Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-92-RC2, 2018

Geoscientific
Model Development
Discussions
Open Access

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

Anonymous Referee #2

Received and published: 4 July 2018

This study presents an 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. 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.

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 captures 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?

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.

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?

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.

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.

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).

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