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

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
Email: repository@iiasa.ac.at **Telephone:** +43 (0)2236 807 342

YSSP Report
Young Scientists Summer Program

Testing two data fusion methods for multiscale and multiclass land-use/land-cover maps to improve fractional information at medium resolution

Caterina Barrasso,
caterina.barrasso@idiv.de

Approved by

Supervisors: Myroslava Lesiv, Juan Carlos Laso Bayas
Program: Advancing Systems Analysis (ASA)
30.09.2021

This report represents the work completed by the author during the IIASA Young Scientists Summer Program (YSSP) with approval from the YSSP supervisor.

It was finished by 30.09.2021 and has not been altered or revised since.

This research was funded by IIASA and its National Member Organizations in Africa, the Americas, Asia, and Europe.

Table of Contents

Abstract

High uncertainty is found during inter-comparison of land-use/land-cover (LULC) maps derived from remote sensing imagery. Among the reasons for classification mismatch, especially in coarse maps and heterogeneous areas characterized by mixed pixels, is that the landscape heterogeneity is ignored by providing only the LULC class covering the largest portion of a pixel. Pixels are arbitrary spatial units determined mainly by the sensor's properties and can have little relation to natural units on the ground. In fact, the use of class proportions in ground-truth training data, that better depict reality, proved to decrease the thematic accuracy of traditional LULC maps characterized by one LULC class per pixel. Because high-resolution LULC maps upscaled to coarser resolutions provide higher accuracy than natively-coarse maps, and because, except from creating new maps, integration of available ones can increase the final accuracy, during this project the potential of two data fusion methods for multi-scale (from high to coarse resolution) and multi-class maps to derive more accurate ones with fraction information at medium resolution (100m) was explored. Two data fusion models were tested in four study areas characterized by both mixed and pure-pixels by using seven LULC maps as input and a ground-truth sub-pixel database as response variable. The models' output was then validated and compared against each individual input map, in both mixed and pure-pixels, by using the sub-pixel thematic accuracy matrix. To make more robust predictions and better answer the research questions of the study improvement of the goodness of fit of the data fusion models is needed. Despite the need of the models' amelioration, it was observed that multiscale and multiclass data fusion improved the sub-pixel accuracy of some LULC classes compared to some of the maps used as input specially in mixed-pixels.

Acknowledgments

The research was developed in the Young Scientists Summer Program at the International Institute for Applied Systems Analysis, Laxenburg (Austria) with financial support from the Barry Callebault Group.

About the author

Caterina Barrasso is PhD candidate at the German Center for Integrative Biodiversity Research (iDiv) in Leipzig, Germany. (Contact: caterina.barrasso@idiv.de)

1.0 INTRODUCTION

Land-use/land-cover (LULC) is an important driver in many of the studies involving the Earth surface, such as climate, food security, hydrology, nutrient cycling, soil erosion, atmospheric quality, conservation biology, plant functioning and ecosystems assessment. The possibility of land monitoring has increased in the last decade, and multiple global LULC maps have emerged to map the current global LULC, as well as to measure its change. To better understand LULC conversions, it is important to have an accurate understanding of the nature and distribution of LULC at high spatiotemporal scale (Verburg et al. 2011). Despite the continuous increased scale of remote sensing observations, problematic levels of uncertainty are associated with derived global LULC maps, and discrepancies are observed during maps inter-comparison (Fritz and See 2008, Fritz et al. 2011). Their spatio-temporal uncertainty affects downstream applications via propagation through models, therefore diminishing the reliability of their predictions (Seebach et al. 2012 from Schepaschenko et al. 2015, Estes et al. 2018). Therefore, there is a need for improvement of global LULC maps that can support scientific and policy applications (Szantoi et al. 2020).

Despite efforts in advancing mapping approaches, the thematic accuracy of LULC maps has not improved significantly and continues to be around 70% (Tsendbazar et al. 2016). Except from creating new maps, one way in which the different LULC maps can be used (and the spatio-temporal discrepancies reconciled) is by using data fusion methods. Data fusion is a domain based on the integration of data coming from a variety of sources, and the derived product could have a resulting thematic accuracy higher than any of the individual sources used as input. Some authors have already demonstrated the efficacy of data fusion methods and explored their utility to reconcile discrepancies between maps. For example, Jung et al. 2006 developed a fuzzy agreement scoring method to determine the synergies between global LULC products for modelling the carbon cycle. Fritz et al. 2011 used this synergy concept, in combination with expert knowledge, to rank LULC products at global and national scales, along with national crop statistics, and combined them into a single cropland layer for Africa. Tuanmu et al. 2014 created a global 1km consensus LULC product by using an accuracy-based integration approach centered on 4 global LULC maps. Schepaschenko et al. 2015 developed a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. The author showed where individual input datasets contributed to the final product, and demonstrated that all were needed to produce a consensus product with higher accuracy. See et al. 2015 reconciled a global hybrid LULC map with crowdsourcing and geographically weighted regression to obtain consensus information from 3 individual LULC maps. Tsendbazar et al. 2015 showed that data fusion of existing maps, especially when accounting for their relative merits, can improve the thematic accuracy. The results, therefore, demonstrated the added value of using reference datasets and geo-statistics to improve LULC maps. Lesiv et al. 2016 compared several data fusion methods using crowdsourced data as reference and available forest products as input. The author found that geographically-weighted regression (GWR) was the best performing method in areas of high disagreement between the inputs.

These authors made use of multi-scale co-existing maps in the data fusion. Upscaled high-resolution maps can provide higher accuracy than natively-coarse ones (Sun et al. 2018), but higher resolution products are not always more accurate than maps with coarser resolution (See et al. 2015). Thus, integrating information at multiple resolutions is essential for providing the most accurate information possible to data fusion approaches. To date, data fusion methods developed ultimately delivered another traditional discrete map via what is called ‘hard’ classification where each pixel unit is represented by the single LULC class covering the largest portion of the pixel (Li et al. 2014). Many areas on the ground are composed of a diverse mosaic of multiple LULC classes (mixed pixel or

heterogeneous area). Among the reasons of classification mismatch, especially in coarse maps and heterogeneous areas characterized by mixed pixels (Stehman and Foody 2019, Fritz et al. 2011), is that the landscape heterogeneity is often ignored by reporting only the LULC class covering the largest portion of a pixel. The magnitude of the problem is a function of the relationship between the image spatial resolution and the landscape mosaic on the ground. The 'hard' classification, therefore, simplifies reality and has implications for both the accuracy assessment and the applications derived from LULC maps, with a decrease in accuracy when fractional information is considered as reality (Fonte et al. 2020). To date only few LULC maps, developed via 'soft' classification methods, report the fractional information of the co-existing LULC classes per pixel unit. Copernicus provides the Copernicus Global Land Cover Layers (CGLS-LC100) at 100 m resolution with fractions for every major LULC class between the time-period 2015-2019 (Buchhorn et al. 2019). The global layers (VCF) provide fractional information at 5000 m during a longer time-period (1982-2016) but only for 3 major LULC classes (Hansen et al. 2018).

Yet, the potential of data fusion to harness the complementary information from multiresolution LULC maps, and derive more accurate ones with fractional information, remains underexplored. LULC at high spatio-temporal scale is important in order to have an accurate understanding of LULC changes (Verburg et al. 2011), but medium resolution products are also still extremely useful from a modelling and assessment point of view, when the issue in not to improve the resolution but simply to improve the accuracy. In this study, the overarching goal is therefore to test the efficacy of data fusion, for multiscale and multiclass LULC maps, to improve fractional information at 100 m. The tested models can potentially be used for future improvement of traditionally (hard) classified global maps meantime there is a lack of fractional classified global maps since they offer low spatial resolution or temporal coverage to date.

This study addresses the following research questions:

I. Can data fusion methods, for multiscale and multiclass LULC maps, be used to improve fractional information at medium resolution?

II. Can the models improve the accuracy assessment in both homogeneous and heterogeneous areas?

To answer the above-mentioned research questions we selected 4 study areas where to test two data fusion models for the base year 2015 (when there was availability of a training dataset), namely Dirichlet and Beta distributed linear regression models. The successful achievement of the overarching goal relied on the availability of a quantitative training dataset with fraction information for each target LULC class. We used Copernicus (Buchhorn et al. 2019) ground-truth fractional information at 100 m as reference database and seven multiscale (10-5000 m resolution) derived LULC maps as predictors. The fitted models were then used to predict LULC fraction information for cropland, grassland, shrubland, forest, urban and built-up area, bare land, water body and 'other'. The accuracy of the predictions obtained with data fusion were then compared to the corresponding input maps used as predictors in both homogeneous areas characterized by pure-pixels (one LULC class per pixel unit) and heterogeneous areas to understand if the models can work with both types of landscape complexity.

2.0 METHODOLOGY

The workflow used during the study is reported in Figure 1. Four study areas were selected to test the models (described in section 2.1). The input data used are the ground-truth database with fraction information (described in section 2.2) and the seven derived remote sensing LULC maps (described in

section 2.3). The input data were pre-processed as described in the sections 2.2 and 2.3, and then used for models' training detailed in section 2.4. The best models were selected to make fraction predictions for each LULC class as described in 2.4, and the results compared to each input map used in the data fusion by using the sub-pixel accuracy matrix described in 2.5

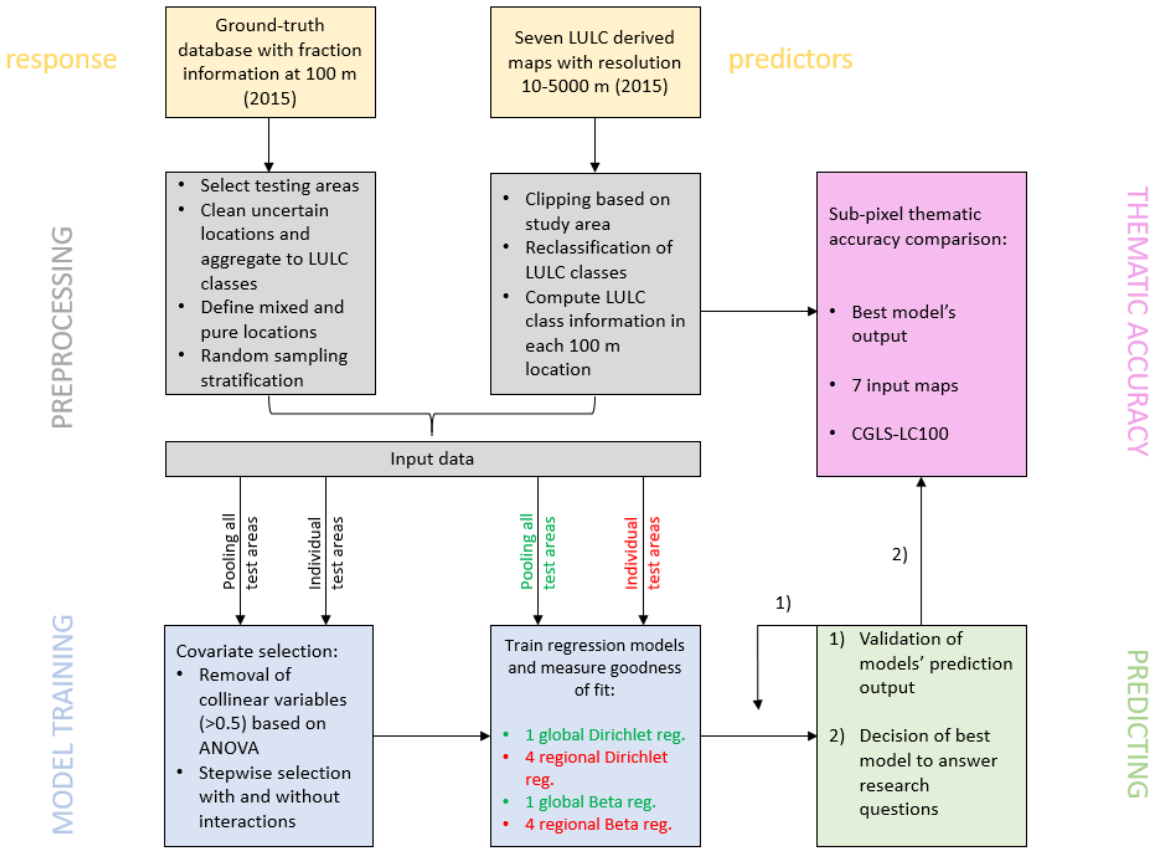

Figure 1. Data fusion workflow diagram. Input data are shown in yellow (response and predictors). Pre-processing of the input data, that were pooled together for all testing areas and used individually, is shown in gray and the models training steps in blue. In green are indicated the global Dirichlet and Beta regressions fitted by pooling together all testing areas, while in red are indicated the regional Dirichlet and Beta regressions fitted in each individual study area. The module for prediction is shown in green, at the end of which model selection was performed. The selected model was used to make predictions and compare the sub-pixel thematic accuracy between the model's output and each of the predictors used in the data fusion (module in pink).

2.1 Study areas

Four testing areas (Figure 2, panel A) were selected by taking into account both, the spatial density of locations available in the ground-truth sub-pixel database and the two different types of landscape (homogeneous and heterogeneous areas) (section 2.2). Preferred areas were those in which some of the locations were also checked for LULC change during the time-period 2015-2019. The reason for this choice is because the work may be extended to improve the fractional information not only spatially but also in time.

2.2 Ground-truth sub-pixel database

A 10 m ground-truth sub-pixel database was used as reference information during the data fusion. The database reports fractional information of LULC classes at each location, where the sum of the co-occurring proportions is always 100. It was developed by expert judgement on the Geo-Wiki platform

(http://www.geo-wiki.org/) to build and validate the 100 m CGLS-LC100 layers (Buchhorn et al. 2019). LULC information was then aggregated in each 100 m grid to obtain the per class fraction information.

At each available location we superimposed a 100x100 m grid based on the pixels of CGLS-LC100. LULC fractional information for fallow, bare land, cropland, lichens, shrubland, grassland, snow and ice, forest, urban and built-up, water body, wetland and burnt area was available. The LULC sub-pixel information of lichens, wetland and snow and ice was aggregated into a LULC class called 'other' as they were deemed rare classes and barely represented in derived remote sensing LULC maps. While the locations with fallow and burnt area greater than zero where removed from the database because the former may be confused with cropland, while the latter could be any of the LULC classes. Subsequently to the above pre-processing steps and cleaning of uncertain locations the LULC fraction information for bare land, cropland, shrubland, grassland, forest, urban and built-up, water body and 'other' class was available at 13,254 locations pooled across the four testing areas.

Based on the sub-pixel database, each location was categorized as pure or mixed-pixel. Pure pixels were those in which one of the LULC classes covered 100% of the 100 m grid. Mixed pixels were those in which none of the LULC classes covered 100% of the 100 m grid. In Western North-America, South-western Europe and Eastern South-Africa more than 50% of the locations were mixed-pixels (Figure 2, panel B). South-western Europe was the study area with the lowest proportion of pure-pixels (still above 10%). While in Tropical South East Asia the fraction of pure and mixed-pixels was 50-50.

