Supplement of "Global hydro-climatic biomes identified via multi-task learning"

Christina Papagiannopoulou[1], Diego G. Miralles[2], Matthias Demuzere[2], Niko E. C. Verhoest[2], and Willem Waegeman[1]

[1]Depart. of Data Analysis and Mathematical Modelling, Ghent University, Belgium,
[2]Laboratory of Hydrology and Water Management, Ghent University, Belgium

Correspondence: Christina Papagiannopoulou (christina.papagiannopoulou@ugent.be)

S1 Importance of a higher-level representation of features

To illustrate the non-linear response of vegetation and explain our choice to use high-level feature representation in our framework, we compare the model performance with and without the use of this high-level representation. Figure S1a shows the predictive performance of the Alternative Structure Optimization multi-task (ASO-MTL) method when the raw variables as well as the corresponding 6-lagged values are included in the model, i.e., the cumulative variables and the extreme indices are not included as predictors. Figure S1b visualizes the difference in predictive performance of the ASO-MTL model with and without the cumulative variables and the extreme indices as predictors. As one can observe, in regions such as Europe, North America, southern and northern parts of Asia and parts of South America, the model performance substantially decreases if these higher-level features are not used in the data representation. In these regions, more than 10% of the variability in NDVI anomalies is explained by this more complex (non-linear) representation, illustrating the non-linear nature of the relationship between climate and vegetation dynamics.

S2 Number of bio-climatic biomes

Figure S2 shows the median of the predictive performance (R^2) for all tasks when the value of the parameter h varies. Note that for these experiments, the λ parameters remain constant in order to assess only the effect of parameter h on the model performance. As one can observe in Fig. S2, the maximum median value R^2 is achieved when h = 11. However, the differences in the predictive performance for $h = 6, .., 15$ are marginal. Therefore, we can conclude that the method gives robust results as the strongest predictive structures are captured for the first most important components given by the singular value decomposition.

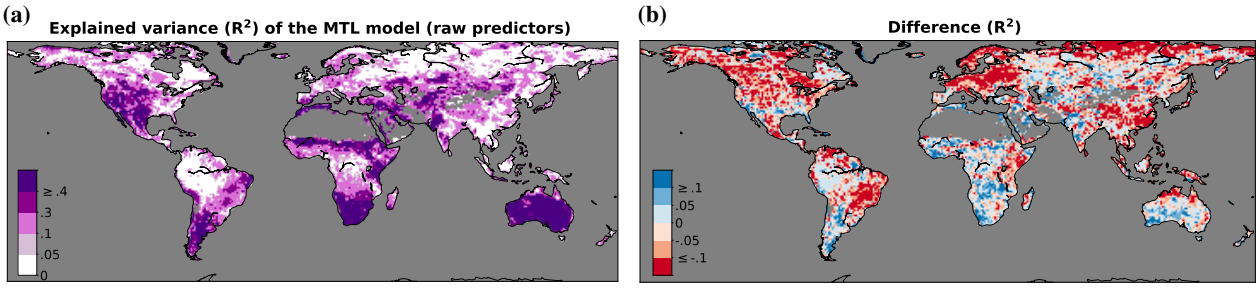

Figure S1. Comparison of the predictive performance in terms of R^2 of the model which does not include the cumulative variables and the extreme indices with the model which is trained with the full collection of higher-level features (Papagiannopoulou et al., 2017a). (a) Explained variance (R^2) of NDVI anomalies based on the raw data of the climatic variables as well as their 6-lagged values (cumulative variables and the extreme indices are not included as predictors to the model). (b) Difference in terms of R^2 between the model without cumulative and extreme predictors and the model which includes all the higher-level feature representation in Fig. 3a of the manuscript.

