

Geosci. Model Dev. Discuss., author comment AC1
<https://doi.org/10.5194/gmd-2021-21-AC1>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.



Reply on RC1

Tobias Conradt

Author 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-AC1>, 2021

Many thanks to the anonymous reviewer for pointing out important weaknesses of the paper in its current state. Here are my replies to the four points of concern:

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. [...] 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. [...]

A colleague of mine also wondered whether so many technical details are of interest to the reader. I just thought a "model description paper" should include a comprehensive description of the code, but I have no problem to adapt.

Suggested solution: Subsections 2.1.1, 2.1.2, 2.2.3, and the methods section 3 will be stripped from technical details of the implementation such as package versions, file names, etc., and they will eventually be merged or restructured. Figure 1 (the flowchart) should be kept, though, but in a simplified manner. To retain the technical information, a "directions for use" leaflet containing these details will be added to the code and data repositories.

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 paper can largely be reduced.

I don't agree here. As explained in the introduction, there are always so many more decisions or steps made in statistical crop yield modelling that most applications deserve a name. Method naming is probably not so common with statistical modelling compared to other disciplines, but following your argument even dynamical climate models could be downplayed to just a lot of coupled differential equations.

I really would like to call my approach ABSOLUT which not only means exhaustive search for predictors in multiple linear regressions but also refers to all the other specifics like

using two-to-six-month aggregates of weather data from the twelve months ahead of harvesting, the two-step approach of the predictor selection using binomial probabilities, or the decision to apply the locally best combinations of predictors out of a small set of globally selected combinations.

Suggested solution: No change regarding the name, but I will try to give a more concise explanation of what it comprises besides the regressions instead of rhetorically asking: "But what is so special about a couple of linear regressions [...] that justifies [...] naming it ABSOLUT?"

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 subtle thing, and many people do such a mistake. I would like to have your opinion on it, and maybe a solution.

Hit and sunk. You may have noticed that I was principally aware of the blotted mixing of training and testing data which only applied to the selection of the predictive features while the resulting regressions were fitted to the training data only (leave-one-out validation). Therefore I wrote "quasi-out-of-sample", e.g. in line 373 on page 16, but I even did not do so consistently, cf. "out-of-sample errors" in line~490 on page 24.

My approach to deal with the resulting errors was sloppy, too: "[... T]he interval should be ± 1.96 RMSE, but here a factor of 2.00 was used to account for the fact that all data have been used to determine the weather aggregate combinations in the regressions so that not a pure out-of-sample approach had been applied [...]" (page 16, line 365ff.) I am still convinced that the possible overconfidence resulting from the incomplete separation between training and testing data is minimal, but you are right, it should not be done this way.

Of course I would like to learn about the solution you found recently, but I see the problem regarding your anonymity. Here comes my

Suggested solution: I will change the algorithm to determine also the predictor combinations separately for each single forecast/hindcast using only data from the remaining years (training set), this shall become version 1.1 of ABSOLUT. I will then repeat all the testing with the German data and adjust the paper accordingly (provided the nice validation performance does not completely vanish).

I actually think that modeling crop yield with a statistical model from a very small database of samples is a true challenge. Crop experts 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.

Yes, the historical records are always too short or incomplete. ABSOLUT does however attempt to compensate for just that by considering many parallel realisations of the process in spatially distributed units (districts). It is not unique in that point, simple panel models are for instance based on the same idea of combining both spatial and temporal dimensions to broaden the data basis.

Finally, ABSOLUT delivers regressions with just five inputs, but I agree that using only three or four may be better in some cases. See Table~4, especially the winter wheat predictions show practically no performance gains from more than three weather inputs.

Suggested solution: The fixation to time plus exactly four weather aggregates as input variables shall be relieved in ABSOLUT v1.1. As the validation performance indicators R^2 and RMSE won't suffer from overconfidence biases any more the most favourable, potentially smaller number of predictors can then be determined straightforward.