Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-53-RC2, 2018

Geoscientific
Model Development
Discussions

Interactive comment on "Topological Data Analysis and Machine Learning for Recognizing Atmospheric River Patterns in Large Climate Datasets" *by* Grzegorz Muszynski et al.

S. Thao (Referee)

soulivanh.thao@lsce.ipsl.fr

Received and published: 9 October 2018

General comments:

The paper presents a method to detect atmospheric rivers (ARs) in climate datasets. Unlike most existing methods, this one relies on marching learning and learns a classification rule for the detection of ARs based on a training dataset. In my opinion, one novelty of the paper lies in the choice of the features used for the classification. From maps of integrated water vapor, new features are constructed from topological data analysis that could me more suited for the problem.

In general, I think that the paper is well written and I appreciate the pedagogical effort made to clearly explain the methodology as well as the illustrations of cases where the algorithm performs well and not so well. Hence, I don't see any major reasons not to published the paper. I only have a few comments and suggestions that I think could benefit the paper.

Specific comments:

1. I find the use of the term "threshold-free" is maybe not the most appropriate. While I understand that in most of the cases, "threshold-free" means that the method does not rely on a fixed, predetermined, arbitrary threshold for the detection of ARs, thresholds are still used several times during the proposed procedure. Indeed, the goal of the SVM step is still to learn a threshold to separate the ARs from non-ARs from the training set and the topological features. The topological features are also constructed from a set of thresholds. (And to be more provocative, for now, the labels in the training set were also generated by an AR detection methods using thresholds). For me, the value of the paper is that it shows that if a we have a good training dataset, there is more efficient way to build this decision threshold than manually tinkering parameters of the classifier/detector.

2. In the same way, I am not sure I understand the following sentence from the abstract and the conclusion (p17, l-14-15) : "We anticipate that because the method is threshold-free, it can be 15 applied to different climate change scenarios without any tuning". If the statistical relationships between features and the target variable change through time, should not you retrain the SVM as the other methods have to reevaluate their thresholds ?

3. I think the explanation on the SVM could be improved if Figure 7 was split into 2: the first figure would illustrate the (linear) SVM and the different quantities in equations (3) and (4) (see e.g. https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Svm_max_sep_hyperplar

The second figure would focus more on the “kernel trick”. For instance, it would show a case were a linear classifier could not separate the two classes in a 2D space but would managed to do it if data were mapped into a 3D space.

4. (P9, l7) “The kernel function that maps the input space into a higher dimensional space ...”. I think the sentence can be a little bit nuanced. As far as I understand, the kernel function returns the inner product between two points projected into higher dimensional space by a mapping function phi. Each kernel function is implicitly associated with a mapping function phi (which does not need to be known for an actual application and that’s one of the strong point of kernel methods). That’s why the function phi is called a kernel induced implicit mapping.

5. (P9, l12) “ applying loose grid-search and fine grid-search for these two parameters”. Do you use grid search with some kind of cross-validation scheme?

6. I think it should be clearly mentioned in the main text or in a table how many data points were used in the training set and the test sets. We could try to deduce it from confusion matrices but it is not very practical.

7. In the same way, for table 3, 4, etc . . . , the number of snapshots mentioned, is it for the test or training sets ?

8. (P18, l1), Authors compare the computing time of their algorithm with the one of Liu et al. (2016) thats uses deep learning. How do both methods compare in terms of performances ?

9. For the sake of reproducibility, it would be nice to at least provide in supplementary materials, details about the actual implementation of the methods. For instance, the programming language used, the potential external softwares/packages/libraries used and for which step of the method.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-53, 2018.