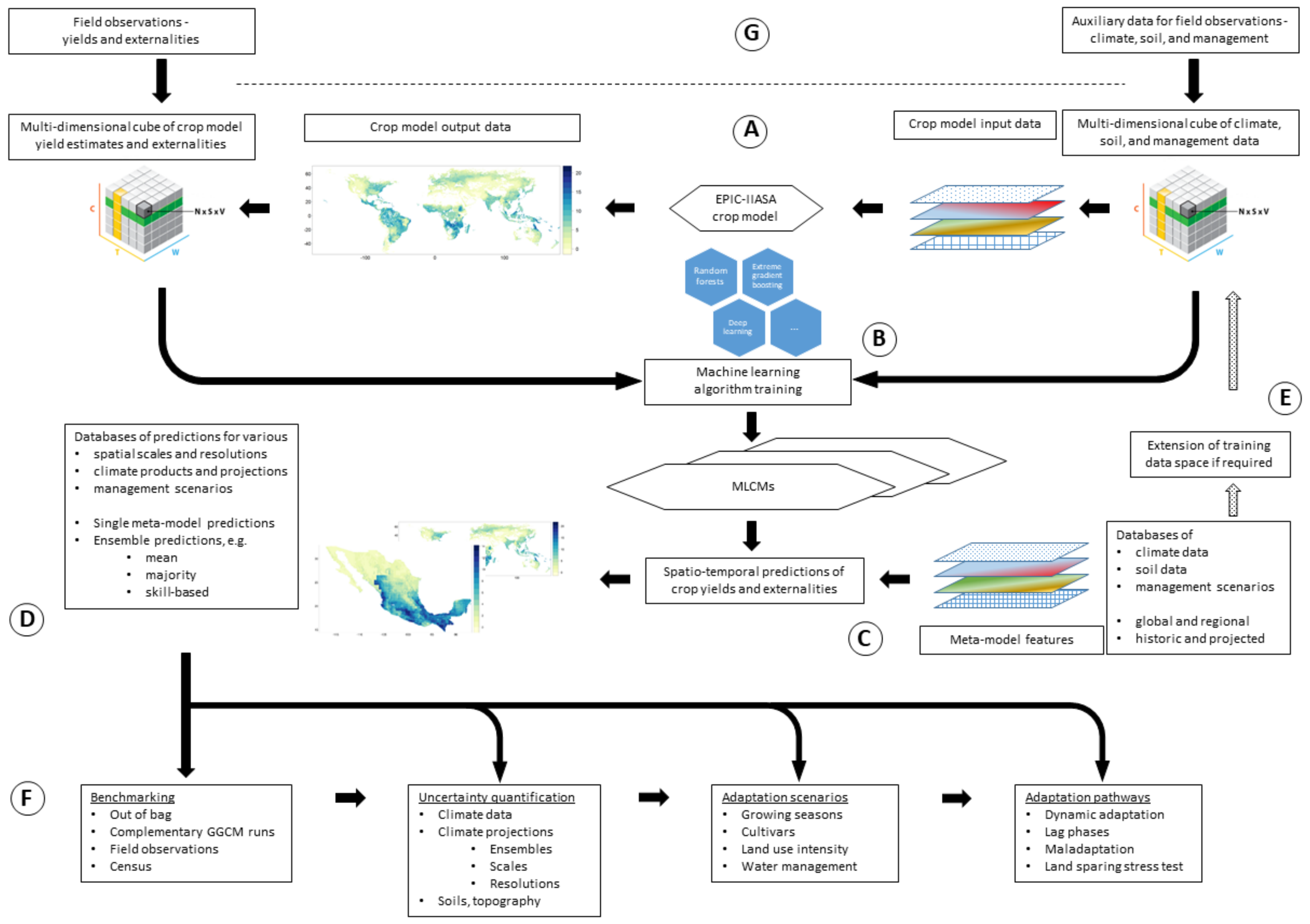
International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

CROMES - A fast and efficient machine learning emulator pipeline for gridded crop models

