Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-391-RC2, 2021

A comprehensive assessment of uncertainty sources in an operational streamflow forecasting system

Anonymous Referee #2

Referee comment on "Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems" by Emixi Sthefany Valdez et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-391-RC2, 2021

This is a fascinating and well presented study of uncertainty estimation in an ensemble streamflow forecasting system. These choices can often be complex and difficult to disentagle, but this study systematically nagivates a clear path through the maze. The assessment of different aspects of the forecasting system is comprehensive and rigorous and tied together very nicely with concrete recommendations. The methods used to generate the forecasts at the leading edge of hybrid dynamical-statistical forecasting systems. The introduction on its own is extremely interesting and gives a really neat guided tour of ensemble quantification in streamflow forecasting. Very nice job! I have a few suggestions for improvement, but really these are more guidelines than hard recommendations. I congratulate the authors on a very nice piece of work.

Major comments

1) Figure 8, and the commentary associated with it, is problematic. The authors are comparing reliability of their hydrological uncertainty quantification methods when they know that their forcings are unreliable (as deftly shown in fig 10). This makes assessments of reliability meaningless: the streamflow forecasts in fact *should* be unreliable, as the forcings are unreliable. So, for example, in Fig 8 System B looks bad next to System D (for reliability), but Fig 13 shows that when forcings are reliable, System B performs similarly to System D. So Fig 8 gives misleading information - indeed, one could argue it shows System B to be superior (despite the appearance to the contrary). There are a few ways around this - e.g. the authors could generate hydrological 'forecasts' that are forced with observations to produce Fig 8 - but the simplest way is to remove this figure and the accompanying discussion, and focus more on Figs 11-14.

2) The paper is admirably comprehensive, but perhaps as a result somewhat long. Longer papers are less likely to be read, and I really feel like this paper deserves to be read widely! So I have a few suggestions for shortening it:

i) Figure 7 replicates information in Figure 9. I suggest omitting Figure 7 and

restructuring the paper to first discuss rainfall forecasts (both raw and post-processed) and then streamflow forecasts.

ii) Fig 8 should be omitted, as noted above

iii) Figs 11-14 could be combined into a single figure. At the very least, I suggest Figs 13 & 14 should be combined - as the authors note, sharpness without reliability is meaningless.

iv) [Optional] The discussion of the hydrological model performance (Section 3.1) and the accompanying figure, while interesting, could be moved to supplementary materials.

3) Figure 15 shows results that appear to be somewhat inconsistent with Figs 12 & 13. For example, in Figure 15 Sys-C looks best (most reliable) across all catchments compared to other systems - much better than Sys-B or Sys-D. This is very much not the case in Fig 13, where Sys-B & Sys-D are more reliable when paired with the CSGD forcing. Further, the RMSE values for the CSGD forcings are often almost identical to the raw forcings, which is not consistent with the CRPS scores presented in Fig 12. I think spread-skill plots are a pretty rough way of looking at reliability (they reduce the 'spread' of the ensemble to root of the average variance, rather than considering the full distribution, e.g. like Probability Integral Transforms/Rank Histograms do), and I suspect the reliability diagram analyses presented by the authors in Fig 13 give a better picture. It may be better to simply use MAE_rd & CRPS in Figure 15 for consistency with Figs 12 and 13. If the authors keep the spread-skill plots in Fig 15 (they may have different views to me on the efficacy of the different reliability assessments) - I think the apparent discrepancies between Fig 15 and Figs 12/13 should be explained.

Specific (minor) comments

L26-27 'The inherent uncertainty of hydrological forecasts stems from three main sources' there is a crucial fourth source the authors have omitted: uncertainty in observations.

L28 "Two main philosophies are generally adopted" I understand why the authors are making this distinction, but I guess I would argue the emergence of a third class of systems, namely hybrid statistical-dynamical forecasting systems - which combine physical models (e.g. ensemble NWP) with statistical processing methods (e.g. NWP calibration) - is worth mentioning. Statistical post-processors (hydrological uncertainty processors, ensemble dressing) are very effective at correcting statistical aspects of forecasts (bias, uncertainty quantification), but usually produce discrete probability distributions in time (e.g. one for each lead time) and space (e.g. one for each gauge) - they are not connected in time and space. Users of hydrological forecasts are often interested in hydrographs, which require coherent information in time (across lead times) and (often) in space (e.g. upstream and downstream gauges). Coherent spatial and temporal information allows users to, for example (i) check flood peaks as well as total volume forecasts for, say, the next week; (ii) when flood peaks are expected to pass by a sequence of gauges. Ensemble systems can present all this information, but when they are based on purely physical/conceptual models they often do not get statistical aspects of forecasts (bias, uncertainty quantification) correct. Hybrid statistical-dynamical systems are in their infancy, but they promise the best of both purely statistical or purely dynamical ensmeble systems.

