egusphere-2025-1891

"Hydrological drought prediction and its influencing factors analysis based on a machine learning model"

This manuscript is the latest contribution to a growing amount of literature that seeks to develop machine-learning predictions for droughts. It is an interesting paper that contains an advance analysis on the interpretation of the predictions. Machine learning is, compared to other hazards, under-researched for droughts, and therefore the results of this paper are very relevant to the wider scientific community. However, there are some significant shortcomings that need to be addressed before the manuscript can be considered for publication in NHESS. These main concerns are listed under "general comments".

General comments:

1) My main concern comprises the lack of focus on the lead time of the predictions. While the abstract starts with "Predicting future drought conditions", it is unclear how this "future" is represented in the paper. This is a recurring pattern throughout the whole paper, and is missing in all sections, including the methods and results (Figure 2, 3, 4 and 5). It is crucial for the scientific quality and relevance of this paper, that it becomes very clear on which lead time (i.e. how far in future) the model is predicting, and for each Figure to be clear for which lead time the results apply. If only the current timestep is used (i.e. no lead times are implemented), this should be clearly mentioned and reflected upon.
2) While the introduction is generally well-written, the article would benefit from a stronger foundation of the research gap. A lot of attention is now focussed on the explainability of the XGBoost algorithm. I do really not agree with the narrative that XGBoost is very explainable, actually it is a very complex algorithm with many hyperparameters, multiple trees and difficult algorithms (.e.g. tree boosting). On line 76, you state that "At present, there are few studies on interpretable machine learning using the SHAP algorithm." This is not true: SHAP is an extremely popular and frequently used ML algorithm, also in the field of (drought) forecasting. Therefore, instead of this narrative, it would be better to give more examples of previous studies deploying SHAP on ML (drought) predictions (including in the discussion, see point 7, below). Moreover, the research gap should be elaborated based on 1) why ML is used, instead of simpler prediction methods (e.g. linear regression), with examples from previous studies showing that ML has a better performance, 2) more regional context: why is an (earlier) prediction of drought needed in the specific case study region, and 3) what is it exactly that we do not yet know about the drivers of drought, which need to be discovered with the SHAP algorithm.

3) The methods (3.1) should start with a crystal clear overview of the modelling setup (training period, test period, validation period, instead of only in the results, line 250), and all data used. It is now insufficiently clear which variables/data are input to the model (features), and which are output (the target variables), and on which lead times. I propose to make one table including all those components. I think this new Section 3.1 and its table should include all information now listed in Section 3.3, Section 2.2 and Table 1, 2 and 3. Furthermore, it is also unclear how the lead times are implemented, and this aspect should be better represented. Therefore, suggestion to add an extra separate column for "Lead time" (including T=0) and add that the unit is "months"
4) The section about SHAP (Section 3.5) should focus less/not on the formulas, but should better explain the basic principle of SHAP. It should include that SHAP values reflect the 1) local feature importance, instead of global feature importance and 2) that the SHAP value is calculated with reference to a SHAP baseline prediction, and it should be clearly stated which baseline prediction is used.
5) A confusion matrix, showing the number of true/false positives and true/false negatives should definitely be included to shed light on the interpretation of the recall/precision. In case of class imbalance (relatively many true positives), even a useless model could generate high skill scores. I suggest to add another skill score to make your conclusions more robust, such as the Heidke Skil Score.
6) The results can be better visualized with the observations (Figure 3) and predictions (Figure 4) next to each other, in one figure. For this figure, I would suggest to select some months, and move the figure(s) with all months to an appendix).
7) The discussion section should be improved, with a stronger reflection on how the results can lead to better "drought mitigation and water resource management" in practice (including the lead times considered and its implications for proactive/anticipatory drought management, and the consequences/reasons for the relatively poor performance for more extreme drought classes). Moreover, it should provide much more insights into how the results found (e.g. strong importance of SPI) compare to the current state-of-the-art literature.

Specific comments:

a. Line 8: suggest to remove "interpretable": XGBoost is not specifically interpretable compared to other decision-tree type algorithms. Whether it is

interpretable, depends on if a thorough analysis has been executed to better understand the predictions

b. Line 9: change “factors” to “features” (also in the rest of the manuscript), and add the 4 drought categories. At a similar note, the “drought impact factors” should be renamed to the “target variable” of the model. The model does not predict real impacts, which is addressed in the general comments section.
c. Line 11: 79.9% accuracy in classifying droughts. On which lead time (i.e. how far in future) are those predictions? This is critical information.
d. Line 132: what do you mean with “Using the interpolation method in array”? Xarray package?
e. Suggestion to put all formulas (except the recall/precision) into the appendix/supplementary material.
f. Line 145: index to indices (plural)
g. Table 1: > 2.0 instead of < 2.0, and abbreviations should be included (as used in the results, e.g. Figure 2).
h. Line 338: More context is needed for the SHAP plot. How can the direction of the relationship of the variables be explained? I see that higher SPI-1 values lead to higher model predictions. Does this mean a higher SPI value leads to stronger drought conditions? Again, it is not entirely clear what the target variable is here.
i. Figure 7: It seems like you can delete this figure, as it is very similar to Figure 6.
j. I really like Table 5 and Figure 8. However, the colors in Figure 8 are not easy to distinguish (e.g. the AMO looks like wind speed..)
k. Figure 11: axis should be labelled.

Technical corrections:

a. References should all be double-checked and aligned with the bibliography (e.g. line 27, American Meteorological Society, 2013, is incorrect and not listed).
b. Suggestion to abbreviate machine learning to “ML”
c. Review the article carefully on typos. I found the following:
    a. Line 104: typo in “large-scale
    b. Line 273, “however”