Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-92-AC3, 2018

Geoscientific
Model Development
Discussions

Interactive comment on "Global hydro-climatic biomes identified via multi-task learning" *by* Christina Papagiannopoulou et al.

Christina Papagiannopoulou et al.

christina.papagiannopoulou@ugent.be

Received and published: 11 July 2018

Response to Anonymous Referee#2

This study presents a new approach for the classification of global biomes. The idea is to focus on the statistical sensitivities of NDVI anomalies to multiple predictors. I do think that it is important to emphasize the "goal" of classification, and therefore the paper is a step in the right direction.

We would like to thank the reviewer for the constructive feedback and thorough assessment. Below we provide a point-by-point response to each comment.

I have, however, doubts if focusing on NDVI anomalies is the right target. In particular for tropical ecosystems NDVI does not tell us much about ecosystem dynamics and the figures show the underlying predictions are indeed not convincing. Hence, I have some doubts about the novelty that this classification can offer. Similar as all classical approaches, also this method fails to reveal the complex spatial patterns in tropical ecosystems. This is why I see this paper more as a methodological contribution that can actually help future studies to realize analogous exercises based on different data sets.

We agree with the reviewer that although NDVI is a commonly-used index, it is known to saturate in tropical ecosystems. As we discussed in our previous work (Papagiannopoulou et al., 2017), the low predictive power of our model in tropical regions can be explained by the fact that in these regions, (i) the uncertainty in the data is larger, and (ii) vegetation might be primarily affected by other factors such as nutrient availability (rather than climate). However, with the proposed data-driven framework, pixels that belong to these tropical regions are grouped together. This means that the learned weight vectors of these pixels are similar and thus the clustering algorithm is able to detect these similarities to conform a coherent biome. Moreover, we also agree that our work can be seen as a methodological contribution, since it can be used in different application scenarios or with an alternative target variable. So, we are willing to explore the applicability of the method to a different target variable. As such, the applicability to microwave Vegetation Optical Depth (VOD) anomalies, instead of the NDVI anomalies, will be explored in the revised manuscript. VOD is known to be less sensitive to saturation in densely-vegetated biomes.

Overall, the approach of the paper is to stack a series of methods. First, "Multi-Task Learning" is used to create a statistical prediction model whose sensitivities (condensed by SVD) later serve as basis for clustering. I applaud the authors for identifying a machine learning method that seems to capture spatial relationships. But my question is if there is no corresponding geostatistical approach out there that could be equally used (e.g. a GWR or so) which deals exactly with such questions? In particular, I believe (but don't know) that the MTL does not consider the fact that lat-lon grid cells represent different geographical distances, or how do the authors considered that a global analysis is executed on a sphere?

As we have described in the manuscript, our approach is purely data-driven. Therefore, we stress that we do not include any prior knowledge about the distances between the different pixels. On the contrary, we let the method learn the relationships between the different pixels. As such, the method may even group together remote pixels in which vegetation might have similar response to climate. Other geostatistical approaches, such as the GWR, assume that neighboring pixels have a similar behaviour with respect to specific variables. In these approaches, similarities between the pixels are learned by defining each time a single pixel as centroid and tuning the parameter of relatedness between this particular pixel and the surrounding pixels. In our work, we prefer to avoid this kind of neighborhood assumptions and focus on the discovery of relationships between the pixels based on the similarity in climate–vegetation interaction. We are also interested in methods that can be applied on large data sets with global coverage. However, we think that the suggested literature (about geostatistical approaches) is relevant to our study. We will refer to it in the revised manuscript.

The paper is neatly written, but I still had trouble finding my way through the paper. One aspect is that it is difficult to follow the paper without knowing the author's previous papers. In addition, I spent most of my time understanding Multi Task Learning. In particular section 2.4. was hard to understand. At this crucial point I would ask the authors to consider rewriting the paper in a way that can be understood intuitively by

environmental scientists who are not familiar with the method. Likewise the link to clustering is a bit opaque. What is a "hierarchical agglomerative clustering approach"? Etc.

We will expand section 2.4 to make the method more intuitive for the broad audience of GMD. Specifically, we aim to provide additional explanations for the notation used in our model. This way, environmental researchers that are not familiar with certain machine learning terminology will be able to have a better understanding of the proposed data-driven method. In the manuscript, it is mentioned that the clustering technique that we use is the agglomerative hierarchical clustering (with Euclidean distance measure) which is a well-known clustering method in Statistics (see Sect. 2.5 and 3.2 of the manuscript). As we mentioned in our response to the Referee#1, we will include in the revised manuscript that we use the hierarchical clustering python implementation of scikit-learn, and add a specific reference.

