Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-62-RC2, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "An adaptive two-stage analog/regression model for probabilistic prediction of local precipitation in France" *by* Jérémy Chardon et al.

Anonymous Referee #2

Received and published: 1 June 2017

doi:10.5194/hess-2017-62

Big Picture

The authors present and explore a methodology to simulate precipitation intensities. Yet, neither time series and/or spatial fields of simulated precipitation intensities are shown nor compared to observations (in a probabilistic manner as the title might suggest). While the methodology might be beautiful, I think this is the biggest missing thing in this paper.

I am not a specialist in analog methods. I did my best to understand what is done here. Ideally my potential failings help to detect shortcomings in the paper and lead

to improvements. Besides the analog part, I tried to help with general statistical - hydrological comments.

"Hybrid" Approach

The authors want to predict a variable (e.g., precipitation) for a given day (say, for the example of this review, May 30th 2018) at a given location (within France). Then they look at all 30-Mays in the past when precipitation amounts were recorded.

- Where exactly do the authors look? - at the closest measurement station? Is an interpolation performed? What kind of spatial dependence between observations (and simulated values) is assumed / considered?

- On p22 l1ff you write that "the predictors and regression coefficients of the regression models vary from one day to the other? – How much do they vary? And how much do they vary in neighbouring cells? Is there some kind of relationship between the variations in neighbouring cells? Can you show this?

- What if the observed time-series is not stationary? Are there any checks performed? Is stationarity assumed? How strong of an assumption is it?

- The authors claim that values outside the range of observations can be simulated via "extrapolation" (p2 line 20ff.) – some background / assumptions / limitations of this extrapolation methodology is required.

- The previous statement seems contradictory to what is said on p2 lines 29ff.:

- the author's method is able of extrapolation? - is there any evidence of the quality of the extrapolation?

- p2 line 28: I am not sure how a linear model can be "extended" to non-Gaussian data. If this is not to be a reference to what Maraun et al. (2010) did, but the authors rather claim that their method is capable of simulating non-Gaussian data, then there is some more extensive explanation required: What kind of non-Gaussian-ness is observed in

the data and how can linear models mimic this kind of non-Gaussian data? How and where is this non-Gaussianness seen in the data and how is the model describing it?

- From the abstract it did not become clear to me, what is meant with an _hybrid_ ("having two kinds of components that produce the same or similar results") approach – the title is worded more suitably. On the other hand "local" could be confused with "small scale"

Setup and Language

At various places within the paper (see comments below) parts of the methodology are explained. I suggest that the introduction is reworded and a section of the introduction is established that clearly and concisely explains what is done in one paragraph. This should also include an explicit statement of the goal and the novelty of the research.

Major Comments

Section 2 - Data

- Here, there is a distinction between "analog stage" and "regression stage" – are these two stages what is mean when the authors refer to as a _hybrid_ approach? This gets back to my original question: In the analog stage, are the authors looking for all May-30's in the past or only those May-30's where the pattern of the geopotential field was similar on the May-29's? How was this similarity determined?

- why 13 predictors? Is this enough? For what goal?

3 - The hybrid analog/regression model - the approach of using a distribution function with a portion of zeros is clear.

- what is not so clear, is how the parameters are estimated and why this is treated independently?

- should the amount of precipitation not be a random variable drawn from the distribution depicted in Figure 1? - It could then be either zero or some precipitation intensity

other than zero.

- why is \pi estimated separately from the parameters of the distribution function? (I am assuming parameters, even though Figure 1 suggests the use of an empirical distribution) Can those parameters not be estimated jointly?

- now, it seems like currently \pi is estimated via a GLM, which seems to be an improved multiple regression with the secondary variables going into xˆo (Eq.2).

- it is not clear what the difference between superscript o and superscript q is in Eqs 2 and 3.

- How does the Gamma distribution come into the game? Are you using this type of distribution to model the non-zero part of the distribution? Why Gamma? Also, the logic in p6 lines 13,14 is off. I think you should use a distribution that fits somewhat well to the data and then fit its parameters to the data.

- what determines how "near" an analog is to the predicted day? (likely this is answered in Sect. 3.2).

- why is the threshold for precipitation 0.1mm?

- p6 lines 23 ff. are difficult to understand. Say again you are trying to predict May-30 2018 in one grid location of France. Then you are searching for the "nearest" geopotential conditions for all May-29 in the past and then estimate \pi based on the precipitation occurrences in those days. The "nearest conditions" could be different for a neighbouring cell? What does this say about consistency and spatial dependence structure of precipitation fields. Also for Jun-1 2018, again a potentially very different set of "nearest conditions" could be used? Or am I understanding this wrongly, and there are more constraints?

- Why are you using the BIC (and not another criterium)?

- I would suggest a more careful wording when the word "significance" is employed.

Arguably, a predictor can be significant at a certain level, but not plainly not significant (p6 line 26ff) – what level of significance did you choose?

- p8, l21 you start to use a differently typeset "P" after the abbreviation "ESP" – please explain.

- Figure 2: top right panel: should there not be dots on the black line? At least for the part "within" \pi?

4 - Results

- p11 l12ff: you write that the BSS gain is "very sensitive to topography". The coasts along the Mediterranean (E portion of southern coast of France) and the Atlantic (W portion of northern coast of France) have opposite BSS gains (Fig 3b). How does that fit to your explanation?

- p11 l32: what do you mean by "greatly and thus significantly"?

5 - Discussion

Generally, this section reads as a strung together explanation of what is shown on several figures. What does it mean remains more unclear than the authors probably think...

- can the selection of structures (what is visualised by Figures 8 and 9) be done in a more quantitative way (contribution of each variable to the prediction)?

- p19 l1: Please describe first what your point is, then what is visualised on Figure 10).

- p22 l4ff: you write that the gain is "non-negligible". Then you write that it is "up to 0.1" – can you quantify how much of a gain this really is?

Minor Comments

- p 16, last word: necessarily

- p2 line 35: remove "obviously" or explain how this is obvious.

- throughout the paper: frequent use of "classic" and its funnily sounding adverb. What is classic in the sense of analog hydrological methods?

- The authors mention multiple times (e.g. p3 l7, p3 l14) a relation of the presented methodology to physical (maybe deterministic?) - Is the goal of the presented approach to be "physically realistic"?

- How / in what sense does this lead to something "more relevant and robust" (p3 l14/15)?

- Table 1: H, W, PV: of what variables?

- p11 l4: "two" or "four" or something else?

- p13 l19: what does "next too low" mean?

- Style: there are many abbreviations, and it's easy to forget what they all mean.

- sometimes you write "metropolitan French territory", sometimes "France". It seems like you never looked at Paris or the major cities specifically, hence I suggest to use "France" everywhere.

- p20 l2: I don't think the interest has been explored. Rather, the model itself has been explored?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-62, 2017.