Anonymous referee #3

The manuscript provides an important contribution to improve current statistical crop modelling practices and clearly illustrates potential pitfalls of most common approaches. The revised manuscript is well structured and the content well presented, however, it would be of great value if the English language could be revised to facilitate a smooth reading.

Thank you for your appreciation. We have made changes to improve the manuscript. We also asked a native English scientist to proofread the manuscript. We hope that our improvement will make the manuscript clearer now.

Anonymous referee #4

General comments:

This publication presents a highly relevant validation technique for statistical crop models, i.e. a possibility to select the input variables and validate the model independently of the testing data set, which improves the robustness of the model. The LTO validation is presented in two case studies. Results show that the LTO validation leads to more robust results and enables a more realistically assessment of the forecasting performance. Because rigorous validation remains rare in the statistical crop modelling community, this paper is of high interest to other scientists. Even though the proposed approach has not often been applied yet, it is also not a new approach. Laudien et al. (2020 and 2022) and Meroni et al. (2021) present examples in which an independent variable selection has been applied to forecast crop yields. However, the explicit comparison of the influence of a different number of input variables (either inputs or potential predictors) on the model performance is – to our best knowledge - new and interesting for a wider audience. The paper is well-structured and has a clear, but partly colloquial language.

We thank the referee for his/her highly constructive comments and suggestions. Please find below our responses (in blue) to each of your comments.

Major comments:

- The comparison of the selection of the best number of predictors and inputs between LOO and LTO does not include the results for fewer than 3 variables. As the RMSE for the validation and the testing increases with the number of inputs, the question arises whether the optimal value is even lower than 3. Also, the paper title suggests that the optimal number of inputs is found. The risk of too simplistic models is not explored in the paper, which would be an interesting addition to the presented results.

Thanks for this constructive comment. In fact, in Figs. 5 and 8, we compared different LIN models, in which the number of inputs increases from two to six.
We now improve the comparison by adding the LIN1 model (i.e linear model with only one input) in both Figs. 5 and 8. Corresponding comments on these two figures are also added:

Section 4.1: "... In the case of a too simplistic model, i.e. LIN model with one input, underfitting occurs as the errors are high in the training, validation, and testing datasets (shown in Fig. 5(b)). These errors decrease gradually with the number of inputs, i.e. from one to three. However, the testing errors do increase when the model has more than three inputs. The LTO procedure indicates that a simple model---with only three inputs---is optimal."

Section 5.1.: "... Again, in both Bas-Rhin and Landes examples, underfitting occurs when models are too simple, for example, with one input. With a higher number of inputs, the LTO procedure shows a similar behaviour as previous examples (Sect. 4.1): the validation/training errors decrease gradually, while the testing errors show an opposite trend. …"

- Laudien et al. (2020): "Robustly forecasting maize yields in Tanzania based on climatic predictors"; Meroni et al. (2021): "Yield forecasting with machine learning and small data: What gains for grains?"; Laudien et al. (2022): "A forecast of staple crop production in Burkina Faso to enable early warnings of shortages in domestic food availability" provide examples of an independent variable selection in a statistical crop model to forecast yields. Whereas Laudien et al. call it "level 2 LOOCV", Meroni et al. (2021) also call it nested oos validation. The statement that the proposed LTO approach has never been used before (Line 50-51) is therefore not correct. Please rephrase this sentence.
Thanks for the comment. We rephrased the sentence and added your suggested references.
"We found very few applications of this approach in the literature on statistical crop modelling (Laudien et al., 2020, Meroni et al., 2021, Laudien et al., 2022)."

- As the authors state, the number of input variables, the number of potential input variables and the model type influence the robustness of the model/forecast and the inclusion of all of these aspects in the study is relevant. The terms model selection, model complexity or "finding the best model" are often used to describe the selection of inputs, potential inputs and model type even though these terms encompass also other aspects, as pointed out in section 2.3. These terms therefore do not adequately describe what the authors are examining, which leads to confusion. Also, the terms are used ambiguously, e.g. in Line 266 it says "Here, the model complexity is considered as a representative example of model selection." The paper would benefit from a clearer language concerning what is actually investigated, i.e. either input selection, potential input selection or selection of the model type.
In this study, we considered all three factors: the inputs, the potential inputs, and the model types. It is true that we did not clear enough in the manuscript. We rephrased Sect. 2.3 to better introduce the problem and factors that we investigated. Sect. 3.2.3 is also rephrased and it shows only one

example of input selection instead of using the confusing term “model complexity”. We hope that all the changes made especially in Sect. 2.3 and Sect. 3.2.3 will resolve this comment.