L71-73 "Hydrological post-processing ... cannot compensate for weather forecasts that are highly biased" Yes, it can - though it depends on the method employed. Hydrological uncertainty processors (e.g. Biondi and Todini 2018) lump all sources of uncertainty -

including biases in rainfall forecasts - into the model, and they are effective at correcting these issues. They are not so good if you need each ensemble member to represent a temporally coherent hydrograph, however.

L84 "Considering several sources of uncertainty in ensemble hydrometeorological forecasting inevitably implies an increase in the number of simulations" This is not 'inevitable' - e.g. Bennett et al. (2016), who add hydrological uncertainty to rainfall forecast uncertainty without increasing the number of ensemble members. Suggest this be changed to "Considering several sources of uncertainty in ensemble hydrometeorological forecasting often implies an increase in the number of simulations".

L87 "However, there is also additional uncertainty associated with the assumptions made in creating a larger ensemble." Please be explicit about why increasing ensemble size is a problem.

L137 "modeling uncertainty" should this be "hydrological modeling uncertainty"?

L211-213 "As such, it relaxes the constraint of parametric assumptions (difficult for variables like precipitation) and high dimensional cases (Gneiting, 2014). This technique uses the raw ensemble forecasts trajectories to identify the dependence template. Therefore, the post-processed ensemble should have the same number of members as the raw ensemble." I didn't find this explanation very clear or helpful. ECC enforces rank correlations (both temporal and spatial) from the raw forecasts in the calibrated forecasts, much like the Schaake shuffle. I would simply say something like this.

L219 "provide model state" The authors use a range of models (nwp, statistical, hydrological), and it would be helpful to clearly identify (here and elsewhere) which ones they are talking about in each case, rather than referring to generic 'models'. Here I think the authors are discussing hydrological model states.

L315 "Root Mean Square Error" is this calculated on the ensemble mean? (And if not, how is it calculated?)

L333-334 "Accordingly, we decided to use the reliability diagram to evaluate precipitation for thresholds of different exceedance
probabilities (EP), namely 0.05, 0.5, 0.75, and 0.95". Could the authors please confirm 1) that to construct the reliability diagrams, they have converted probabilistic forecasts to binary predictions for exceeding thresholds based on these quantiles and 2) these quantiles are taken from observations and 3) that they have calculated MAE_rd across all these thresholds? These things weren't entirely clear to me from the description.

L338 IQR (interquartile range) usually denotes the range between two quartiles (25%ile and 75%ile). So 'IQR' is not the correct term to use for a different interval. Suggest using the term 'average width of prediction interval' (AWPI).

L362 Figure 4 - this is a really nice figure!

L386 Figure 5 - would be helpful to have abbreviated names of the models along the x-axis rather than M1, M2, etc, to avoid having to track back to Table 2 to interpret this figure. These only need to appear on the bottom panel.

L507 "This is particularly observed in the streamflow forecasts and may be related to the use of lumped hydrological models, which tend to smooth errors over larger modeling areas" This is kind of saying the same thing, but I'd shift the emphasis to the properties of the catchments themselves, rather than lumped hydrological models. Something like "For a given time step, smaller catchments tend to show more variance and are thus more difficult to forecast than larger catchments." (I.e., it's the larger catchments themselves that smooth, e.g., spatial and temporal variability in rainfall; the models just try to replicat this.)