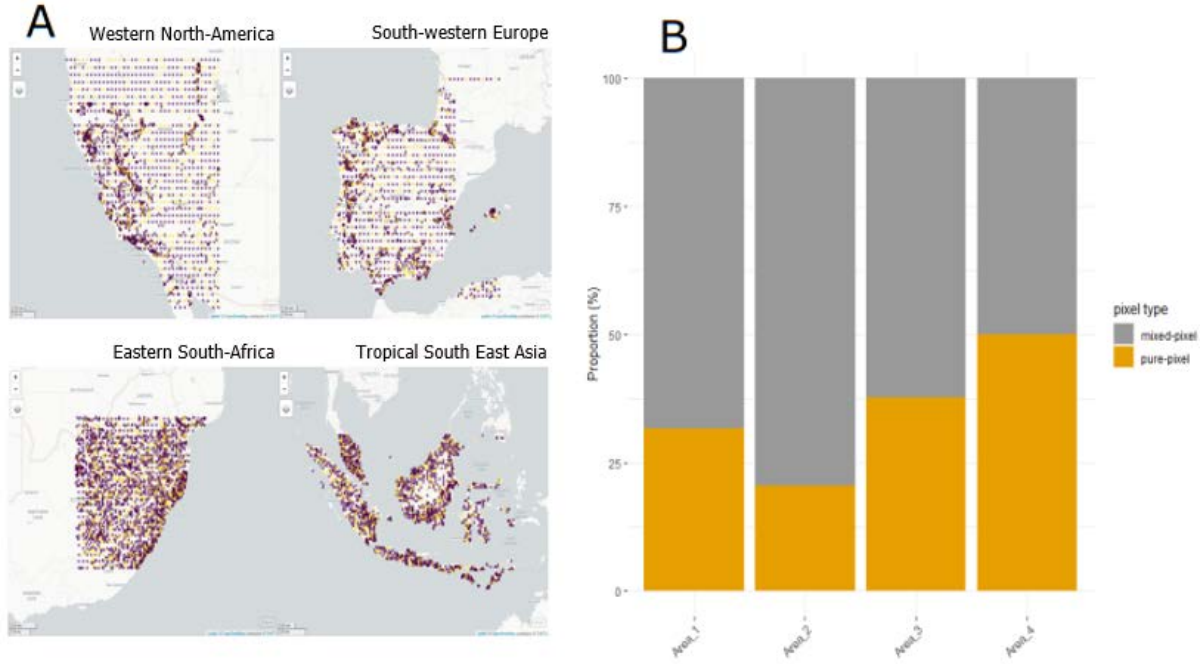

Figure 2. Characterization of the ground-truth sub-pixel database. Panel A shows the locations of the database that were split between training in purple (70%) and validation in yellow (30%). The four selected areas were: Western North-America (upper left – area_1), area_2 (upper-right, South-western Europe), area_3 (bottom-left, Eastern South-Africa) and area_4 (bottom-right, Tropical South East Asia). Panel B shows the proportion of pure (orange) and mixed (grey) pixels in each of the testing areas.

The sub-pixel database was used for training and validation of the models. Thus, it was split between training and validation locations by doing random sampling with the function *createDataPartition* in the 'caret' (version 6.0-88) R package. For each testing area, 70% of the locations were retained for training

the models while 30% for their validation. In the latter locations the models were used to make predictions (as described in 2.4.3) and measure the sub-pixel thematic accuracy.

A common issue with LULC fractional data as input in regression models is data imbalance. The more LULC classes are mapped, the more likely it is that some of the classes are not present in a given location (have 0% fraction). The problem is shown in Figure 3 with fractional information's distribution per LULC class pooled in all 4 testing areas. How to deal with the high presence of zeros in the data is treated in section 2.4.1.

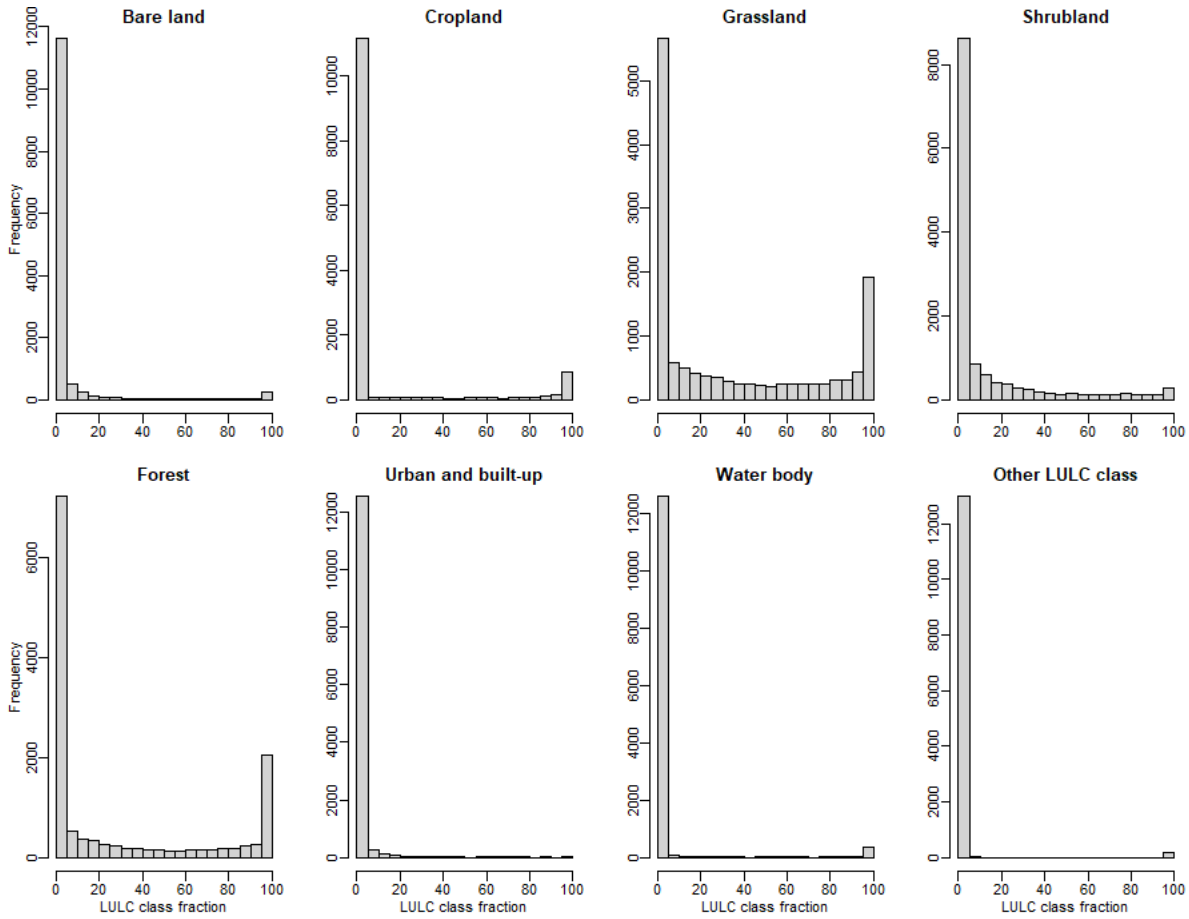

Figure 3. Fraction distribution. Fraction distributions for each of the 8 LULC classes in the ground-truth sub-pixel database.

The most abundant classes were cropland, grassland, shrubland, forest and bare land (Table 1). The rare ones were water body, urban and built-up area and 'other' LULC class. Exceptions were observed for cropland in Western North-America and bare land in Tropical South East Asia where not many locations were available with fraction greater than zero (Table 1). The spatial distribution of the LULC classes used during the study is reported in Appendix A (FigureA1-A4).

Table 1. Abundance* of LULC classes per study area in mixed and pure pixels

Region	cropland	grassland	shrubland	forest	water	bare	urban	other
North-America	112	2080	1591	1220	183	742	129	69
Europe	609	1975	1489	1669	149	730	393	33
North-Africa	914	3062	1853	1491	206	742	380	118
Asia-tropical	567	1223	1031	2726	227	307	282	61

*when fraction greater than zero

2.3 Predictors

Seven maps derived from remote sensing images were used as predictors in the tested models. To exploit the useful information coming from different scales, the spatial resolution of the predictors used was ranging between 10 and 5000 m. The number of LULC classes was different per predictor as reported in Table 2.

For each LULC map used as predictor we selected the year 2015 to have the same LULC information in time between the reference database and the predictors. Their LULC classes were reclassified to the 8 LULC classes of interest as reported in Appendix A Table A1. The Hansen map was the only one with fraction information for forest class. Thus, the values were not converted to a 'hard' classification and each pixel value greater than zero was considered as the fraction of forest present in the 30 m pixel of the derived map.

Table 2. Predictors used in the study

map	url	LULC classes	resolution (m)
GLASS	https://doi.pangaea.de/10.1594/PANGAEA.913496	bare land, cropland, grassland, forest, shrubland	5000
ESA-CCI	https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=form	bare land, cropland, grassland, forest, shrubland, water body, urban and built-up, other LULC	300
GLC-FCS30	https://zenodo.org/record/3986872#.YN7ZPkyxWUk	bare land, cropland, grassland, forest, shrubland, water body, urban and built-up, other LULC	30
GFSAD	https://e4ftl01.cr.usgs.gov/MEASURES/	cropland, water body	30
Hansen	http://earthenginepartners.appspot.com/science-2013-global-forest	forest	30
Pekel	https://global-surface-water.appspot.com/download;https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_2_YearlyHistory#terms-of-use	water body	30
WSF	https://springernature.figshare.com/articles/dataset/World	urban and built-up	30

_Settlement_Footprint_WSF_2015/10048412

To prepare each predictor for the data fusion we derived their information at 100 m resolution by following two different methods. In both methods, the 100 m extent of the reference locations was used to crop the predictors. The LULC maps with coarser resolution than 100 m (ESA-CCI and GLASS) were prepared as dummy variables (0-1) depending on the presence/absence of each LULC class as shown in Figure 4 (panel A). Because more than one LULC class could potentially be intersected, the one covering the largest portion of the grid was considered as presence (value 1) and the remaining LULC classes as absence (value 0) at each location. The LULC maps with resolution higher than 100 m (GLC-FCS30, GFSAD, Hansen, Pekel and WSF) were prepared in each location as fractions with a value between 0 and 1 indicating the percentage of the 100 m grid covered by the co-existing LULC classes (Figure 4, panel B).

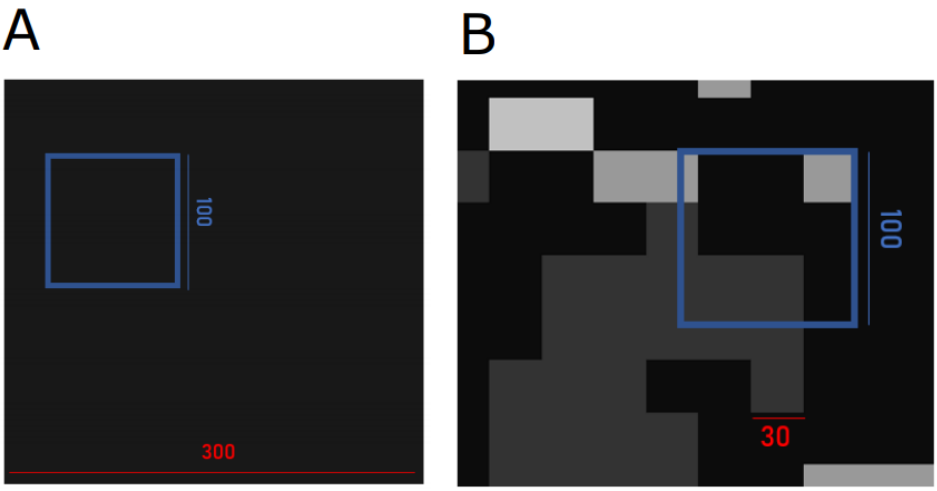

Figure 4. Predictors preparation. Blue grid is one location extent of the ground-truth sub-pixel database. Red grid is the spatial resolution of the predictors whose information is prepared at 100 m for the data fusion. Maps coarser than 100 m (panel A) were prepared as dummy variable (0-1). Maps with resolution higher than 100 m (panel B) were prepared as fractions indicating the percentage of the 100 m blue grid covered by the co-existing LULC classes (value between 0 and 1).

2.4 Modelling

2.4.1 Models

Statistical analysis of proportions can present numerous difficulties. By definition, the observations are limited to numerical values between, and including, 0 and 1 (or 0 and 100). A common issue with the use of LULC fraction data as input into regression models is data skewness, and the more classes are mapped the more likely it is that one or more classes are not present in a given pixel, leading to zero inflation (Figure 3). These properties of proportional data mean that the standard techniques of statistical analyses are usually not appropriate. A common recommendation is to apply a data transformation to meet the normality assumption and proceed with ordinary linear models, but the solution has important drawbacks with respect to interpretability and the validity of the resulting inference.

Generalized linear models (GLMs) extend linear regression to many types of response variables. When modelling proportional data coming from non-count data, two more flexible solutions than

transformation-based methods are available, namely Beta and Dirichlet regressions. The Beta regression is used to model univariate data, meaning that each LULC class is modelled independently and final rescaling is needed to ensure the unit-sum constraint. The second more sophisticated approach, the Dirichlet regression, is used to model multivariate data where the unit-sum constraint is maintained by an alpha parameter in the distribution that establishes the negative correlation between the LULC classes. For the Beta regression, the R package 'betareg' was used. For the Dirichlet regression, the R package 'DirichletReg' was used instead. Because of their implementation in a GLM-like setting, where a logit-link function establishes a linear relationship between the predictors and the response variable, the regressions cannot deal with extract zero and ones in the data. Thus, the following transformation was done:

$$p^* = \frac{p(n-1) + \frac{1}{C}}{n} \quad \text{Equation 1}$$

with *p* being the proportion of a LULC class, *n* the total number of observations in a dataset and *C* the number of categories (Smithson and Verkuilen 2006). The data is compressed symmetrically around 0.5, thus extreme values are more affected than values lying close to 0.5. With a large number of samples, the compression vanishes.

2.4.2 Model fitting

For both regressions we fitted one global model (non-spatial case) by pooling all study areas together (indicated in green in Figure 1), and 4 regional models (spatial case) for each of the study areas (indicated in red in Figure 1). Thus, a total of 10 models were fitted as shown in Appendix B (Table 1B). The models were trained by using the 70% of the sub-pixel database locations prepared as described in section 2.2, and indicated in purple in Figure 2A.

