We warmly thank the editor and the referees for carefully reading the manuscript and their valuable comments. The author's answers are provided point by point in blue characters.

Comment by the editor

I have reviewed the latest revised version of the paper, the reviewers' comments. The authors have addressed most of the reviewers' comments during the first review round. However, I would agree with the comments made by reviewer#3 during the second review round. The revisions requested by the reviewer are quite moderate and valid. Reviewer #2 has accepted the paper for publication. I suggest that the authors address the revisions required by reviewer#3 and send the revised paper for a final review by the editor and the final decision.

Report #1 by Anonymous Referee #3

The paper deals with an important issue in the field of changing climate and the use of machine learning methods to identify drought conditions. In my opinion, the paper has been prepared in a good manner and presents adequate technical and experimental details. The paper is a novel study and is generally well-structured as it explains the methodology, the mathematical framework and the assumptions used, and the justification of the results and the conclusions. However, the application research part needs improvements and corrections to verify the novelties of the method employed in the study area. Furthermore, there are few critical points that should be addressed in the revised manuscript. Addressing these comments will improve the quality of the paper and help the general reader of the paper. The paper could be accepted for publication considering the following revisions.

1. Why the authors use the parametric Pearson correlation (ρ) and not a non-parametric test (like Kendall tau and/or Spearman rho)? In case of nonlinear correlation between climatic signals and local drought values, statistics such as mutual information (MI) could be more informative than the conventional correlation coefficient. Therefore, I suggest to present mutual information as another statistic in Table 1. The authors state (lines 116-117) "The correlation coefficients reveal that out of the full list of 42 variables 14 are sorted out as being redundant". Hence, nonlinear and/or non-parametric statistic values should be added to verify this conclusion.

The authors thank the reviewer for the helpful suggestion. An explanation for the necessity of variable subset selection is added in the L. 107-108. The authors prefer not to include mutual information for the following reason:
As this calculation step aims to omit redundant variables, we are interested in the upper bound of the value. By definition, Mutual Information (MI) has no upper bound (Strehl et al., 2002). Therefore, MI is not comparable between the different variables. A normalized version of MI by Strehl et al. (2002) then approximates MI akin to the Pearson correlation coefficient.

2. Justification of the selected timescale of SPI and the selected forecasted lead-time. Why the authors use the SPI-1month?
Due to the overall complexity of the input dataset with 28 atmospheric and soil variables and 2500 model years that are used for the analysis, our goal was to find a robust setup in terms of lead-time to explore the influence of the prediction variables on the prediction. As shown by previous studies, shorter prediction lead times are usually more robust than longer periods (Belayneh et al., 2012). To calculate SPI3/SPI6, precipitation values for the preceding three/six month months are used. As noted in Yoon et al. (2012), when performing a prediction of a lead-time less than the accumulation period of the SPI value, the skill of the forecast can largely be explained due to this relationship.

Therefore, the authors argue that the accumulation period should not be chosen any lesser than the lead time to evaluate the effects besides the explained relationship. Therefore the authors chose to use a lead-time and accumulation period of one month. One month lead-time was also used in previous studies by Yoon et al. (2012) and Deo et al. (2017). The authors have added the explanation to the revised version of the manuscript.

How reliable is the SPI-1 for drought prediction at the study sites?

The two study sites have hugely different meteorological conditions, especially in terms of precipitation averages throughout the year. While Lisbon has a Mediterranean climate, Munich has a continental one. As noted by Zargar et al. (2011), SPI is essentially a measure to compare the precipitation departure from normal, and therefore it is a measure that applies to highly different climates and makes them comparable. As the complexity and related uncertainties of the prediction rise with extended lead times, the authors chose a lead time of one month. Therefore, the accumulation period of SPI had to be also chosen one month, as explained in the previous answer. Unfortunately, no comparable studies exist for the two domains to be able to evaluate the prediction performance.

Why the previous 12 months are used as input for predicting the SPI of the next month (please provide scientific evidence for this assumption and why not testing up to 36 months before)?

The twelve months before the event are chosen following the study by Morid et al. (2007), which found that the best performing drought prediction model was the one including the value up to twelve months before the predicted one. The explanation has been added to the revised version of the manuscript in L. 150-152. The authors are aware that experimenting with enlarging the input period might improve the algorithm's performance. However, given the huge amount of variables and the fact that the overall analysis includes 2500 model years, it would require substantial computational resources.

3. The results clearly show (see Accuracy and F1-score) that the final selected models cannot be used for forecasting purposes for the selected drought index. The Threat Score (TS) or the Critical Success Index (CSI) could be used to verify if the proposed method should be used (https://www.cawcr.gov.au/projects/verification/Hewson/DeterministicLimit.html). Furthermore, I have the impression that the selected timescale of SPI (1-month) leads to unstable results. If the authors use for example the SPI-6 month how different would be the derived results?

As described in comment 2), the authors would expect the performance to improve for the prediction of SPI6 with a lead time of one month due to the fact that five out of six values needed for a numerical calculation of SPI6 would be given as input to the algorithm. Therefore, the study's approach was to display the dependencies beyond the ones given by the definition of the drought index.
The authors thank the reviewer for the useful suggestion to introduce additional verification by calculating the Threat Score (TS)/Critical Success Index (CSI). However, the authors would argue in accordance with Jolliffe et al. (2002) that the score is less suitable for our problem, as it is highly dependent on the frequency of the event and therefore biased. Instead, the Heidke Skill Score (HSS) is proposed as an unbiased version given the high class imbalance. For the best-performing models, HSS equals 0.06 for Lisbon and 0.04 for Munich. These results confirm that the obtained prediction is better than the random forecast and therefore show a weak prediction skill. The HSS is added to the revised version of the manuscript.

Minor comments

1. A flow chart of the proposed method may also be added. The authors are requested to ensure that international readers/scientists will be able to apply this methodology on their data sets by following the flow chart.

Authors thank the reviewer for the useful suggestion. A flowchart is added to the revised version of the manuscript.

References

Belayneh, A., Adamowski, J., Khalil, B., and Quilty, J.: Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction, Atmospheric Research, 172, https://doi.org/10.1016/j.atmosres.2015.12.017, 2016.

Deo, R. C., Kisi, O., & Singh, V. P. (2017). Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmospheric Research, 184, 149-175.

Jolliffe, I. T., & Stephenson, D. B. (2012). Forecast verification: A practitioner's guide in atmospheric science. Chichester, West Sussex: Wiley-Blackwell.

Morid, S., Smakhtin, V., and Bagherzadeh, K.: Drought forecasting using artificial neural networks and time series of drought indices, INT J CLIMATOL, 27, 2103–2111, https://doi.org/10.1002/joc.1498, 2007.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, 3(Dec), 583-617.

Yoon, J. H., Mo, K., & Wood, E. F. (2012). Dynamic-model-based seasonal prediction of meteorological drought over the contiguous United States. Journal of Hydrometeorology, 13(2), 463-482.

Zargar, A., Sadiq, R., Naser, B., & Khan, F. I. (2011). A review of drought indices. Environmental Reviews, 19(NA), 333-349.