Final replies to the reviewers for the preprint https://doi.org/10.5194/egusphere-2025-4244, by Saint Fleur et al. (2025)

General statement

We would like to thank Reviewers #1 and #2 for their thoughtful and constructive comments on this manuscript. Their feedback led us to include substantial additional analyses compared to the initial version, particularly through a reconsideration of the train/test split and the use of additional ensemble weather forecast products.

These two major revisions resulted in complementing the ensemble-based analysis using the BoM hindcast product available over the US territory for the test period (1989–1991). However, we were not fully convinced by this product which appear in the considered test period. Consequently, we decided to extend the analysis to another dataset—CAMELS-FR (Delaigue et al., 2025)—which enabled us to test the proposed approach using two recent ECMWF forecast products (hindcast and forecast archives) over a new test period 2018–2021. This extension is limited to DA1 and to a regional LSTM model developed and trained from scratch on the CAMELS-FR dataset. This additional work is presented in Section 4 of the revised manuscript. The findings from this extended work are consistent with those reported in the initial version of the manuscript and reinforce its conclusions.

Moreover, since the assimilation strategies were applied exclusively to discharge data, we revised the manuscript title by replacing “data assimilation” with “discharge assimilation.”

These revisions also raised several additional questions that fall beyond the scope of the present study. In particular, rank diagrams for all weather forecast products at lead times of 1–7 days reveal systematic biases in the ensemble meteorological forcings, especially for precipitation and potential evapotranspiration, whereas this issue is not observed for the no-skill climatological ensembles.

We hope that the concerns raised in the initial version have been adequately addressed through these substantial revisions.

In the following, we provide our detailed answers to the comments of both reviewers, along with a detailed description of the corresponding modifications.

Responses to the comments of reviewer #1

The authors indicate that they are benchmarking against Kratzert2019 and Newman2017, but they have different training / testing periods as the benchmarks. In section 2.1 you indicate that you are training from 1989-2006 and testing in 2006-2008. However, Kratzert used 1999-2008 for training and 1989-1999 for testing. This means you are training much more than them, and testing in only a 2-year period. Why this difference?
Moreover, you are also indicating that you re-simulated the 1989-2008 period with the pretrained models from Kratzert and Newman, but again, you are not respecting the training/testing split that the authors did in their original study.
Using benchmarks is an extremely valuable technique, because it automatically places your method in existing literature, however the conditions of the original studies need to be respected. One should, if possible, adapt the new experiment to the existing benchmark, otherwise it does not make any sense to do a benchmark.

We thank the reviewer for this very important comment. To ensure relevance and consistency of our analysis the train/test split has been rethought in depth. We implemented the discharge assimilation (DA) strategies exclusively over the 1989-1999 period, which corresponds to the test period of the benchmark models (LSTM and SCA-SMA). Within this framework, the orchestrator (MLP) was trained on the 1992-1999 period, while the evaluation was conducted exclusively on the 1989-1991 period. This setup ensures that the 1989-1991 period serves as a common evaluation window for all tested approaches, including the benchmark models.

Also, why are you using only two years of testing?

Our initial choice was to maximize the size of the training set, following common practice in AI-based studies, even at the expense of a shorter testing period. Under the revised configuration, a two-years evaluation period is a balanced choice, as it captures at least two wet and two low-flow seasons while preserving a sufficient amount of data to train the orchestrator. This setup is also consistent with the commonly used 80/20 train/test split in many machine learning frameworks. In a way, the reduced length of the test period is also counterbalanced by the large data set of test basins.

Lastly, you should benchmark your study against other studies that used data assimilation methods. In CAMELS US there is the study of Nearing (2022), Feng (2020) or more recently Yang (2025).

We thank the reviewer for bringing these studies to our attention. The approach developed in this study can be applied to any model that does not incorporate any discharge assimilation, as we mainly assess the added value of the DA strategies from an operational perspective. Studies such as Nearing et al (2022) may serve as a relevant benchmark, as they provide perform DA also.

While maintaining the focus on both primary choices, the manuscript now includes a brief comparison with Nearing et al. (2022) in terms of the performance improvement achieved through DA strategies. Overall, the magnitude of improvement is of the same order both works. The implementation results of the methods on the CAMELS-FR are also compared.