L515-516 "These findings confirm that large improvements in precipitation forecasts do not necessarily lead to improvements in streamflow forecasts." I don't wholly agree with this. First, Sys-B *does* show marked improvement in reliability when it is forced with reliable inputs. So there is improvement there. The 2 best peforming methods, to my eye, are Sys-B and Sys-D when forced with Q_CSGD. Sys-B is marginally superior: at Day 6, Sys-B with Q_CSGD is sharper than Sys-D, despite being similarly reliable. (When you consider the additional complexity of Sys-D, Sys-B to my mind is clearly preferable for operational forecasting.) Further, the measures of reliability the authors have chosen may not be all that sensitive. I personally prefer the use of probability integral transforms (PIT), or their summary statistics (e.g. Renard et al.'s (2010) 'alpha' statistic - very similar to the authors' MAE_rd). Because PIT treats forecasts as continuous variables, there's no need to simplify forecasts to binary predictions of thresholds. I'm not suggesting any changes here - the analysis the authors have presented is easily thorough enough, but I think they are understating the value of the CSGD in the system, and should probably soften their wording here.

L537 "In system D, the application of precipitation post-processing does not lead to reliable streamflow forecasts: the RMSE and spread curves in red in Fig. 15 are not aligned, while the blue curves (for raw forecasts) are for lead times greater than 3 days" The reason seems clear: Sys-D ensembles are wider than they need to be at Day 6, as shown cf Sys-B in Fig 14. So while they are reliable, they are not as sharp as they could be. In short, Sys-D overestimates uncertainty at long lead times.

L574 Fig 16 - this is a fascinating figure, probably worth a paper on its own. Great stuff!

L578 "(Fig 15 system C)" As already noted, I have misgivings about Fig 15. In other figures (Figs 11-14) Sys-C is clearly outperformed by both Sys-B and Sys-D. And the reason is not really surprising: hydrological models, while somewhat diverse in structure (in some respects they can be quite similar, too), often perform very similarly once calibrated (as the authors themselves show in Section 3.1). So the uncertainty explained by a range of calibrated models is likely to be quite small, explaining why Sys-B can outperform Sys-C even though it uses a single hydrological model.

L581 "If the priority is to achieve a reliable and accurate system, then system B with precipitation post-processor presents a better alternative than a system like D." Yes, I fully agree.

L586 "When post-processing is not applied to precipitation forecasts... systems with multimodel provide better reliability" I agree that this is what Figs 8 and 15 show, but as discussed: Fig 8 shouldn't be expected to produce reliable forecasts, and Fig 15 appears to be at least partly inconsistent with Figs 12 & 13. These conclusions can only be drawn with certainty if Fig 8 was constructed from streamflow forecasts generated with 'perfect' (observed) rainfall forcings to isolate the effects of the hydrological uncertainty generation methods. I don't think this paper needs this additional analyses - it has plenty of material already - so I suggest omitting this paragraph.

L589-590 "System B, with 2,500 members, performed similarly to system C (350 members) in terms of MCRPS as the EnKF effect fades with increasing lead times." I hadn't realised the difference in ensemble size was so stark. It's likely that the discrepancy in ensemble sizes would have impacted on all their calculations (though I accept the authors' contention that the effect was small) - however for future work I suggest subsampling from the large ensembles to control for ensemble size, or using unbiased estimators that control for ensemble size (e.g. Ferro et al. (2008)) so any potential artifacts of ensemble size are totally removed.

L630 "Since model C presented a good option, it would be interesting to determine if this system with a hydrological post-processor that corrects the models' bias would improve the forecasting performance without resorting to sophisticated precipitation post-processor techniques." To reiterate, the conclusion that System C is a good option is really based only on Fig 15, about which I have misgivings. If these misgivings are borne out, I would suggest that System C would not necessarily be considered a 'good option'.

Typos etc.
L38 "parameters uncertainty" should be "parameter uncertainty"

L101 "directly use NWPs outputs" should be "directly use NWP outputs"

References

Bennett JC, Wang QJ, Li M, Robertson DE, Schepen A. 2016. Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model. Water Resources Research 52: 8238–8259. DOI: 10.1002/2016wr019193.

Biondi D, Todini E. 2018. Comparing Hydrological Postprocessors Including Ensemble Predictions Into Full Predictive Probability Distribution of Streamflow. Water Resources Research 54: 9860-9882. DOI: 10.1029/2017wr022432.

Ferro, C.A.T., Richardson, D.S., Weigel, A.P., 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol. Appl. 15 (1), 19–24. https://doi.org/10.1002/met.45.