Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-47-RC3, 2020

Earth System Dynamics
Discussions

Interactive comment on "Identifying meteorological drivers of extreme impacts: an application to simulated crop yields" *by* Johannes Vogel et al.

Flavio Pons (Referee)

flavio.pons@lsce.ipsl.fr

Received and published: 4 November 2020

1 Overview and major comment

The paper is concerned with a relevant problem in a two-fold way: developing a methodology to detect which meteorological variables drive impacts on a certain socioeconomic variable, and giving an in-deep practical application to bad crop years.

I found the paper overall well written, and the problem is clearly stated and understandable even to a reader not familiar with crop modelling. The crop model remains quite a black box to the reader, but I feel like the main point of the article is to propose a

methodology and show its performance, rather than focusing on technical details of crop simulation.

The methodology itself is simply based on the application of a logistic lasso regression: the model is fed meteorological variable as covariates, to predict whether the yield at a certain grid point will result in a bad crop year (Y=0) or a good crop year (Y=1). The 'lasso' formulation allows to include a large number of predictors (in some cases even larger than the sample size) and only select - or, in some of its variants, group - a subset of predictors that reduces the problem dimensionality while maximizing forecasting performance. The model is tested against two competitors, a generalized linear model (I suppose binomial with logistic link, it would be nice to specify this detail in Section 2.5) and a random forest run in binary classification mode. The authors find comparable performances between lasso regression and random forest, however the latter is way less interpretable, making lasso a feasible yet effective way to model impact drivers.

As a major comment, which doesn't necessarily imply need for major revisions in the paper, I would like to stress that this greater interpretability is still quite limited by the nature of the lasso model. This is designed to select the variables that produce the best forecasting performance with minimal number of covariates in a linear model that may be a strong approximation of the real world phenomenon. This means that the selected variables are surely the ones that provide better explanation of crop failure *in the considered crop simulation model* and *in terms of prediction*. This does not necessarily imply selecting variables that directly physically drive the crop failure, just like the resulting regression coefficients are not estimates of a real linear law existing in nature, but of an approximation that optimizes forecasting.
In all fairness, results in the presented case study appear to be physically reasonable, and I found the discussion in Section 3.2 convincing in this sense. However, it is possible that in different problems, where processes are less understood, results can provide indications useful for forecasting but not really provide physical insights,

making the methodology not necessarily effective in all fields of application. I would explicitly stress this in the main body and in the conclusions, because a reader not familiar with the shortcomings of applied statistical modelling may over-generalise these findings to a problem where it is not possible to do so. Also, I would add a warning that critical interpretation of the results is always necessary, especially in cases with smaller or non gridded datasets, where the hints coming from spatial coherence (which in this paper play a role in making results more solid) may not be available.

2 Minor/technical comments

A general consideration: the notation calling "positive" years with a good crop may be a bit confusing when trying to interpret results. While a good yield is surely positive news, the model is designed to detect drivers of impacts leading to bad years: it would be more coherent with traditional terminology to address the non-baseline case under investigation with this term. I do not think that this is worth modifying the phrasing in the whole article, but maybe I would stress this, especially readers with a statistical rather/other than physical background may not pick up on this immediately (I didn't!).

1. (line 14) "both between" should read "of both"
2. (line 115) the authors state that they normalize all the variables to be in [-1,1]. I understand rescaling/normalizing variables when they take values that differ by several orders of magnitude, but I do not understand the choice of squeezing them into a close interval, as logistic regression handles continuous real valued covariates.
3. (line 150) the authors state that lasso is superior in handling correlations in the covariates better than standard GLMs. This is certainly true for correlation among

covariates, but I am not so sure about autocorrelation. In particular, meteorological data display a strong seasonality, which introduces long range autocorrelation in the data. Can the author provide some reference specific to this aspect?

4. (lines 168-175) I am not sure if I understand correctly the choice of λ_{1se}: is it because, using $\lambda_{min}+1se$ falls almost exactly in the middle of the 95% confidence interval that would require $2se$? If so, it makes sense but it should be explained more explicitly.
5. it seems that the authors choose a priori $s^* = 5\%$ and try also 2.5 and 10% to test the sensitivity as a threshold to define bad crop years. If so, does it make sense to define s^* as the argmin of C(s) as in line 205?
6. (lines 219-222) not sure about these lines: it is a good idea to check for significant interactions and report it, but then I would explain in larger detail what interactions are in regression models, because the reader may not be familiar with the concept. Also, which one did they try, and did they have an a priori idea about possible meaningful interactions?
7. (line 231) "eastward" $\longrightarrow$ "westward"?
8. (line 327) the authors say that their analysis is based on a time series model, but maybe they mean that the dataset is constituted by gridded time series data.
9. (line 380) "With our approach with" should be "With our approach we"

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-47, 2020.