Comment 2.3

In section 2.3 you indicate that a specific model should be calibrated for each lead time, and that the alternative is inefficient. But this is not true. The LSTM has a linear layer at the end that transforms the hidden states to discharge.

Your model is a simple extension of this, but instead of a linear layer that go from hidden states to discharge, you have a feed forward neural network (so a couple of linear layers), in which, besides

the simulated discharge you concatenate some past meteorological variables and discharge. I do not have anything against the simplicity of the model, because if it is simple and it works then great, but there are multiple things that need to be considered.

You can run the LSTM as a seq-seq model, roll it over the forecast horizon (so if you are predicting 7 days forecast just do seq-7) and in each step concatenate the hidden states with whatever you want and pass it through the feed forward neural network. This way you have a consistent and generalizable model that does not require a different embedding for each case.

The alternative we had in mind is the recursive forecasting approach. The reviewer is correct, the seq-to-seq architecture is another alternative which also introduces additional challenges, particularly the risk of obtaining intermediate performance across lead times. Training may become more complex and prone to sub-optimal solutions for individual lead times. Nonetheless, as shown in Nearing (2022), seq-to-seq models remain an interesting option in our perspective. We did nevertheless not want in this revised version to modify to significantly the tested approaches. Comparison with seq-to-seq models or Kalman filters as suggested by reviewer 2 could be the topic of future works or publications.

Also why are you evaluating only on days 1, 3 and 7? Why not all seven days?

This choice was initially motivated by time saving reasons and by expectations that results at intermediate lead times would evolve monotonically. In the revised version, the full range of 1 to 7 days is included in the analysis. Nevertheless, for ease of visualization, and given the monotonic evolution of the scores, some results are shown for only 1-, 3- and 7 days lead times.

Furthermore, you are using climatological ensembles to create a possible forecast, but I do not believe this is the best strategy to do that. Climatology is used, normally, in medium to seasonal range (so couple of weeks to some months) where the forecast models of the meteorological variables are no longer reliable, and for variables that present a cyclic pattern (temperature, radiation…) but you are using it in precipitation 1 or 3 days ahead, which I do not believe is a good practice. How is the precipitation of the 1st of November for the last 18 years related to the 1st of November of this year? I do not think there is a strong relationship between these values that can be used for short-term forecasting, especially for precipitation, which is the most important variable to drive the forecast. If you want to use it for temperature or radiation, that can be an (non-ideal but defendable) option, but I would highly recommend to not use it for precipitation.

This comment is valid, and it was difficult not to address it, despite the significant additional work it required. The revised manuscript now includes evaluations based on BoM ensemble weather hindcasts. Given the poor quality of these ensembles over the US for the selected test period 1989-1991, we decided to extend the train and test work to a new dataset (CAMELS-FR) and two ECMWF ensembles (hincasts and forecast archives) for a test period 2018-2021 : section 4 of the manuscript. This clearly enriches the final version of the manuscript and the extrapolation of the methods to another dataset shows the robustness of the conclusions initially drawn. The results based on the climatological ensembles have been maintained in the manuscript as a no-skill benchmark for meteorological forecasts. This helped identifying some limitations of the available weather forecast ensemble.

In this section, you also indicate the absence of operational weather forecast archives, but this is also not fully correct. Shalev (2024) release a historical weather forecast for CARAVAN, which includes the CAMELS US dataset. It is true that not all the products are available in the testing periods that you have but: One option is to use CHIRPS-GEFS which is only precipitation but is available in your period of interest. Another, better option, is to benchmark a simple LSTM against Kratzert 2019, and once that is working well, you evaluate your model and the new LSTM in the

periods where you have a historical forecast (2016-2024) from Shalev. This is more work but would actually give you a robust study to evaluate your model under real forecast conditions.

We thank the reviewer for mentioning this study and the associated datasets. However, the use of the meteorological forecasts (and nowcast) products provided by Shalev and Kratzert (2024) in the present study poses several major challenges, some of which were correctly identified by the reviewer.