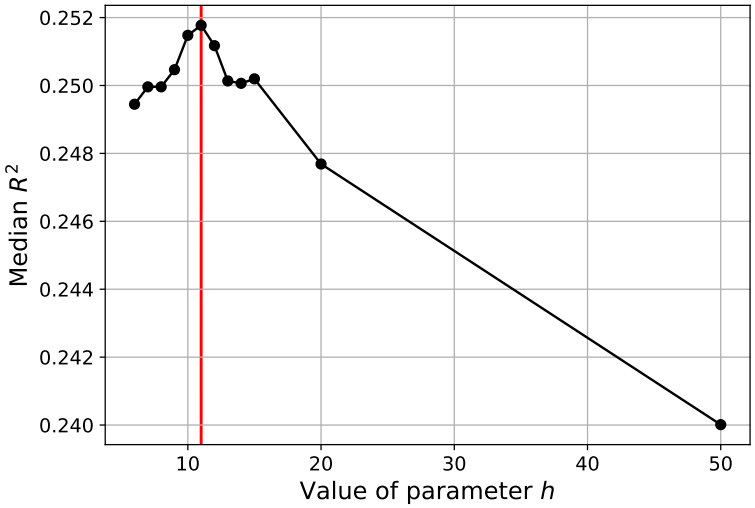

Figure S2. Assessing the number of biomes: Median of the predictive performance of the ASO-MTL model in terms of R^2 when the value of the h parameter varies. For $h = 11$ the model scores the maximum value of R^2. However, the differences in the predictive performance for $h = 6, [...], 15$ are marginal.

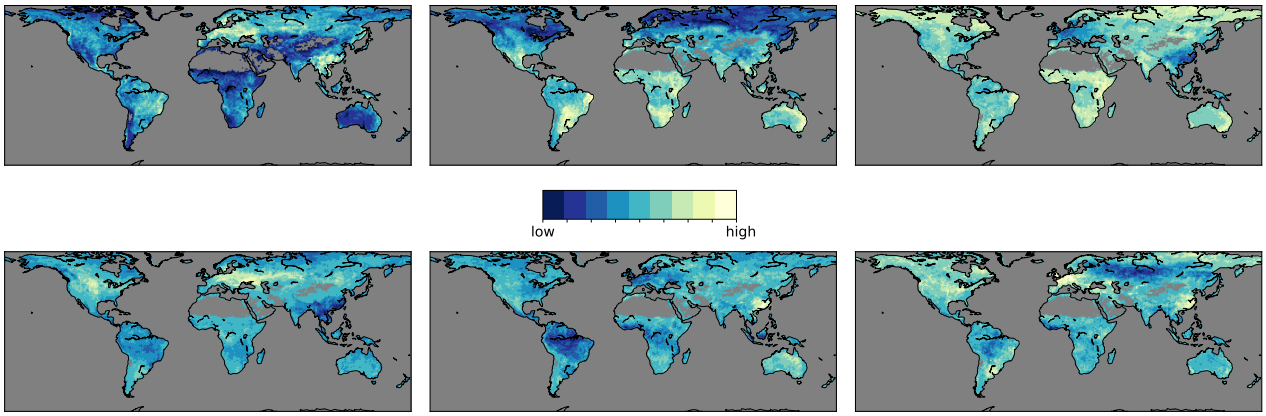

Figure S3. Visualization of the first 6 “principal components" of the predictive structures. The classification of the land surface into the hydro-climatic biomes is based on the importance of these structures for each location. The color intensity in the map indicates the value magnitude of each pixel in a particular predictive structure.

S3 Visualization of the most important predictive structures

In Sect. 2.5 of the manuscript, we describe the steps of the SVD-based ASO algorithm, which learns a low-dimensional feature representation for our tasks based on the relationships between them. The learned matrix $\mathbf{\Theta}$ maps the high-dimensional space to a (lower) h-dimensional space, storing the loads of the original weights to the “highly predictive structures”. Thus, the task models are also projected to this shared lower-dimensional space. This information is stored in the matrix $\mathbf{V}$ on which the clustering approach is performed. Figure S3 presents the values of the tasks in the first 6 components of the matrix $\mathbf{V}$. Similar pixel values to the same components mean similar climate-vegetation dynamics. There are several remarks considering Fig. S3: (1) all the 6 components are able to distinguish specific regions according to different criteria such as regions with temperate and dry climate, regions with cold and dry climate, tropical and dry climate, etc.; (2) pixels which are grouped into the same region in the final clustering result (Fig. 4a of the manuscript) tend to have similar values in a particular predictive structure, and (3) the differences in the values across regions are intense, and in some cases one can recognize the boundaries of the regions depicted in Fig. 4a of the manuscript.