A problem which has long been acknowledged in regression modelling is that of collinearity among the predictor variables. Diagnostics to investigate the nature of collinearity should always be conducted and collinearity reduced in order to: i) delete redundancy between covariates, ii) avoid a biased assessment of variable importance (Strobl et al. 2008). Thus, the selection procedure for the predictors consisted in the following two steps: correlation analysis and step-wise selection. The correlation analysis was carried out as an initial step with the R package 'corrplot' to find the global and regional collinearity between predictors (Appendix B, Figure1B-5B). A quantified correlation higher than 0.5 was considered for further investigation with the step-wise selection. In the step-wise selection, for each set of LULC class predictors, the pair of covariates correlated above the 0.5 threshold were compared against the full model using ANOVA and the ones that gave the best model were used. Interaction terms between predictors were also checked, and the significant ones retained. The final list of selected predictors per fitted model is reported in column 2 of Table1B (Appendix B). For each LULC class the same set of predictors were used in the global and regional models, thus to compare the performance of Dirichlet versus Beta regression.

Location data often implies spatial dependence (spatial autocorrelation) and spatial heterogeneity (non-stationarity). Spatial heterogeneity was checked by using the function *gwr.sel* in the R package 'spgwr'. The spatial autocorrelation for Dirichlet regression was tested by using the Mantel test with the R package 'ade4' and Moran's I test for Beta regression using the R package 'moranfast'. Significant ($p<0.05$) spatial autocorrelation can bias model selection because spatially autocorrelated variables may be picked up as having a significant contribution to the fitted model. The spatial autocorrelation was detected, but not yet addressed in the study.

2.4.3 Goodness of fit and model selection

For each model fitted, the accuracy and the goodness of fit were measured for the observed and predicted values using the indicators: mean absolute error (MAE), root mean square error (RMSE) and R-squared (R^2) (Appendix B, Table 2B). The R^2 was calculated as the square of the correlation between the observed and predicted values. The MAE is not very influenced by a small number of large misclassifications, thus is more indicative of the overall model accuracy. RMSE is more indicative of the presence of large errors. R^2 is used to measure the goodness of fit of a model. This metric shows how far the predicted values are from a 1:1 line with the observed values and it was used for model's selection (see section 3.0). The indicators were measured separately per LULC class, but also for all LULC classes with the overall indicator that was calculated by taking the mean of the per-class means.

The best performing regression (with highest R^2) was then selected for making predictions in the locations chosen for validation (30% per testing area indicated in yellow in Figure 2A). Here the predicted values were used to measure and compare the sub-pixel thematic accuracy between the model's output and each of the LULC maps used as input in the data fusion.

2.5 Thematic accuracy

The model's predictions were post-stratified in mixed and pure-pixels to measure if the thematic accuracy obtained per study area was improved with the data fusion model in both types of landscape complexities. The thematic accuracy was compared versus each input map used as predictor and the CGLS-LC100 layer. The latter was not used as predictor during the modelling because the reference database used in the study was used to develop the CGLS-LC100 maps.

To measure the thematic accuracy, we used the sub-pixel confusion-uncertainty matrix (SCM) for multi-class classification developed by Silvan-Cardenas et al. 2008 and implemented in code by Masiliunas et al. 2021. The SCM is an adaptation of the classical confusion matrix to fractional data. We used the MIN-PROD operator as recommended by the authors and proposed by Pontius and Cheuk (2006). When using this operator, the diagonal of the matrix expressed the maximum overlap (minimum fraction - MIN) between the target and the predicted class fractions. The off-diagonal, instead, is an expression of which classes the non-overlapping fractions should belong to and it is calculated as the product (PROD) between the reference and the predicted class fraction.

The overall accuracy was compared between the predicted values and the maps ESA-CCI, GLC-FCS30 and Copernicus since these are the only maps with all the 8 LULC classes of interest. The per-class accuracy (user's accuracy) allowed comparison with more input maps. E.g. for the class forest the accuracy comparison was done between the predicted values and the maps: ESA-CCI, GLC-FCS30, Copernicus, GLASS and Hansen.

Significant differences between sub-pixel accuracies were shown by fitting a GLM from the R package 'glm' with family quasibinomial and logit as link function. The least squares means were compared with the R package 'lmerTest' and the letters display for the pairwise comparisons were computed with the R package 'multcomp'. The results of the analysis are shown in section 3.0.

3.0 RESULTS

When comparing the global models, Dirichlet regression resulted in slightly better fit than Beta regression for the LULC classes forest and water body (Table 3) but, for the overall and remaining LULC classes, higher goodness of fit was obtained with Beta regression. Exceptions are the classes cropland and 'other' for which the two regressions gave the same goodness of fit (Table 3).

When comparing the regional models, in Eastern South-Africa the Dirichlet regression gave slightly higher R^2 when modelling the fraction of grassland (Table 3). In Western North-America higher R^2 was observed with Dirichlet regression for the classes bare land, grassland, shrubland, forest and water body. In Tropical South East Asia Dirichlet regression did not give higher R^2 than Beta regression with any of the LULC classes (Table 3). While in South-western Europe higher R^2 was observed with Dirichlet regression for the classes cropland, shrubland, forest, urban and built-up and water body.

Although with some LULC classes the R^2 was slightly higher when fitting Dirichlet regression, the higher goodness of fit of some classes did not compensate the overall indicator. The same result was observed with the overall indicators MAE and RMSE that were lower (thus indicating better accuracy) when using Beta regression in both global and regional contexts (Appendix B, Table 2B). The result indicated that the best model to use for the data fusion, and obtain the predicted fractions per LULC class, was the Beta regression (Table 3).

When comparing the R^2 obtained by fitting one global and 4 regional Beta regressions it was observed that higher overall R^2 was reached with the global model (Table 3), but regional models showed higher R^2 for some LULC classes. For example, in Western North-America higher R^2 compared to the global model was observed for the classes bare land and cropland. For the LULC class urban and built-up higher R^2 was observed by fitting regional models in all testing areas except Tropical South East Asia (Table 3).

Table 3. R^2 of the fitted models per LULC class and overall

	Bare land	Cropland	Grassland	Shrubland	Forest	Urban and built-up	Water body	Other LULC	Overall
Global Dirichlet	0.28	0.26	0.20	0.04	0.50	0.44	0.43	0.01	0.27
Dirichlet_Africa	0.01	0.37	0.15	0.05	0.28	0.66	0.54	0.01	0.26
Dirichlet_America	0.33	0.21	0.16	0.05	0.50	0.53	0.51	0.01	0.29
Dirichlet _Asia	0.01	0.20	0.10	0.04	0.31	0.35	0.27	0.01	0.16
Dirichlet _Europe	0.03	0.22	0.12	0.04	0.34	0.28	0.60	0.01	0.20
Global Beta	0.29	0.26	0.36	0.25	0.49	0.46	0.40	0.01	0.31
Beta_Africa	0.04	0.38	0.11	0.21	0.28	0.66	0.54	0.03	0.28
Beta _America	0.32	0.32	0.11	0.03	0.49	0.61	0.43	0.01	0.29
Beta _Asia	0.01	0.23	0.28	0.22	0.34	0.47	0.34	0.01	0.24
Beta _Europe	0.05	0.21	0.33	0.03	0.32	0.27	0.55	0.01	0.22

*in red R^2 below 0.10, in yellow R^2 between 0.10 and 0.39, in green R^2 equal or greater than 0.40

For the LULC classes bare land, cropland, grassland, shrubland and 'other', the R^2 values were very low (indicated in red and yellow in Table 3). Therefore, during the study seemed feasible to use the Beta regressions to make predictions only for the LULC classes forest, urban and built-up and water body. The fractions of the remaining LULC classes were thus aggregated into the 'other' LULC class. Because for some LULC classes, as urban and built-up, higher R^2 was observed when fitting the regional regressions (Table 3) we decided to select the regional Beta linear models to test the potential of data fusion and answer the research questions of the study.

By restricting the analysis to the 4 LULC classes (forest, urban and built-up, water body and 'other' class) higher goodness of fit for the 'other' LULC class (compared to Table 3) was observed, in both pure and mixed-pixels, with the regional Beta regressions (Appendix B, Figure 6B). When looking at the validation locations, higher R^2 was always observed in pure-pixels compared to mixed-pixels (Appendix B, Figure 6B) except in Tropical South East Asia where slightly higher R^2, with the 'other' LULC class, was observed in mixed-pixels. In the validation locations, after post-stratification of the predictions between pure and mixed-pixels, low R^2 was observed in Tropical South East Asia for the LULC classes forest, water body and 'other' in mixed-pixels and for the 'other' LULC class in pure-pixels (Appendix B, Figure 6B). In Western North America low R^2 was observed in mixed-pixels for the LULC classes water body and 'other'. While in South-western Europe low R^2 was observed in mixed-pixels for the classes forest and 'other' (Appendix B, Figure 6B).

The overall sub-pixel accuracy of the Copernicus layer, pooled across the four testing areas, was significantly higher compared to the predictors ESA-CCI, GLC-FCS30 and the predictions obtained with the Beta regression (Figure 5, panel A). While no significant difference was observed between the Beta regressions' output and the predictors ESA-CCI and GLC-FCS30. The result was expected since the Copernicus layer was developed and validated by using the reference database described in section 2.2. Same result was observed for the overall accuracy in mixed-pixels (Figure 5, panel B). In pure-pixels, instead, significantly higher overall accuracy was observed for the maps ESA-CCI, GLC-FCS30 and Copernicus although non-significant difference was observed between the Beta regressions' output, ESA-CCI and GLC-FCS30. The result indicated that slightly better overall accuracy with the data fusion model was obtained in mixed-pixels.

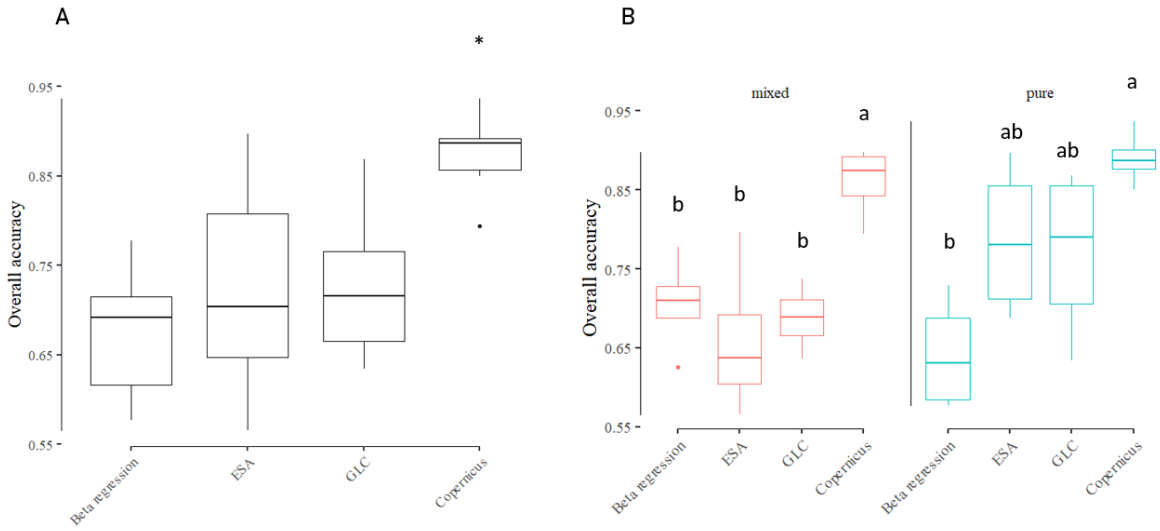

Figure 5. Overall accuracy comparison (pooled study areas). Panel A shows the overall accuracy with non-distinction between mixed and pure-pixels. Panel B shows the overall accuracy in mixed-pixels (pink box-plots on the left) and pure-pixels (blue box-plots on the right).

For the forest class significantly higher accuracy, compared to the Beta regressions' output, was observed with the maps Hansen and Copernicus in mixed-pixels (Figure 6) although the Hansen layer did not show the forest accuracy to be significantly higher than the Beta regressions' output and the maps ESA-CCI, GLC-FCS30 and GLASS. While in pure-pixels the accuracy of forest was the same across all maps.

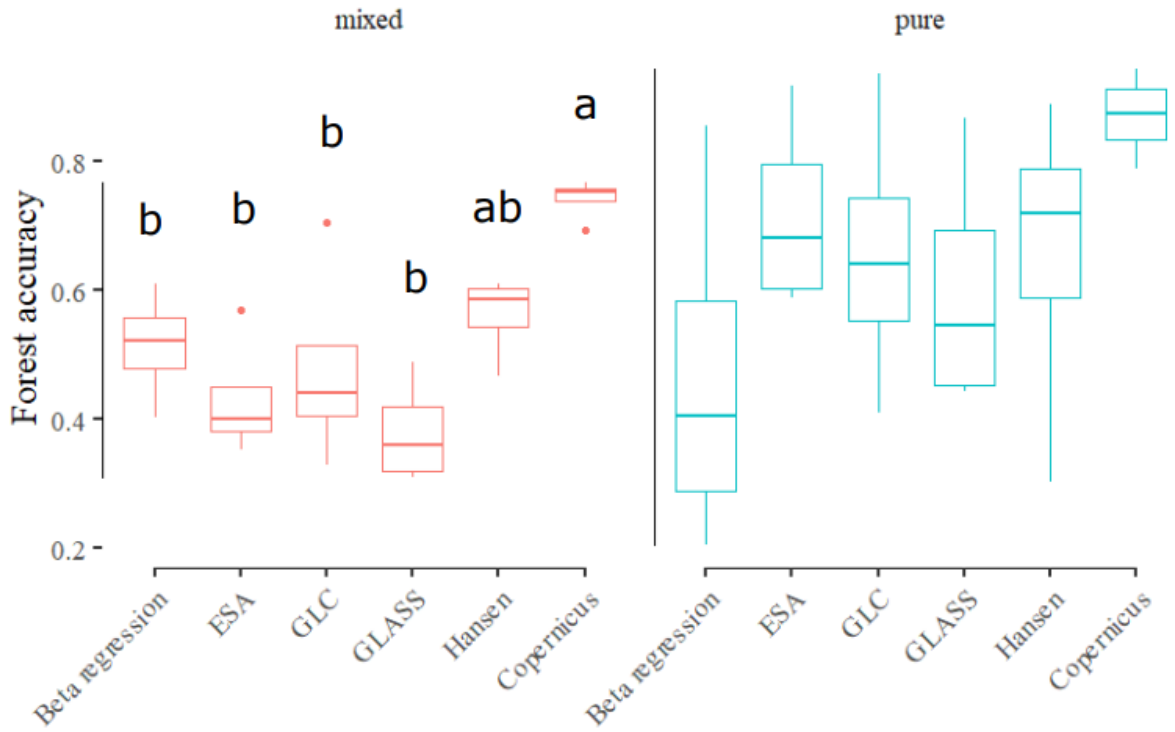

Figure 6. Forest accuracy comparison (pooled study areas). Forest accuracy in mixed-pixels (pink box-plots on the left) and pure-pixels (blue box-plots on the right).

Significantly higher urban accuracy was observed in mixed-pixels only with the map WSF (Figure 7). As with the forest accuracy, in pure-pixels difference in urban accuracy between maps was not observed. Significantly higher water body accuracy, compared to the Beta regressions' output, was observed with the Copernicus map in mixed-pixels (Figure 8). While in pure-pixels the maps GLC-FCS30, GFSAD and Copernicus showed significantly higher water accuracy compared to the Beta regressions' output. The 'other' LULC class accuracy was not significantly different between any of the maps in both mixed and pure-pixels (Figure 9).