First, CHIRPS-GEFS provides only precipitation forecasts and covers the training but not the testing period that should be used in the revised version of the manuscript. Its integration would require generating realistic forecast scenarios for the remaining forcing variables, applying an appropriate bias-correction procedure to precipitation and rerunning the entire training/testing workflow process for all tested models. In addition, no overlap period exists between the meteorological forcing series used in the present study (Maurer extended version) and the historical forecasts products available on the 2016-2024 period. **Figure A** summarizes the availability of the variables between the products suggested by the reviewer and the datasets used in the present work.

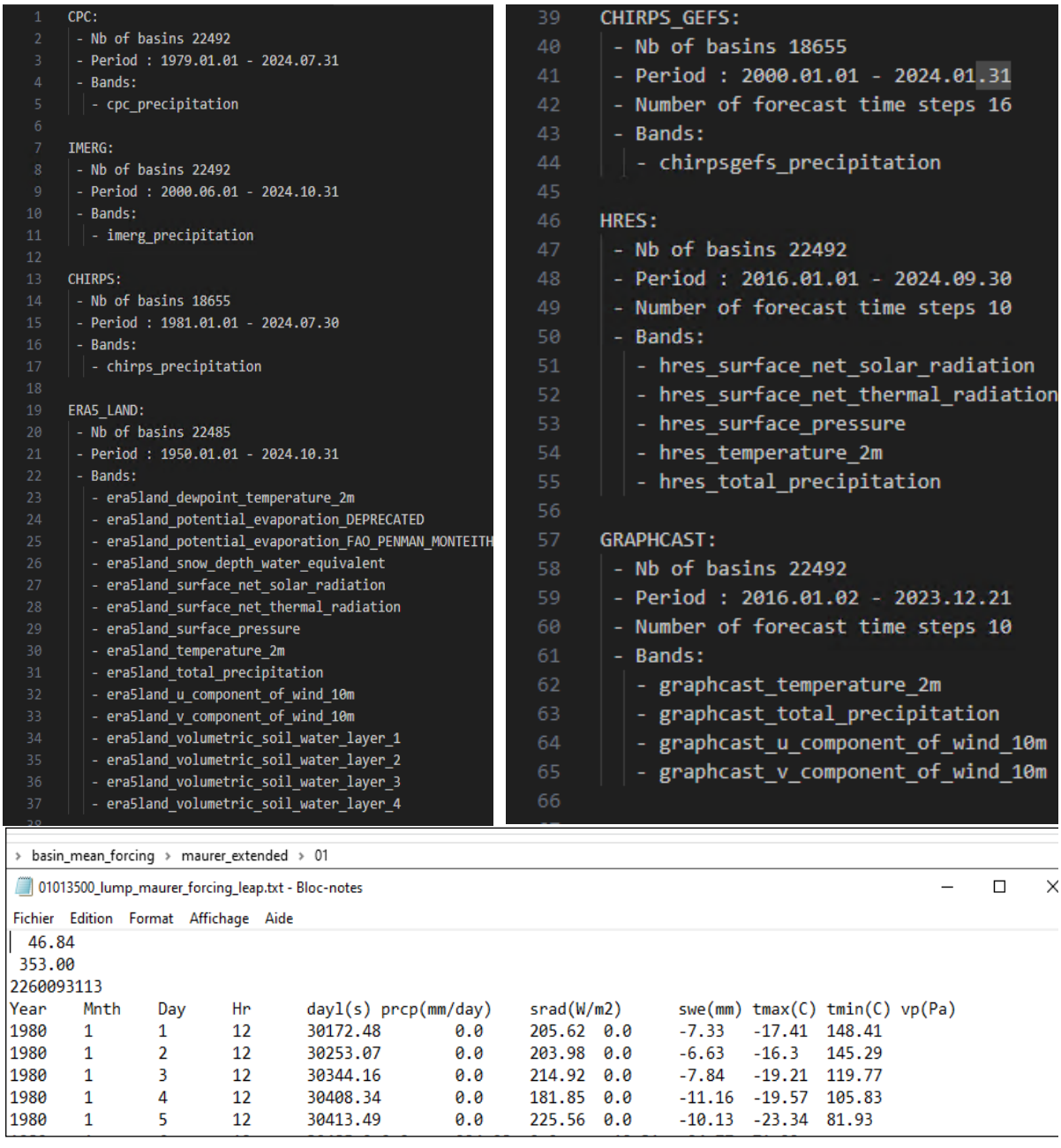

Figure A: Availability between the needed archives and those provided in Shalev and Kratzert (2024). Screenshot are shown for the nowcast products (top-left), forecast products (top-right) up to our covered lead times, and an example of the Maurer extended forcing variables (bottom) which stops in 2008.

