RC2: 'Comment on hess-2022-252', Anonymous Referee #2, 22 Jul 2023

https://doi.org/10.5194/hess-2022-252-RC2

#	Comment	Reply
1	In this manuscript, the authors employ data-driven techniques to predict rice crop yields in India. The paper's objective is clear; however, the methodology is not rigorously employed, the novelty is limited, and the document's structure could be enhanced. In order to improve the study, the authors could consider the following points:	We appreciate your comments. The document has been updated, improving the structure and writing. We mention in the comments below the novelty of the work and also indicate what is related to the methodology, we hope that in its new form, this new manuscript will be of your approval.
2	1. In lines 64-68 you mention that ML techniques have already been tested to predict crop yield but that “the use of spatial characteristics of drought such as its spatial extent has not been fully explored to crop yield prediction”. Does this mean that the only conceptual novelty of this work is that it considers a new variable?	We introduce an innovative approach to predict crop yield using spatiotemporal changes in drought areas. Most of the previous work has been focused more on the development of indicators and on the use of multivariate methodologies to improve crop yield prediction. In this research, we proved that changes in drought areas are a good indicator of how drought negatively impacts crop yield. Another novel element is the conceptualisation of the approach, we used two types of ML models: polynomial regression (PR) and artificial neural network (ANN) as integrated tool. Lines 69-81
3	2. The authors write in the Modeling Limitations section that insufficient crop yield data is an issue, however, the last year for which crop yield data is available is 2015, is it possible to increase the dataset? Much more importantly, the basis of data-driven techniques (of which ML algorithms are part) is that a lot of information is available, and the algorithm can learn from the data. If you don’t have enough information, how can you justify the application of a ML algorithm?	It is possible to increase the period of the data but we foresee that the conclusions are still valid. The use of ML is justified when using not only a single time series of drought areas but many (see Methods and data). We are using monthly drought areas calculated with various aggregation periods of the drought indicator. Moreover, in drought monitoring, the variable aggregation is often done in different periods to try to monitor different types of droughts. In our research, we used the drought indicator with 3, 6, 9, and 12 months aggregation period.

4	3. Some of the plots presented in Figure 7 show a serious problem. Your predictions present a lag of one year (the red curve is shifted one year to the right). This usually indicates that an auto-regressive algorithm (like the one that you are using) is not capable of learning and that the prediction of year t+1 is strongly influenced by the crop yield of year t.	Thanks, we included the change in crop yield of the previous season which considerably improves the prediction compared to not using it. In future research, the best order of the crop yield (i.e. t+2, t+3) can be investigated.
5	4. Go through the entire document and check English usage and typos.	Thanks, we have checked the entire manuscript.
6	5. I suggest that the authors revisit the document and avoid repeating information (unless strictly necessary) and avoid presenting graphs with excessive information.	Thanks, the entire document was updated and restructured to address this comment.
7	6. You need to improve the description of your work in the introduction. As it is right now, it is unclear. What do you mean by “the crop yield calculation is clear”? What do you mean by “is not as clear”? What does “The ANN is expected to be used with the final input data” mean?	The logic of this integrated tool is as follows. PR provides the prediction where the crop yield calculation is easy-going to the performer (the end-user) because she/he has access to the equations that have a straightforward interpretation, and calculations can be done with early and preliminary input data. For its part, ANN is used as the most accurate model, although the output calculation is not always easy to follow, as in the case of PR, due to the difficulty of interpreting the structure of the resulting ANN. The ANN model is used with the final and more accurate input data. We have updated the Introduction section.
8	7. Did you evaluate the cross-correlation between input variables? Is it possible that you provide redundant information to the algorithm?	We calculated the correlation between inputs and crop yield and based on the described procedure we selected the variables to build the ML models. As we mentioned, in future applications the exploration of best inputs, models, and scales (spatial and temporal) could be done to improve the approach we introduced in this research.
9	8. In the results section you write sentences using terms like “perhaps” and “may”. However, the results should be able to prove or reject a hypothesis. I strongly recommend that you avoid that type of sentences in the work	Thanks, we have rewritten the Results and Conclusion section to avoid those terms.