Christian Folberth[1], Artem Baklanov[2], Nikolay Khabarov[2], Thomas Oberleitner[1], Juraj Balkovič[1], Rastislav Skalský[1]
[1] Biodiversity and Natural Resources Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[2] Advancing Systems Analysis Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Background

Global gridded crop models (GGCMs) have become state-of-the-art tools in large-scale climate impact and adaptation assessments. Yet, these combinations of large-scale spatial data frameworks and plant growth models have limitations in the volume of scenarios they can address due to computational demand or complex software structures. Emulators mimicking such models are therefore emerging as an attractive option to produce reasonable predictions of GGCMs' crop productivity estimates at much lower computational costs. However, such emulators' flexibility is thus far typically limited in terms of crop management flexibility and spatial resolutions among others.



G

Field observations - yields and externalities

Auxiliary data for field observations - climate, soil, and management

A

Multi-dimensional cube of crop model yield estimates and externalities

Crop model output data

Crop model input data

Multi-dimensional cube of climate, soil, and management data

EPIC-IIASA crop model

Random forests

Extreme gradient boosting

Deep learning

...

B

Machine learning algorithm training

E

Extension of training data space if required

Databases of predictions for various
- spatial scales and resolutions
- climate products and projections
- management scenarios

- Single meta-model predictions
- Ensemble predictions, e.g.
  - mean
  - majority
  - skill-based

MLCMs

Spatio-temporal predictions of crop yields and externalities

Databases of
- climate data
- soil data
- management scenarios

- global and regional
- historic and projected

D

C

Meta-model features

F

Benchmarking
- Out of bag
- Complementary GGCM runs
- Field observations
- Census

Uncertainty quantification
- Climate data
- Climate projections
  - Ensembles
  - Scales
  - Resolutions
- Soils, topography

Adaptation scenarios
- Growing seasons
- Cultivars
- Land use intensity
- Water management

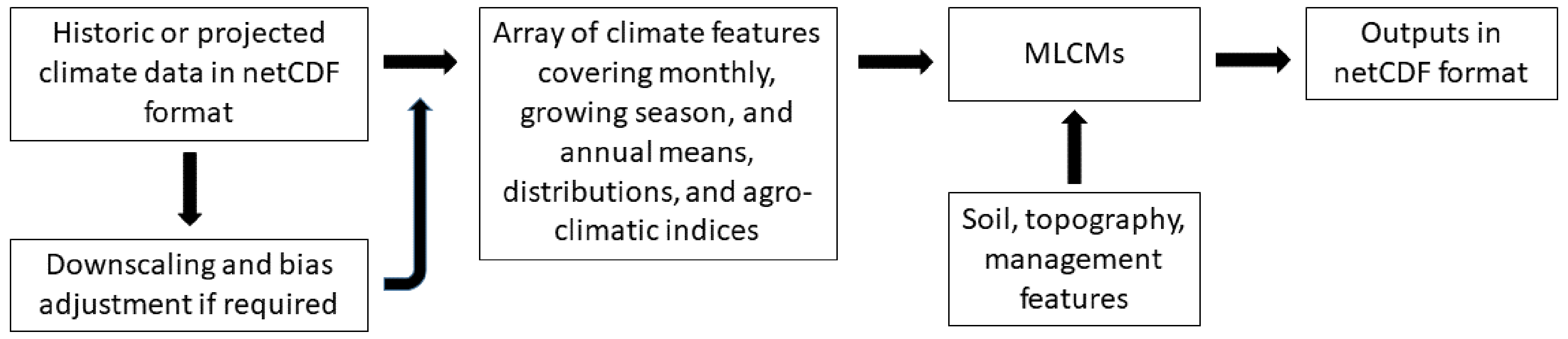
Adaptation pathways
- Dynamic adaptation
- Lag phases
- Maladaptation
- Land sparing stress test

