Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-21-RC1, 2021

Comment on gmd-2021-21

Anonymous Referee #1

Referee comment on "Choosing multiple linear regressions for weather-based crop yield prediction with ABSOLUT v1.1 – Initial tests for the districts of Germany and an over-confidence trap in statistical modelling" by Tobias Conradt, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-21-RC1, 2021

Review of paper « The multiple linear regression modelling algorithm ABSOLUT v1.0 for weather-based crop yield prediction and its application to Germany at district level » by Tobias Conradt

Major comments:

- I don’t know if GMD journal is different from the journals I usually work with, but to me, this paper is too much oriented towards the computer implementation of your codes. The organisation of you codes is not of special interest for the reader, description of the readers, etc. One line should be enough to tell the language (R), machine on what it was done, and this is enough. Because of this orientation of the paper towards a specific implementation of the code, the paper describes steps like describing a code. But in a paper, we want a synthesis, much shorter, with the methods that are used, some formulas if needed, and explain the general ideas and comments. In the current state, the paper is too much oriented towards a report I am afraid. But if the GMD is OK with this, then this is the decision of the editor.

- The main point of this paper is that a complete test of all the combinations of predictors is accomplished, where a traditional step-wise regression only test some of them. I don’t think that this justifies a name for a particular method, you just need to say that you have an exhaustive search of the combinations of the predictors. This is enough. So the papier can largely be reduced.

- It is claimed that it is important to test all the combinations of predictors in order to find

the best model. I actually think that the approach that is used has a major problem. We actually run into similar difficulties and found recently the solution. Basically, since you are limited in your database (only 20 years for instance) then you use leave-one-out procedure to train in the database minus one sample, and test it on the left sample. This could be considered legitimate, but… by doing so, you actually chose your model (with a particular combination of predictors) based on the testing score of the LOO. So you use the testing base to select and estimate the generalisation ability of your model. We have shown that this is not correct, because you are overtraining, and your generalisation score is not reliable. When using such a procedure, your results push you to use more inputs, and more complex models. It is a good thing that you are using only a linear model, but still, your assessment of the generalisation is not reliable to my understanding. This is a very subtile thing, and many people do such a mistake. I would like to have your opinion on it, and maybe a solution.

- I actually think that modeling crop yield with a statistical model from a very small database of samples is a true challenge. Crop expert actually think that many variables are important for the development of the plant, but actually, samples are just not enough to calibrate a complex statistical model (in terms of complexity or number of predictors). A true assessment of the generalisation of the errors should show you that very simple models (linear with 2 or 3 inputs) are actually what we can do the best. The search for complexity is flawed, because no large historical record of crop yield is available.