Review of "Optimising Ensemble Streamflow Predictions with Bias-Correction and Data Assimilation Techniques" by Maliko Tanguy and collaborators.

Principal criteria	Excellent (1)	Good (2)	Fair (3)	Poor (4)
Scientific significance: Does the manuscript represent a substantial contribution to scientific progress within the scope of Hydrology and Earth System Sciences (substantial new concepts, ideas, methods, or data)?		X		
Scientific quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?	X			
Presentation quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?	X			

Thank you for the opportunity to review this manuscript. I fully support the body of work presented in the Hydrological Outlook UK (HOUK), which offers a novel, brave, transformative, and impactful approach to comparing bias-correction (BC) and data assimilation (DA) techniques for improving hydrological forecasts using the Ensemble Streamflow Prediction (ESP) method in the UK. I applaud the authors for their significant accomplishments in this field.

For full disclosure, I acknowledge a potential bias in my review as I am the author of the Flow Duration Curve Quantile Mapping Bias Correction Method used in this study. I am honored that this method is contributing to operational hydrological forecasting.

In my view, this paper holds great value for publication in *Hydrology and Earth System Sciences (HESS)*. Based on my reading and review, I believe the manuscript would benefit from some "minor revisions." I recommend a few amplifications to enhance clarity for the reader. My specific comments are provided in the order they appear in the document.

In Line 48 or Line 55, it would be helpful to include a brief description of how the operational ESP currently works. For example, in 2024, which historical years are used to generate forecasts? Additionally, how are the initial hydrological conditions (IHCs) calculated for each month in the operational setting? I suggest also including a comparison with the methodology used in this paper to highlight any key differences

In Line 85, Figure 1a is referenced, showing the absolute percent bias (absPBIAS) in streamflow simulations using the GR4J model. Could you clarify whether the absPBIAS data presented in this figure is based on results from this study or if it references a previous study, such as Smith et al. (2019)? If the figure represents data from the current study, it may be more appropriate to introduce or elaborate on it in the Results section to align with the manuscript's flow.

In Line 86, you reference Figures A4 and A5 in the appendix. While the figures provide valuable insights, the specific types of biases (e.g., percent bias, raw bias) illustrated in these figures have not been clearly defined. To enhance clarity, I suggest adding brief definitions or explanations of these biases directly in the appendix where Figures A4 and A5 are presented. This would help readers better understand the data without needing to refer back to earlier sections of the manuscript.

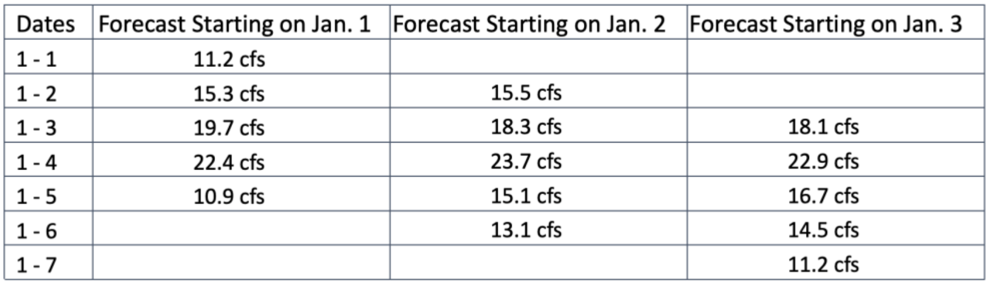
The following point is philosophical, and the authors are free to disagree or rebut, but I would appreciate your consideration. In the paragraph from Lines 96 to 108, you introduce the concept of Quantile Mapping Bias Correction (QM-BC). While Quantile Mapping (QM) typically involves adjusting the cumulative distribution function (CDF) of simulated data to match that of observed data, the method you describe uses Flow Duration Curves (FDCs) instead of CDFs. FDCs and CDFs, while similar in that they both involve sorting data and calculating probabilities, serve different purposes and represent different types of probabilities. FDCs focus on the exceedance probability of flow rates, making them particularly valuable in hydrological studies, whereas CDFs provide a broader statistical tool for analyzing any data distribution. Given this distinction, the bias correction method you are using might be more accurately described as a 'Flow Duration Curve Quantile Mapping Bias Correction Method (FDCQM-BC).' I suggest considering this terminology in the revised version of the paragraph to better align with the methodology you are applying. It's also important to note that some of the papers you cite (e.g., Chevuturi et al., 2023; Farmer et al., 2018) refer specifically to FDCQM-BC, while others (e.g., Usman et al., 2022; Li et al. 2017; Hashino et al., 2007; Wood and Schaake, 2008) discuss the typical QM method.