Figure 2: Framework of the project modified from (Folberth et al., 2019) starting from a multi-dimensional cube of crop model input data at the top right and resulting in climate impact and adaptation projections at the bottom. Letters A-F indicate key steps in the methodology: A: Global gridded crop model simulations for a multidimensional cube of input data to generate training data for machine learning algorithms, B: training of machine-learning crop meta-models (MLCMs) based on various machine learning algorithms and the global GGCM training sample, C: predictions for comprehensive sets of GCMs and derived high-resolution climate data, other agro-environmental features, and management trajectories, D: storage and combination of MLCM predictions, E: extension of training data and GGCM simulation space if required, F: processing and interpretation of outcomes from benchmarking to adaptation pathway development, G: inclusion of field observations and associated data in MLCM training data space as an add-on to A-F. Dimensions of the training data cube are: C=atm. CO_2 concentration, W=precipitation (incl. sufficient water supply), T=temperature, N=mineral nitrogen input, S=soil type, V=crop variety, MLCMs=machine-learning crop meta-models

FWF Der Wissenschaftsfonds.
Supported by the Austrian Science Fund under grant agreement no. P 36220-N

Project design

Here we present a new emulator pipeline CROp model Machine learning Emulator Suite (CROMES) that serves for processing climate features from netCDF input files, combining these with site-specific features (soil, topography), and crop management specifications (planting dates, cultivars, irrigation) to train machine learning crop meta-models (MLCMs) and subsequently produce predictions (Figure 1).



Historic or projected climate data in netCDF format

Downscaling and bias adjustment if required

Array of climate features covering monthly, growing season, and annual means, distributions, and agro-climatic indices

MLCMs

Soil, topography, management features

Outputs in netCDF format

Figure 1: Processing and computational pipeline to produce machine learning crop meta-model (MLCM) predictions in a fully integrated framework. The same design is used for MLCM training.

As a first strategy to emulator development, training data have been generated in a synthetic cube of variations in weather data and CO_2 following the approach from Franke et al. (2020) as a forcing for the GGCM EPIC-IIASA (Figure 1). Crop model outputs were fed into machine-learning algorithms combined with features derived from crop model input data including climate, soil, topography, and crop management (Figure 2). These resulted in a reasonable performance in reproducing yield estimates from the GGCM for climate projections but indicated bias towards values for quasi factorial features, i.e., atm. CO_2 concentrations (not shown).

To provide a training sample that is less structured and has a wider range of feature combinations, a new approach was developed using GGCM simulation outputs from actual climate forcings (e.g., GCM X) to subsequently predict yields for unseen forcings (GCM Y).

