Modelling global biomass using deep learning from land use and earth system variables in the ESDL data cube

Karina Winkler

Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMK-IFU), KIT Campus Alpin, Karlsruhe Institute of Technology (KIT), mail: karina.winkler@kit.edu

Project of the Earth System Data Lab (ESDL) early adopters call

July 18, 2019

Abstract

People have increasingly been shaping the surface of our planet. Land use change is not only cause but also consequence of global environmental and socioeconomic change. As one of the main contributors to greenhouse gas (GHG) emissions, it is key for current sustainability debates and climate change mitigation. For better understanding its processes and environmental effects, I developed a data-driven reconstruction of global land use change from 1960-2015 by exploiting potentials of available observational data (remote sensing and statistical inventories) as part of my PhD project. Here, the land use fractions from this preliminary work and multiple satellite-derived variables from the Earth System Data Lab (ESDL), particularly, evaporation, root moisture, temperature, albedo, net ecosystem exchange, respiration, gross primary productivity, leaf area index and precipitation, are used for modelling above ground biomass at large scales for different world regions. The objective of this project is to investigate the capabilities of the ESDL data cube for modelling the biomass distribution of 2010 by using deep learning. Further, the role of land use on explaining the dynamics of biomass is analysed by changing the set of descriptor variables. The modelling approach shows the potential of reconstructing changes of above ground biomass for longer time spans at a global scale. This can contribute to a re-estimation of carbon stocks and their dynamics on the land surface and a more accurate quantification of the "carbon footprint" of land use change.

1 Introduction

Human activity has shaped the Earth‘s surface for many centuries. Rapidly growing world population and changing food consumption have increasingly placed demands on land to supply food, animal feed and fuel. Consequently, managed land has been expanding continuously around the globe while natural forests have been declining (Alexander et al., 2016, Keenan et al., 2015). Anthropogenic land change significantly alters climate and ecosystem processes. It is a major contributor to global greenhouse gas (GHG) emissions and can be either a sink or source of carbon (Le Quéré et al., 2013; van der Werf et al., 2009). Understanding land change dynamics and its role for the carbon budget is crucial for overcoming today's pressing global sustainability challenges such as climate change, biodiversity loss, and food security (Liu and Yang, 2015). As part of my PhD project, I developed a data-driven, spatially explicit reconstruction for global land use change, covering the period between 1960 and 2015 (see figure 1). Exploiting the potential of multiple available observational data such as remote sensing-based land cover classifications and statistical databases on national land use inventories (FAO), spatio-temporal patterns of land use change was derived on a 1 km grid resolution (publication in planning). The resulting yearly maps on land use distribution comprise the following land use/land cover categories:

- urban (built-up areas, settlements and construction/mining areas)

- cropland
- pasture (grasslands under grazing, rangelands)
- forest (incl. tree plantations)
- shrub-/grassland (shrubs and natural grasslands)
- other land (non-vegetated areas: bare, rock, snow/ice)

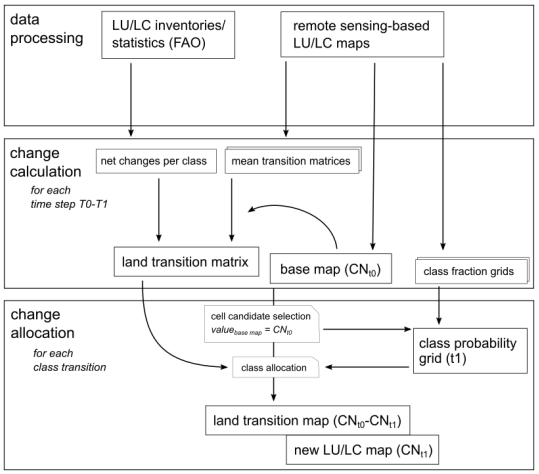

Figure 1: *Processing steps for iterative land use/land cover (LULC) reconstruction as preliminary work*

In this study, land use fractions from the preliminary work on land use reconstruction and multiple satellite-derived variables from the Earth System Data Lab (ESDL) are used for modelling above ground biomass at large scales for different world regions. The objective of this project is to investigate the capabilities of the ESDL data cube for modelling the biomass distribution of 2010 by using a deep learning framework. Further, the role of land use on explaining the dynamics of biomass is analysed by changing the set of descriptor variables. The modelling approach will be tested for a potential reconstruction of biomass changes for longer time spans at a global scale. This can be essential for re-estimating of carbon stocks and their dynamics on the land surface and more accurately quantifying the "carbon footprint" of land use change.