Although the data fusion with the regional Beta regressions seemed to not give significant improvement compared to any of the predictors used as input, in Eastern South-Africa higher forest accuracy was observed with the data fusion in mixed-pixels compared to the predictors ESA-CCI, GLC-FCS30 and GLASS (Appendix B, Figure 7B). In pure-pixels instead, the forest accuracy was never higher than any of the predictors. In the same study area, the urban accuracy obtained with the regional Beta regression in mixed-pixels was higher compared to all predictors except WSF, while lower than all predictors in pure-pixels (Appendix B, Figure 7B). The water body accuracy obtained with the regional Beta regression was lower than any of the predictors in both mixed and pure-pixels (Appendix B, Figure 7B). While the 'other' LULC class accuracy obtained in mixed-pixels with the regional Beta regression was higher than the maps ESA-CCI, GLASS and Hansen, but never higher in pure-pixels (Appendix B, Figure 7B).

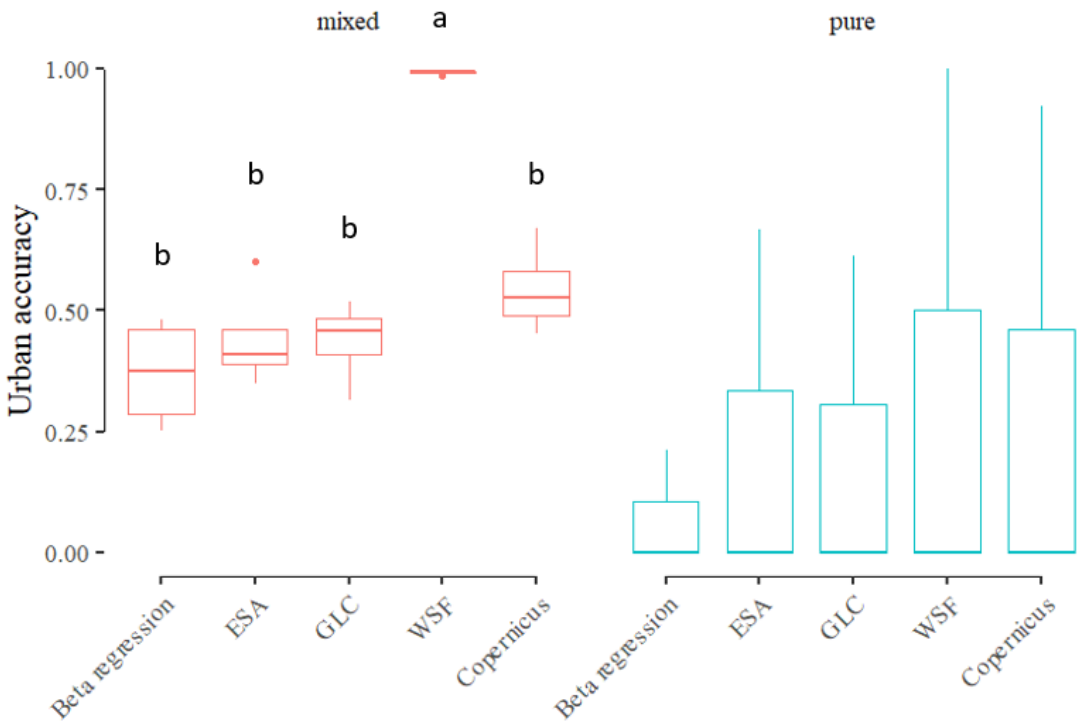

Figure 7. Urban and built-up accuracy comparison (pooled study areas). Urban and built-up accuracy in mixed-pixels (pink box-plots on the left) and pure-pixels (blue box-plots on the right).

In Tropical South-East Asia higher forest accuracy was observed with the data fusion in mixed-pixels compared to the predictors ESA-CCI, GLASS and Hansen (Appendix B, Figure 7B). In pure-pixels instead, the forest accuracy was never higher than any of the predictors. In the same study area, the urban accuracy obtained with the regional Beta regression in mixed-pixels was higher compared to the predictors ESA-CCI and GLC-FCS30, while always zero in pure-pixels (Appendix B, Figure 7B). The same result in pure-pixel was observed in Western North-America and South-western Europe. This is happening when the fraction of a class is always zero (as shown in Appendix B, Table 3B). In these cases, the sub-pixel thematic accuracy considers zero as maximum overlap between reference and predicted class fraction causing the accuracy to be zero. In Tropical South-East Asia the water body accuracy obtained with the regional Beta regression was lower than any of the predictors in pure-pixels, but higher than the map GFSAD in mixed-pixels (Appendix B, Figure 7B). While the 'other' LULC class accuracy obtained in mixed-pixels with the regional Beta regression was higher than the maps ESA-CCI, GLASS and GLC-FCS30, but never higher in pure-pixels as in Eastern South-Africa (Appendix B, Table 3B).

In Western North-America higher forest accuracy was observed with the data fusion in mixed-pixels compared to the predictors ESA-CCI, GLC-FCS30 and GLASS as in Eastern South-Africa (Appendix B, Figure 7B). In pure-pixels instead, the forest accuracy was never higher than any of the predictors. In the same study area, the urban, water body and 'other' class accuracies obtained with the regional Beta regression in both mixed and pure-pixels were never higher than any of the predictors (Appendix B, Figure 7B).

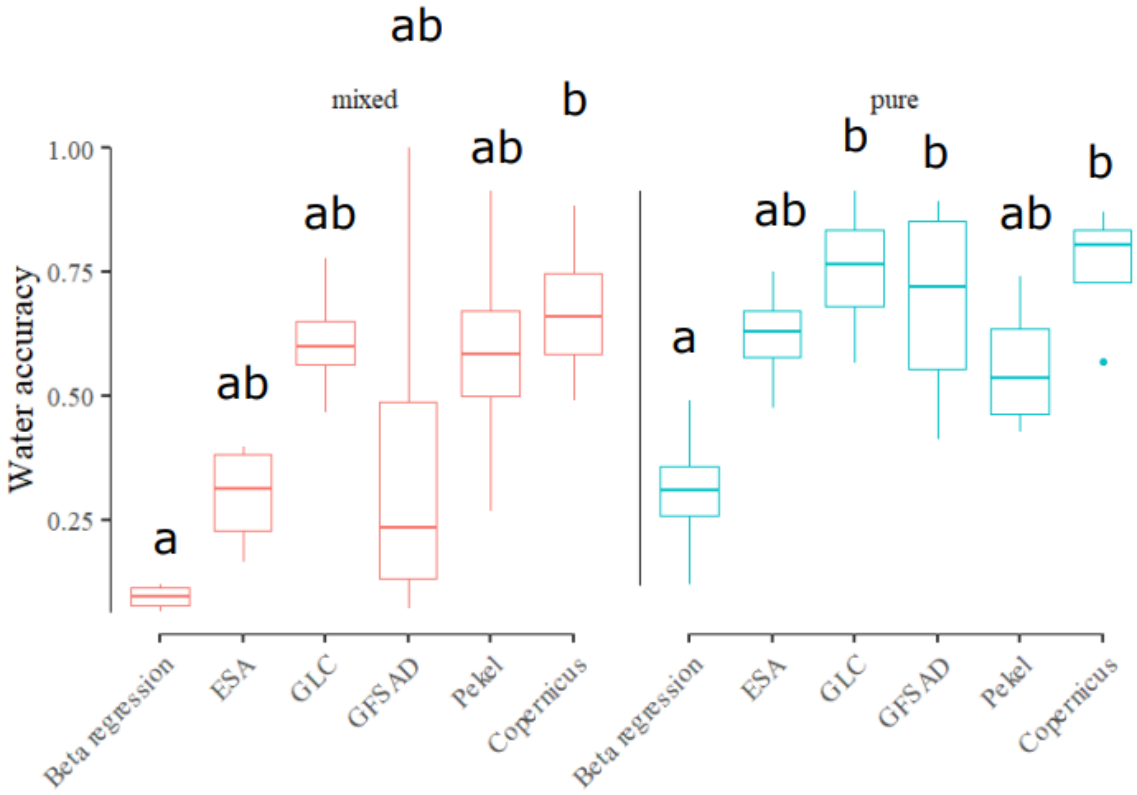

Figure 8. Water body LULC class accuracy comparison (pooled study areas). Water body accuracy in mixed-pixels (pink box-plots on the left) and pure-pixels (blue box-plots on the right).

In South-western Europe higher forest accuracy was observed with the data fusion in mixed-pixels compared to the predictors ESA-CCI, GLC-FCS30 and GLASS as in Eastern South-Africa and Western North-America (Appendix B, Figure 7B). In pure-pixels instead, the forest accuracy was only higher than the GLASS map. In the same study area, the urban and water body accuracies obtained with the regional Beta regression in both mixed and pure-pixels were never higher than any of the predictors (Appendix B, Figure 7B). While the 'other' accuracy was higher than the Hansen and GLASS maps in mixed-pixels and lower than any map in pure-pixels.

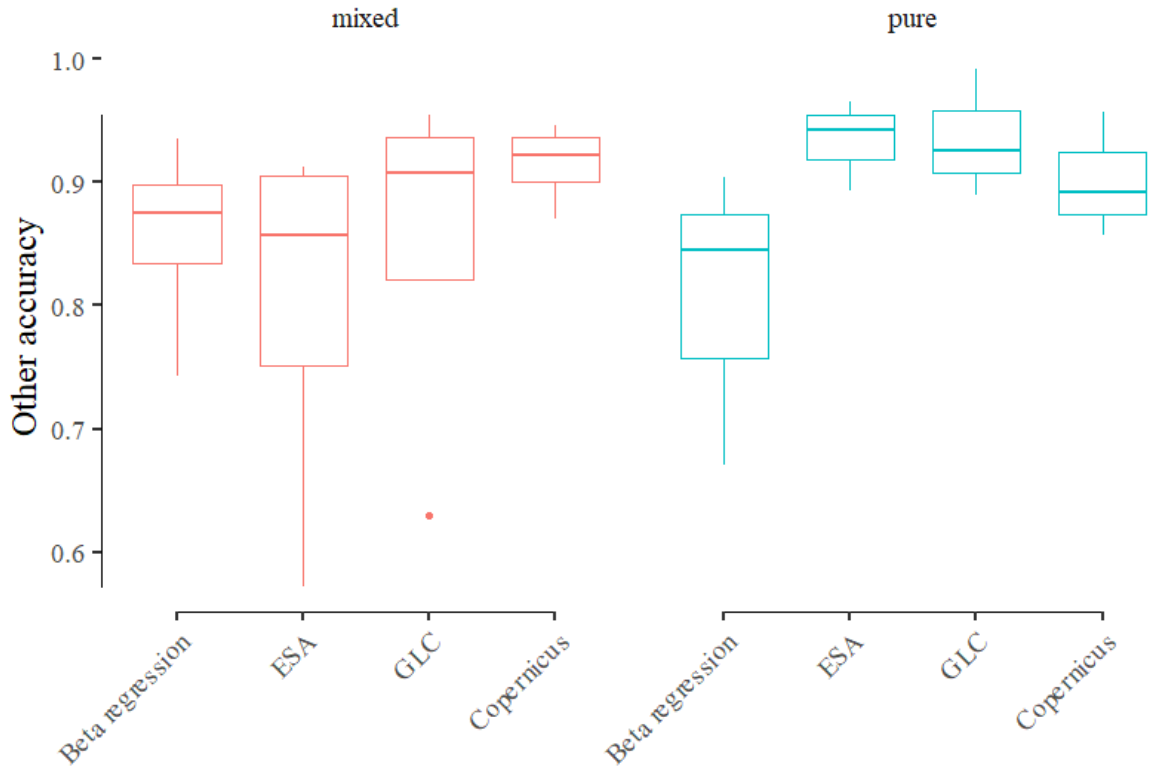

Figure 9. 'Other' LULC class accuracy comparison (pooled study areas). 'Other' LULC class accuracy in mixed-pixels (pink box-plots on the left) and pure-pixels (blue box-plots on the right).

4.0 DISCUSSION

Several points of improvement (detailed below) are needed to better answer the research questions of the study, but it was possible to observe that data fusion of multiscale and multiclass LULC maps can improve the fractional information of certain LULC classes compared to certain LULC maps used as input in the data fusion, especially in mixed-pixels. For example, in Eastern South-Africa, higher forest accuracy was reached with the data fusion's output in mixed-pixels compared to the predictors ESA-CCI, GLC-FCS30 and GLASS (Appendix B, Figure 7B), although it was found not significantly different (Figure 6). In Figure 6 the forest accuracy observed with Hansen's layer was also not significantly higher than ESA-CCI, GLC-FCS30 and GLASS, despite the Hansen's map is globally accepted as the best available product to map forest. The result indicates that improvement of forest accuracy was obtained in mixed-pixels with the data fusion in comparison to the maps ESA-CCI, GLC-FCS30 and GLASS (Figure 6).

While Dirichlet regression may seem a natural choice for modelling data with a unit-sum constraint, the implicit pairwise correlations established between all classes by the alpha parameter described in section 2.4.1 can be overly restrictive. Additionally, the fact that the mean of the Dirichlet distribution determines the covariance structure between all classes at each location is also restrictive. The model indeed has been criticized by authors such as Leininger et al. 2013. In this study we found that the LULC class fractions were better modelled by using the Beta regression followed by a post-rescaling of the predictions (Table 3).

Forest, urban and built-up and water body are easier LULC classes to be mapped than bare land, cropland, grassland and shrubland via remote sensing images. The confusion of their classification is linked, for example, to the rotation between bare land and cropland and the similar spectra signal between cropland and grassland. Depending on the used (and available) remote sensing images, as well as the time of collection of ground-truth data during the year, classified maps can disagree more easily on these LULC classes more difficult to be mapped. In this study was not possible to fit good models for the LULC classes bare land, cropland, grassland and shrubland (Table 3). In order to improve the goodness of fit to model the fractions of these classes, in the future, more accurate LULC maps will be used as predictors. In addition, localized weighting of the predictors could potentially help achieving higher goodness of fit.

Although non-significant higher overall (Figure 5) as well as forest, urban and built-up, water body and 'other' accuracies (Figure 6-9) were obtained with the Beta regressions' output, in both mixed and pure-pixels, higher forest, urban and built-up and 'other' accuracies were observed in some testing areas as described in section 3.0, especially in mixed-pixels. The worst accuracy obtained with the Beta regressions' output was observed in mixed-pixels for the class water body (Appendix B, Figure 7B). A reason can be the abundance of the LULC class with a fraction equal 1. As shown in Appendix B, Table 3B, the number of locations with a fraction greater than 0 for the water body class, in both pure and mixed-pixels, is similar in all testing areas. The result indicates that there are as many locations in which the fraction of water body is 1 as there are mixed-pixels. This data imbalance leaves little training data in the middle for the regression model to learn from, causing non-good predictions in mixed-pixels. The use of zero/one-inflated models may help solve the data skewness since the zeros and ones are here modelled independently.