S4 Visualization of the predictive structures with the different land surface classifications

As in Zscheischler et al. (2012), we conduct a dimensionality reduction to the matrix V which contains the clustering data. We separately present the results for the Northern and the Southern Hemisphere (ibid.) – see Figs. S4 and S5, respectively. The data is projected onto the first 2 principal components and visualized based on the Köppen-Geiger clustering (Köppen, 1936) (Fig.S4c and S5c), the IGBP clustering (Loveland and Belward, 1997) (Fig.S4d and S5d), the Granger-causality result which illustrates the main climatic drivers of each region presented in Papagiannopoulou et al. (2017b) (Fig.S4b and S5b) and the hydro-climatic biomes (Fig.S4a and S5a). We use the same color representation as in Fig. 4a of the manuscript. That way we can assess if the learned predictive structures match well the classes of the different classification schemes.

Considering Fig. S4, one can see that the best-formed clusters are depicted in Fig. S4a, as the clustering has been performed on this dataset (as expected). Figure S4a is mostly similar to Fig. S4d which represents the IGBP land use classification; the tropical regions are well-detected as well as the forest- and the cropland-covered regions. This means that the learned predictive structures are highly relevant to the vegetation type of each region. On the other hand, Fig. S4c indicates that the cold, the arid and the tropical regions can be well distinguished by the learned structures whereas the temperate climate is scattered among the others and is thus harder to identify. Finally, the clustering based on the main drivers, in Fig. S4b, is more scattered than the others. However, one can clearly distinguish the water-driven regions from the radiation/temperature-driven ones. This result comes from the fact that the two latter drivers (radiation and temperature) are highly correlated.

Figure S5 depicts the same plots for the Southern Hemisphere. As in Zscheischler et al. (2012), overall, the classes identified by the various classification schemes show a worse match than for the Northern Hemisphere. However, Fig. S5a shows that the predictive structures can clearly distinguish the sub-tropical water-driven region and the transitional energy/water-driven regions as well. In addition, the Köppen-Geiger climate classes (Fig. S5c) of the tropic and the arid regions are also identified in a certain degree. The IGBP classes, in Fig S5d, do not form clear clusters in the plot. Finally, one can notice that the water-driven class based on Granger causality is well separated since water is the most dominant factor in the Southern Hemisphere while energy/temperature- driven regions are rather limited.