2 Methods

2.1 ESDL infrastructure set-up

In this early adopters project, a deep learning modelling approach was used in the Python 3 framework within the ESDL jupyterhub. As an addition to the available python libraries, tensorflow was installed.

2.2 External input data

2.2.1 Biomass in 2010

For the modelling approach, above-ground biomass derived from Earth Observation data in the framework of ESA's GlobBiomass project (Santoro et al., 2018) was used as a response variable. The data set covers above ground biomass (AGB, unit: tons/ha i.e., Mg/ha), as the the mass, expressed as oven-dry weight of the woody parts (stem, bark, branches and twigs) of all living trees excluding stump and roots for the year 2010. In preliminary preprocessing steps, the gridded dataset was downloaded, resampled and reprojected from the original extent and 100 m grid size to match the spatial extent and resolution of the ESDL data cube of 1/12 degree. The global map (GeoTiff format) was uploaded into the workspace on the ESDL jupyterhub and later accessed and loaded as DataArray using *openrasterio* from *xarray* package (see figure 2).

2.2.2 Land use in 2010

Land use fractions for urban, cropland, pasture, forest, shrub and other land were used as descriptor variables for the modelling approach. Land use maps for the year 2010, which were derived from the preliminary land use reconstruction model HILDA+ (publication in planning), were reprojected and resampled from their original spatial reference and 1 km grid size to the settings of the ESDL data cube. Up-scaling was carried out by averaging the original pixel values in order to derive the fractional area under each land use category for each pixel. After this preprocessing steps, the layers were uploaded