Since the initial motivation for using the climatology-based data was to enable a probabilistic (ensemble-based) evaluation, the deterministic (single member) forecasts provided in Shalev (2024) are not suitable for this purpose. This limitation is compounded by the lack of temporal overlap between the datasets (2000-2024 vs 1989-1999).

This issue is ultimately addressed by using the BoM hindcast products (Hudson et al., 2020), which covers the 1989-1991 evaluation period. In addition, the extension to the CAMELS-FR dataset allows for the use of both hindcast and forecast archives from the ECWMF over the more recent 2018-2021 period.

Section 2.2 and 2.5.2

Here you indicate that you use an ensemble of 60 runs. Why 60? Most studies use between 5 and 10. Do you get a significant different with 60? If you want to use 60 that is your choice, but in section 2.5.2 you indicate that because of the 60 ensembles you have unreasonable computational cost, and that you are only going to use a subset of basins. If 5 to 10 ensembles give you the same as 60, then you can reduce the computational costs, and then do the study in the full region, which would produce more robust result.

Also, the 18 ensembles members can be accommodated in the batch dimension of the tensor and the different seeds can be run in parallel (even in a single GPU) so most of the computational overhead you are reporting can be overcome with some technical tweaking.

We thank the reviewer for this insightful remark. Our initial choice to use a large number of seeds was intended to provide a thorough representation of the model uncertainty while ensuring large ensemble for forecast analysis. However, this choice ultimately proved computationally demanding, particularly for the ensemble-based evaluation. Following the reviewer's recommendation, minor tests, as well as insights from related studies (e.g., Darras, 2014), the number of seeds has been reduced to 20 in the revised version. Nevertheless, we retain the same sample of 56 basins for the ensemble-based evaluation to limit the computational costs as it remains sufficiently representative of the full set of 531 basins.

Line 38: I do not agree that discharge simulation and discharge forecasting are fundamentally different tasks. You are trying to model the same system and the same rainfall-runoff response. In forecasting mode, you have the increased uncertainty of the meteorological input, however that is more of a limitation and not a fundamentally different task. Multiple operational models are calibrated with observed data in pseudo-forecast mode and later incorporated in forecasting pipelines, and they work well. Data-driven methods give the advantage that, if trained with real forecast, they can learn to compensate for systematic biases, but again, this is more of training strategies to compensate for data quality limitations and not because the task is fundamentally different.

We thank the reviewer for this remark. We have revised the paragraph accordingly by removing the segment suggesting that simulation and forecasting are fundamentally different tasks.

Line 66. The tittle of Fig 1 should be more self-explanatory.

Following the suggestion of Reviewer #1 & #2, we have removed this figure from the revised manuscript. The format of similar figures has been changed from ECDF to boxplot for improved readability. The figure legends have also been updated accordingly.

Line 127-128: What do you mean by: “The direct forecasts from the benchmark models were assumed to be unchanged for the tested lead time; therefore, no further running was necessary”.

This paragraph has been simplified. The key point is that the forecast discharge for a given date is the same for all lead times for the benchmark models under the perfect meteorological forecast assumption.

Figure 7. You should explain the colors also on the legend of the figure and not only on the text above. The figure plus the legend should be self-explanatory. Also, as a suggestion the message of this figure would be better explain by a boxplot per lead time graph. The boxplot would give the distribution along the basins and because you have one for each model and for each lead time it can be easily compared. Something similar to your Figure 8, but for the different lead times (you can also see Figure 3 from Nearing 2024).

Figure 7 has been replaced by Figure 8, and the ECDF plots have been replaced by boxplots which are much easier to read. The legend has also been adapted for easier reading: colors now denote the different approaches, with the benchmark model shown red, DA1 in green, DA2 in dark-orange, and DA3 in gold.

Section 3.2.2: Can you explain in detail how did you constructed this figure? How did you construct the 10 classes? What does a lower or a higher rank indicates?

