Authors reply to anonymous referee #1 are provided point by point in blue characters.

Comment on nhess-2021-110

Anonymous Referee #1

Referee comment on " Applying machine learning for drought prediction 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

Title: Applying machine learning for drought prediction using data from a large ensemble of climate simulations
Author(s): Elizaveta Felsche et al.
MS No.: nhess-2021-110
MS type: Research article: First review
Special Issue: Recent advances in drought and water scarcity monitoring, modelling, and forecasting (EGU2019, session HS4.1.1/NH1.31).

RC1: 'Comment on nhess-2021-110',

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 are using a large ensemble of 50 members for the study. This means that the period 2000 – 2005 is available 50 times, resulting in 250 model years for the test dataset. The same is true for the training dataset: There the period of 1957-1999 is to be multiplied by 50, resulting inf 43*50=2150 model years. The authors apologize for the misleading description and will work on clarifying this in the manuscript.

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

The authors consistently used “unseen” test data to evaluate the performance of the algorithms. During the training of the algorithms an additional validation set was selected from training data, to prevent overfitting and monitor the performance. A similar methodology was used e.g. in Morid et al. (2007).

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%”.

Authors thank the reviewer for the comment. The accuracy on both classes is at least 50%, however, due to the class imbalance the marginal F1 Score is low. An explanation to this behavior is given in P.

9, L. 180-182: “Due to the class imbalance within the dataset we require that the accuracy on each class is at least 50%. In that case given the distribution of the test dataset of 1803 non-drought events to 387 droughts for Lisbon and 1848 non-drought events to 352 drought events for Munich a marginal F1-score of 0.26 for Lisbon and 0.24 for Munich is given.”

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

The model architecture consists of overall seven layers; two of those are Dropout Layers, which are setting in a random way half of the neuron outputs to zero. Five layers are Dense neuron layers. The architecture of the layers is given in Table 5. For example the architecture for the model mentioned in the first line of Table 5 is the following:

1. Dense layer with 4000 neurons
2. Dropout Layer randomly setting 50% of weights to zero
3. Dense layer with 1000 neurons
4. Dropout Layer randomly setting 50% of weights to zero
5. Dense layer with 500 neurons
6. Dense layer with 100 neurons
7. Dense layer with 5 neurons

The authors apologize for the misleading description and will work on clarifying this in the manuscript.

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”.

Authors thank the reviewer for the comment and will add it to the limitations in the revised manuscript version.

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.

The analysis was performed on the test dataset. This will be added to the revised version of the manuscript.

A series of twelve months of each variable was taken as input to the machine learning model. For the calculation of Shapely values each month of each variable was considered individually, resulting in 27 atmospheric variables *12 + 6 teleconnection indices * 12 = 396 variables. This means that the cumulative contribution of the top 15 variables out of 396 variables (not 27) amounts to 20%/5% for

Lisbon/Munich. Authors thank the reviewer for the comment and will add the above explanation to the revised manuscript version.

The authors are aware of the limitations and would argue that although there is a comparably weak performance the obtained results can be of huge value for the development of a forecasting model.

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

The stated performance was seen on the test dataset, which was not used for model training, therefore the authors argue that the result can be used for the forecasting performance evaluation. In P. 10 L. 210-212 the authors state that only the results on the training set are shown: “. Training results are displayed in this particular case [L2 regularization] since the regularization is introduced to prevent overfitting. Generally the performance on the test dataset is more important and will be inspected in following experiments.” The authors apologize for the misleading description and will work on clarifying this in the manuscript.