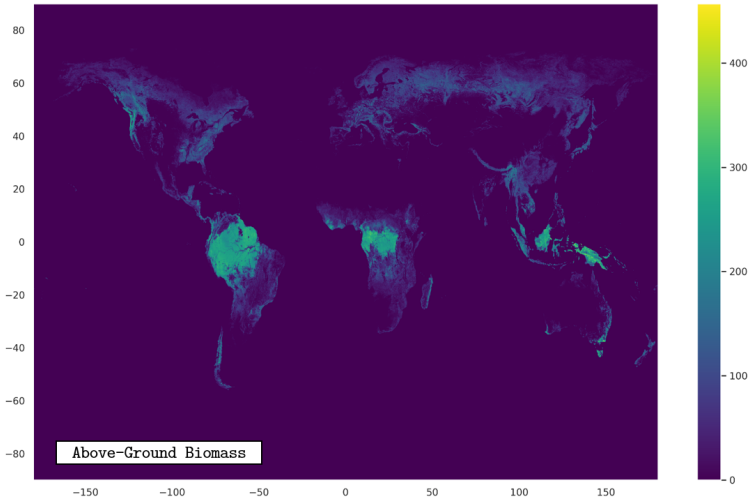

Figure 2: *Above ground biomass in 2010*

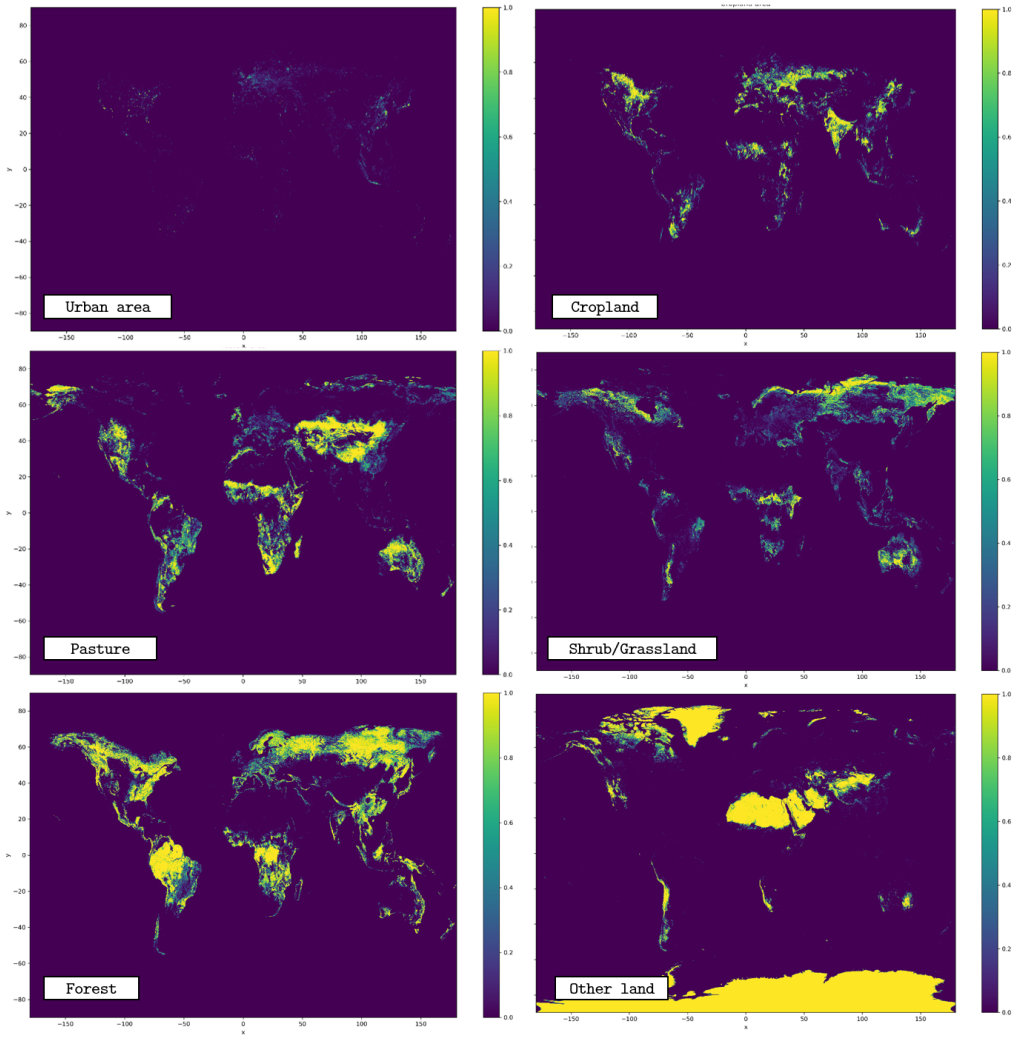

Figure 3: *land use fractions used as descriptors for biomass modelling*

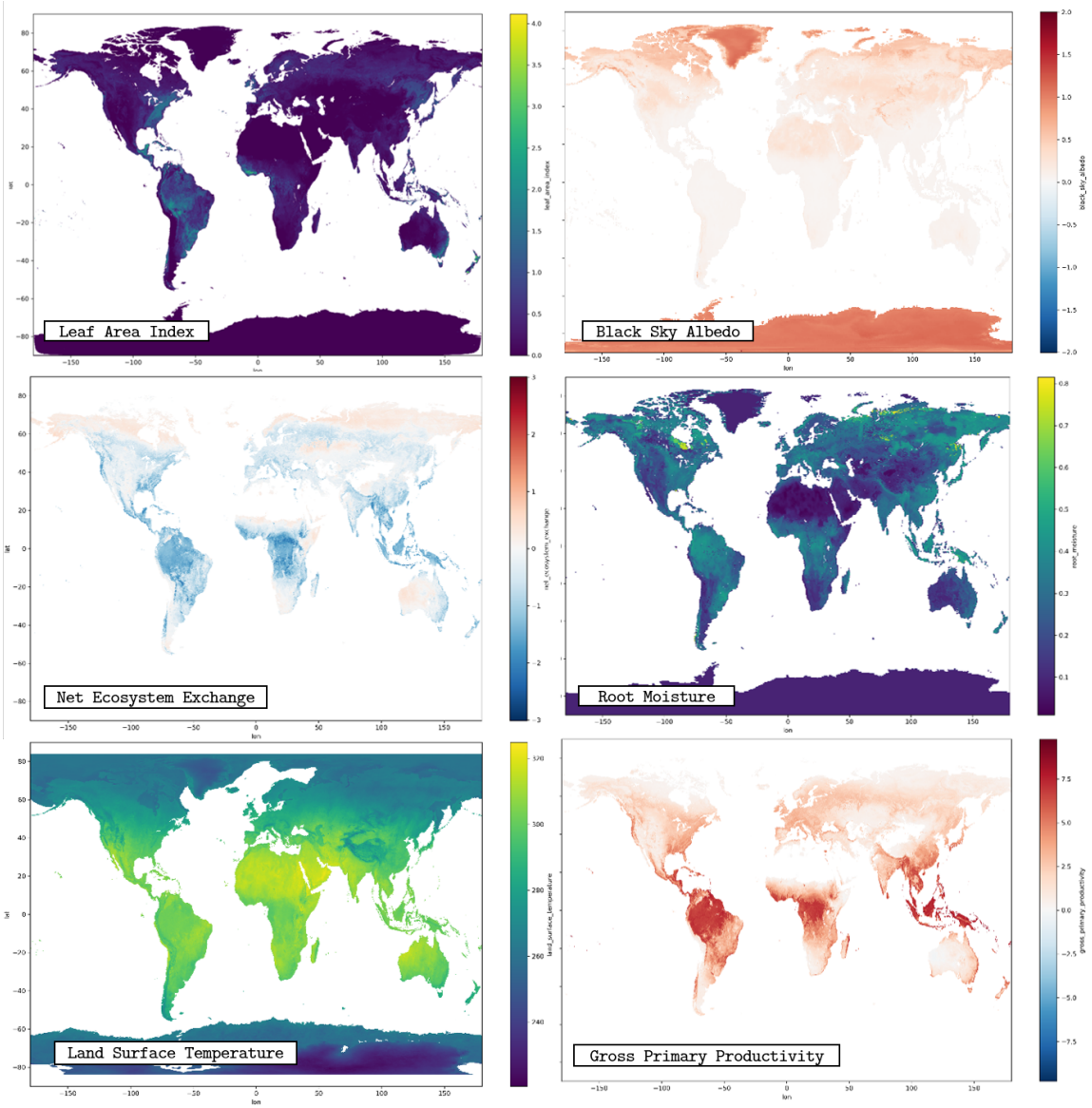

Figure 4: *Selection of ESDL variables (annual mean of 2010) used as descriptors for biomass modelling*

into workspace on the ESDL jupyterhub and loaded as DataArray with *xarray.openrasterio*.

2.2.3 Country mask

Further, a gridded country mask was generated based on GADM 3.6 (`https://gadm.org/download_world.html`). Values were converted into FAO country codes. After resampling and reprojecting, the gridded data set was uploaded and opened as DataArray with *xarray.openrasterio* and later used to separate inland from ocean pixels.

2.3 ESDL variables

The following ESDL parameters were selected as explanatory variables for modelling biomass:

- evaporation
- bare soil evaporation
- surface moisture
- root moisture
- land surface temperature
- black sky albedo
- net ecosystem exchange
- terrestrial ecosystem respiration
- gross primary productivity
- leaf area index
- precipitation

After defining a world region and subsetting the data cube according to the respective boundary coordinates, temporal means for the year 2010 were calculated for each variable (see figure 4).

2.4 Setting-up the feature table

The modelling procedure was carried out for different world regions and various ensembles of descriptor variables (all variables, ESDL variables only, land use variables only).
For both descriptor and response variables of each scenario, values on land were extracted and loaded into a *pandas* data frame with latitude and longitude information. All entries which contain missing data were excluded from the analysis. For better comparability, feature values were scaled to a range between 0 and 1 using *MinMaxScaler* from the *preprocessing* module of *sklearn* library. Further, a random sample of 80 % of the data frame were put aside for training while 20 % were used for testing the model.