The use of zero/one-inflated models may also improve predictions in pure-pixels. As discussed in section 3.0 the predictions obtained with the regional Beta regressions showed slightly higher overall accuracy in mixed-pixels (Figure 5). One reason can be that the pure-pixels are penalized because the Beta regressions fitted via the R package 'betareg' cannot model exact zeros and ones and the data transformation with Equation 1 was needed.

In addition, to better answer the research questions of the study, higher goodness of fit of the data fusion model is needed. Besides the above mentioned zero/one-inflated models, model stratification can help in achieving more localized effects for each map used as predictor. Agro-ecological zones introduced as an additional fixed or random effect can help to further stratify the testing areas and potentially improve the modelling of the difficult LULC classes bare land, cropland, grassland and shrubland. As shown in Table 3, fitting the global Beta regression improved the R^2 for the LULC class bare land in Eastern South-Africa, Tropical South East Asia and South-western Europe and for the LULC class shrubland in Western North-America and South-western Europe. Thus, for the LULC classes difficult to map, the use of the entire sub-pixel database in one model may be advantageous.

Another future improvement will be to model the spatial autocorrelation detected, but not addressed, during the study. The reduction of the significant spatial autocorrelation identified with the Moran's I test will help in estimating less biased model parameters and comply with the GLM assumption of independence of residuals. The Beta linear regression model will, thus, be formulated to control for spatial autocorrelation.

In addition, modification of the sub-pixel thematic accuracy matrix to correctly cover cases of urban and built-up areas with pure-pixels characterized only by zero fraction is needed. As discussed in section 3.0, in these cases the sub-pixel thematic accuracy considers zero as maximum overlap between reference and predicted class fraction causing the accuracy to be zero despite the high R^2 shown in

Appendix B, Figure 6B. To adjust the sub-pixel thematic accuracy to correctly cover these cases, the MIN operator (see section 2.5) can be substituted by the difference in fractional information between the reference and the predicted class fractions.

REFERENCES

Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N.-E., Herold, M., Fritz, S. Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe. Zenodo. URL: https://doi.org/10.5281/zenodo.3243509.

Estes, Lyndon, Peng Chen, Stephanie Debats, Tom Evans, Stefanus Ferreira, Tobias Kuemmerle, Gabrielle Ragazzo, et al. 2018. "A Large-Area, Spatially Continuous Assessment of Land Cover Map Error and Its Impact on Downstream Analyses." Global Change Biology 24 (1): 322–337. https://doi.org/10.1111/gcb.13904.

Fonte, C. C., L. See, J. C. Laso-Bayas, M. Lesiv, and S. Fritz. 2020. "ASSESSING THE ACCURACY OF LAND USE LAND COVER (LULC) MAPS USING CLASS PROPORTIONS IN THE REFERENCE DATA." ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences V-3–2020 (August): 669–74. https://doi.org/10.5194/isprs-annals-V-3-2020-669-2020.

Fritz, Steffen, and Linda See. 2008. "Identifying and Quantifying Uncertainty and Spatial Disagreement in the Comparison of Global Land Cover for Different Applications." Global Change Biology 14 (5): 1057–1075. https://doi.org/10.1111/j.1365-2486.2007.01519.x.

Fritz, Steffen, Linda See, Ian McCallum, Christian Schill, Michael Obersteiner, Marijn van der Velde, Hannes Boettcher, Petr Havlík, and Frédéric Achard. 2011. "Highlighting Continued Uncertainty in Global Land Cover Maps for the User Community." Environmental Research Letters 6 (4): 044005. https://doi.org/10.1088/1748-9326/6/4/044005.

Hansen, M., Song, X. (2018). *Vegetation Continuous Fields (VCF) Yearly Global 0.05 Deg* [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2021-08-17 from https://doi.org/10.5067/MEaSUREs/VCF/VCF5KYR.001

Jung, Martin, Kathrin Henkel, Martin Herold, and Galina Churkina. 2006. "Exploiting Synergies of Global Land Cover Products for Carbon Cycle Modeling." Remote Sensing of Environment 101 (4): 534–553. https://doi.org/10.1016/j.rse.2006.01.020.

Leininger, Thomas J., Alan E. Gelfand, Jenica M. Allen, and John A. Silander. 2013. "Spatial Regression Modeling for Compositional Data With Many Zeros." Journal of Agricultural, Biological, and Environmental Statistics 18 (3): 314–34. https://doi.org/10.1007/s13253-013-0145-y.

Lesiv, Myroslava, Elena Moltchanova, Dmitry Schepaschenko, Linda See, Anatoly Shvidenko, Alexis Comber, and Steffen Fritz. 2016. "Comparison of Data Fusion Methods Using Crowdsourced Data in Creating a Hybrid Forest Cover Map." Remote Sensing 8 (3): 261. https://doi.org/10.3390/rs8030261.

Li, M., Zang, S., Zhang, B., Li, S., & Wu, C. (2014). A Review of Remote Sensing Image Classification Techniques: The Role of Spatio-contextual Information. European Journal of Remote Sensing, 47(1), 389–411. https://doi.org/10.5721/EuJRS20144723

Masiliūnas, Dainius, Nandin-Erdene Tsendbazar, Martin Herold, Myroslava Lesiv, Marcel Buchhorn, and Jan Verbesselt. 2021. "Global Land Characterisation Using Land Cover Fractions at 100 m Resolution." Remote Sensing of Environment 259 (June): 112409. https://doi.org/10.1016/j.rse.2021.112409.

Pontius, R. G., Jr., & Cheuk, M. L. (2006). A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. International Journal of Geographical Information Science, 20(1), 1−30.

Schepaschenko, Dmitry, Linda See, Myroslava Lesiv, Ian McCallum, Steffen Fritz, Carl Salk, Elena Moltchanova, et al. 2015. “Development of a Global Hybrid Forest Mask through the Synergy of Remote Sensing, Crowdsourcing and FAO Statistics.” Remote Sensing of Environment 162 (June): 208–20. https://doi.org/10.1016/j.rse.2015.02.011.

See, Linda, Dmitry Schepaschenko, Myroslava Lesiv, Ian McCallum, Steffen Fritz, Alexis Comber, Christoph Perger, et al. 2015. “Building a Hybrid Land Cover Map with Crowdsourcing and Geographically Weighted Regression.” ISPRS Journal of Photogrammetry and Remote Sensing 103 (May): 48–56. https://doi.org/10.1016/j.isprsjprs.2014.06.016.

Seebach, L., McCallum, I., Fritz, S., Kindermann, G., Leduc, S., Böttcher, H., et al. (2012). Choice of forest map has implications for policy analysis: A case study on the EU biofuel target. Environmental Science & Policy, 22, 13–24. http://dx.doi.org/10.1016/j. envsci.2012.04.010.

Silván-Cárdenas, J.L., and L. Wang. 2008. “Sub-Pixel Confusion–Uncertainty Matrix for Assessing Soft Classifications.” Remote Sensing of Environment 112 (3): 1081–95. https://doi.org/10.1016/j.rse.2007.07.017.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods, 11(1):54–71.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests, BMC Bioinformatics, 9, 307, https://doi.org/10.1186/1471-2105-9-307

Sun, Peijun, Russell G. Congalton, and Yaozhong Pan. 2018. “Improving the Upscaling of Land Cover Maps by Fusing Uncertainty and Spatial Structure Information.” Photogrammetric Engineering & Remote Sensing 84 (2): 87–100. https://doi.org/10.14358/PERS.84.2.87.

Szantoi, Zoltan, Gary N. Geller, Nandin-Erdene Tsendbazar, Linda See, Patrick Griffiths, Steffen Fritz, Peng Gong, Martin Herold, Brice Mora, and André Obregón. 2020. “Addressing the Need for Improved Land Cover Map Products for Policy Support.” Environmental Science & Policy 112 (October): 28–35. https://doi.org/10.1016/j.envsci.2020.04.005.

Tsendbazar, N.E.; de Bruin, S.; Mora, B.; Schouten, L.; Herold, M. Comparative assessment of thematic accuracy of GLC maps for specific applications using existing reference data. Int. J. Appl. Earth Obs. Geoinf. 2016, 44, 124–135

Tsendbazar, Nandin-Erdene, Sytze de Bruin, Steffen Fritz, and Martin Herold. 2015. “Spatial Accuracy Assessment and Integration of Global Land Cover Datasets.” Remote Sensing 7 (12): 15804–21. https://doi.org/10.3390/rs71215804.

Tuanmu, Mao-Ning, and Walter Jetz. 2014. “A Global 1-Km Consensus Land-Cover Product for Biodiversity and Ecosystem Modelling: Consensus Land Cover.” Global Ecology and Biogeography 23 (9): 1031–1045. https://doi.org/10.1111/geb.12182.

Verburg, Peter H., Kathleen Neumann, and Linda Nol. 2011. “Challenges in Using Land Use and Land Cover Data for Global Change Studies: LAND USE AND LAND COVER DATA FOR GLOBAL CHANGE STUDIES.” Global Change Biology 17 (2): 974–89. https://doi.org/10.1111/j.1365-2486.2010.02307.x.

APPENDIX

Appendix A

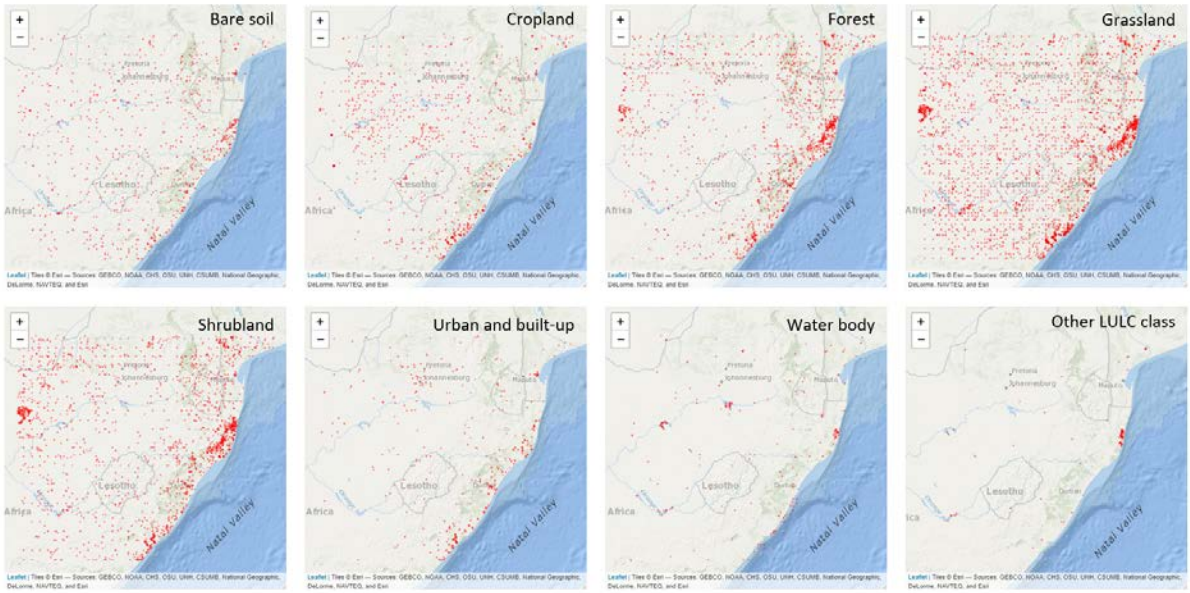

Figure A1. Spatial distribution of LULC classes in South-Africa. The 8 LULC classes (bare land, cropland, forest, grassland, shrubland, urban and built-up, water body and other LULC class) are plotted when their fraction is greater than zero.

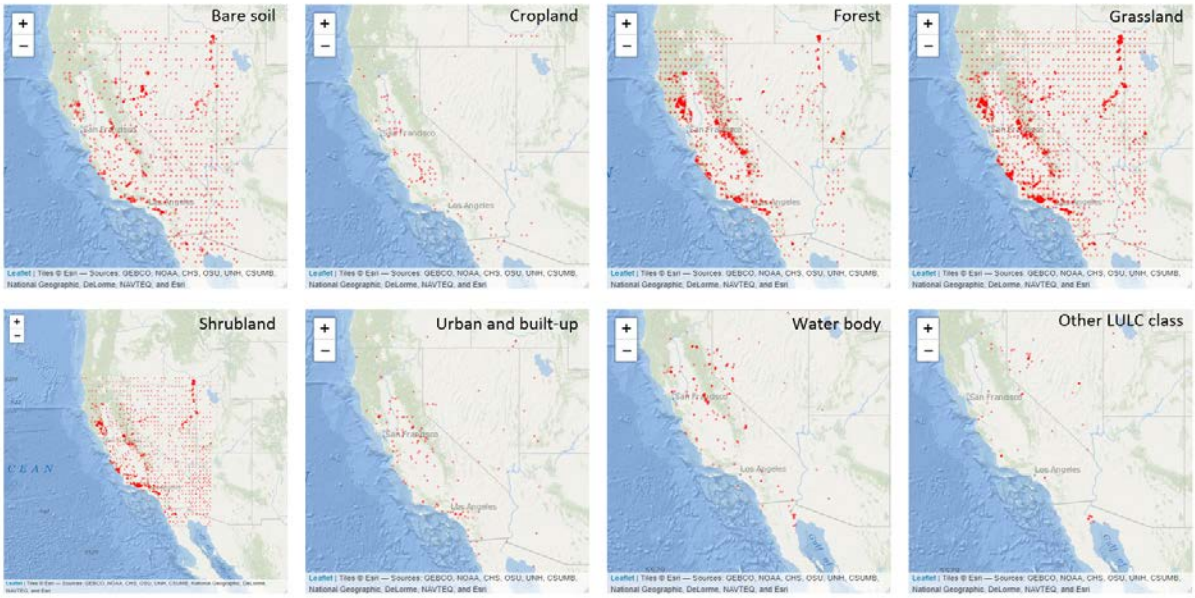

Figure A2. Spatial distribution of LULC classes in North-America. The 8 LULC classes (bare land, cropland, forest, grassland, shrubland, urban and built-up, water body and other LULC class) are plotted when their fraction is greater than zero.

