Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-72-RC1, 2022

Comment on hess-2022-72

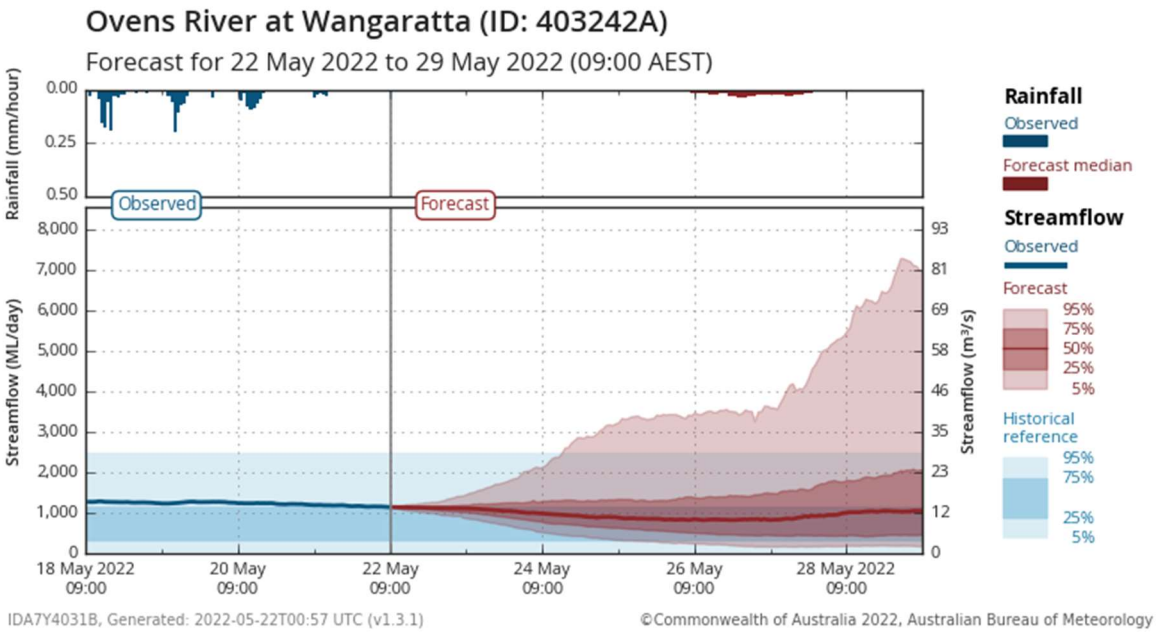
Anonymous Referee #2

Referee comment on "Development of a national 7-day ensemble streamflow forecasting service for Australia" by Hapu 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?

Author Response: We thank the reviewer for commending the use of ERRIS for error correction and understand the concern of its use in operational setting. Note that we do not add any noise to the simulation. The ERRIS model initial state is carried forward, and therefore the simulation and the forecast are smooth. As in the example given below, every day, we initiate the model run from 4 days prior to the forecast time. This allows us to capture most recent observed data. The model runs with observed data for 4 days and then smoothly switch to the forecast mode. If the observed data is missing, the impact of ERRIS gradually declines (switch to forecast mode) with lead-time and the corrected hydrograph overlaps with the raw simulated flow. Therefore, model simulation to forecast mode is smooth. Sometimes if poor quality observed data is ingested to the model, we find jumpiness in the forecasts. If this occurs, the model is taken temporarily out of the service, and the user community is notified through the website. The Bureau has a dedicated monitoring team to do this. To address the reviewer's question, we can elaborate on this issue in Section 6 and the last dot point of Section 5.4.



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.

Author Response: We thank the reviewer for the comment. In section 2.3 we describe the methodology for deciding the optimal ensemble size and the results are presented in Section 4.1. As the reviewer correctly pointed out, we generated 1000 rainfall ensemble members and resampled them to find the optimal ensemble size. Alternatively, one could independently generate a different number of ensemble members and analyse them. We will elaborate the Section 2.3 to make it clear for the reader.

Acceptance criteria. You mention here 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 skilful forecast can still be very useful.

Author Response: We agree with the reviewer's comment, and a similar opinion was expressed by the other reviewer. The forecast skill criteria are one set of measures for selecting a forecasting location for the service. As the reviewer suggested, we consult our stakeholders and identify forecast locations critical for their decision making and add to the service. Sometimes, the forecast locations with poor skill are only made available to registered users (accessed with a username and password). This is to reduce possible miscommunication with the public and to keep the reputation of the service. To address reviewer's question, we will elaborate Sections 4.5 and 5.4.

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.

Author Response: Yes, we agree. Through stakeholder consultation we found that they need forecast information from hours-to-days-to-months-to-years-to-decades depending on their planning and management needs. We will refine the text accordingly.

- Reference for EFAS is missing

Author Response: We will include EFAS in revised manuscript.

- L94-95. This sentence could be split to increase readability

Author Response: Yes, we will split the sentence 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.

Author Response: We agree with the reviewer and will elaborate the hydrological modelling concept in the paper.

