Response to the review #2 of the manuscript: AC (in classic black) and RC (in red))

"Testing data assimilation strategies to enhance short-range AI-based discharge forecasts" by Saint-Fleur at al.

Scope:

The manuscript is in the scope of the journal.

Summary:

The authors exploring two MLP-based data assimilation strategies to improve the discharge forecasts of two different hydrological models, an AI based LSTM and the mechanistic SAC-SMA model. The resulting forecasting setups are applied for three different lead times within two forecast scenarios, a deterministic optimal forecast and an ensemble forecast based on a climatological ensemble approach.

In my opinion the manuscript needs improvements in content and structure.

On behalf of all co-authors, we thank the reviewer for the time and effort devoted to reviewing this preprint including the insightful comments and questions provided.

General comments:

The manuscript focuses on DA strategies. But why do the authors choose exclusively MLP-based strategies? This choice is not sufficiently justified in the introduction. How would establish DA strategies such as Ensemble Kalman filters perform in comparison? Are there already studies that address this question? And what are the advantages of using MLP-based strategies over established DA strategies?

We thank the reviewer for these insightful questions. A comparison between the MLP-based strategies and the Ensemble Kalman Filter (EnKF) (Clark et al., 2008) would indeed be highly informative. However, implementing such a comparison in a consistent manner would require substantial additional methodological developments that go beyond the scope of the present study. The EnKF approach relies on perturbations of inputs and/or state variables, and potentially of model outputs, whereas the used DA strategies considered here are restricted to observed and simulated discharge only. Nevertheless, we consider this comprehensive comparison as promising direction for future work.

In this regard, it would also be necessary to examine the required amount of training data. Is 18 years really necessary for this? And are there any catchment-specific differences?

In light if this comment, including that of RC1, the training period is being reconsidered. Since the MLP assimilates the output of the benchmark models, only their evaluation period will be used for training. The analysis will mainly focus on the added value of the MLP as an orchestrator (or assimilator). A catchment-wise analysis is also a highly relevant perspective; however, this is seen for now as a promising investigation for future work.

The chosen training method also seems rather random. Why are 60, 8 and 10 seeds used? Would smaller numbers also be suitable? Why are the reference models treated differently?

The initial motivation for using such a large number of seeds was to obtain a sufficiently large ensemble for forecast analysis. However, preliminary results indicate that approximately 20 seeds are sufficient, and it will be reduced in the revised version of the manuscript.

In terms of structure: the authors use result figures in the introduction and conclusion, which in my opinion is not good style. On the one hand, they refer to content introduced later on, and on the other hand, they add new content when a summary is required. Figure 1 is unnecessary, as its content is repeated in Figure 7. Figure 14 shows examples of the analysed data and should be presented at the beginning of the results section to give an impression of the data. A similar figure should be presented for the SAC-SMA model at least in the appendix.

We thank the reviewer for this comment and will revise the structure of the manuscript accordingly.

Minor comments

Chapter 1 intro

- L38: I wouldn't argue that forecasting is "a fundamentally different task" since hydrologic discharge simulation models are used for forecasting but in a wider framework.

This sentence will be removed in the revised version, as the process does not differ between forecasting and simulation.

- L49: “model structures”: of what?

By “model structures”, we refer to imperfections related to the design and configuration of the model (e.g. architectural choices and parameterization) which may lead to structural model errors.

- L55-65: manuscript structure, see comment above

Chapter 2 Materials and methods

2.1 dataset

- style issue: repeating citations are not necessary

Thank you for this remark, we will remove the repeated citations accordingly.

2.2 data assimilation

- L99-104: to be placed and elaborated in the introduction

We thank the reviewer for this suggestion, and we will restructure the manuscript to improve clarity and coherence in the revised version.

- L129-L134: difficult to understand, “In both forecasting approaches --> this is not clear. You describe three DA approaches. But is MLP alone a DA approach or rather an alternative ML approach compared to LSTM?