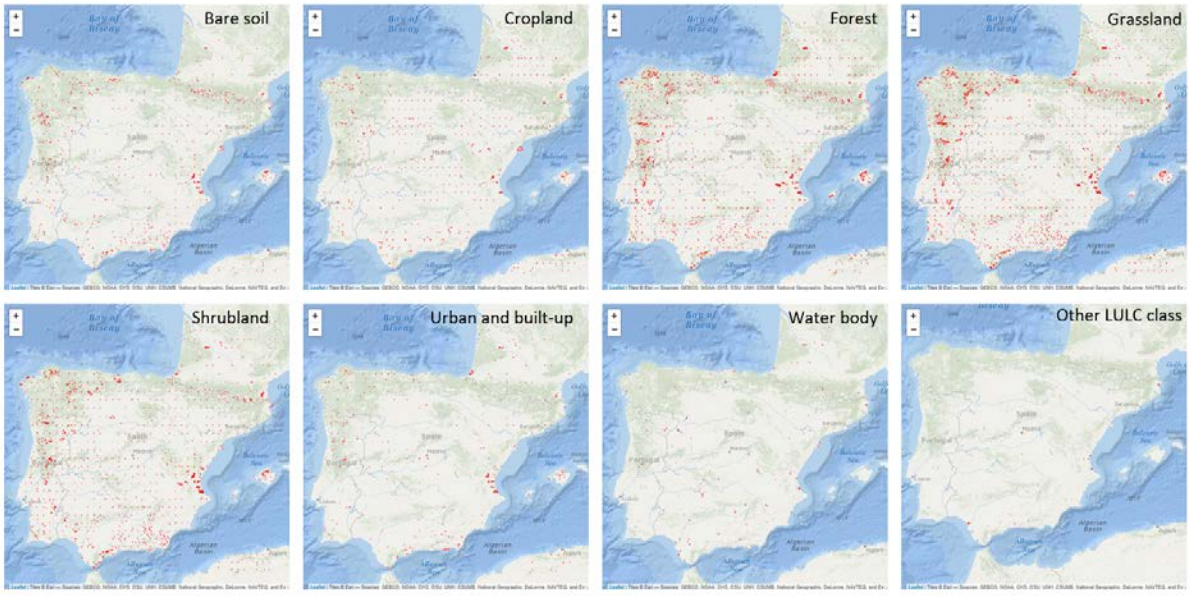

Figure A3. Spatial distribution of LULC classes in Europe. The 8 LULC classes (bare land, cropland, forest, grassland, shrubland, urban and built-up, water body and other LULC class) are plotted when their fraction is greater than zero.

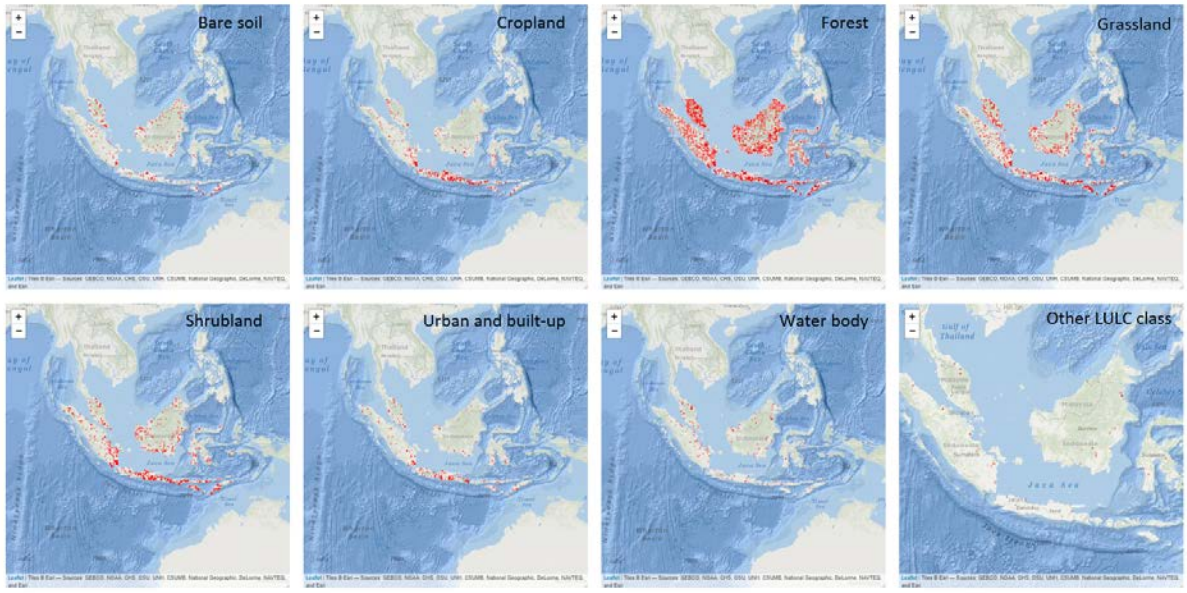

Figure A4. Spatial distribution of LULC classes in Asia-tropical. The 8 LULC classes (bare land, cropland, forest, grassland, shrubland, urban and built-up, water body and other LULC class) are plotted when their fraction is greater than zero.

Table A1. **Legend reconciliation (between original and translated terms) for the LULC classes of predictors used as input**

original_class	translated_class	translated_name	predictor
10	5	cropland	ESA-CCI
11	5	cropland	ESA-CCI
12	5	cropland	ESA-CCI
20	5	cropland	ESA-CCI
30	5	cropland	ESA-CCI

40	10	other	ESA-CCI
50	7	forest	ESA-CCI
60	7	forest	ESA-CCI
61	7	forest	ESA-CCI
62	7	forest	ESA-CCI
70	7	forest	ESA-CCI
71	7	forest	ESA-CCI
72	7	forest	ESA-CCI
80	7	forest	ESA-CCI
81	7	forest	ESA-CCI
82	7	forest	ESA-CCI
90	7	forest	ESA-CCI
100	7	forest	ESA-CCI
160	7	forest	ESA-CCI
170	7	forest	ESA-CCI
110	1	grassland	ESA-CCI
130	1	grassland	ESA-CCI
180	9	wetland	ESA-CCI
190	8	urban	ESA-CCI
120	2	shrubland	ESA-CCI
121	2	shrubland	ESA-CCI
122	2	shrubland	ESA-CCI
140	10	other	ESA-CCI
150	10	other	ESA-CCI
151	7	forest	ESA-CCI
152	2	shrubland	ESA-CCI
153	1	grassland	ESA-CCI
200	3	bare	ESA-CCI
201	3	bare	ESA-CCI
202	3	bare	ESA-CCI
210	6	water	ESA-CCI
220	4	snow	ESA-CCI
0	0	no data	GLC_FCS30
10	5	cropland	GLC_FCS30
11	10	other	GLC_FCS30
12	2	shrubland	GLC_FCS30
20	5	cropland	GLC_FCS30
50	7	forest	GLC_FCS30
60	7	forest	GLC_FCS30
61	7	forest	GLC_FCS30
62	7	forest	GLC_FCS30
70	7	forest	GLC_FCS30
71	7	forest	GLC_FCS30
72	7	forest	GLC_FCS30
80	7	forest	GLC_FCS30
81	7	forest	GLC_FCS30
82	7	forest	GLC_FCS30
90	7	forest	GLC_FCS30

120	2	shrubland	GLC_FCS30
121	2	shrubland	GLC_FCS30
122	2	shrubland	GLC_FCS30
130	1	grassland	GLC_FCS30
140	10	other	GLC_FCS30
150	10	other	GLC_FCS30
152	10	other	GLC_FCS30
152	2	shrubland	GLC_FCS30
180	9	wetland	GLC_FCS30
190	8	urban	GLC_FCS30
200	3	bare	GLC_FCS30
201	3	bare	GLC_FCS30
202	3	bare	GLC_FCS30
210	6	water	GLC_FCS30
220	4	snow	GLC_FCS30
250	0	no data	GLC_FCS30
0	6	water	GFSAD
1	10	other	GFSAD
2	5	cropland	GFSAD
0	10	other	WSF
255	8	urban	WSF
1	1	grassland	GLASS_GLC
2	2	shrubland	GLASS_GLC
3	3	bare	GLASS_GLC
5	5	cropland	GLASS_GLC
7	7	forest	GLASS_GLC
0	0	no data	GLASS_GLC
NA	10	other	Pekel
1	10	other	Pekel
2	6	water	Pekel
3	6	water	Pekel
0	10	other	Hansen
>0	7	forest	Hansen

Appendix B

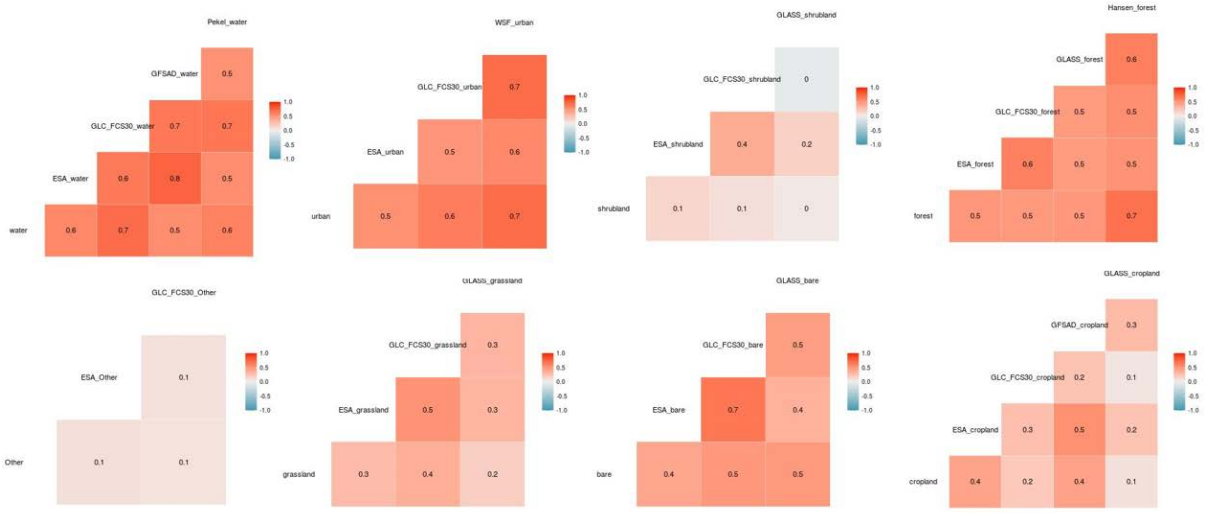

Figure 1B. Global correlations. Global correlations for each of the LULC classes among the predictors and between the predictors and the ground-truth sub-pixel database (variables: water, urban, shrubland, forest, Other, grassland, bare and cropland). Value 1 (red) high positive correlation. Value -1 (blue) high negative correlation.

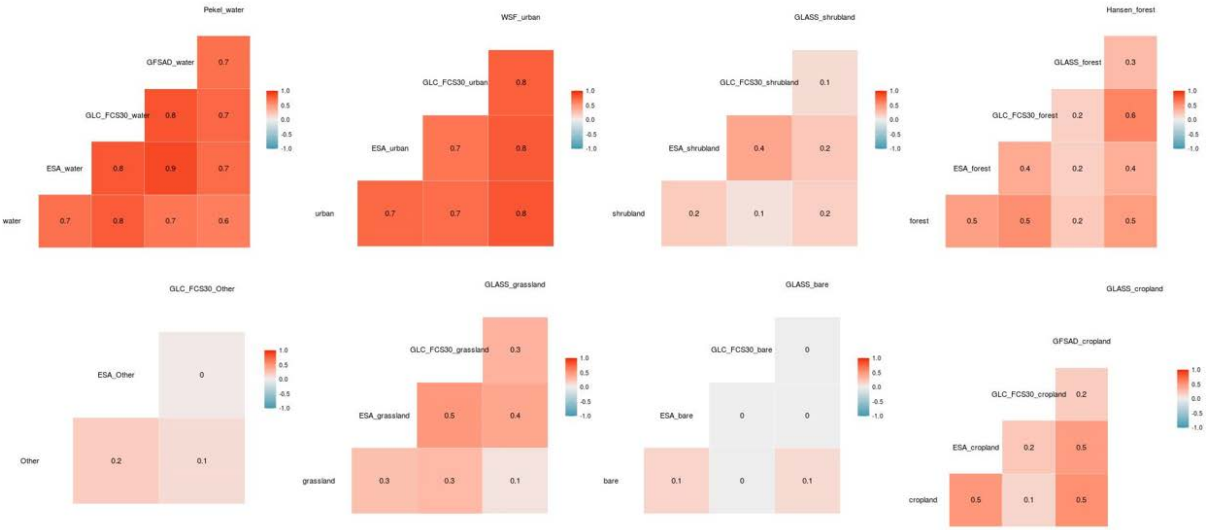

Figure 2B. South-African correlations. Regional correlations for each of the LULC classes among the predictors and between the predictors and the ground-truth sub-pixel database (variables: water, urban, shrubland, forest, Other, grassland, bare and cropland). Value 1 (red) high positive correlation. Value -1 (blue) high negative correlation.

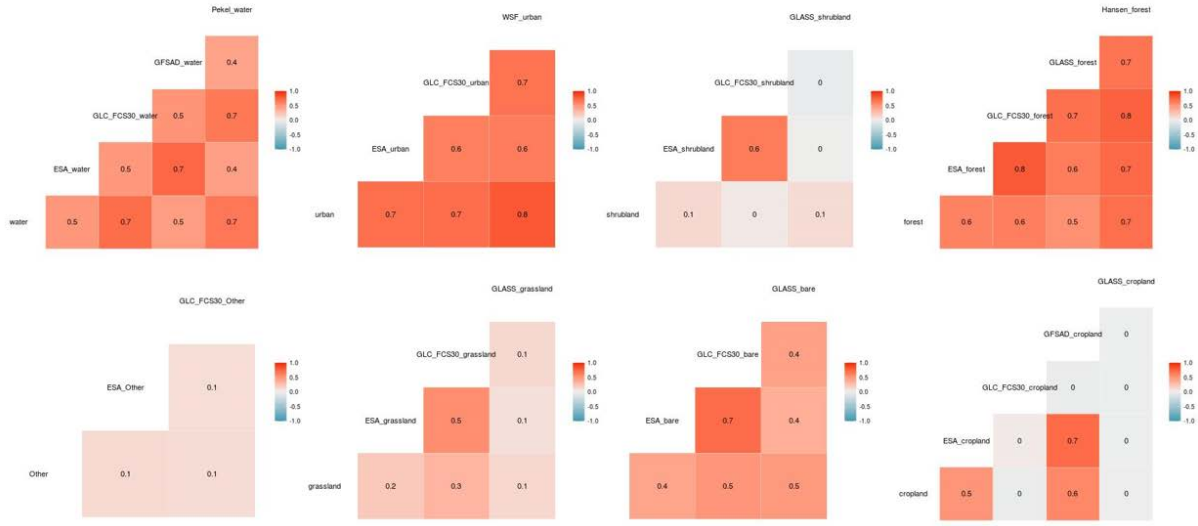

Figure 3B. North-American correlations. Regional correlations for each of the LULC classes among the predictors and between the predictors and the ground-truth sub-pixel database (variables: water, urban, shrubland, forest, Other, grassland, bare and cropland). Value 1 (red) high positive correlation. Value -1 (blue) high negative correlation.