- The reference to Dinh et al. (2022) is made about 10 times in the paper – also as a way to justify some methodological decisions – even though this paper is not yet published and has the same first author. Please consider finding alternative literature sources.
We added several alternative literature sources to replace this reference.

Minor comments:

- L5: It is not impossible to split the data set into 3 in statistical modelling. Please rephrase.
We changed it to: “Splitting the overall database into three datasets is often impossible in crop yield modelling due to the limited number of samples.”

- L33-34: “not an easy task” is a rather subjective statement. Sentence should be rephrased.
We changed it to: “Splitting a small number of samples into three datasets is not easy.”

- L 43: This statement should be supported by references.
We added related references.

- L 46: Instead of “from” – it should be “for”.
Change made.

- L74, 76, 79: Do you mean reproductive stage?
For coffee, this stage is called the “productive stage”.
(please see, for instance, Valeriano et al., 2018: Estimation of Coffee Yield from Gridded Weather)

- L87: Please provide a reference.
We added a related reference of Mathieu and Aires, 2016: Statistical weather-impact models: An application of neural networks and mixed effects for corn production over the United States.

- L99: Please reformulate the phrase. “change in time” is misleading. Do you mean changes in the timing of phenological stages?
We changed the sentence to: “Although the sowing time varies for different regions (Olesen et al., 2012), the average growing season of French grain maize ranges from April to September …”

- L115, L197: There is no comma after “i.e.”. Please, also check at other parts of the paper.
Thanks. Change made.

- L116: It should be “at” the equator, not “in” the equator.
Change made.

- L117-118: How has the data being matched to administrative levels? This should be specified, as it influences the results.

We added more details to explain this: "In detail, the gridded data have been aggregated over district or department shapes: (1) if the shape is smaller than the cell, the gridded value will be representative of the region; (2) if the shape includes several cells, the weather data will be averaged based on the area of cells inside the shape."

- L148: The error term in the regression equation is missing.

We added this term.

- L179: The regression does not necessarily require an intercept (In case the dependent variable was demeaned beforehand, the intercept is no longer needed.). Therefore, it is not always n_input +1. Please rephrase, e.g. the LIN model "usually" requires n_input +1 inputs.

We corrected the sentence.

- L189-192: This statement does not belong to the section on Methods. Findings are presented later in the paper and should not already be mentioned at this point.

We removed this statement as the idea is mentioned in the Results and Conclusions.

- L166-67: Please rephrase. Also many other factors can influence the model performance, not only complexity and potential input variables.

As mentioned earlier, we rephrased the whole section (Sect. 2.3) to better introduce the problem. "Model selection is the process of selecting one model—among many candidate models—that best generalises (Hastie et al., 2009). This process can be applied across models of the same types with varying model hyperparameters or across different model types. Here we investigate some practically important factors of the model selection: … "

- L219: This is "often" the case in crop modelling studies, however there are also studies with big samples (e.g. Schauberger et al. (2022): French crop yield, area and production data for ten staple crops from 1900 to 2018 at county resolution, Lobell (2008): Prioritizing Climate Change Adaptation Needs for Food Security in 2030 or Renard, Tilmann (2019): National food production stabilized by crop diversity).

We rephrased the sentence: "If the database is small (as often in crop modelling tasks), the model selection can be too specific for the particular samples of the testing dataset …"

- L276: Applicability of a model is not only defined by its skill – please rephrase.

We rephrased the whole sentence to: "The goal is to find a model that makes the most robust predictions of crop yield anomaly as a function of weather variables."

- L296: It should be "models" not "model".

Change made.