- 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

*Author Response: We agree that 'merge, is a better word than 'concatenate'. We will change the text accordingly. 3-hourly rainfall data are disaggregated to hourly using linear interpolation. Kindly refer our paper, Bennett et al., 2016 (*doi:10.1016/j.envsoft.2015.11.006*), which shows that even converting daily totals to hourly in this way produces plausible rainfall-runoff model outputs.*

- 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.

Author Response: Agree with the comment. In sub-Section 2.2.1 we present typical size of sub-areas. In our modelling approach, the smallest unit is a sub-area. A collection of sub-areas makes a sub-catchment. We can elaborate the sub-catchment and sub-area sizes in the sub-

section. If needed, we can provide a figure showing catchment area (x-axis) and number of sub-catchments (y-axis).

- 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.

Author Response: We agree with the reviewer that a forecast cannot in principle be verified since there is no absolute truth. This is a very literal interpretation of the term 'verification' - which relates to truth - sometimes we have seen this argument made by philosophers of science who argue models can only be 'validated' not 'verified'. However, the words do not have fixed meanings as given in dictionaries. In streamflow forecasting (and forecasting more generally) the term 'verification' is widely used (Kunnath-Poovakka and Eldho, 2007; Anctil and Ramos, 2019; Wu et al., 2020) to describe what is presented in our paper. We could change it to 'benchmarking' to satisfy the reviewer, but then our target audience - fellow forecasters - would not be clear on what we mean. Our preference is to keep the existing word 'verification' in the manuscript.

- 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?

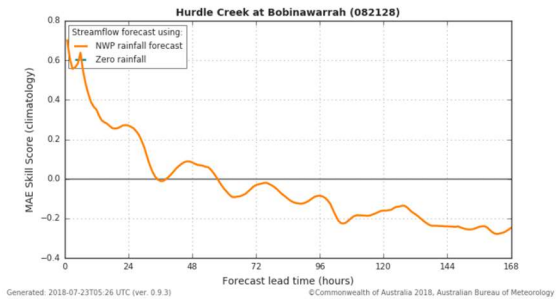
Author Response: We thank the reviewer for the comment. In Section 2.3 we describe the methodology followed for deciding the optimal ensemble size and results are presented in Section 4.1. As the reviewer correctly pointed out, we generated 1000 rainfall ensemble members and resampled it to find the optimal ensemble size. Alternatively, one could independently generate different number of ensemble members and analyse them. We wish to ensure that the forecasts are true ensembles - i.e. each ensemble member can be summed across time to produce reliable forecasts of accumulations (e.g. 7-day streamflow totals). 'Ensemble dressing' methods that simply add noise to a given lead time are not suitable for this type of calculation. We will elaborate on this in Section 2.3 to make it clear for the reader.

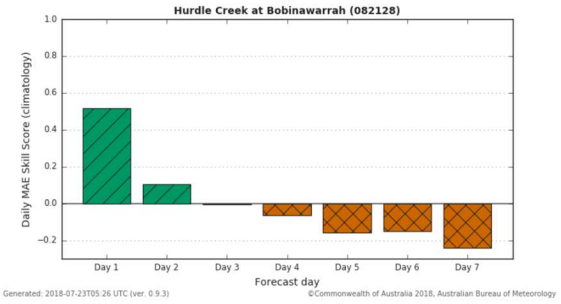
- Here you mention hourly forecasts from ACCESS-GE2 and ECMWF, in table 1 you mentioned 3-hourly forecasts?

Author Response: We generate hourly rainfall forecasts after calibration using CHyPP at sub-area scale. However, the raw NWP data are 3-hourly. We acknowledge reviewer's point and will modify the text to accommodate it.

- 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 with the daily forecasts.

Author Response: A good question. We have tested hourly and 3-hourly skill of the forecasts. However, for most of the forecast locations, it is noisy and hard to summarise the skill of the forecasts. Please find below an example of mean absolute error skill score for hourly (fig 1 – left) and daily (fig 2 – right) data for a selected location.
The daily forecast skill plots provide a consistent message, and they assist the users to make decisions with the forecasts. As expected, the daily forecast skill (generally) declines with lead-time. Stakeholders we communicated with prefer daily skill plots, and so these are provided to the public via the website as reference (expected skill). For consistency with the service, here we presented daily skill plots.





- This is a personal preference, but I would suggest to change the order of chapter 2 and 3.

Author Response: We are happy to accommodate this request. However, we prefer the methodology first so that the reader is clear on what data is necessary to trial the methodology.

- You use NSE here as a metric, but it is nowadays the standard to use Kling-Gupta Efficiency.

Author Response: We agree that KGE is commonly used in the literature nowadays, but NSE is also still very common. Among the operational forecasting community, NSE is more widely used and understood by the Bureau's stakeholders (as opposed to in the academic literature), which is why we have used it here for this service.

- 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 into its components, or at least show also the bias alongside CRPSS

Author Response: We agree with the reviewer, and we will merge 2.5.5 with 2.5.3.

- You mention here a threshold value of 0.6. is there any particular reason this is used?

Author Response: Similar question was asked by Reviewer 1. We adopted NSE of 0.6 from Chew and McMahon (1993) in consultation with the stakeholders. We will elaborate this section and will include the reference.