The tested (DA) approaches differ in ensemble sizes, with ensemble sizes varying from 10 to 50, and seeds from 8 to 20. For the ease of comparison, all ensembles were reduced into 10 equiprobable classes for the rank diagrams.

Technically, the ranks of the observed discharge relative to the ensemble forecast were calculated using the *rank_histogram()* function from the **xskillscore** library, which is based on Hamill (2001). The so-obtained rank values were then grouped into 10 bins using the *groupby_bins()* method applied to the corresponding *xarray.DataArray* object, with the **bins** argument set to 10. Each bin was assigned a categorical label (1-10), and the frequency of occurrences in each category was computed over the evaluation period for each basin.

The lowest class represents the frequency of the observed discharge falling lower than the ensemble’s 10^{th} expected percentile, whereas the highest class corresponds to observations exceeding the 90^{th} percentile.

To assess the potential impact of differences in ensemble size across approaches, we constructed an illustrative example based on a synthetic ensemble drawn from a Gaussian distribution N(0,1). Using 1000 realizations and 30 independent resampling, we compared the 10-class rank histograms obtained with ensembles of 100 and 2000 members. As shown in **Figure B**, and as expected, both ensemble sizes produce uniform distributed rank diagram with 10 classes.

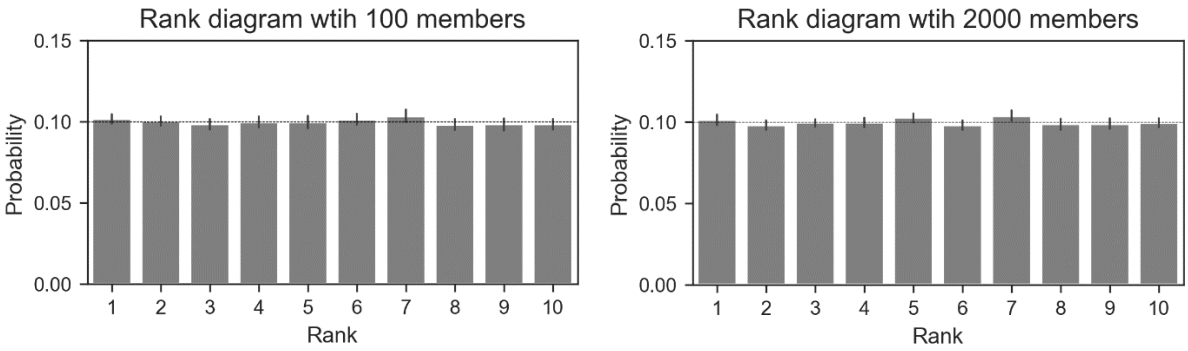

Figure B: Example of a 10-class rank diagram constructed from a Gaussian distribution (N (0,1)) generating an ensemble of 100 members (left) and 2000 members (right), with 1000 examples each. In each case, the last member (100^{th} or 2000^{th}) is used as the verification value. Error bars indicate

variability across 30 independent random realizations, while uniformity line is shown with the horizontal dotted black line.

Line 330-335: Can you explain in detail how are you evaluating the LSTM here to produce these results? Also, you are indicating that "This result suggests that the LSTM model is insufficiently responsive to recent meteorological inputs". Can it be that because the LSTM is driven only by meteorology, and because the climatological forecasts are not good (see my comment above about that) then the predictions are biased? The other models have the advantage of having discharge, which is a highly autoregressive variable, so they somehow compensate. However, if all you have is meteorological forecast and these are non-sense, how can the model perform well? I think this point is important and can biased the results you are presenting.

In this revised version, the analyses have been nuanced. A U-shaped rank diagram suggests, but is not necessarily related to under-dispersion. It can be the consequence of other types of biases in the forcast ensembles. This being said, the analyses of all ensemble-based forecasts reveal consistently across the two cases studies (U.S. and French basins) that biases in the ensemble forecasts produced by the LSTM benchmark (Fig. 11, formerly Fig. 10 and Fig. 18) may reduce its skill (CRPS, AUC, Brier). The conclusion that is drawn is that ensemble bias correction techniques may further improve the performance of LSTM forecasts in the future.