2.5 Neuronal network

2.5.1 Building the model

The model architecture was defined with *keras.Sequential* using a sequential neural network with two densely connected hidden layers of 64 units and rectified linear unit (relu) as an activation function, and an output layer that returns a single, continuous value. Parameters which define the optimisation behaviour of the neural network were set. Here, stochastic gradient descent (SGD) was used as optimiser *keras.optimizers.SGD*, with a learning rate of 0.01, a momentum value of 0, Nesterov value set to True and a decay value of 1e-6.

2.5.2 Training

The generated training data set is used to train the model. Here, *keras.callback_ early_ stopping* is used to stop the training procedure when a monitored quantity, here val_loss, has stopped improving (minimum delta value of 0.001) with a patience of 100 epochs. For the training procedure, a maximum of 1000 epochs was set. 10 % of the training data set was used for validation. The training progress, a decrease of the mean square error after each iteration, is depicted in figures 5.

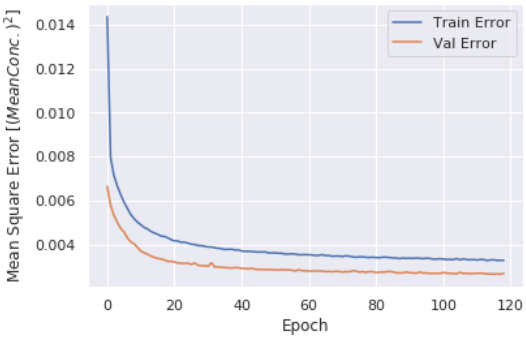

Figure 5: *Training progress for modelling biomass in Europe with all descriptor variables*

2.5.3 Testing and running

After training, the model was used for predicting above-ground biomass on the test data set and, subsequently on the whole data set. Performances were compared.

3 Results

3.1 Model performance

The performance of the neural network for modelling biomass was proven good. However, accuracies differ for each world region. In Europe, the value of R^2 was around 0.89, in northern and central America 0.84, in Asia 0.86, in South America 0.96, in Africa 0.93, in Southeast Asia and Australia 0.93 (see figures 8 - 13).
Figures 14 and 15 show the spatial distribution of the predicted biomass and its respective difference to the reference data set for the different world regions. While large continental areas with higher biomass tend to be slightly underestimated, coastal regions with high biomass show signals of overestimation.

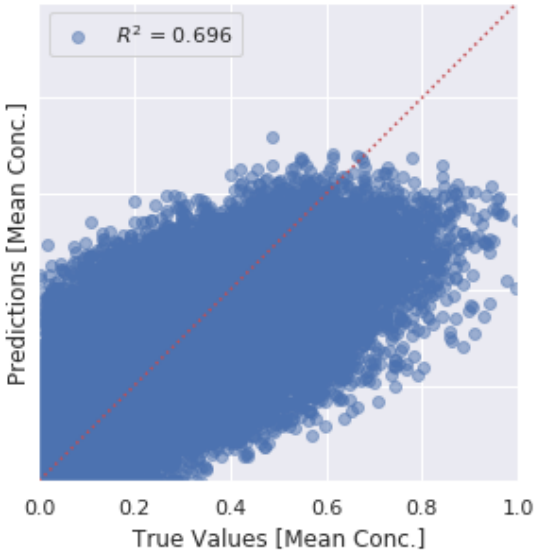

Figure 6: *Model performance for Europe using only ESDL variables as descriptors*

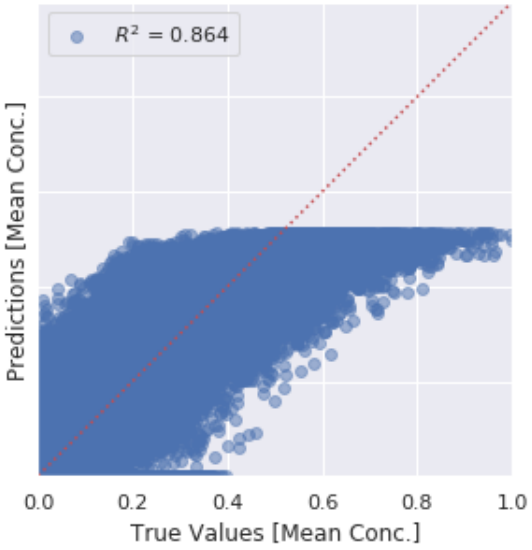

Figure 7: *Model performance for Europe using only land use variables as descriptors*

3.2 Role of land use parameters

By applying separate model runs with different parameter settings, the influence of land use for biomass modelling could be analysed. In Europe for instance, model prediction using only the ESDL variables could achieve an R^2 value of around 0.69, whereas another run, using only land use parameters, led to a R^2 of ca. 0.86 (see figures 6 and 7). By using all variables as descriptors, a maximum accuracy of $R^2 = 0.89$ was reached (see figure 8).