I would appreciate some clarification in section 2.2.2 on how the different forecasts are established. Here are the specific elements I would like to understand better

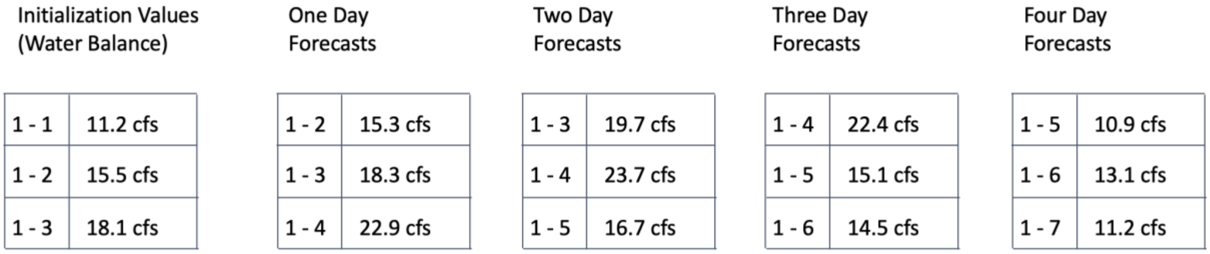
1. **Forecast Initialization and Horizon:**
   My understanding is that after running the hydrological model to obtain the 'Simulated Observed River Flows' dataset, the first three years are used to warm up the model. The OR-ESP is then calculated each month with this data (initialized on the 1st day of the month obtained from the 'Simulated Observed River Flows') and with the calibrated parameters with a 1-year forecast horizon. For example, the first forecast, starting on 1964-01-01, extends to 1964-12-31, the second forecast, starting on 1964-02-01, extends to 1965-01-31, and so on, until the last forecast starting on 2015-12-01 and

extending to 2016-11-30. Based on this, it seems there would be a total of 624 forecasts over this period. Could you confirm if this interpretation is correct? Additionally, I would appreciate an explanation that defines these details in the forecast datasets.

2. **Number of Ensembles in ESP Datasets:**
   The different ESP datasets (OR-ESP, BC-ESP, DA-ESP) contain 51 ensembles. Considering that these datasets span from 1964 to 2015. Using historical climate sequences from 1961 to 2015, I would expect 55 years of data. After applying the leave-three-years-out cross-validation (L3OCV), this would leave 52 ensembles. On the other hand, using historical climate sequences from 1964 to 2015, I would expect 52 years of data. After applying the leave-three-years-out cross-validation (L3OCV), this would leave 49 ensembles. Could you clarify how the number 51 was determined? Specifically, was either 1961 or 2015 excluded in generating the ESP data?
3. **Construction of Time Series for CRPSS:**
   I am unclear on how the time series used to calculate the Continuous Ranked Probability Skill Score (CRPSS) were constructed, especially for different forecast horizons like 1-day, 3-day, 1-week, etc. For example, assuming a 4-day forecast which is launched daily, we can build time series for Initialization, 1-day, 2-day, 3-day, and 4-day forecast as shown in the schema below:



Dates	Forecast Starting on Jan. 1	Forecast Starting on Jan. 2	Forecast Starting on Jan. 3
1 - 1	11.2 cfs		
1 - 2	15.3 cfs	15.5 cfs	
1 - 3	19.7 cfs	18.3 cfs	18.1 cfs
1 - 4	22.4 cfs	23.7 cfs	22.9 cfs
1 - 5	10.9 cfs	15.1 cfs	16.7 cfs
1 - 6		13.1 cfs	14.5 cfs
1 - 7			11.2 cfs



Initialization Values (Water Balance)		One Day Forecasts		Two Day Forecasts		Three Day Forecasts		Four Day Forecasts	
1 - 1	11.2 cfs	1 - 2	15.3 cfs	1 - 3	19.7 cfs	1 - 4	22.4 cfs	1 - 5	10.9 cfs
1 - 2	15.5 cfs	1 - 3	18.3 cfs	1 - 4	23.7 cfs	1 - 5	15.1 cfs	1 - 6	13.1 cfs
1 - 3	18.1 cfs	1 - 4	22.9 cfs	1 - 5	16.7 cfs	1 - 6	14.5 cfs	1 - 7	11.2 cfs

   I would appreciate a detailed description of how these time series were built. It may also be helpful to include more than one graphical schema to clarify this process. From my understanding, the 6-month forecast will include time series starting on 1964-06-01 (from the forecast launched on 1964-01-01) to 2016-05-31 (from the forecast launched on 2015-12-31 and would also include the months from 1964-07 (from the forecast launched on 1964-02-01), 1964-08 (from the forecast launched on 1964-03-01), and so on, until the last forecast month in May 2016. I assume, in a similar way, the time series for 30-day, 1-month, and 3-months were built. However I am not certain if this applies 1-day, 3-day, 1-week, and 2-week forecast horizon.
   Additionally, could you clarify what the 'initialization month' in Figure 6 refers to? Does it represent the month when the forecast was launched, the start of the forecast time

series, or the specific month within the forecast time series data? Does this change the forecast time series construction?