What irritated me about the results is that the prediction method does not manage to explain more than 40% of the variance (why else would the scale in Fig. 3 a otherwise be cut off at $\geq$ 0.4?). This is actually a bit disappointing and suggests that the regression model was not the right choice, or?

In our study, the seasonal cycle from the NDVI time series is removed. Therefore, the task of predicting the NDVI anomalies is more difficult than just predicting the raw NDVI time series. This is due to the fact that the presence of autocorrelation in the NDVI anomalies time series is much lower. Note that if we target the raw NDVI time series (which includes the seasonal component), the R^2 is close to 1 in most of the regions (Papagiannopoulou et al., 2017). In addition, it is worth noting that there are other factors – such as fires, harvesting, etc. – that affect vegetation dynamics but are not included in the data set. Therefore, we should be aware that we focus on explaining the variance of the NDVI

anomalies, taking into account only climatic variables.

Minor remarks:

The introduction does not provide a systematic overview of alternative approaches. Rather, we find here a rather random selection of climate and land cover classifications and the wording is not always correct. For example, the paper speaks of “big data" approaches, but I did not find any of the referenced studies really dealing with big data topics (“volume", “diversity", “speed", ...). I think we are talking here about (sometimes semi-heuristic), but essentially classical data exploration and machine learning methods. So, I think it would be nice to revise this part a bit to have a smooth start.

In general, we would like to stress that the goal of our study is to provide a new data-driven methodology that can identify coherent regions in which vegetation responds to climate in a similar way. To the best of our knowledge, there are no other works that study this particular problem at global scale, with the arguable exception of the article pointed to by Referee#1 (Ivits et al., 2014). In addition, in the manuscript, we describe the most naive approach that one could follow by using single-task learning techniques (and by learning one model per pixel). In the Introduction, we provide an overview of the most related works to our study that indeed use machine learning methods and/or prior knowledge. The term “big data” is used to explain that data-driven methods have been applied on climate data sets, which are inherently characterized by their volume, diversity, etc. We think that our work builds upon and goes one step further from previous efforts, such as the ones described in the Introduction, since it combines information from climate and vegetation and models the relationship between them. We will clarify these aspects in the revised manuscript and include literature related to geostatistical approaches

used in modelling climate–vegetation interactions (see e.g. Zhao et al., 2015).

The paper is full of shortcuts such as “detrended seasonal NDVI anomalies", which are not as clear as they appear at first glance. I can think of a large number of possibilities for robustly estimating (linear/non-linear) trends and a further variety of methods for estimating seasonal cycles. It would be nice if such statements were more precise.

We agree that these terms are not clearly described in the manuscript, and understand that the article should stand alone without the need of prior knowledge with regards to our previous work. We will add additional statements to briefly describe this terminology.

The same comment applies to the selection of predictors e.g. seasonal anomalies, detrended seasonal anomalies, time delayed variables, and cumulative variables etc. look like a very arbitrary selection of predictors. In a paper that has a strong affinity to data-driven methods, I would expect a more formal variable selection following a clearly defined cost function. Maybe this is too late now, but still one question can be answered: why are these predictors all regarded as “non-linear"? In most cases, they read like fairly linear transformations (maybe with the exception of cumulative variables)

We refer the reviewer to our previous answer for the first part of the comment. In addition, we would like to stress that our choice to use this set of predictors is based on the previous literature, as it has been analytically described in Papagiannopoulou et al. (2017). These constructed predictors are regarded as “non-linear”, because their derivation from the raw data is not linear (see e.g. calculation of extreme indices).

References

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., 2017, 10, 1945-1960

Ivits, E., Horion, S., Fensholt, R., Cherlet, M. Global Ecosystem Response Types Derived from the Standardized Precipitation Evapotranspiration Index and FPAR3g Series. Remote Sensing, 2014, 6, 4266-4288.

Zhao, Z., et al. Exploring spatially variable relationships between NDVI and climatic factors in a transition zone using geographically weighted regression. Theoretical and Applied Climatology 120.3-4, 2015, 507-519.