4 Concluding remarks

In this experimental study, the ESDL data cube infrastructure was used to exploit the synergies of multiple available earth system variables in an easily accessible data frame structure. A neural network was trained on multiple parameters, including additional land use data, which proofed to be extremely relevant for explaining the global patterns of biomass. Overall, a good modelling performance was obtained.
Regarding the early adopters state of this project, some recommendations for future usage of the ESDL data cube can be made:

- adding of land use/land cover variables into

Figure 8: *Performance for biomass modelling over Europe*

Figure 9: *Performance for biomass modelling over North and Central America*

Figure 10: *Performance for biomass modelling over Asia*

the data cube to expand the scope of possible analysis for human-environment interactions
- updating the country mask (e.g. FAO coded grid) within the data cube and including land/water masks
- including quality layers (when available) for the ESDL variables for uncertainty assessments

The framework of the ESDL data cube provides benefits for large-scale analysis which comprise multi-source and multi-temporal data. Since more and more observational data from remote sensing are becoming available, "big data"-adapted infrastructures as the data cube are getting increasingly relevant.
Following-up on this study, the modelling approach will be tested for multi-temporal prediction in order to reconstruct biomass changes on a yearly scale. This may contribute to a better quantification of global carbon stock dynamics.

References

Alexander, P., Brown, C., Arneth, A., Finnigan, J., Rounsevell, M.D.A., 2016. Human appropriation of land for food: The role of diet. Glob. Environ. Chang. 41, 88-98. https://doi.org/10.1016/j.gloenvcha.2016.09.005

Keenan, R.J., Reams, G.A., Achard, F., de Freitas, J. V., Grainger, A., Lindquist, E., 2015. Dynamics of global forest area: Results from the FAO Global Forest Resources Assessment 2015. For. Ecol. Manage. 352, 9-20. https://doi.org/10.1016/J.FORECO.2015.06.014

Le Quéré, C., Andres, R.J., Boden, T., Conway, T., Houghton, R.A., House, J.I., Marland, G., Peters, G.P., Van Der Werf, G.R., Ahlström, A., Andrew, R.M., Bopp, L., Canadell, J.G., Ciais, P., Doney, S.C., Enright, C., Friedlingstein, P., Huntingford, C., Jain, A.K., Jourdain, C., Kato, E., Keeling, R.F., Klein Goldewijk, K., Levis, S., Levy, P., Lomas, M., Poulter, B., Raupach, M.R., Schwinger, J., Sitch, S., Stocker, B.D., Viovy, N., Zaehle, S., Zeng, N., 2013. The global carbon budget 1959-2011. Earth Syst. Sci. Data 5, 165-185. https://doi.org/10.5194/essd-5-165-2013

Liu, T., Yang, X., 2015. Land Change Modeling: Status and Challenges, in: Monitoring and Modeling of Global Changes: A Geomatics Perspective. Springer, pp. 3-16. https://doi.org/10.1007/978-94-017-9813-6_1

Santoro, M., Cartus, O., Mermoz, S., Bouvet, A., Le Toan, T., Carvalhais, N., Rozendaal, D., Herold, M., Avitabile, V., Quegan, S., Carreiras, J., Rauste, Y., Balzter, H., Schmullius, C., Seifert, F.M., 2018, GlobBiomass global above-ground biomass and growing stock volume datasets, available on-line at http://globbiomass.org/products/global-mapping

van der Werf, G.R., Morton, D.C., DeFries, R.S., Olivier, J.G.J., Kasibhatla, P.S., Jackson, R.B., Collatz, G.J., Randerson, J.T., 2009. CO2 emissions from forest loss. Nat. Geosci. 2, 737-738. https://doi.org/10.1038/ngeo671