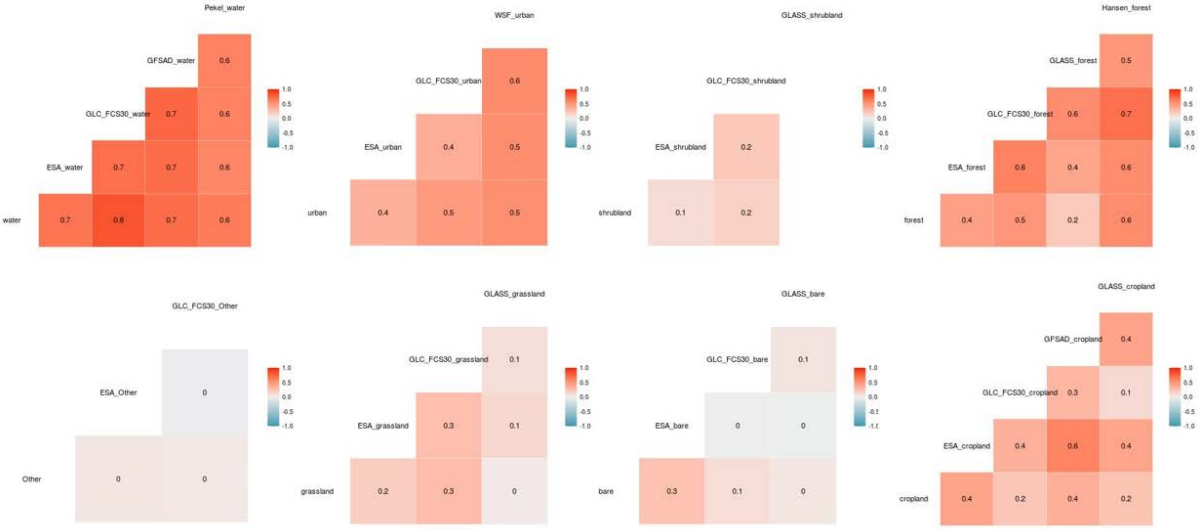

Figure 4B. European correlations. Regional correlations for each of the LULC classes among the predictors and between the predictors and the ground-truth sub-pixel database (variables: water, urban, shrubland, forest, Other, grassland, bare and cropland). Value 1 (red) high positive correlation. Value -1 (blue) high negative correlation.

Figure 5B. Asian-tropical correlations. Regional correlations for each of the LULC classes among the predictors and between the predictors and the ground-truth sub-pixel database (variables: water, urban, shrubland, forest, Other, grassland, bare and cropland). Value 1 (red) high positive correlation. Value -1 (blue) high negative correlation.

Table 1B. List of 10 fitted models

I.	Global_DirichletRegression	Model 1: DirichReg(formula = AL ~ GLC_FCS30_bare + GLASS_bare | ESA_cropland * GFSAD_cropland + GLC_FCS30_cropland + GLASS_cropland | ESA_grassland + GLC_FCS30_grassland + GLASS_grassland | ESA_shrubland * GLC_FCS30_shrubland + GLASS_shrubland | Hansen_forest + ESA_forest | WSF_urban | GLC_FCS30_water | ESA_Other + GLC_FCS30_Other, data = ALake, model = "common")
II.	Global_BetaRegression	
	Forest	Call: betareg(formula = AL[, 5] ~ Hansen_forest + ESA_forest, data = db)
	Urban and built-up	Call: betareg(formula = AL[, 6] ~ WSF_urban, data = db)
	Water body	Call: betareg(formula = AL[, 7] ~ GLC_FCS30_water, data = db)
	Shrubland	Call: betareg(formula = AL[, 4] ~ GLC_FCS30_shrubland + ESA_shrubland + GLASS_shrubland, data = db)
	Cropland	Call: betareg(formula = AL[, 2] ~ ESA_cropland + GFSAD_cropland + GLC_FCS30_cropland + GLASS_cropland, data = db)
	Grassland	Call: betareg(formula = AL[, 3] ~ ESA_grassland + GLC_FCS30_grassland + GLASS_grassland, data = db)
	Bare land	Call: betareg(formula = AL[, 1] ~ GLC_FCS30_bare + GLASS_bare, data = db)
	Other LULC class	Call: betareg(formula = AL[, 8] ~ GLC_FCS30_Other, data = db)
III.	DirichletRegression_SouthAfrica	Model 1: DirichReg(formula = AL ~ ESA_bare | ESA_cropland * GFSAD_cropland | ESA_grassland + GLC_FCS30_grassland + GLASS_grassland | ESA_shrubland + GLASS_shrubland | Hansen_forest + ESA_forest | WSF_urban | GLC_FCS30_water | ESA_Other, data = ALake, model = "common")
IV.	DirichletRegression_AsiaTropical	Model 1: DirichReg(formula = AL ~ GLASS_bare | GFSAD_cropland + ESA_cropland + GLC_FCS30_cropland | GLASS_grassland | GLC_FCS30_shrubland + GLASS_shrubland | Hansen_forest * ESA_forest * GLC_FCS30_forest | ESA_urban + WSF_urban | ESA_water * Pekel_water | GLC_FCS30_Other, data = ALake, model = "common")
V.	DirichletRegression_NorthAmerica	Model 1: DirichReg(formula = AL ~ GLASS_bare + GLC_FCS30_bare | GFSAD_cropland | GLC_FCS30_grassland + ESA_grassland | ESA_shrubland + GLASS_shrubland | Hansen_forest | WSF_urban | ESA_water + GLC_FCS30_water | GLC_FCS30_Other, data = ALake, model = "common")
VI.	DirichletRegression_Europe	Model 1: DirichReg(formula = AL ~ ESA_bare | GFSAD_cropland + ESA_cropland + GLASS_cropland | GLC_FCS30_grassland + ESA_grassland + GLASS_grassland | ESA_shrubland + GLC_FCS30_shrubland | Hansen_forest + GLASS_forest | GLC_FCS30_urban + WSF_urban | GLC_FCS30_water | ESA_Other, data = ALake, model = "common")
VII.	BetaRegression_SouthAfrica	
	Forest	Call: betareg(formula = AL[, 5] ~ Hansen_forest + ESA_forest, data = db)
	Urban and built-up	Call: betareg(formula = AL[, 6] ~ WSF_urban, data = db)

Water body	Call: betareg(formula = AL[, 7] ~ GLC_FCS30_water, data = db)
Shrubland	Call: betareg(formula = AL[, 4] ~ ESA_shrubland + GLC_FCS30_shrubland + GLASS_shrubland, data = db)
Cropland	Call: betareg(formula = AL[, 2] ~ ESA_cropland + GFSAD_cropland, data = db)
Grassland	Call: betareg(formula = AL[, 3] ~ ESA_grassland + GLC_FCS30_grassland + GLASS_grassland, data = db)
Bare land	Call: betareg(formula = AL[, 1] ~ ESA_bare, data = db)
Other LULC class	Call: betareg(formula = AL[, 8] ~ ESA_Other, data = db)
VIII. BetaRegression_AsiaTropical	
Forest	Call: betareg(formula = AL[, 5] ~ Hansen_forest + ESA_forest + GLC_FCS30_forest, data = db)
Urban and built-up	Call: betareg(formula = AL[, 6] ~ WSF_urban + ESA_urban, data = db)
Water body	Call: betareg(formula = AL[, 7] ~ ESA_water + Pekel_water, data = db)
Shrubland	Call: betareg(formula = AL[, 4] ~ GLC_FCS30_shrubland + GLASS_shrubland, data = db)
Cropland	Call: betareg(formula = AL[, 2] ~ GFSAD_cropland + ESA_cropland + GLC_FCS30_cropland, data = db)
Grassland	Call: betareg(formula = AL[, 3] ~ GLASS_grassland, data = db)
Bare land	Call: betareg(formula = AL[, 1] ~ GLASS_bare, data = db)
Other LULC class	Call: betareg(formula = AL[, 8] ~ GLC_FCS30_Other, data = db)
IX. BetaRegression_NorthAmerica	
Forest	Call: betareg(formula = AL[, 5] ~ Hansen_forest, data = db)
Urban and built-up	Call: betareg(formula = AL[, 6] ~ WSF_urban + ESA_urban, data = db)
Water body	Call: betareg(formula = AL[, 7] ~ ESA_water + GLC_FCS30_water, data = db)
Shrubland	Call: betareg(formula = AL[, 4] ~ ESA_shrubland + GLASS_shrubland, data = db)
Cropland	Call: betareg(formula = AL[, 2] ~ GFSAD_cropland, data = db)
Grassland	Call: betareg(formula = AL[, 3] ~ GLC_FCS30_grassland + ESA_grassland, data = db)
Bare land	Call: betareg(formula = AL[, 1] ~ GLASS_bare + GLC_FCS30_bare, data = db)
Other LULC class	Call: betareg(formula = AL[, 8] ~ GLC_FCS30_Other, data = db)
X. BetaRegression_Europe	
Forest	Call: betareg(formula = AL[, 5] ~ Hansen_forest + GLASS_forest, data = db)
Urban and built-up	Call: betareg(formula = AL[, 6] ~ GLC_FCS30_urban + WSF_urban, data = db)
Water body	Call: betareg(formula = AL[, 7] ~ GLC_FCS30_water, data = db)
Shrubland	Call: betareg(formula = AL[, 4] ~ ESA_shrubland + GLC_FCS30_shrubland, data = db)
Cropland	Call: betareg(formula = AL[, 2] ~ GFSAD_cropland + ESA_cropland + GLASS_cropland, data = db)
Grassland	Call: betareg(formula = AL[, 3] ~ GLC_FCS30_grassland + ESA_grassland + GLASS_grassland, data = db)

Bare land	`Call:` `betareg(formula = AL[, 1] ~ ESA_bare, data = db)`
Other LULC class	`Call:` `betareg(formula = AL[, 8] ~ GLC_FCS30_Other, data = db)`

Table 2B. Accuracy and goodness of fit obtained with the fitted models

Global_DirichletRegression		**MAE**	**RMSE**	**R^2**
	Bare soil	0.11	0.17	0.28
	Cropland	0.17	0.28	0.26
	Grassland	0.30	0.39	0.20
	Shrubland	0.17	0.25	0.04
	Forest	0.24	0.32	0.50
	Urban and built-up	0.08	0.09	0.44
	Water body	0.09	0.14	0.43
	Other LULC	0.09	0.13	0.01
	Tot	0.16	0.22	0.27
Global_BetaRegression				
	Bare soil	0.11	0.16	029
	Cropland	0.23	0.28	0.26
	Grassland	0.31	0.37	0.36
	Shrubland	0.19	0.25	0.25
	Forest	0.25	0.29	0.49
	Urban and built-up	0.03	0.06	0.46
	Water body	0.03	0.13	0.40
	Other LULC	0.08	0.13	0.01
	Tot	0.16	0.21	0.31
BetaRegression_SouthAfrica				
	Bare soil	0.04	0.07	0.04
	Cropland	0.27	0.31	0.38
	Grassland	0.36	0.40	0.11
	Shrubland	0.15	0.21	0.21
	Forest	0.17	0.22	0.28
	Urban and built-up	0.02	0.05	0.66
	Water body	0.07	0.11	0.54
	Other LULC	0.11	0.16	0.03
	Tot	0.15	0.16	0.28
BetaRegression_AsiaTropical				
	Bare soil	0.04	0.08	0.01
	Cropland	0.2	0.26	0.23
	Grassland	0.23	0.28	0.28
	Shrubland	0.16	0.22	0.22
	Forest	0.32	0.36	0.34
	Urban and built-up	0.02	0.06	0.47
	Water body	0.09	0.13	0.34
	Other LULC	0.06	0.10	0.01
	Tot	0.14	0.19	0.24
BetaRegression_NorthAmerica				
	Bare soil	0.23	0.27	0.32
	Cropland	0.13	0.17	0.32
	Grassland	0.32	0.37	0.11

	Shrubland	0.23	0.28	0.03
	Forest	0.21	0.26	0.49
	Urban and built-up	0.01	0.04	0.61
	Water body	0.11	0.15	0.43
	Other LULC	0.08	0.13	0.01
	Tot	0.16	0.21	0.17
BetaRegression_Europe				
	Bare soil	0.06	0.10	0.05
	Cropland	0.24	0.29	0.21
	Grassland	0.29	0.33	0.33
	Shrubland	0.23	0.28	0.03
	Forest	0.25	0.30	0.32
	Urban and built-up	0.05	0.09	0.27
	Water body	0.05	0.10	0.55
	Other LULC	0.03	0.07	0.01
	Tot	0.15	0.20	0.22
DirichletRegression_SouthAfrica				
	Bare soil	0.09	0.10	0.01
	Cropland	0.21	0.32	0.37
	Grassland	0.38	0.45	0.15
	Shrubland	0.15	0.21	0.05
	Forest	0.15	0.22	0.28
	Urban and built-up	0.08	0.09	0.66
	Water body	0.09	0.13	0.54
	Other LULC	0.10	0.16	0.01
	Tot	0.16	0.21	0.26
DirichletRegression_AsiaTropical				
	Bare soil	0.08	0.10	0.01
	Cropland	0.15	0.25	0.20
	Grassland	0.17	0.27	0.10
	Shrubland	0.13	0.21	0.04
	Forest	0.35	0.41	0.31
	Urban and built-up	0.08	0.09	0.35
	Water body	0.09	0.14	0.27
	Other LULC	0.08	0.11	0.01
	Tot	0.14	0.20	0.16
DirichletRegression_NorthAmerica				
	Bare soil	0.16	0.28	0.33
	Cropland	0.18	0.17	0.21
	Grassland	0.30	0.38	0.16
	Shrubland	0.20	0.28	0.05
	Forest	0.20	0.27	0.50
	Urban and built-up	0.08	0.09	0.53
	Water body	0.10	0.16	0.51
	Other LULC	0.09	0.14	0.01
	Tot	0.16	0.22	0.29
DirichletRegression_Europe				
	Bare soil	0.09	0.12	0.03
	Cropland	0.18	0.29	0.22

Grassland	0.27	0.35	0.12
Shrubland	0.20	0.29	0.04
Forest	0.24	0.31	0.34
Urban and built-up	0.09	0.11	0.28
Water body	0.09	0.12	0.60
Other LULC	0.08	0.10	0.01
Tot	0.16	0.21	0.20

Summary 1B. Selected models' summary
South-Africa: regional beta regression

```
Call:
betareg(formula = AL[, 1] ~ Hansen_forest + ESA_forest, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.9986 -0.3285 -0.0912  0.4247  2.6787

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.00861    0.03982  -50.44   <2e-16 ***
Hansen_forest  1.99421    0.16502   12.09   <2e-16 ***
ESA_forest     0.72435    0.06877   10.53   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.94301    0.02871   32.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.069e+04 on 4 Df
Pseudo R-squared: 0.2251
Number of iterations: 18 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 3] ~ GLC_FCS30_water, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-5.3753 -0.0138 -0.0138 -0.0138  4.7490

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.80219    0.04409  -63.55   <2e-16 ***
GLC_FCS30_water  4.21775    0.14931   28.25   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.08115    0.09144   22.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.635e+04 on 3 Df
Pseudo R-squared: 0.5641
Number of iterations: 24 (BFGS) + 4 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 2] ~ WSF_urban, data = db)

Standardized weighted residuals 2:
     Min       1Q   Median       3Q      Max
-13.3325  -0.0735  -0.0735  -0.0735  13.5775

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.42508    0.04387 -100.87   <2e-16 ***
WSF_urban    4.94316    0.09195   53.76   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  14.2556     0.6716   21.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.608e+04 on 3 Df
Pseudo R-squared: 0.375
Number of iterations: 29 (BFGS) + 4 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 4] ~ GLC_FCS30_Other + ESA_Other, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.7587 -0.4009  0.1034  0.1045  2.3691

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.13123    0.05943  -2.208   0.0272 *
GLC_FCS30_Other  1.12289    0.06645  16.898   <2e-16 ***
ESA_Other        0.78479    0.06664  11.777   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.85310    0.02302   37.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  8370 on 4 Df
Pseudo R-squared: 0.3392
Number of iterations: 23 (BFGS) + 2 (Fisher scoring)
```