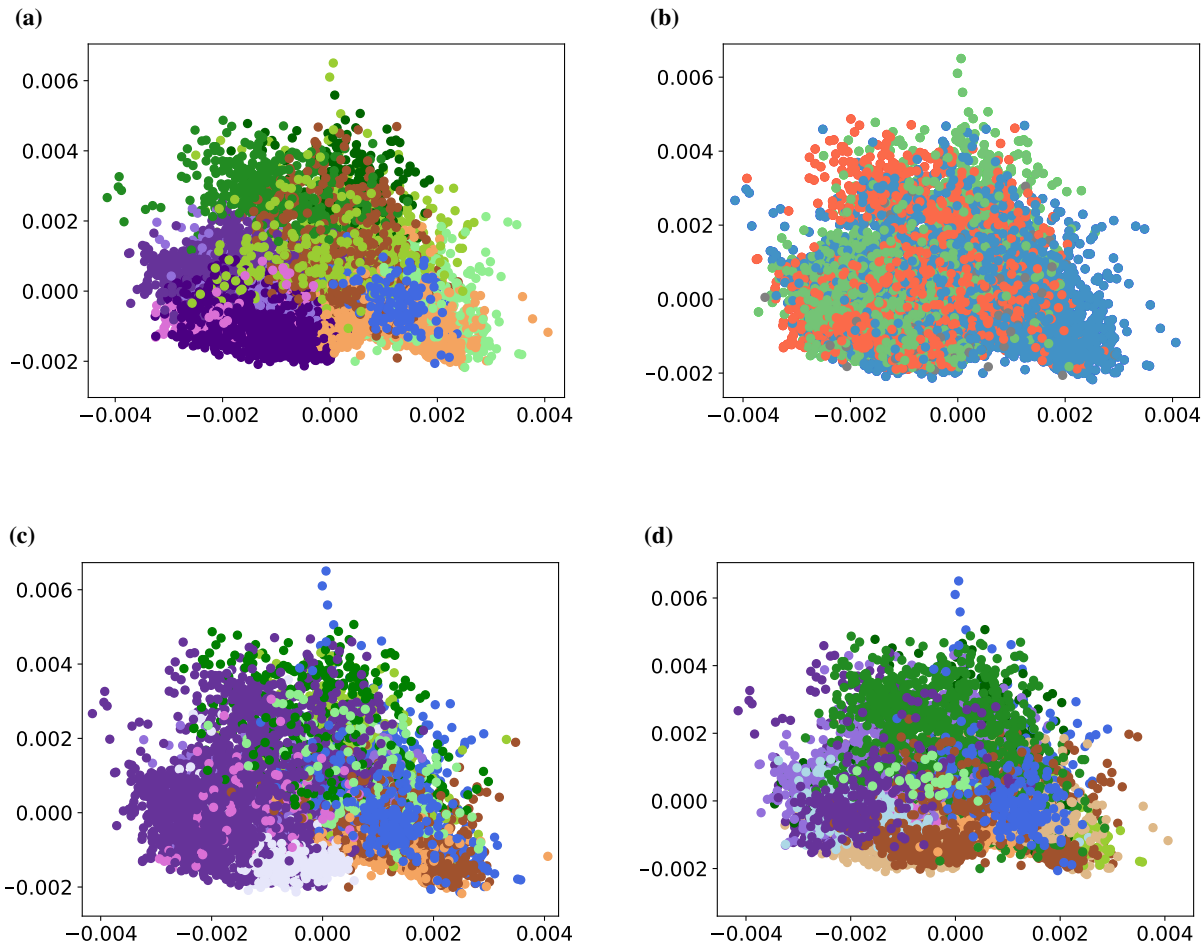

Figure S4. Data projection to the first 2 principal components for the Northern Hemisphere. Each point represents one pixel of the global grid and it is colored based on (a) the hydro-climatic biomes (b) the Granger-causality classification, (c) the Köppen-Geiger climate classification, and (d) the IGBP land use classification. For the color-class mapping see Fig. 4 of the manuscript.