Figure 11: *Performance for biomass modelling over South America*

Figure 12: *Performance for biomass modelling over Africa*

Figure 13: *Performance for biomass modelling over Southeast-Asia and Australia*

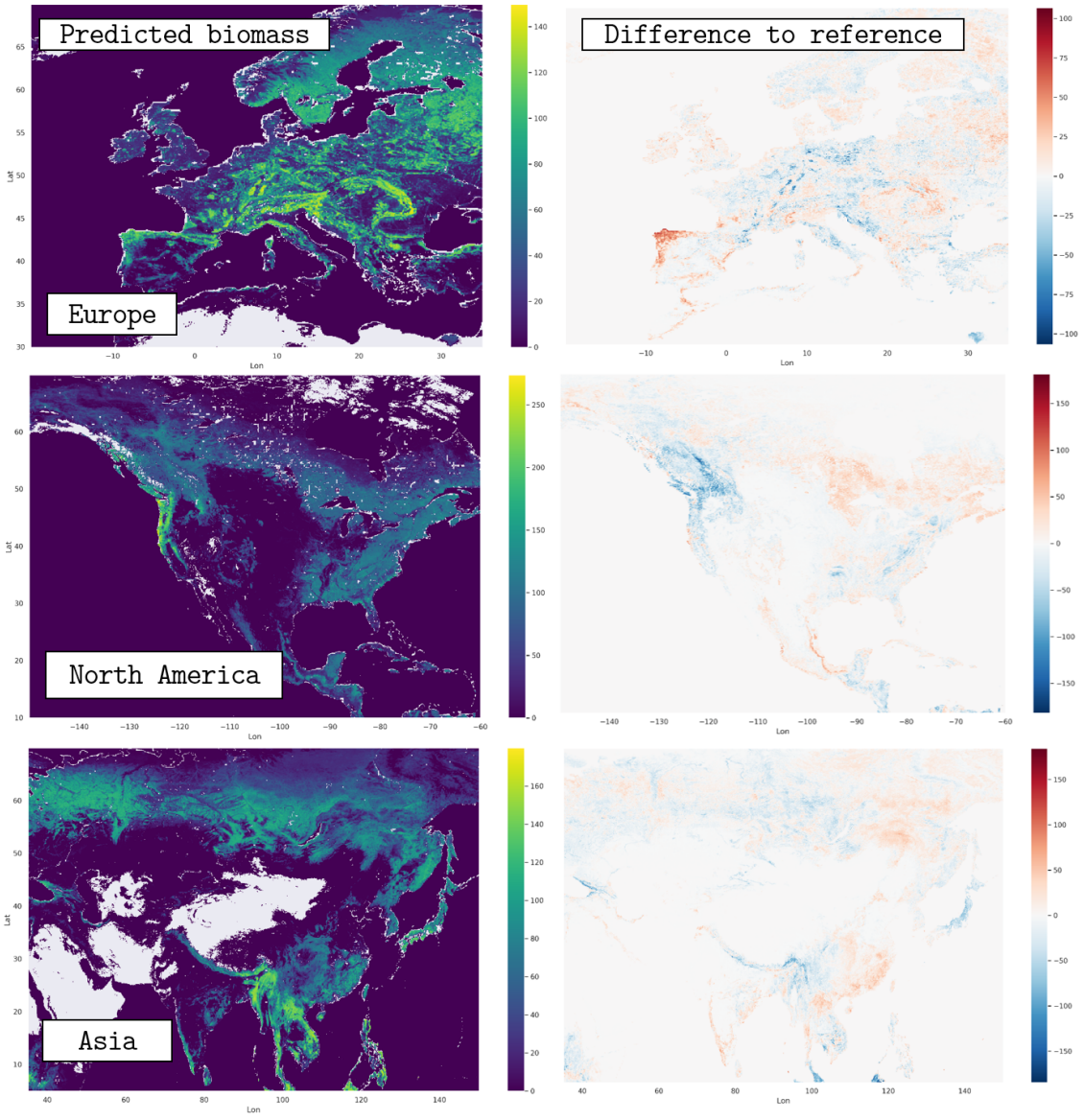

Figure 14: *Maps on predicted biomass and difference to reference biomass for different world regions (1)*