North-America: regional beta regression

```
Call:
betareg(formula = AL[, 1] ~ Hansen_forest, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.6913 -0.1508 -0.1508  0.4408  2.4063

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.82802    0.04251  -43.01   <2e-16 ***
Hansen_forest  2.92306    0.09737   30.02   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.85877    0.02786   30.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  6848 on 3 Df
Pseudo R-squared: 0.4915
Number of iterations: 22 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 3] ~ ESA_water + GLC_FCS30_water, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.0961 -0.0188 -0.0188 -0.0188  3.6159

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.44354    0.04875 -50.123  < 2e-16 ***
ESA_water        0.36610    0.10734   3.411 0.000648 ***
GLC_FCS30_water  4.25053    0.20224  21.018  < 2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  1.46327    0.06884   21.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.124e+04 on 4 Df
Pseudo R-squared: 0.5498
Number of iterations: 28 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 2] ~ WSF_urban + ESA_urban, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-7.2836 -0.0398 -0.0398 -0.0398 12.5102

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.74712    0.05007  -94.82   <2e-16 ***
WSF_urban    3.81316    0.21221   17.97   <2e-16 ***
ESA_urban    2.21641    0.17527   12.65   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   21.386      1.168   18.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.186e+04 on 4 Df
Pseudo R-squared: 0.3583
Number of iterations: 56 (BFGS) + 4 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 4] ~ GLC_FCS30_Other, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.0295 -0.3561  0.1531  0.1545  1.6853

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.13340    0.04792  -2.784  0.00538 **
GLC_FCS30_Other  1.50706    0.06763  22.283  < 2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.59967    0.01636   36.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  5367 on 3 Df
Pseudo R-squared: 0.333
Number of iterations: 18 (BFGS) + 1 (Fisher scoring)
```

Asia-tropical: regional beta regression

```
Call:
betareg(formula = AL[, 1] ~ Hansen_forest + ESA_forest + GLC_FCS30_forest, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-1.7932 -0.4818  0.0739  0.4645  1.6327

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.85202    0.05898 -14.447  < 2e-16 ***
Hansen_forest    1.17219    0.09884  11.859  < 2e-16 ***
ESA_forest       0.40852    0.07149   5.715 1.10e-08 ***
GLC_FCS30_forest 0.62951    0.09139   6.888 5.64e-12 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.50092    0.01225   40.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  5877 on 5 Df
Pseudo R-squared: 0.3314
Number of iterations: 20 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 3] ~ ESA_water + Pekel_water, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.7005 -0.0385 -0.0385 -0.0385  4.0544

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.57905    0.04675 -55.161  < 2e-16 ***
ESA_water    2.53248    0.22748  11.133  < 2e-16 ***
Pekel_water  1.51209    0.20571   7.351 1.97e-13 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  1.71574    0.07887   21.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.252e+04 on 4 Df
Pseudo R-squared: 0.3163
Number of iterations: 27 (BFGS) + 2 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 2] ~ WSF_urban + ESA_urban, data = db)

Standardized weighted residuals 2:
     Min       1Q   Median       3Q      Max
-12.2099  -0.0521  -0.0521  -0.0521   6.3451

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.34058    0.04841 -89.670  < 2e-16 ***
WSF_urban    3.32112    0.11851  28.024  < 2e-16 ***
ESA_urban    0.71267    0.17235   4.135 3.55e-05 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  13.0860     0.6771   19.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 1.281e+04 on 4 Df
Pseudo R-squared: 0.3683
Number of iterations: 49 (BFGS) + 3 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 4] ~ GLC_FCS30_Other, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-1.4132 -0.3857 -0.1494  0.4812  1.5166

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.25980    0.05080  -24.80   <2e-16 ***
GLC_FCS30_Other  1.27261    0.06984   18.22   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.44918    0.01068   42.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  6451 on 3 Df
Pseudo R-squared: 0.2178
Number of iterations: 13 (BFGS) + 1 (Fisher scoring)
```

Europe: regional beta regression

```
Call:
betareg(formula = AL[, 1] ~ Hansen_forest + GLASS_forest, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.4627 -0.5462 -0.0053  0.5415  2.2015

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.36930    0.05127 -26.706  < 2e-16 ***
Hansen_forest  2.48249    0.12366  20.076  < 2e-16 ***
GLASS_forest  -0.32650    0.06913  -4.723 2.32e-06 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.78763    0.02522   31.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  3455 on 4 Df
Pseudo R-squared: 0.2745
Number of iterations: 16 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 3] ~ GLC_FCS30_water, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-6.6958 -0.0225 -0.0225 -0.0225  5.7152

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.1000     0.0579  -53.54   <2e-16 ***
GLC_FCS30_water   5.1781     0.2152   24.07   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   3.1775     0.1884   16.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  8907 on 3 Df
Pseudo R-squared: 0.5881
Number of iterations: 34 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 2] ~ GLC_FCS30_urban + WSF_urban, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.5606 -0.0904 -0.0904 -0.0904  2.5205

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.29535    0.05628 -58.549  < 2e-16 ***
GLC_FCS30_urban  0.85041    0.23963   3.549 0.000387 ***
WSF_urban        2.66336    0.25280  10.535  < 2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   4.2829     0.2488   17.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  8129 on 4 Df
Pseudo R-squared: 0.2219
Number of iterations: 34 (BFGS) + 1 (Fisher scoring)
```

```
Call:
betareg(formula = AL[, 4] ~ GLC_FCS30_Other, data = db)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.1768 -0.4340 -0.0272  0.4374  1.7429

Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.01801    0.05006   -0.36    0.719
GLC_FCS30_Other  1.19304    0.07815   15.27   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.67494    0.01951   34.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  2502 on 3 Df
Pseudo R-squared: 0.1725
Number of iterations: 13 (BFGS) + 1 (Fisher scoring)
```

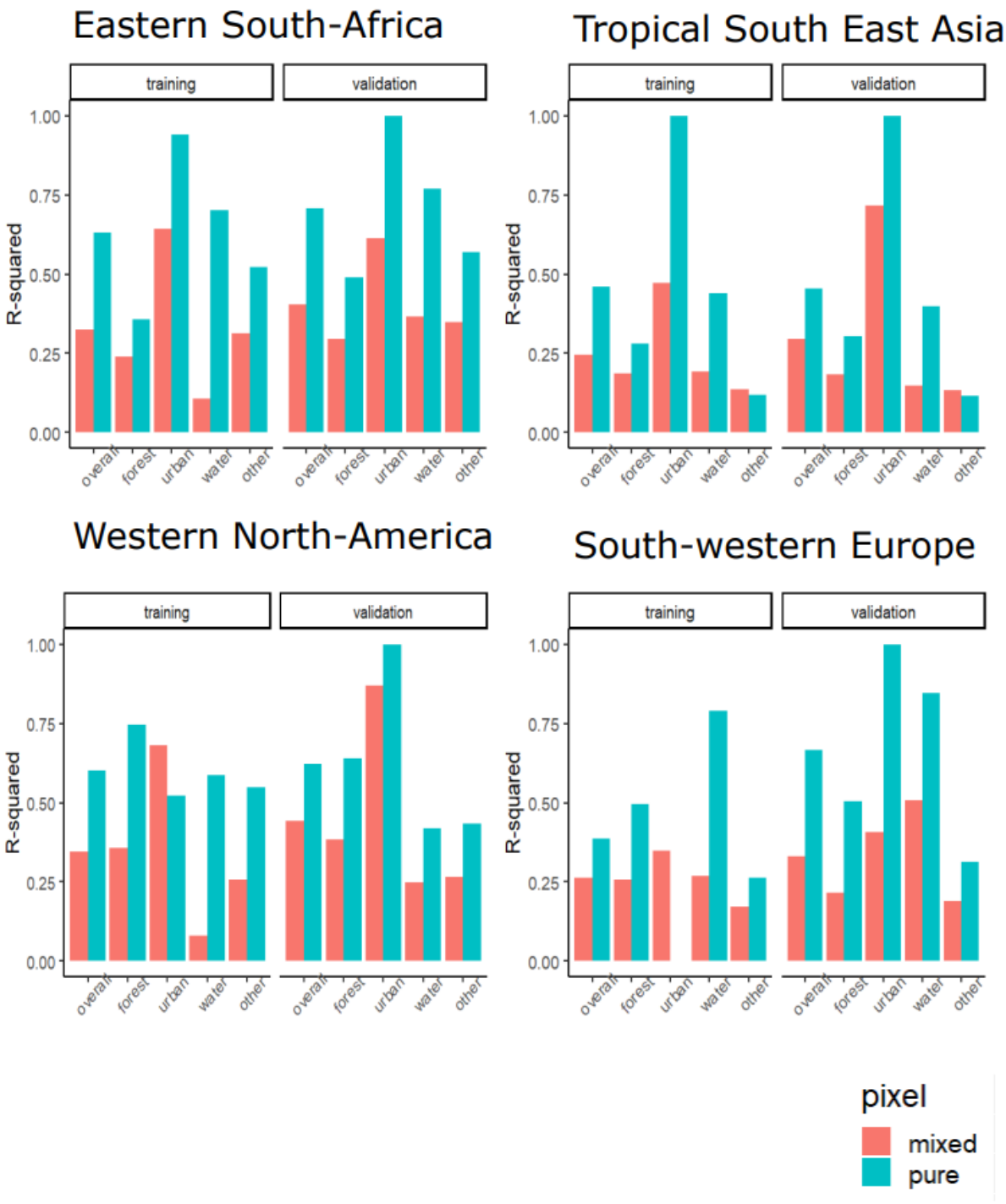

Figure 6B. Selected models' goodness of fit. R-squared in each study area obtained with the selected regional beta regressions for the selected 4 LULC classes (and overall) in mixed (pink) and pure (blue) pixels with both fitted (training) and predicted (validation) locations.

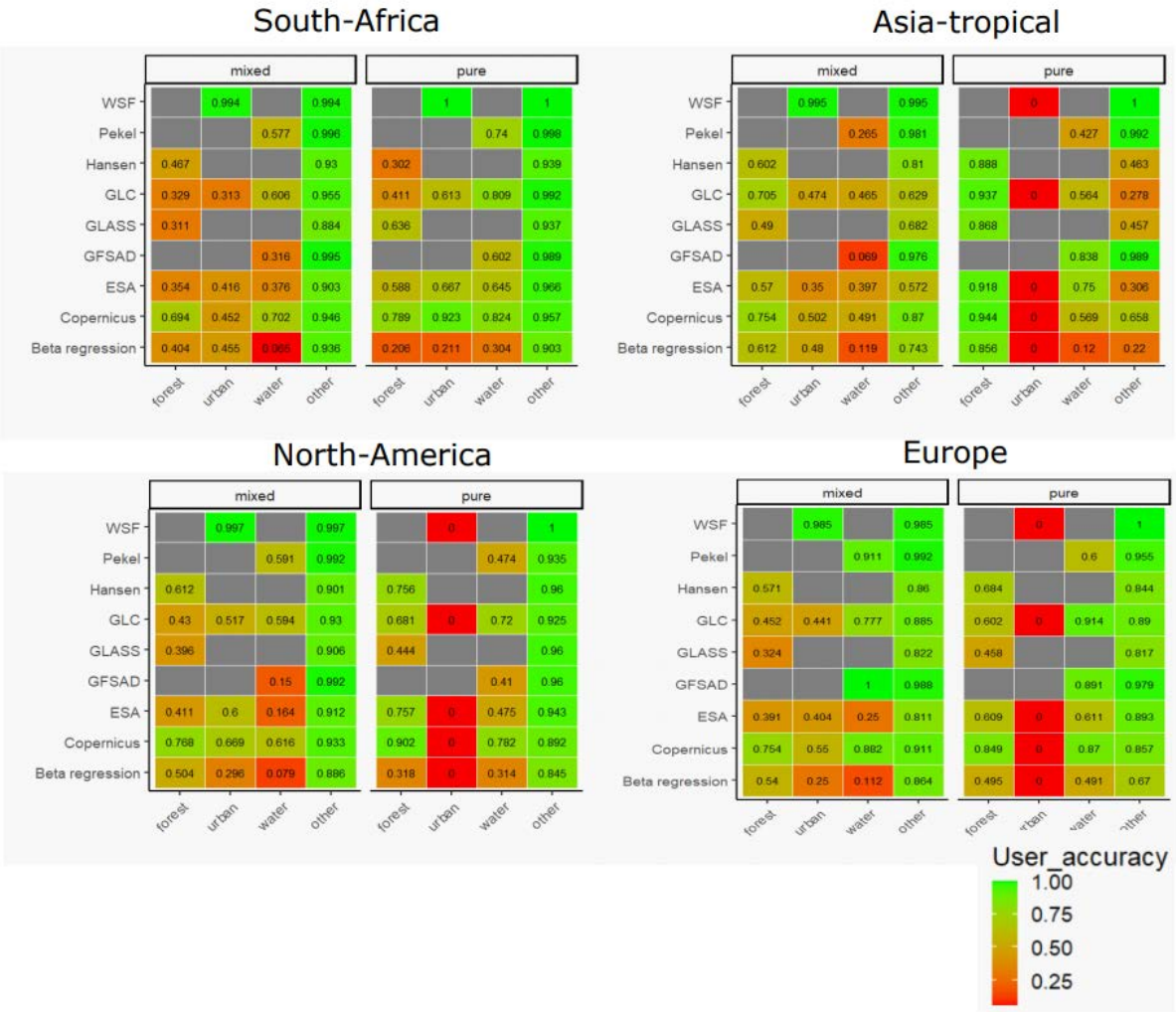

Figure 7B. Sub-pixel thematic accuracy comparison in each study area. User accuracy is shown for each of the 4 LULC classes in the study areas. Value 1 (green) means high accuracy. Value 0 (red) means low accuracy. Grey boxes indicate that the accuracy wasn't possible to be calculated for some classes of the maps.

Table 3B. Number of validation locations with fraction greater than zero for each LULC class in pure and mixed-pixels.

		Pure	Mixed
South-Africa			
	Water	21	25
	Urban	3	98
	Forest	40	403
	Other	327	745
North-America			
	Water	25	33
	Urban	0	30
	Forest	32	304
	Other	195	607
Asia-tropical			
	Water	13	43
	Urban	0	75

	Forest	386	409
	Other	76	464
Europe			
	Water	11	25
	Urban	0	99
	Forest	49	370
	Other	83	559