In this paper, the MLP is presented as an orchestrator, it is rather a DA approach than a model to be directly compared with the LSTM. The term “both forecasting approaches” refers specifically to the “deterministic” and the “ensemble” forecasting strategies. We will clarify this point in the revised version of the manuscript.

2.3. forecasting setup

- L154: climatological ensemble --> unusual for daily data, provide discussion and reference

The “climatological ensembles” approach employed in this study was considered as an alternative to the optimal “deterministic forecast” investigated. We have deliberately tested a “poor man’s ensemble” by sampling all the meteorological variables on a date-

to-date basis from past observations. While this approach is unusual and may appear counterintuitive for daily rainfall data, it nonetheless represents a valuable alternative. Furthermore, it is conceptually similar to the Ensemble Streamflow Prediction (ESP) approach proposed by Day (1985), which assumes that any past observed event is equally probable in the future for a given date. The main differences between our approach and ESP, including the implementation by Crochemore et al. (2017), lies in the lead times range: ours are less than one week, whereas ESP typically considers longer horizons. While this approach deserves further investigations, we consider it a no-skill baseline ensemble that represents the opposite extreme from the perfect rainfall forecast assumption.

2.4. evaluation metric

Metric choice: You use CRPS. Would CRPSS also be an option for comparison with the reference?

We thank the reviewer for this question. The Continuous Ranked Probability Skill Score (CRPSS) is indeed a valid option for relative performance assessment, using the climatology as reference. In this work, we focused on the absolute quality of the forecasts and therefore used the CRPS. Nevertheless, we will consider the CRPSS in the revised version to complement this analysis.

- L186: Please provide reference for this statement.

This sentence will be revised and completed with references as follows: “An ensemble forecast is considered reliable (or statistically consistent) when the ensemble spread adequately reflects forecasts uncertainty, such that the observations are statistically indistinguishable from the ensemble members (Buizza et al., 2005; Hamill, 2001; Talagrand et al., 1997; Whitaker & Loughe, 1998).

- I am not familiar with rank diagrams. How are they created? Why does the number of ranks differ between Figures 4 and 10?

We use the rank histogram to assess the reliability of the ensemble forecast with respect to a target. In Fig.4, it reflects the distribution of the evaluation period (2006-2008), as target, compared to the remain period (1989-2006) for the rainfall and the PET observations. In Fig.10, it depicts the distribution of the forecasted discharge through the forecasted ensemble. Regarding the difference in number of ranks between the two figures, there is no particular methodological reason, except a technical adaptation

related to the size of the ensembles. In Fig.4, each individual year (out of 18) was retained as a member, so the number of ranks matched the ensemble size. For Fig.10, because the members resulted from a combination of N seeds and M years, and for clearer visualization, the ensemble was binned into 10 classes using quantiles.

- L197: You can also provide a formula for the spread skill ratio

We will include the related formula in the revised version of the manuscript. The spread-skill ratio (SSR) (Whitaker & Loughe, 1998) provides insights into the consistency between the ensemble uncertainty and the actual forecast error. Specifically, given a time step t (from a full evaluation period T), y the observed target value, x_i the forecast member i (from the ensemble N), and $\bar{x}$ the mean of the ensemble, the SSR can be formulated as:

$$SSR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}{\sigma_t}^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - y_t)^2}} \qquad \text{Spread: } \sigma_t = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{t,i} - \bar{x}_t)^2}$$

Best values are expected to be close to 1, and indicate well-calibrated ensemble. Values below (or above) 1 indicate under-dispersion (or over-dispersion) of the ensemble.

- L215: I am not sure what you mean by “ensemble rank”. Do you mean each ensemble member or a quantile of the ensemble?

The ranks (or classes) have been constructed using quantiles of the ensemble. Therefore, a more accurate label would be “ensemble quantile” or simply “rank”.

