

Ann. Geophys. Discuss., referee comment RC2
<https://doi.org/10.5194/angeo-2021-33-RC2>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on angeo-2021-33

Anonymous Referee #2

Referee comment on "Unsupervised classification of simulated magnetospheric regions" by Maria Elena Innocenti et al., Ann. Geophys. Discuss., <https://doi.org/10.5194/angeo-2021-33-RC2>, 2021

In the manuscript under review the authors have applied unsupervised machine learning algorithms to analyse global magnetospheric simulation data obtained from OpenGGCM-CTIM-RCM code. The automated classification process uses principal component analysis for input data dimension reduction, self-organizing maps for training of an artificial neural network and K-means for cluster extraction from the trained neural network. These unsupervised machine learning algorithms offer an automated way to determine clusters in the physical simulation data. The results shown in the paper are surely of great interest to space physics. The paper also displays the advantages and performance of the unsupervised clustering algorithm.

The paper contains interesting results, but contains no new method development, rather it presents an automated algorithm comprised of multiple unsupervised clustering algorithms. The application of methods is in overall well executed and explained.

There are points for revision that need to be addressed before advising for publication and minor suggestions for consideration. They are listed as following.

Major comments:

1. The authors do not discuss the stochastic nature of artificial neural networks, their sensitivity to initial conditions and convergence to local minima. Clarification for this issue would be needed to the section discussing Self Organizing Maps.

2. The nature of the hyper-parameters of the Self Organizing Maps are only briefly discussed in Section 3, more clarification on their influence to the algorithms performance

is needed. A case study is done in appendix A, but a more theoretical description of the nature of the hyper-parameters should be added.

3. In Section 3 the authors do not mention what distance metric is used for the matching rule in Equation (1) and updating rule in Equation (2) of the SOM.

4. Clarification should be added about how the authors initialized the neural map of the SOM.

5. The influence of sampling of input data during learning needs to be discussed.

6. The authors should describe how much confidence they have in the result of the SOM.

7. In Section 4.2 the model validation is done only by visual inspection. The colors of similar clusters differ from image to image. This fact also makes the readability of the figure very hard, as the eye automatically matches colors. A quantitative measure should be introduced for the robustness of the SOM. To discuss Figure 7 data, one could simply calculate for each result comparison pair for the same data a percentage of similarly labelled data points. Human labelling of the 7 clusters is already previously done (Figure 4 panel (d)). The same labeling system could be used for the K-means classification results displayed on Figure 7.

8. Clarify what time snapshots and data is analysed on Figure 8. Do the panels on the figure correspond to the same dataset with different training features? Clusters depicted on Figure 8 follow different coloring schemes, which diminish the readability of the results. To better quantify the comparison of independent algorithm results on the same data a quantitative measure should be added. Similarly to the previous comment 7, Figure 4 panel (d) labeling system can be applied to all results. One could count the fraction of similarly labelled input data points for each compared pair of independent results for the same data.

9. Similar quantitative robustness measure should be added for the comparison of independent results for the same data in Figure 9.

10. Use the same coloring scheme for all cluster analysis results (Fig 6 – Fig 9) as different colors lose information in Figures. In Figure 4 a labelling system with colors is created, which could be used for all figures. Only visual comparison of clusters over multiple images throughout the paper with different color coding is tiresome.

11. The comparison of panels between Figure 4, Figure 6, Figure 7, Figure 8 and Figure 9 would be easier if all the captions of the panel names would contain the time snapshot information. Panels on the Figure 6 have different sizes, panel sizes on a Figure should be uniform.

Minor suggestions and questions:

L167: Does the dimension of the input data influence the training time of the map considerably? Are there other metrics that could possibly influence it more? Can the PCA be used to initialize the neural map of the SOM? Is the main motivation for input data dimensionality reduction to have more reliable results from training of the SOM?

Fig 3: the name of panel (a) is in different size.

Fig 4: the style of panel descriptions differ in the Caption. A more uniform style would increase the readability of the figure.

Fig 7: the name of panel (b) is in different size.