Hydrology and Earth System Sciences
Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2022-72-RC2, 2022

Comment on hess-2022-72

Anonymous Referee #2

Referee comment on "Development of a national 7-day ensemble streamflow forecasting service for Australia" by Hapu Arachchige Prasantha Hapuarachchi et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-72-RC2, 2022

Major comments

Error correction vs consistency. The application of ERRIs is quite impressive in terms of taken care of the errors and producing the best reliable forecast estimate. However, I am a bit concerned about the methodology in an operational setting. You state that observed discharge is used if available, and if not the post-processed streamflow is used instead. Is there not a risk that the forecast becomes jumpy if it is initialised differently from one forecast to the other? How is this information relayed to the forecaster, and how can they take this into account when taking decisions?

Evaluation and calibration of the ensemble forecasts. Maybe I am missing something in the methodology, but it is not clear to me exactly how the optimal ensemble forecast is derived. In Section 2.3 you describe something that sounds more like a resampling from the available data than actually expanding the ensemble size (see specific comment). Later, it is mentioned that CHyPP generates 400 bias corrected forecasts. The calibration of forecast is mentioned but since no closer description of the method is given it is not clear to me how the optimal ensemble size is achieved. I suggest the authors to be clearer on these points.

Acceptance criteria. You mention her skill criterion for releasing forecasts to the public, but would not the value of the forecast be a more informed measure? In areas with high risk, even a not so skillful forecast can still be very useful.

Minor comments

- You state that the forecasters need information on the longest possible lead time, but I would argue it depends on the action needed.
- Reference for EFAS is missing
- L94-95. This sentence could be split to increase readability
- You start here by describing how you created the area-averaged rainfall, but I miss some information on the size of these sub-catchments. I would suggest at least introduce the hydrological modelling concept to better understand why this step is necessary.
- L129-136, Table 1. The description of the Super-ensemble is a bit confusing to me. When you say concatenate, I assume you mean that the ensembles are added to create a larger ensemble. I might use merge here, since concatenate to me suggests they are stitched together in time. Also, how do you create the hourly temporal resolution from the 3-hourly. There might be some feature in CHyPP method, but it is not clear
- Here you describe how sub catchments are created, but I still miss information on the typical sizes. I would recommend a table or figure to show the distribution of sub basin sizes to put it into context with the resolution of the NWP models.
- In the evaluation framework you use the terms validation of the calibration, but forecast verification. I think the term validation is good, but the term verification is very often used a bit misleading in meteorology. A forecast cannot in principle be verified since there is no absolute truth, and we are not looking for the absolute truth. We are looking for a forecast that can pass certain criteria, so the term benchmarking is to me a better term to use.
- Section 2.3 is interesting. Normally this is not how you determine the optimal ensemble size. If I understand correctly your method you are sampling randomly from the hindcast period, thus choosing forecasts from a random starting date. The forecast skill is however very varying from time to time, so I am not sure that it is the best way of deciding the optimal skill. Would it not be better to dress the ensembles to create more members for each forecast time, than reducing the number of ensembles taking the whole hindcast period into consideration?
- Here you mention hourly forecasts from ACCESS-GE2 and ECMWF, in table 1 you mentioned 3-hourly forecasts?
- What is the reason for averaging over 24h before making the skill assessment? Is that not blurring the skill assessment? You will have better results, but you might miss some important information for example on timing errors in the forecast. I would suggest to also look at 3 or 6-hourly scores to see how they compare witht the daily forecasts.
- This is a personal preference, but I would suggest to change the order of chapter 2 and 3.
- You use NSE here as a metric, but it is nowadays the standard to use Kling-Gupta Efficienc.
- Section 2.5.5, I would suggest to merge this with the description of CRPS. I would suggest to always use CRPSS since it standardises the values automatically. CRPS(S) is

very sensitive to bias, therefore it does makes sense to decompose it inot its components, or at least show also the bias alongside CRPSS

- You mention here a threshold value of 0.6. is there any particular reason this is used?
- In section 4.2 you discuss the effect of calibration on the bias of the forecasts. That is all good, but I would like to see how the spread is affected by the calibration.
- In the same section you also show the relative CRPS of the rainfall. I would suggest ot instead show the CRPSS here as a measure of skill, alternatively other scores which are more targeted towards the skill of precipitation.
- In section 4.3 you show the effect of error correction on the streamflow, and it is clear that removing the bias improves the forecast. What is not clear to me is if the calibration of forecasts is applied as well?
- The acceptance criteria of 0.6 of NSE seems to me a bit contrived. All values above zero carries some values, so it would still be useful for the users?
- L507-514. You discuss there the value of the calibration and I agree that the method is most likely very beneficial to the users, but in the acceptance criteria you did not weigh in the users perspective (value). To be consistent I would suggest to actually add that to the acceptance criteria
- In the same section you mention the fact that the calibration worsen CRPS(S) for longer lead times but you do not give an explanation to this behavior. Could you say something about that?
- Section 5.2 I am a bit confused why you have this section. It is names uncertainties in forecast, but you almost only talk about the uncertainties in observations. I do not see the real relevance of this discussion with regards to this paper? I would suggest reducing this bit, or at least not into so much details regarding observations.
- Section 5.3. I really like this section and the very important discussion of the complexity of correcting forecast errors. It should also be mentioned that data assimilation has a potentially negative effect for hydrology since the water budget is compromised, which in turn can lead to long term biases in variables such as soil moisture runoff and discharge.
- Section 5.4 This list of challenges is good, but can you state which of these are specifically important for Australia?
- Section 6. I do not understand why this section comes here, this should have been presented at the beginning of the paper. Am I to understand that the CHyPP model “generates” 400 ensemble members form ECMWF’s 51? I would need more detail or at least a very good reference to this method to understand it better.