I would appreciate some clarification in Section 2.4 regarding the Data Assimilation process. My understanding is that each time a new forecast is launched, the hydrological model needs to be adjusted to accurately reflect the initial hydrological conditions (IHCs). You mention that the Particle Filter (PF) method is used, which simulates potential scenarios with different sets of parameters. Once new observed data becomes available, the set of parameters most likely to describe the initial state of the forecast is determined, and these parameters are then used to run the model with the corresponding 51 historic sequences to calculate the DA-ESP. Could you clarify the following:

1. Is the period for applying the Particle Filter 4 years prior to the forecast launch? Is this understanding, correct?
2. Is there any specific error metric used to determine the new model parameters during the Data Assimilation process? If so, could you elaborate on which metrics are used and how they influence the selection of parameters?
3. If as supposed in the previous point, 624 forecasts were launched, it means 624 adjustments for initial conditions were made. Then, how the Data Assimilated (DA) 'Simulated Observed River Flows' time series is calculated?

In Table 2, Section 2.5, if Figure 1a corresponds to the results obtained in this study and includes the absPBIAS metric, I suggest adding absPBIAS to the list of performance metrics in the table.

In Table2, Section 2.5 the metric bias ratio $\beta = \frac{\mu_{\sqrt{q}}}{\mu_{\sqrt{Q}}}$ is presented. My understanding is that this ratio cannot be negative; it should range from 0 to +∞, with 1 being the optimum value. However, in Figures A4 and A5, it appears that negative bias values are presented. Could you clarify whether the metric in these figures refers to this bias ratio (β), percent bias (Pbias), or another bias metric? How do the negative values arise in these figures if they are indeed based on the bias ratio?

In Section 2.6, I would appreciate it if you could include the mathematical formulation for the CRPSS calculation, detailing both the Continuous Ranked Probability Score (CRPS) and the skill score. This would help in understanding the specific methodology used for evaluating forecast skill. Additionally, I would like more details regarding the sentence: *'The Ferro et al. (2008) ensemble size correction for CRPS was applied to account for differences between the number of members in the hindcasts (51 members, corresponding to the historic period from 1961-2015 with the L3OCV approach) and the benchmark (47 members, corresponding to the period of 1965-2015 with the L3OCV approach and four years removed for the spin-up period), as done in the evaluation of hydrological ensemble forecasting elsewhere (e.g., Crochemore et al., 2017).'* While I understand the concept of hindcasts, the benchmarks are being mentioned here for the first time. Furthermore, in Line 280, it is mentioned that the performance metrics from Section 2.5 were calculated for OR-ESP, BC-

ESP, and DC-ESP. However, these metrics do not appear in the results (as Section 3.1 seems to focus on 'Simulated Observed River Flows'). Similarly, Lines 282-284 reference additional metrics such as MAESS and MSESS, but these are also not shown in the results. Would it be possible to include these results in supplementary material? This would allow interested readers to explore these metrics further.

In Section 3.1, could you clarify whether the discussion is related to the 'Simulated observed river flows'? If it is, the process of obtaining the BC dataset is clear. However, I am unsure how the DA was calculated. From my understanding, 624 simulations were adjusted to fit the IHCs before each forecast launch, and with the 4-year spin-up period for initializing every forecast, wouldn't there be overlapping periods? Could you clarify how this was handled? If Section 3.1 is not related to the 'Simulated observed river flows' and instead pertains to OR-ESP, could you specify which forecast lead time was used? Additionally, was the mean ensemble used to calculate the performance metrics?

The following point is philosophical, and the authors are free to disagree or rebut, but I would appreciate your consideration. I suggest that Sections 3.2 and 3.3 could be rewritten to explain Figures 4 and 5 independently. While the results descriptions are clear, the current structure can be a bit confusing as it requires the reader to frequently refer back to the plots.

In Figure 6, I recommend placing the legend at the bottom to allow for a wider and taller plot, which would make it easier to see more details. Additionally, as mentioned earlier, could you clarify what the term 'initialization month' in Figure 6 refers to? Does it represent the month when the forecast was launched, the start of the forecast time series, or the specific month within the forecast time series data?

Lines 373-375 "*It is also interesting to note that there are cases where OR-ESP is better than both DA-ESP and BC-ESP (magenta points in Figure 7), especially in autumn, winter and beginning of spring (October to March) in the western part of the country for short lead-times (Figure 7a); and in spring for longer lead-times (Figure 7b) with no clear spatial pattern*." Could you provide any hypotheses or insights into why this situation might occur?

Personally, I loved the discussion and conclusions. I found them to be completely clear, precise, and pertinent

Thank you once again for the opportunity to review this manuscript. I hope the authors find my comments helpful, and I appreciate their understanding if any of my suggestions stem from a misunderstanding on my part