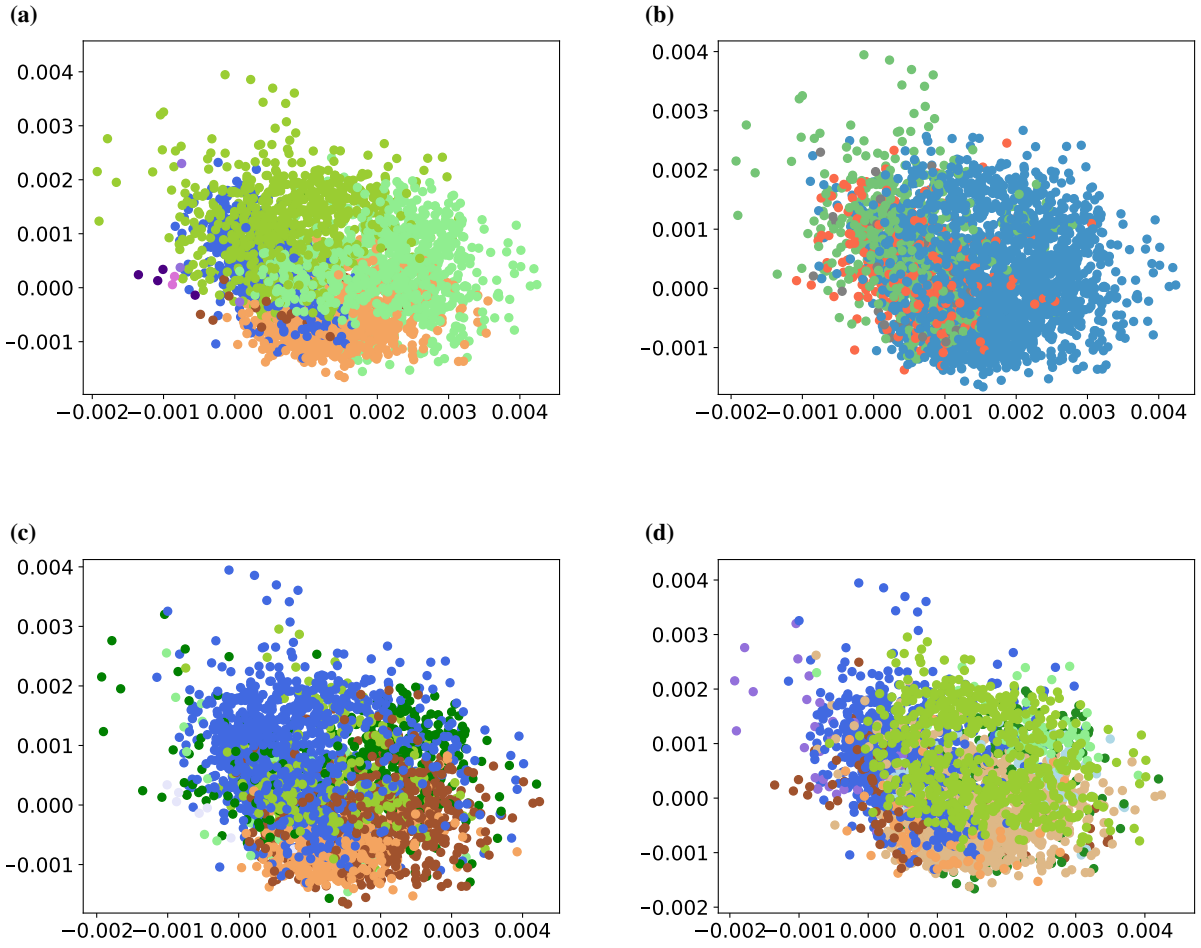

Figure S5. As Fig. S4 but for the Southern Hemisphere.

References

Köppen, W.: Das Geographische System der Klimate, Handbuch der klimatologie, 1, 1936.

Loveland, T. and Belward, A.: The IGBP-DIS global 1km land cover data set, DISCover: first results, Int. J. Remote Sens., 18, 3289–3295, 1997.

Papagiannopoulou, C., Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E. C., Dorigo, W. A., and Waegeman, W.: A non-linear Granger-causality framework to investigate climate–vegetation dynamics, Geosci. Model Dev., 10, 1945–1960, https://doi.org/10.5194/gmd-10-1945-2017, 2017a.

Papagiannopoulou, C., Miralles, D. G., Dorigo, W. A., Verhoest, N. E. C., Depoorter, M., and Waegeman, W.: Vegetation anomalies caused by antecedent precipitation in most of the world, Environ. Res. Lett., 12, 074 016, https://doi.org/10.1088/1748-9326/aa7145, 2017b.

Zscheischler, J., Mahecha, M. D., and Harmeling, S.: Climate classifications: the value of unsupervised clustering, Procedia Comput. Sci., 9, 897–906, 2012.