Regarding the evaluation procedure we are not sure to understand the question. For the ensemble evaluation, the LSTM model is run N*M times every day t, with input sequences corresponding to observed values for the past and to one member of the ensemble over the forecasting horizon *[t,t+hp]* (N being the number of member of the meteorlogical ensembles and M the number of seeds). This produces an ensemble of forecasts for each day of the period *[t,t+hp]* corresponding to forecasting lead times form 1 to hp. This is repeated for all days of the test period to build forecast ensembles for the lead times 1 to hp for all the days of the test period. The ensembles are than compared to the observed discharges at the same day. This approach is applied consistently across all DA strategies, as well as for the benchmark SAC-SMA model.

The introduction of new meteorological ensembles led to a complete revision of the presentation of the ensemble results and their analysis.

Responses to the comments of reviewer #2

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.

We thank the reviewer for the time and effort devoted to reviewing this preprint including the insightful comments and questions provided.

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 reviewer 1, the training period has been reconsidered in the revised version. Since the orchestrator (MLP) assimilates the output of the benchmark models, only their evaluation period is used in the DA framework for training, validation, and testing. The analysis then focuses mainly on assessing the added value of the DA strategies. Catchment-wise analysis also represents a highly relevant perspective; but may be complex due to the size of the data sets. Illustrations such as figure 9 show that at least results are consistent across catchments.

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 both forecast analysis and model uncertainty. However, in line with studies such as (Darras et al., 2014), as well as a preliminary analysis conducted in this work, approximately 20 seeds were found to be sufficient to obtain stable rank diagrams (i.e. provide ensembles reflecting modelling uncertainties in a reliable way) and are therefore used in the revised version.

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.

Following the reviewer's comment, Figure 1, previously included in the introduction, is removed from the revised manuscript. In addition, the results are now introduced using illustrative hydrographs rather than as a summary description.

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.

We thank the reviewer for this remark. We have revised the paragraph accordingly by removing the sentence suggesting that simulation and forecasting are fundamentally different tasks.

- L49: "model structures": of what?

By "model structures", we refer to imperfections related to the design and configuration of the orchestrator (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

We thank the reviewer for this remark, we have adapted these citations, limiting repetitions.

2.2 data assimilation

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

We thank the reviewer for this suggestion, we have moved and integrated this part into the introduction.

- 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?

We have clarified this point as follows: "In all considered DA strategies and for each basin, the MLPs were trained (i.e., calibrated) 20 times, accounting for the random initialization (seeds) of their parameter values, resulting in 20 distinct trained models." The MLP alone is indeed a DA approach, and is named DA1 in the revised version.

2.3. forecasting setup

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

In the revised version of the manuscript, this no-skill ensemble-based analysis has been complemented by several weather ensemble forecast products: a hindcast product from the Bureau of Meteorological (BoM) (Hudson et al., 2020) and forecast archives and hincasts from ECMWF platform. The climatological ensemble has also been kept as a no-skill ensemble baseline.

The “Climatology-based ensembles” approach employed in this study was considered as an alternative to the optimal “deterministic forecast” assumption. We had 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 unconventional and may appear counterintuitive, in particular for daily rainfall data, it nonetheless represents a usable 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 calendar date. The main differences between our approach and ESP, including the implementation by Crochemore et al. (2017), lies in the lead times range: our tested lead times are limited to one week, whereas ESP typically considers longer horizons. Although this approach deserves further investigations, we consider it here as a no-skill ensemble baseline.

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 relevant metric for relative performance assessment, particularly when using the climatology as reference. While this option was not fully feasible in the initial version, it is now included in the appendix of the revised manuscript to facilitate comparison between the climatology-based and the other forecast products.

- L186: Please provide reference for this statement.

This sentence has been 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 ensemble forecasts with respect to a given target. In Fig.4, the histogram represents the distribution of the evaluation period as the target, compared to the remaining period for both rainfall and PET observations. In the revised version, it is complemented by its equivalent for the hindcast-product. In Fig.10 (revised as Fig. 11), it illustrates the distribution of the forecasted discharge across the ensemble members.

