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

Review 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 included a clarification in the revised version of the manuscript (see L. 133-135).

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 only the training dataset is used as input information to the model. 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. 183-185: “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. A clarification is added in lines 234-240 of the revised 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".

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 is added to the revised version of the manuscript e.g. in lines 250-251.

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 28 atmospheric variables *12 + 6 teleconnection indices * 12 = 408 variables. This means that the cumulative contribution of the top 15 variables out of 408 variables (not 27) amounts to 20%/5% for Lisbon/Munich. Authors thank the reviewer for the comment and added the above explanation to the revised manuscript version in lines 251-252. In first version of the submitted document it was claimed erroneously the total number of variables to be 41, instead of 42. This error is corrected in the revised version of the manuscript.

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 and therefore the model has not seen the information in the test dataset. The authors argue that the result can be used for the forecasting performance evaluation. In P. 10 L. 214-215 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. A similar methodology was used e.g. in Morid et al. (2007).

Authors reply to anonymous referee #2 are provided point by point in blue characters.

Review on nhess-2021-110

Anonymous Referee #2

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-RC2, 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).

RC2: 'Comment on nhess-2021-110',

This study present a methodology for drought prediction at seasonal scale using machine learning algorithms. The study is treating a highly relevant subject with the usage of a novel methodology based on machine learning for droughts predictions. However, the issue with machine learning and climate is the length of the climate records, which does not allows to build AI models. Therefore, this study proposes a methodology fully based on a down scaled ESM. The study is interesting however, I have major comments listed bellow, in particular, at that stage it is very hard to evaluate properly the manuscript since the data/method is not clear enough:

1) the method (if I understood it correctly) is fully based on model data, therefor it is not really a study about prediction but according to me it is only potential predictability, since this study does not demonstrate any skill in predicting observed past climate in the two regions of interest, but only the ability to forecast the model climate. The paper should be much clearer about this, for example the title and the abstract should use the term “perfect model framework” and/or potential predictability.

Authors thank the reviewer for his/her useful suggestion. The title of the revised version of the manuscript is:
“Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations”

2) The method description is very unclear about the prediction aspects. What are the target month analyzed? From which start date? For example, it is really confusing to me to predict SPI1 with one month lead time for different seasons. What do you mean here? Do you mix all together the start date of March (to predict the SPI1 of April), April (to predict the SPI1 of May) and May (to predict the SPI1 of June)? Or do you predict SPI1 integrated over MAM, but in this case, to my understanding it is not SPI1 but SPI3. In any case the methodology should be much clearer about this point, at this stage I cannot evaluate properly the manuscript without this clarification.

The authors apologize for the misleading description and added a clarification this in the manuscript in lines 141-144. The study predicts SPI1 with a lead time of one month. To predict e.g. SPI1 in April of 2000, the data for twelve months before the event is used as input, this is SPI1 and other variables

for the period April 1999 – March 2000. For the calculation of SPI1 in April only precipitation for the month of April is used.

3) "The data from the years 1957 - 1999 was used as training data, the years 2000-2005 were used for the testing purpose." Do you mean that the score calculation is performed only for 6 years from 2000 to 2005? This is a far too short period for any skill assessment. Usually, in seasonal prediction the skill is assessed over the whole hindcast period (1957-2005), using cross validation to construct the prediction.

We 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. In order to obtain the total number of available datapoints those numbers need to be multiplied by 12, as we are using the value for each months as individual input to the model. The authors apologize for the misleading description and added the details to the description in lines 133-135 of the revised manuscript. The authors are aware that cross validation would add value to the study, however given the huge amount of variables and the fact that the overall analysis includes 2500 model years, it would require huge computational resources, that are not available.

4) The discussion does not mention at all the main limitation of this study according to me: at that stage the authors have demonstrated some ability to predict a model using AI, but we don't know how to use such method for real prediction. Would it be possible to apply your model on observation and then verify its skill? If yes, it should be included in the study and if not this should be clearly mentioned.

Authors thank the reviewer for his/her useful suggestion. The authors would argue that the immediate application of the framework on observation is not possible, due to the fact that observational data usually lacks a multitude of variables which were used as input in this study e.g. Heat Fluxes, radiation, etc. The objective of this study was not to develop a framework that can be applied on observation, but to use the large amount of events provided by the large ensemble for prediction. The results obtained by shapely value calculation are of high importance for the choice of variables for a development of a model which could be applied to observational data.

Typos:
This study uses the monthly sea level pressure (pr)

The stong influence of ps/psl and NAO shows the influence of the atmospheric pressure

Typos are fixed in the revised manuscript version.