- L300: Robust statistical models can also be based on smaller samples than 19. Please make the sentence more general by e.g. saying: “when having a limited sample.”
Change made.

- L306f: It is not illusionary to model complex weather-yield relations with a sample of 19 observations - many papers show that it is possible. The choice of input variables should also account for more complex weather-yield relations (i.e. only studying monthly mean temperature or precipitation sum might not be sufficient). Rather refrain from this statement.
We removed this statement.

- L312 and L430: Do you mean key phenological phases in plant development by moments of coffee?
We changed “key moments” to “key phenological phases”.

- L312-316: This is a very interesting discussion as it explains why the selected variables potentially show a good performance in the model. However, this should be supported by literature.
Thanks for the comment. We added related references to this paragraph.

- L322: Weather is only one factor among other factors. However, the examples are not well-chosen, i.e. by omitting the yield trend you deliberately omit the influence of e.g. agricultural practices (e.g. irrigation) that usually only change gradually over time. Also, one could argue that diseases are indirectly covered in statistical models. Please, refer to literature at this point to support your examples.
Thanks for your comment. We changed the sentence and added several supporting references (Miao et al., 2016: Responsiveness of Crop Yield and Acreage to Prices and Climate; KC et al., 2020: How climatic and sociotechnical factors influence crop production: a case study of canola production; Liliane and Charles, 2020: Factors Affecting Yield of Crops).
The sentence became: “This value is reasonable as the weather is among several factors (e.g. prices, sociotechnical factors, managerial decisions) affecting coffee yield (Miao et al., 2016; KC et al., 2020; Liliane and Charles, 2020).”

- L323-324: As pointed out earlier, even with smaller samples, the model can capture complex and robust weather-yield relationships. The model quality depends on many other factors such as the quality of the input data, the choice of potential predictors, the accuracy of the defined growing season etc. Please delete this sentence or support it with literature.
We removed these sentences.

- L346: Please provide an explanation of why the validation and test errors show so much variability.

We added the explanation: “These fluctuations imply that the model is overfitted, and thus, random error or noise appear. ”

- L359 and L84: The reason for the selection of the case study regions should be made explicit (the selection of Cu M’gar as one district in 4 major coffee producing regions is based on a paper that is not yet published and the selection of the 10 maize producing regions in France is not explained at all).

For coffee, we removed the reference and we added the production statistic to explain our selection of the Cu M’gar district: “We focus on Cu M'gar district as it is a leading coffee-producing district in Vietnam, accounting for about 10 % of Vietnam's total coffee production (i.e. 76400 tons for the 2000-2018 average).”

For the ten maize producing regions, we mentioned in Sect. 2.1.2:

“Some specific tests (in Sect. 5) will focus on ten departments (as presented in Fig. 1(d)) where the average grain maize production is higher than 4 × 10^5 tons (or the area is higher than 40 thousand hectares).”

- L425: The remaining variability could also stem from other factors (see comment to L322) and change this sentence accordingly.

We changed the sentence to: “The remaining variability is rather large, and may be explained by non-climatic factors (e.g. prices, sociotechnical factors, managerial decisions, or political and social context).”

- L427: A possibility is also that the input variables do not sufficiently cover crop sensitive climatic drivers as only mean temperature and precipitation sum are considered in this study.

We changed the sentence to: “It could also come from climate; however, the model would require more detailed variables (e.g. at a daily scale) or more samples to go into deeper details of the climate-crop yield relationship.”

- L434: The sentence does not make sense. Please rephrase.

We rephrased these sentences to: “LTO indicated that a simple linear model is preferable because it has a lower testing error. ”

- L444: What do you mean with “other crops will be investigated”? Afterwards you cite papers that already studied these crops I suppose. L446-447: The sentences are not easy to understand in terms of language. Please rephrase.

We rephrased these sentences to: “In addition, by using a similar approach presented here, other crops will be investigated, for instance, over France (Ceglar et al., 2016; Schauberger et al., 2018; Ceglar et al., 2020), over Europe (Ceglar et al., 2017; Lecerf et al., 2019) or globally (Bunn et al., 2015). Furthermore, these types of statistical crop models can be used to refine the potential adaptation and mitigation strategies.”