Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-238-RC1, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Exploring hydrologic post-processing of ensemble stream flow forecasts based on Affine kernel dressing and Nondominated sorting genetic algorithm II" *by* Jing Xu et al.

Anonymous Referee #1

Received and published: 4 October 2020

This paper explores the post-processing of ensemble forecasts using two methods: Affine kernel dressing (AKD) and Non-dominated sorting genetic algorithm II (NSGA-II). The paper concludes on, in general, a better performance of the NSGA-II method. The topic is relevant and interesting to the forecasting community. However, the paper is sometimes too concise and needs some additional clarifications. The presentation and interpretation of the figures is too brief, and many aspects seem to remain undiscussed or unexplored. Overall, the paper is well written, but, in some places, it reads a bit strangely, and I would suggest a review of the use of the English language.

General questions and remarks:

- The aim of the paper should be more clearly stated already (and earlier) in the Introduction. My impression is that we discover the aim of the study while reading the methods and results (for instance, line 262). I also struggled to find out what the novelty of the paper is, with regards to other existing post-processing techniques in the literature. What is the additional (scientific or operational) value of the paper?

- Concerning the Introduction, I found it very difficult to follow the argumentation, since I could not see the direct links between paragraphs, and, most importantly, why the authors were raising, and long discussing, the issue of "sources of uncertainty": if a statistical post-processor is going to be applied, what difference does it make if one, previously, in the raw ensemble, quantified all sources of uncertainty, or, for instance, all but one source of uncertainty? Wouldn't the post-processor work equally well if we had 50 ensemble members from each hydrological model instead of 50x50 members?

- Also in the Introduction, overall, I think the key concepts are not introduced very clearly and just loosely thrown in the sentences. For a reader not used to the techniques, it becomes uncomprehensive. For instance, the whole paragraph on lines 49-65 reads very confusing to me. We read about "bias-corrected ensemble member", "normally distributed data", "predictive weights", or "other dressing parameters", without much explanation about what these terms mean.

- Then from line 66 onwards, it is not clear why it is novel to apply NSGA-II and compare it to a kernel-based dressing method. What are the advantages of using NSGA-II? Line 70: what "different conceptualizations" are we talking about? Line 74: what do you mean by "credibility"?

- I think the authors should completely re-write the Introduction, and think about better presenting the literature, the novel aspect of the paper and the questions the paper wants to answer (i.e., its aim). Some review of the literature presented in the "methods" section 3.2 (page 9, lines 179-204) should go to the Introduction to better explain the

reader why using NSGA-II could be considered a novel aspect in this paper.

- The paper investigates post-processing of ensemble forecasts based on 5 hydrological models and 5 sub-catchments in Canada. However, there is nothing in the paper that discusses the differences in performance among models and sub-catchments? What drives a better/worse performance of the post-processors used in the study? I missed some reflexions about this issue, which would certainly increase the value of the paper. Without this reflexion, and without aggregated (averages) results, I do not understand very well the usefulness of carrying out the study over 5 models and 5 sub-catchments. What does this diversity of applications bring to the analysis?

- I found the distinction between training and validation datasets and criteria very confusing . For instance, we present MCRPS as a validation criteria (section 3.3), but it is then said it is used in calibration (line 269). It is also not clear to me why we do not have a calibration for each lead time. What is the impact of using one unique lead time for calibration?

- Much of the justification for the selection of the study area comes from its operational role in reservoir management. However, the post-processing application presented in the paper is based on a "non-operational" context: the parameters of the post-processor are calibrated over the entire data available (not over a split sample) for a given lead time (4 days) and validated over different lead times. Operationally, though, a forecaster would have to calibrate the post-processor over a long series of past pairs of forecasts and observations, and apply it to a different set of real-time forecast (for which the observations are not yet available). What are the implications of the method proposed for an operational service? Would the operational service be fine with a post-processing that is optimized for a 4-day lead time? Is that the lead-time that most count for the service when forecasting over these catchments? Maybe some lines of discussion would be interesting in the final section of the paper.

Specific questions and remarks:

- lines 23-24: these sentences are not very clear to me.

- line 38: what are the three main sources mentioned?

- line 47-48: what are the implications of autocorrelation in the post-processing? Besides, aren't meteorological forecasts also auto-correlated? Why is it specifically a problem to hydrological forecasts?

- Fig. 2: I understand these are daily streamflow (it is written: mm/day) averaged over each month, and not monthly streamflows. Is that so? The caption should state the period over which the averages were obtained.

- Table 1: I do not understand the data on reservoir area: why it is important to this paper? Furthermore, I do not understand all these physical and climatic data provided: if the results are not going to be interpreted according to the characteristics of the catchments, why are these characteristics presented in the table? In what do they influence the results?

- Line 122: why have you chosen 5 models and why not work with the 20 models? If this is a matter of computational time, could you explain it to the reader? How long it takes to post-process one single model H-EPS?

- Line 136: I am used to forecast post-processing, but not with the term "ensemble interpretation method" or "interpreted ensemble (line 157). I would be happy with more explanations here.

- Line 147: correct English

- Line 169: what is this rule of thumb? Please, clarify.

- Equations, overall: it seems to me that not all terms are always defined, explained after the equations where they are presented. r1, r2, s1, s2, etc. zi is lower case in equation 10 but upper case in equation 11. a is alpha (line 169)? Please, check the equations and the way terms are presented.

- Line 175: “Eq. (6) can be further defined”, should maybe be replaced to “can be re-written”

- Line 205: X(t) was already defined in line 143. Please, check.

- Line 191, 192: I tend not to agree with the authors. I think “accuracy” is what is first of all searched when issuing a forecast at a given day for a short lead time such as 7 days. This is specially the case for flood events, for instance. Please, explain your arguments.

- Lines 195-196: not very clear to me. Hydrologists may rely on NSE, but for simulations (long time series), not necessarily for forecasters. Please, clarify.

- Line 200: I do not understand “elitist”. Please, clarify.

- Line 215: the concept of crowding distance was not clear to me. Please, clarify.

- line 235: was the MCRPS calculated using empirical distributions or a fitted theoretical distribution? Please, clarify.

- line 238: why do you need both, MAE and MSE?

- line 248-249: I do not understand why the Taylor diagram is mentioned here. Did you use it? How? Can you explain it?

- Figure 3: where do we find “w” in the text (output of NSGA-II in the figure)?

- lines 287-288: it is not unexpected that forecast performance decreases with lead time. I do no understand why it is “revealed” here. Please, clarify.

- lines 293-294: check for the English language.

- line 324: delete “In the meanwhile,”

- line 335: I understand that “error growth” is usually depicted as an increase in spread with lead time and decrease in accuracy. Why should it be maintained for a single model H-EPS if the post-processor was calibrated for 4 days of lead time only and

applied to other lead times? Please, clarify.

- Figure 9 is not explained in the text (notably the number of lines in each graph). Also, why AKD seems to work well with M05?

- line 343-344: not clear; please, revise it.

- line 345: figure 10 shows much more than spread. Please, clarify when presenting (fully) the figure.

- Figure 10 is very difficult to read. It is not clear (B&W print) which graph is AKD, which is NSGA-II. We can barely see what is inside the figure. I think it needs to be re-designed.

- Overall, terminology could be uniformed (ex., use of AKD)

- It is a pity that the paper does not have a discussion section. I would suggest the authors to introduce one, commenting further the results obtained, comparing post-processing performance among catchments (i.e., geographic location) and hydrological models in a summarized way. This piece of work is missing in the paper and would better justify the use of several catchments and models in the analysis.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-238, 2020.