Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2020-220-RC2, 2021

Natural Hazards and Earth System Sciences
Discussions
Open Access

Interactive comment on "The potential of big data and machine learning for weather index insurance" *by* Luigi Cesarini et al.

Anonymous Referee #2

Received and published: 23 March 2021

General comments

This paper explores whether satellite and reanalysis data for rainfall and soil moisture can be combined using machine learning methods to assess, in an objective way, whether floods or droughts are happening or have recently occurred. This is placed in the context of improving index insurance. The paper is extremely detailed in terms of how the machine learning models are constructed, and validation metrics.

1) My main comment is: the paper is very heavy on text-book style review of methods (which isn't a bad thing), and very heavy on technical detail (which isn't a bad thing), but lacks any exhibits that show clearly whether the methods actually work or not. There are masses of technical validation metrics. But what I personally would like to see are

some results along the lines of: a) we took the data shown in figure 7 (predicting this data is what the whole thing is about in the end) b) we split that data in half, trained the models on one half, chose the best model, and tested it on the other half c) and for the single best model, here's a picture that shows the results of that side by side with the actual floods and droughts that occurred in the validation period. Did it capture them all, or half of them, or none of them? d) then I'd be able to look at that and make a judgement as to whether the method works or not.

Specific comments

2) There's a whole discussion about training and validation data, but then in the end it's not clear how the data is actually split into training and validation data (relates to point 1 above), in relation to Figure 7. The construction of the validation is critical for us to be able to understand whether there's anything in this or not, especially since a large part of the scientific community associates the word ‘machine learning’ with ‘overfitting’, and will be sceptical.

3) With such a small amount of data, and after testing so many models and configurations (line 341: ‘almost boundless domain of model configurations’), it seems to me that overfitting is quite likely. Could the authors elaborate on why testing so many configurations doesn't lead to overfitting? And if you are evaluating the models against each other using the validation dataset, of course one model will do best. How do we know that the model that does best would genuinely do best in a true out of sample sense? Don't you need another level of cross-validation?

4) Line 18 says $3.3B. This is wrong by several orders of magnitude. Individual events during that period were in excess of $50B (since at this point you are talking globally).

5) The word ‘loss’ is used with two different meanings, as far as I can tell. Line 105=loss in the usual sense of damages, vs line 249 in a technical sense. This is a bit confusing. Different terminology should be used, somehow, to avoid this.

6) I think it should be made clear that the runoff model – flood intensity relationships are simplistic relative to start of the art runoff and flood modelling as practised by hydrologists

7) Line 176 refers to loss data. What is this loss data?

8) Line 319, there is a comment that TensorFlow allows ‘embedding the validation process into the construction of the model’. That sounds like overfitting to me. Please explain how this is consistent with the claim that the data is really being split in order to do out of sample validation.

9) Is reanalysis data really available soon enough to be useful? I thought it usually appears at least a year or two later, but maybe I’m wrong.

10) There should be a bit more discussion about the problems with satellite data and re-analyses (i.e., talk about the reasons why these data-sets aren’t really used at present for index insurance purposes, even after 20 years of academics suggesting that they should be).

11) As far as I understand it, there has been no comparison here with standard methods for assessing whether an event has occurred, which are based on rain gauges, levels of river flow, etc. That should be pointed out.

12) Are there any further diagnostics that could be produced to help show that the model is really doing something sensible, to help allay the suspicion that some readers may have that it’s all just over-fitted?

Technical corrections

line 171: you say T_t, but don’t you mean Y_t?

line 205: SP1, 3 etc need to be defined. I can guess what they are, but they should be defined.

line 367: is that citation really correct? Is the person’s name just M?

i.e. and e.g. are usually followed by commas I believe

the plural of reanalysis is reanalyses

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-220, 2020.

Printer-friendly version

Discussion paper