Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-600-AC1, 2022

Reply on RC1

Vitali Diaz et al.

Author comment on "Machine-learning approach to crop yield prediction with the spatial extent of drought" by Vitali Diaz et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-600-AC1, 2022

Response to the reviewers

Thank you very much for the review and valuable suggestions. We describe below how we addressed each comment. Our responses begin with the word '[Reply]'. We have numbered each reviewer's comment for ease of reading, and this numbering is indicated between brackets.

Reviewer 1
General comments:
The paper is very interesting and I think it is related to the main objectives of the journal. The authors use statistical models to try to adjust the annual rice yield with the area affected by droughts. They use SPEI at different scales and with a threshold of -1 to define droughts.
However, I have observations. Among them is the difficulty to follow the reading due to the order in which the information is presented. On the other hand, I think there are many assumptions in the methodology and in the way results are shown that should be further discussed. The methodology used is not 100% justified, as it leaves many options and questions open.
I elaborate on these issues below.
Specific comments:

[1] I recommend reordering the information to make it easier to read. For example, section 2.1 explains a lot of information that is formally presented in section 3. Information that could be presented in the introduction (for example) or by arranging the text in a more coherent way.

[Reply] Thank you very much. We have reordered (swapped) Sections 2 and 3 to facilitate the reading and understanding. They are now Sections 2. Data and 3. Methodology. We believe that in this order, the reader when approaching the 'Sect. 3 Methodology' will have more information and details on the case study/data that are previously presented in Sect. 2.

[2] The title and abstract I think should emphasise the area of study, because while the methodology is appropriate and interesting, a broader study at a global level is needed to generalise it. Therefore, I believe that specifying the region of the research is necessary.

[Reply] Thanks for your suggestion. Although the results are indeed of great interest for the analysed regions, we consider that the main paper's contribution is the methodology and especially the assessment of the input used for predicting crop yield, i.e., drought area. Drought area has not been investigated fully in previous studies. Therefore, we consider that the title and the abstract suit well.

[3] The paragraph starting on line 101 explains the different sources of agricultural data in India. I recommend without extending the text too much to discuss with bibliographical sources (if available, otherwise make it clear) the different sources with their pros, cons and quantitative differences at different spatio-temporal levels.

[Reply] Thank you very much for your comment. The main advantage between the two systems indicated in that paragraph, i.e., ground-field visits-based and satellite-based systems, is that the latter provides forecasting, that is, information before harvesting the crop. There are differences in the spatial and temporal scales due to the configuration of both systems; we have included a paragraph describing such characteristics based on your suggestion. We have also emphasised that this paper aims to introduce a methodology for crop prediction that may well serve to compare the other two systems.

[4] Saying that using the area affected by droughts could be discussed more in the introductory line, so as to justify its use. Some studies discuss this issue, for example: Araneda-Cabrera, R.J., Bermudez, M., Puertas, J., 2021c. Assessment of the performance of drought indices for explaining crop yield variability at the national scale: Methodological framework and application to Mozambique. Agric. Water Manag. 246. https://doi.org/10.1016/j.agwat.2020.106692

[Reply] Thanks. We have read and included the reference you mentioned in the Introduction Sect. Now have included a text indicating that one of the motivations for using the drought area as an input variable is its high correlation with crop yield, as noted in some previous studies, including Araneda-Cabrera et al. (2021).

[5] Use at least 1 or 2 model fit indicators to get a better picture of the results (e.g. R2).

[Reply] Thanks. We do not show the results, but we also used R2 in our analysis. Using R2 leads to the same results. Although other error metrics can be used, it is not intended to evaluate how the choice of metric influences the results. We believe that this type of analysis could be investigated in future studies. Here in this document, we have set a precedent for the effective use of drought area for crop yield prediction.

[6] In line 270 in the ANN model incorporation, the 85% and 15% data split is arbitrary. Although it is valid in principle, I believe that the possibility of obtaining the fact that different model fit values can be obtained each time it is used should be discussed, due to the randomness that exists in the selection of these variables. Several methods can be used to decrease this randomness (e.g. running the model 10 times, averaging the results or averaging the 10 RMSEs, etc.).

[Reply] Thanks. The 85% and 15% rule for training-validation and verification (testing) is a common practice in model building, as well as the use of iteration-based approaches to calculating the best realisation (model). This splitting rule was only applied to the ANN models. In the case of PR (polynomial regression), Eqs. 5 to 8 were used for the four types of PR, for each set (combination) of inputs (Table 2). Regarding the construction of the ANN models, there is no randomness in the selection of the input variables. The complete procedure was applied to each set of inputs (Table 2), and the errors were also calculated for each set of inputs. The iterations served to evaluate and select the best parameters of the ANN models for each set of inputs, varying the number of hidden layer nodes (from 1 to 10); for this, we applied 100 iterations for each number of nodes, giving a total of 1000 iterations for each set of inputs. We have improved our text to indicate that the iterative procedure, including the calculation of RMSE, was carried out for each set of input variables in the ANN models. We have also included an additional figure in our manuscript to show the general scheme of how the input and output variables are tied to facilitate the reading of our methodology.

[7] Rice is a crop that may depend on water from regulation or other sources. Is it the best example to fit with SPEI? Since this is a meteorological index, which, although it correlates very well with agricultural droughts when using 3–12-month scales, it does not detect the possible regulations that rice may have. Are we sure that it is a rainfed production in the whole area or where there is irrigation? These issues are not even mentioned in the results, which confuses the reading.

[Reply] Thanks. The three regions in which the methodology was tested show various configurations between rain-fed and irrigated production, with different percentages in each case, as shown in Figure 4.d (original manuscript). In regions 1 and 2, the rain-fed and irrigated production percentages are more or less than 50% in each case. Only in region 3 irrigated production reaches 35%. The correlation analysis shows that for the case of region 3, the R values are high between the drought areas and crop yield. The high correlation is due to in region 3, the lack of water resources is less mitigated for irrigation, and the water anomalies are more detected by the drought index that considers precipitation. Our original document pointed out this (lines 408-413). In the case of regions 1 and 2, although only half of the territory is rain-fed, the drought areas correlate well in the months of the growing season, although to a lesser extent than in the case of region 3.

[8] To this last fact I also have two doubts. One is an observation of lack of clarity in figure 3, it could be further improved to appreciate the time series of rice crops. And the second is to discuss the use of a scale of 0.5deg. Is it in accordance with the agricultural data available and for the purposes of using it to calculate the area affected by droughts? Figure 5 could be improved or summarised and shown separately (including as supplementary material and/or tables).

[Reply] Thanks. Figures 3 and 5 have been improved. In the case of Figure 3, the figure shows the de-trended crop yield data. We have also added in Figure 1 the time series of crop yield for each region.
On the other hand, although the 0.5-degree resolution of the drought indicator might seem coarse, it was found to be a suitable resolution for the drought indicator to calculate drought areas according to the results. Finer resolutions could improve results and even build models for smaller regions like the crop district level. We use drought data from the SPEI global monitor to motivate the use of our methodology in other study areas. If there is a lack of data, the SPEI data from this drought monitor could be used as the initial implementation.

[9] The limitations of the work are mentioned, but I think they could be expanded, so that the application of the methodology is justified despite the limitations.

[Reply] Thank you very much for the issues raised. We have extended the discussion in our section on methodology’s limitation (Sect. 4.5).

[10] Are the presented results (equations) suitable for operational use in the region?

[Reply] Yes, the constructed PR equations are suitable for operational use. We have also enhanced the text to indicate so. We also include the description of an example of the PR equations application.

Kind regards,
Vitali Diaz
on behalf of the authors