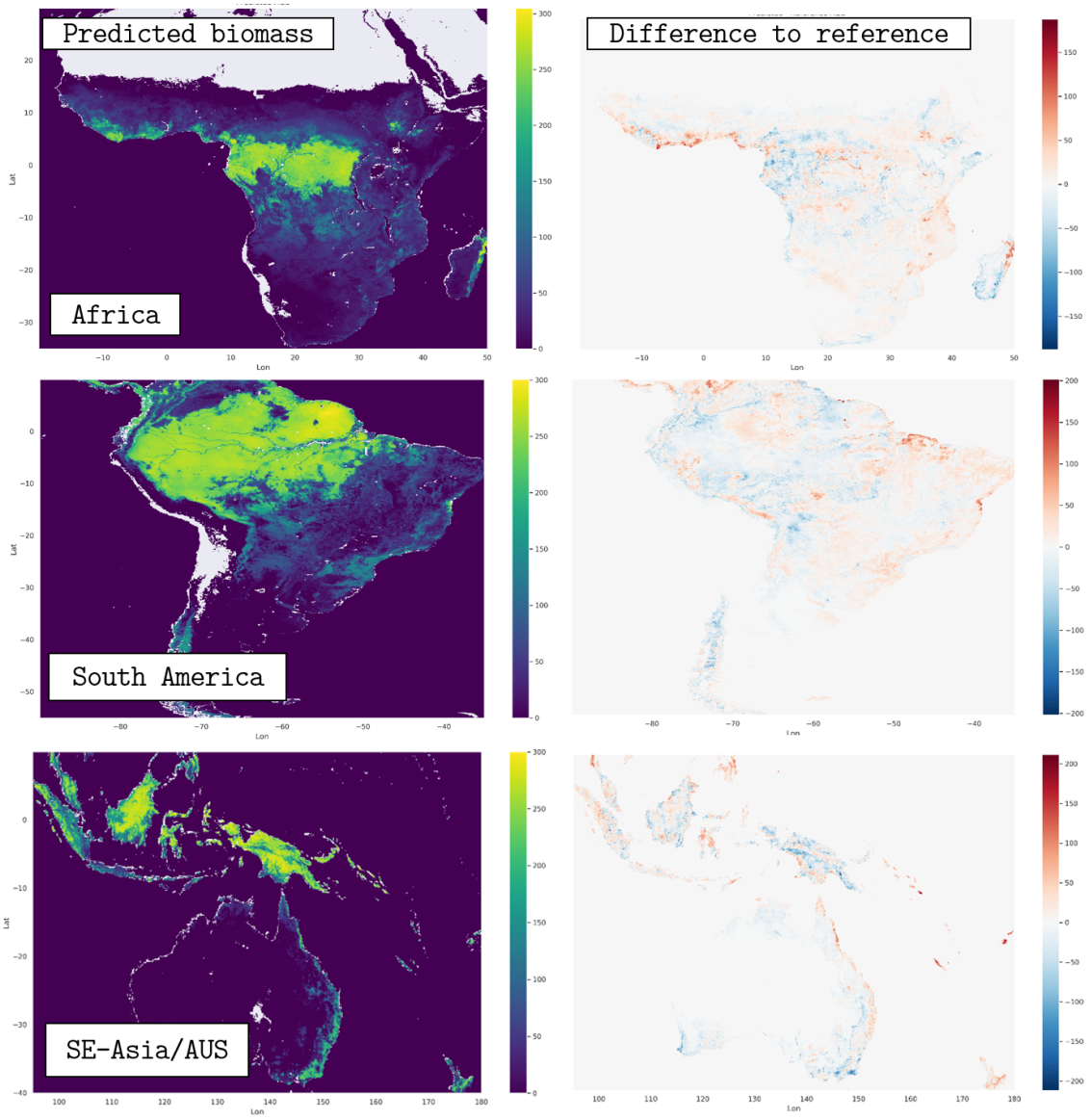

Figure 15: *Maps on predicted biomass and difference to reference biomass for different world regions (2)*