, 2017.

[5] M. Chini, R. Pelich, L. Pulvirenti, N. Pierdicca, R. Hostache, and P. Matgen, "Sentinel-1 insar coherence to detect floodwater in urban areas: Houston and hurricane harvey as a test case," *Remote Sensing*, vol. 11, no. 2, p. 107, 2019.

[6] G. Schumann, L. Giustarini, A. Tarpanelli, B. Jarihani, and S. Martinis, "Flood modeling and prediction using earth observation data," *Surveys in Geophysics*, vol. 44, no. 5, pp. 1553–1578, 2023.

[7] R. Pelich, M. Chini, R. Hostache, P. Matgen, L. Pulvirenti, and N. Pierdicca, "Mapping floods in urban areas from dual-polarization insar coherence data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[8] A. Jia, K. Mallick, D. Upadhyaya, T. Hu, Z. Szantoi, B. Bhattacharya, M. Sekhar, D. Skoković, J. A. Sobrino, L. Ruiz, and G. Boulet, "Deriving a clear-sky soil moisture index from ecostress land surface temperature," *Remote Sensing of Environment*, vol. 329, p. 114945, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425725003499

[9] J. J. Keetch and G. M. Byram, *A drought index for forest fire control*. US Department of Agriculture, Forest Service, Southeastern Forest Experiment . . . , 1968, vol. 38.

[10] J. C. Price, "Thermal inertia mapping: A new view of the earth," *Journal of Geophysical Research*, vol. 82, no. 18, pp. 2582–2590, 1977.

[11] Y. Li, P. Matgen, and M. Chini, "Extraction of built-up areas using sentinel-1 and sentinel-2 data with automated training data sampling and label noise robust cross-fusion neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, p. 104524, 2025.

doi:10.2760/2119408

Author Index

doi:10.2760/2119408

doi:10.2760/2119408

doi:10.2760/2119408

doi:10.2760/2119408

doi:10.2760/2119408

doi:10.2760/2119408

doi:10.2760/2119408

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
Joint-research-centre.ec.europa.eu

Publications Office
of the European Union