- 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.

Author Response: We show that calibration improves the spread (reliability) of the forecasts dramatically in figure 5.

- In the same section you also show the relative CRPS of the rainfall. I would suggest to instead show the CRPSS here as a measure of skill, alternatively other scores which are more targeted towards the skill of precipitation.

Author Response: We thank the reviewer for the comment. Our focus here is to understand the error in rainfall to explore the uncertainty contributing to the streamflow forecasts. Therefore, we thought of presenting the error (relative CRPS), not the skill. We present the streamflow forecast skill because it is useful for users and managers to make decisions. We acknowledge the reviewer's comments and agree that rainfall skill assessment is important for comparing different rainfall forecast products and we will make a note for our future research.

- 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?

Author Response: Thanks for the observation. Calibration is used in Section 4.2 for rainfall and not applied to streamflow. Instead, we use error modelling (ERRIS) to reduce hydrological model errors and quantify uncertainty in hydrological processes. We do not apply a calibration to streamflow, as we wish to ensure that the forecasts are true ensembles - i.e. each ensemble member can be summed across time to produce reliable forecasts of accumulations (e.g. 7-day streamflow totals). Calibration must be applied at discrete lead times, and reassembling temporal properties (e.g., with the Schaake Shuffle) is more difficult for streamflow than for rainfall. We will elaborate this section and make it clear to the reader.

- 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?

Author Response: We thank the reviewer for this question. Similar statement was also made by the other reviewer. We adopted NSE of 0.6 from Chew and McMahon (1993) in consultation

with the stakeholders. We will elaborate this section and will include reference. We agree that additional sites where forecast are not ''scientifically acceptable" still bring benefit to the user communities. We have responded to a similar suggestion above under the reviewer's major comments.

- 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

Author Response: We agree with the reviewer and will elaborate this section to cover users' perspective.

- 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 behaviour. Could you say something about that?

Author Response: We agree with this comment. There is some explanation given L384-386 though. This is an impact from the calibration. The relative error at shorter lead times is low but it is slightly higher than raw rainfall at long lead times. However, the reliability is significantly high at longer lead-times. Normally, calibrated rainfall values are closer to climatology values at long lead times, and there is a trade-off between sharpness and reliability. We will elaborate this section and make this clear to the reader.

- 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.

Author Response: We thank the reviewer for the observation and suggestions. We agree, and will reduce this discussion, in particular the text related to 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.

Author Response: We acknowledge and agree with the reviewer's comments. We will modify this section as suggested.

- Section 5.4 This list of challenges is good, but can you state which of these are specifically important for Australia?

Author Response: We thank the reviewer for this suggestion. We will elaborate this section and make it more relevant to Australian context.

- 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.

Author Response: We thank the reviewer for the opinion. As suggested by the other reviewer, we will swap the Section 6 and Section 5. As discussed in the results, we plan to generate 200 ensemble members from each ACCESS-GE2 and ECMWF (total 400) using CHyPP (Robertson et al., 2013). However, at the time of the operational release of the service, ACCESS-GE product was not operationally available. We tested and decided to generate 400 ensemble members from ECMWF alone in the operational system (Fig. 12) until ACCESS-GE is operationally available. The current operational system runs with these settings and the incorporation of ACCESS-GE is planned for the near future. We will elaborate this section and make it more clear to the reader.

References

Anctil, F., & Ramos, M.-H. (2019). Verification metrics for hydrological ensemble forecasting. In Duan, F. Pappenberger, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), Handbook of Hydrometeorological Ensemble Forecasting. Berlin: Springer. https://doi.org/10.1007/ 978-3-642-39925-1_3

Bennett, J. C., Robertson, D. E., Ward, P. G. D., Hapuarachchi, H. A. P., Wang, Q. J.: Calibrating hourly rainfall-runoff models with daily forcings for streamflow forecasting applications in meso-scale catchments. Environmental Modelling & Software 76: 20-36. doi:10.1016/j.envsoft.2015.11.006, 2016.

Brown, J. D., Wub, L., He, M., Regonda, S., Lee, H., and Seo, D.J. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS):1. Experimental design and forcing verification, http://dx.doi.org/10.1016/j.jhydrol.2014.05.028, J. Hydrol. 519, 2869-89.

Chiew, F.H.S., McMahon, T.A., 1993. Assessing the adequacy of catchment streamflow yield estimates. Australian Journal of Soil Research 31, 665–680.

Kunnath-Poovakka, A. and Eldho, T. I.: A comparative study of conceptual rainfall-runoff models GR4J, AWBM and Sacramento at catchments in the upper Godavari river basin, India, J. Earth Syst. Sci., doi:10.1007/s12040-018-1055-8, 2019.

Robertson, D. E., Shrestha, D. L. and Wang, Q. J.: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, Hydrol. Earth Syst. Sci., doi:10.5194/hess-17-3587-2013, 2013.

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F. and Robertson, D. E.: Ensemble flood forecasting: Current status and future opportunities, WIREs Water, doi:10.1002/wat2.1432, 2020.