Regarding the difference in number of ranks between the figures, there is no specific methodological reason; it mainly reflects a technical adaptation to the ensemble size for visualization purposes. In Fig.4 (left), each individual year (17 in total) could be treated as a distinct ensemble member; however, we have adopted a 10-bin representation to facilitate comparison with the hindcast product (Fig.4, right). In Fig.10 (revised as Fig. 11), the members result from a combination of *n* seeds and *m* members, making this 10-bin representation even more appropriate for readability.

Technically, the ranks of the observed discharge relative to the ensemble forecast were calculated using the *rank_histogram()* function from the **xskillscore** library, which is based on Hamill (2001). The so-obtained rank values were then grouped into 10 bins using the *groupby_bins()* method applied to the corresponding *xarray.DataArray* object, with the **bins** argument set to 10. Each bin was assigned a

categorical label (1-10), and the frequency of occurrences in each category was computed over the evaluation period for each basin.

The lowest class represents the frequency of the observed discharge falling lower than the ensemble's 10th expected percentile, whereas the highest class corresponds to observations exceeding the 90th percentile.

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

We have included the related formula in the revised version of the manuscript (Eq. 9). The spread-skill ratio (SSR) (Whitaker & Loughe, 1998) provides insights into the consistency between the ensemble uncertainty and the actual mean 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

$$SSR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T} {\sigma_t}^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - y_t)^2}}$$

Spread

$$where\ {\sigma^2}_t = \frac{1}{N-1}\sum_{i=1}^{N}(x_{t,i} - \bar{x}_t)^2$$

Best values are expected to be close to 1, and indicate well-calibrated ensemble. Values significantly 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?

A more accurate label would be "ensemble quantile" or simply "rank bins". Given the way these bins are constructed, we have re-labelled the x-axis to "rank bins" in the revised version.

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. Indeed, 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.

We thank the reviewer for this suggestion. The figures have been revised accordingly: overlapping colors have been removed, individual basins are now represented by tiny lines, and separated figures are used for the cross- and auto-correlograms.

- 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 (revised as Fig. 11 and Fig 18), there are no systematic bias for the flows. In Fig. 13 (revised as Fig. 14 and Appendix B.9), there is no particular over- or under-estimation of the low flows.

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. Fig.14 from the initial version has been revised and is now presented as Fig. 7 (with similar figures provided in Fig.16, Appendix B1., B4 and B5) in the revised manuscript.

The spread observed on the left-most side for the deterministic optimal forecast originates from differences in model initializations (i.e., the number of seeds), rather than from forcing uncertainty. While this setup shares some 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 only the variability induced by model initialization (i.e. modelling uncertainties). This spread is less pronounced in the revised manuscript, as the number of seeds has been reduced from 60 to 20.

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

Thank you for this suggestion, we have provided examples of references to support this statement in the revised version, including Kratzert et al. (2018, 2019), Feng et al. (2020, 2024) and Yang et al. (2025)

Cited 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

Darras, T., Johannet, A., Vayssade, B., Kong A Siou, L., & Pistre, S. (2014, March). *Influence of the Initialization of Multilayer Perceptron for Flash Floods Forecasting: How Designing a Robust Model*

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)

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., & Andréassian, V. (2025). CAMELS-FR dataset: a large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking. *Earth System Science Data*, *17*(4), 1461–1479. https://doi.org/10.5194/essd-17-1461-2025

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

Hudson, D., Alves, O., Hendon, H. H., Lim, E.-P., Liu, G., Luo, J.-J., MacLachlan, C., Marshall, A. G., Shi, L., Wang, G., Wedd, R., Young, G., Zhao, M., & Zhou, X. (2020). Corrigendum to: ACCESS-S1: The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1), 393. https://doi.org/10.1071/ES17009_CO

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

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., & Nevo, S. (2022). Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. Hydrology and Earth System Sciences, 26(21), 5493–5513. https://doi.org/10.5194/hess-26-5493-2022

Nearing, G., Cohen, D., Dube, V. et al. Global prediction of extreme floods in ungauged watersheds. Nature 627, 559–563 (2024). https://doi.org/10.1038/s41586-024-07145-1

Shalev, G., & Kratzert, F. (2024). Caravan Multi Met: Extending Caravan with multiple weather nowcasts and forecasts. arXiv preprint arXiv:2411.09459. https://arxiv.org/abs/2411.09459

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