2.5. Experimental settings

- L230: typos: '.', RR-> RQ? In Figure 5 you are using Rainfall - Discharge cross-correlation.

Thank you for these remarks, the “.” is indeed a typographical error. Yes, the R-Q indicates Rainfall-Discharge cross-correlation. We used this information to approximate the input sequence length of the input features in the orchestrator.

- Figure 5: Please avoid overlapping colour areas, which will appear as a new colour. Perhaps just draw the lines.

Thank you for this suggestion. We will revise the figures, we will try with tiny lines highlight individual basins, which should avoid the overlapping issue.

- L239: Why 56 basins? Is the choice only motivated by NSE-values or also by different hydrological regimes?

The motivation behind this choice was primarily computational. Given the time and computation necessary to perform this approach, we did not find it necessary to run the ensemble climatology over all 531 basins. Regarding the selection approach, we used the NSE values, although other approaches such as hydrological regimes could also be considered.

Chapter 3 Results

- L340: Why should the observation be calibrated? Please provide reference.

By stating that the observation should be calibrated, we meant that the observations should be evenly distributed through the ensemble (Talagrand et al., 1997). However, this is more of an ideal expectation than a guaranteed outcome.

- fig 13: Are the low flows over- or underestimated?

As shown in Fig.10, there is no systematic bias for the low flows, except for the benchmark SACSMA (which exhibits overestimation) and to a lesser extent, the benchmark LSTM.

Chapter 4 Conclusions

- fig 14: see comment above

- fig 14: Why is there a spread on the left side for the deterministic optimal forecast? Is this due to the forecast initialization times and different lead times as a 'poor man's ensemble'??

We thank the reviewer for this insightful question. The spread observed on the left side for the deterministic optimal forecast originates from differences in model initializations (number of seeds) rather than from forcing uncertainty. While this setup shares similarities with the "poor man's ensemble", commonly used in meteorological

forecasting, the forcing data were assumed to be perfectly known and identical across all forecasts and lead times. Thus, the spread reflects solely the effect of the model initialization variability.

- L397:" numerous publications": Please provide examples.

Thank you for this suggestion, we will provide examples of references to support this statement, including Kratzert et al. (2018, 2019), Feng et al. (2020, 2024) and Yang et al. (2025)

References

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., & Wei, M. (2005). A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review*, *133*(5), 1076–1097. https://doi.org/https://doi.org/10.1175/MWR2905.1

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., & Uddstrom, M. J. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, *31*(10), 1309–1324. https://doi.org/10.1016/j.advwatres.2008.06.005

Crochemore, L., Ramos, M.-H., Pappenberger, F., & Perrin, C. (2017). Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrology and Earth System Sciences*, *21*(3), 1573–1591. https://doi.org/10.5194/hess-21-1573-2017

Day, G. N. (1985). Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., & Shen, C. (2024). Deep dive into hydrologic simulations at global scale: harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1.0-hydroDL). *Geoscientific Model Development*, *17*(18), 7181–7198. https://doi.org/10.5194/gmd-17-7181-2024

Feng, D., Fang, K., & Shen, C. (2020). Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/https://doi.org/10.1029/2019WR026793

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Talagrand, O., Vautard, R., & Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Workshop on Predictability, 20-22 October 1997*.

Whitaker, J. S., & Loughe, A. F. (1998). The Relationship between Ensemble Spread and Ensemble Mean Skill. *Monthly Weather Review*, *126*(12), 3292–3302. https://doi.org/https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2

Yang, Y., Pan, M., Feng, D., Xiao, M., Dixon, T., Hartman, R., Shen, C., Song, Y., Sengupta, A., Delle Monache, L., & Ralph, F. M. (2025). Improving streamflow simulation through machine learning-powered data integration and its potential for forecasting in the Western U.S. *Hydrology and Earth System Sciences*, *29*(20), 5453–5476. https://doi.org/10.5194/hess-29-5453-2025