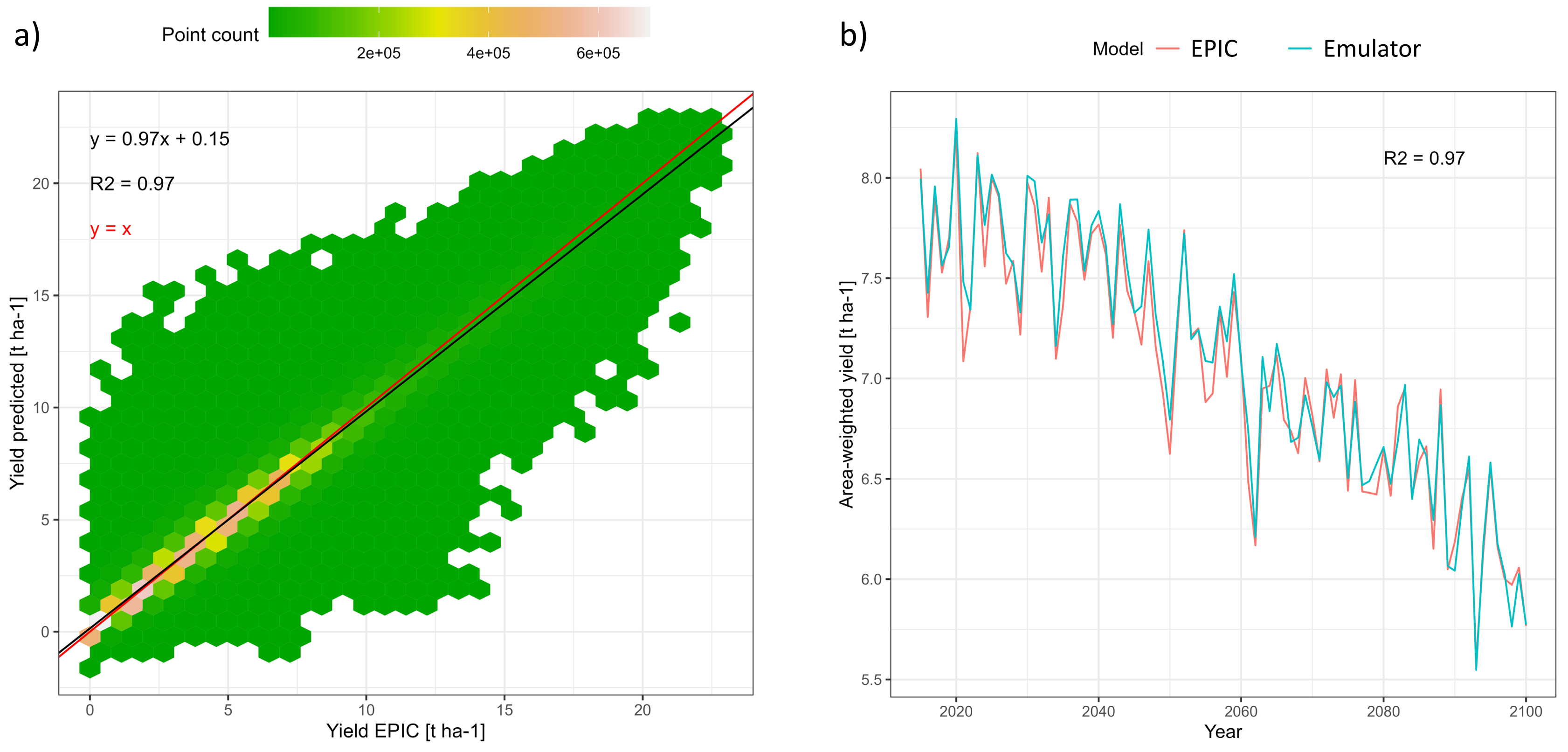


Figure 3: (a) Comparison of global gridded crop yields for rainfed maize from EPIC crop model simulations vs predictions by an ML model that was trained on one GCM and applied to another GCM for RCP8.5 in both cases. (b) Global area-weighted crop yields for the same data over time for the original model EPIC and the emulator.

Contact: Christian Folberth, IIASA, folberth@iiasa.ac.at

Performance

Predictions require for a first used climate dataset about 45 min to convert from netCDF to a faster readable binary file format and 10 min for any subsequent scenario, including climate feature generation and predictions, compared to approx. 14h for a GGCM simulation on the same system.

Prediction accuracy is highest if modeling the case when crops receive sufficient nutrients and are consequently most sensitive to climate. When training an emulator on crop model simulations for rainfed maize and a single global climate model (GCM), the yield prediction accuracy for out-of-bag GCMs is R^2=0.93-0.97, RMSE=0.2-0.7, and rRMSE=8-10% in space and time (Figure 3).

The best agreement between predictions and crop model simulations occurs in (sub-) tropical regions, the poorest in cold and arid climates (Figure 4) where both growing season length and water availability limit crop growth. The performance slightly deteriorates if fertilizer supply is considered, more so at low levels of nutrient inputs.

