Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/nhess-2021-110-RC1, 2021

Comment on nhess-2021-110

Anonymous Referee #1

Referee comment on "Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations" by Elizaveta Felsche and Ralf Ludwig, Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-110-RC1, 2021

As presented by authors this study consists of two parts:

the first part focuses on a systematic **search** for the best performing setup of ANN models for Munich and Lisbon.

the second part focuses on the **analysis** the best performing models using explainable AI methods.

The set-up of this paper seems to have significant problems as both the search and the analysis have been performed using the same data-set (TEST – set) which actually a very small data set (years 2000-2005, only five years of monthly data for SP1 case).

The authors should have used a hold-out set to investigate the actual performance in "unseen" data.

In any case both the architecture selection, loss functions performance appear with really low F1 score values (below 0,3 in test set) – whereas the authors have stated earlier "we require that the accuracy on each class is at least 50%".

The paragraph that presents the model architecture is not clear. How many layers and neurons do we have in the selected model(s)?

This performance cannot and should not be considered as appropriate for a forecasting model. Therefore, both models (Lisbon and Munich) cannot be used for drought prediction.

This is something that the authors actually acknowledge as they state “The precision of the prediction in both cases was rather moderate, as a high percentage of data is misclassified”.

The second half of the study presents the analysis of the performance obtained architectures.

In the 3.2.1 Shapely values section, we do not know which data set has been used – we assume that we are looking at the Test set. For Lisbon the cumulative contribution of the top 15 variables (out of the 27) is 20% which should explain the underperformance of the selected architecture. The case of Munich is even worse as the cumulative contribution of the top 15 variables is less than 5%.

Similar performance can be seen with seasonality analysis.

Last, in conclusion (line 290) the author state “Best performing models obtained accuracies of 57% for the Lisbon domain and 55% for the Munich domain”. This is not true as this performance has been seen in train set , not the test set. Even if it was in the test set it would have been insufficient as the model has already seen the information in the data set and therefore should not be considered for forecasting performance evaluation.