

Geosci. Model Dev. Discuss., referee comment RC3
<https://doi.org/10.5194/gmd-2021-218-RC3>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on gmd-2021-218

Anonymous Referee #3

Referee comment on "Nested leave-two-out cross-validation for the optimal crop yield model selection" by Thi Lan Anh Dinh and Filipe Aires, Geosci. Model Dev. Discuss., <https://doi.org/10.5194/gmd-2021-218-RC3>, 2021

The authors discuss an important aspect of statistical analysis with limited data availability and exemplify the shortcomings of the more frequently used 'leave one out' cross validation compared to the 'leave two out'. Two case studies are used, an annual crop in Europe and a perennial crop in Vietnam. Two different statistical approaches are used, linear regression and neural network. The two case studies convincingly demonstrate how to select the best model for forecasting crop yield with limited data availability, while considering the problem of overfitting. In general the study is well designed and clearly communicated, although substantial English revision is required.

Some general observations:

- Line 132: the model architecture is explained as consisting of i) number of potential predictors, and ii) the number of inputs. Please specify clearer what the difference is between predictors and inputs. The authors also mention model types: Some models require more parameters to estimate even with equal numbers of predictors. This should be described in clearer terms.
- The value of using LIN3 and LIN5 are not clear to me, which is related to the previous comment: the authors explain them as linear regression models with three and five inputs, respectively. Could you explain what the value is of this experiment? What added insights do we gain when comparing LIN3 and LIN5? Do we not already get all the insights from adding the number of potential predictors to LIN3?
- There is quite a bit of redundancy – the authors explain the problem several times, thereby repeating themselves. For example, section 3 (lines 168-181) already has been elaborated in previous sections. More concise presentation of the problem and methods would benefit the manuscript
- Even though the main aim of the study is to compare the LOO and LOT, presenting the chosen predictors and final model would be appreciated. Any reader familiar with the case study crops will be interested to understand what climate descriptors were tested and selected in the final model. This transparency is further necessary for making the study reproducible.