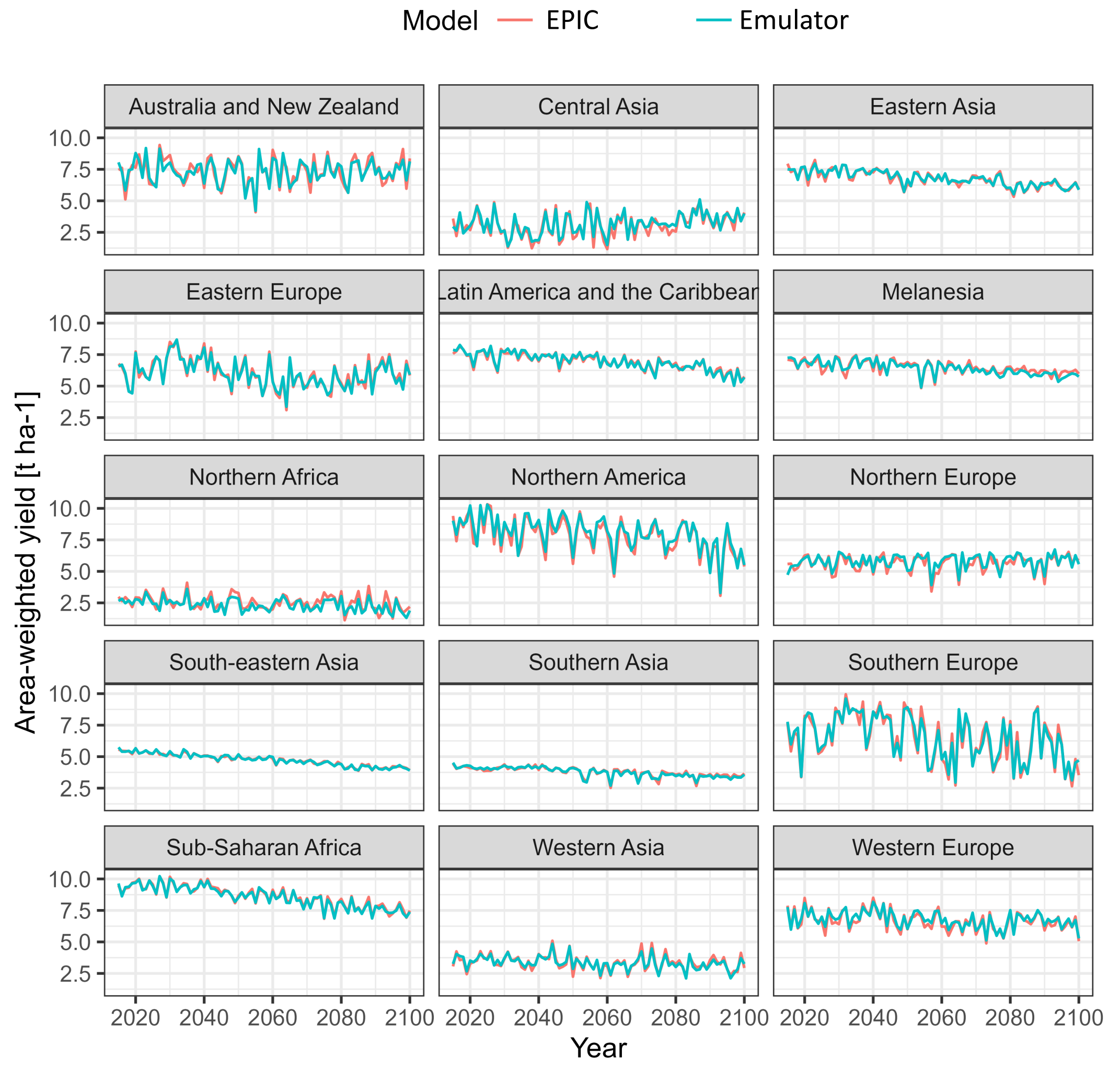


Figure 4: Same as Figure 3b but for macro-regions.

Outlook

Importantly, emulators produced by CROMES are virtually scale-free as all training samples, i.e., pixels, are pooled and treated as individual points without geo-referencing. This allows for applications on increasingly available high-resolution climate datasets or in regional studies for which more granular data may be available than at global scales. Using climate features based on crop growing seasons and cardinal growth stages, also adaptation studies such as growing season and cultivar shifts are facilitated. We expect CROMES to enable explorations of comprehensive climate projection ensembles, studies of dynamic climate adaptation scenarios, and cross-scale impact and adaptation assessments.

References:

Folberth, C. et al., 2019. Spatio-temporal downscaling of gridded crop model yield estimates based on machine-learning. Agr. For. Met. 264, 1–15. https://doi.org/10.1016/j.agrformet.2018.09.021

Franke, J.A. et al., 2020. The GGCMI Phase 2 experiment: global gridded crop model simulations under uniform changes in CO2, temperature, water, and nitrogen levels (protocol version 1.0). Geoscientific Model Development 13, 2315–2336. https://doi.org/10.5194/gmd-13-2315-2020