Response to RC1: 'Comment on egusphere-2025-4978', Anonymous Referee #1

We thank the anonymous referee for the careful reading of the manuscript and the very detailled and interesting comments. Your feedback helps identifying areas where the presentation and analysis could be strengthened, and we believe the paper is significantly improved as a result. Each comment is addressed in turn below, with the referee comments in blue.

This manuscript addresses the challenge of deploying machine-learning hydrological models in operational forecasting by explicitly considering domain shift between reanalysis and forecast meteorological inputs. The authors explore alternative training strategies and LSTM architectures to improve 1-day streamflow forecasts, and the results suggest that architectures combining hindcast and forecast phases, which use reanalysis and forecast data respectively, provide the greatest performance gains. The study tackles an important problem, presents interesting results, and is structured well. Some additional analysis and clarifications would further strengthen the interpretation of the experiments and results.

Thank you for your summary.

General comments

1. The paper would be greatly improved by further analysis on the variability of the skill across basins (please see comments 15 and 24 below). The manuscript emphasizes the impact of domain shift between reanalysis and forecast meteorological datasets; however, this shift is not spatially uniform across basins or meteorological variables. Further analysis of how variability in domain shift relates to basin-to-basin differences in model performance would strengthen the conclusions. In addition, the role of catchment attributes in modulating model skill is not discussed. Given the focus on a 1-day lead time, differences in hydrological response will strongly influence streamflow predictability and should be considered.

We agree with the referee that a deeper investigation of basin-to-basin variability in model skill and its relationship to catchment attributes would strengthen the manuscript. As detailed in our response to comment 13, we add a new Section 3.1.1 "Linking meteorological domain shift to topographical features" to the revised manuscript, and introduce a corresponding methods section 2.3 "Quantifying Domain Shift and Its Physiographic Controls" describing the analytical framework. In Section 3.1.1, we compute the 1-Wasserstein distance between the reanalysis and forecast distributions for each of the five shared meteorological variables from the CrossDomain Evaluation experiment across all 451 basins, and correlate these distributional differences with static catchment attributes to identify which basin characteristics are associated with larger meteorological domain shifts. Additionally, we add a new Section 3.6 "Physiographic Controls on Model Skill", in which we

compute the Pearson correlation between per-basin NSE and all 33 static catchment attributes across all experimental configurations, visualized as a heatmap. We note that while the referee raises the important question of how domain shift variability propagates into basin-level performance degradation, a full causal investigation of this link goes beyond the scope of this study, which focuses on methods to reduce forecast-induced bias. The referee's comments have however been a strong motivation for a dedicated follow-up study, in which we plan to systematically investigate the temporal and physiographic controls on forecast-induced bias.

2. The paper is largely framed in terms of mitigating biases in meteorological forecasts. However, the inclusion of near real-time streamflow observations (and to some extent the use of model architectures with hindcast-forecast phases) are likely reducing uncertainty associated with initial conditions and/or model structure. A slight adjustment to the framing, or an explicit discussion of the different uncertainty sources being addressed, would improve clarity (see comment 21).

We agree with the referee that the proposed strategies address uncertainty sources beyond meteorological forecast bias alone. Most notably, the integration of near real-time discharge observations in the hindcast phase reduces uncertainty associated with the initial hydrological state, which itself contributes to improved forecast skill independently of meteorological bias correction. The framing as bias correction is nevertheless deliberate: meteorological forcing has been identified as a major source of uncertainty in hydrological modelling, with Bárdossy et al. (2022) demonstrating that precipitation uncertainty alone can be responsible for up to 50% of hydrological model output uncertainty. Our cross-domain evaluation further establishes that replacing reanalysis with forecast inputs alone causes substantial performance degradation, directly motivating this framing. We add a brief discussion of the role of initial condition uncertainty in the revised manuscript, acknowledging that the performance gains from discharge integration reflect a combination of meteorological bias correction and improved state initialization.

Specific comments

1. The introduction is well written and covers may aspects of the topic, but some of the motivations are incomplete.

1.1. Lines 51-54: As the authors note, the focus on ECMWF-IFS data implies that the transferability of the method proposed by Han et al. (2021) to other NWP systems still needs to be tested. However, this limitation does not in itself imply that more robust bias-correction approaches are required, as suggested by the authors (“highlighting the need for more robust approaches”). The wording could be revised to reflect this distinction more clearly.

We thank the referee for this comment and agree that the original wording overstates the implication. The focus on ECMWF-IFS data raises questions about transferability to other NWP systems, but this does not necessarily imply that more robust bias correction approaches are needed in general. We revise the sentence to: „However, the focus on ECMWF-IFS data raises important questions about the correction model's transferability to other numerical weather prediction systems, potentially limiting the generalizability of their bias correction framework to broader operational contexts.“

1.2. Line 64: It is my understanding that errors in river discharge measurements are often localised (e.g., by vegetation growth or changes in the river channel, hydraulic disturbances near the gauge, etc)? Potentially the authors mean that river discharge observations are impacted by less spatial representativeness errors? Please clarify.

We agree that the original phrasing is imprecise, as discharge measurement errors can indeed be localized. Our intended point is that discharge observations are less affected by spatial representativeness errors compared to precipitation, since discharge represents an integrated hydrological response over the entire catchment area and is measured continuously at a fixed gauging station, rather than requiring spatial interpolation from a network of point measurements. We revise line 64 accordingly.

1.3. Lines 72-74: The GloFAS forecasts are available across the globe so the comment “the quantile mapping correction is dependent on GloFAS forecasts, meaning it is not applicable for regions where no GloFAS forecast is available” may be misleading and should be clarified. The Hunt et al., (2022) method can also be applied to other distributed hydrological forecasts so is not dependent on GloFAS.

We thank the referee for this clarification. We agree that the Hunt et al. (2022) method is not strictly limited to GloFAS and could in principle be applied to other distributed hydrological forecasting systems. We revise the statement accordingly. However, we maintain that the method requires a hydrological forecast at the specific location of interest, which may not be available for all catchments, particularly smaller or poorly monitored ones. We soften the original wording to reflect this more nuanced limitation.

2. Meteorological datasets:

2.1. Section 2.1: It is not clear to me which meteorological variables and datasets are used. Are the E-OBS, MSWEP, and GLEAM datasets used? If so, in which experiments, and if not, please remove them from the discussion.

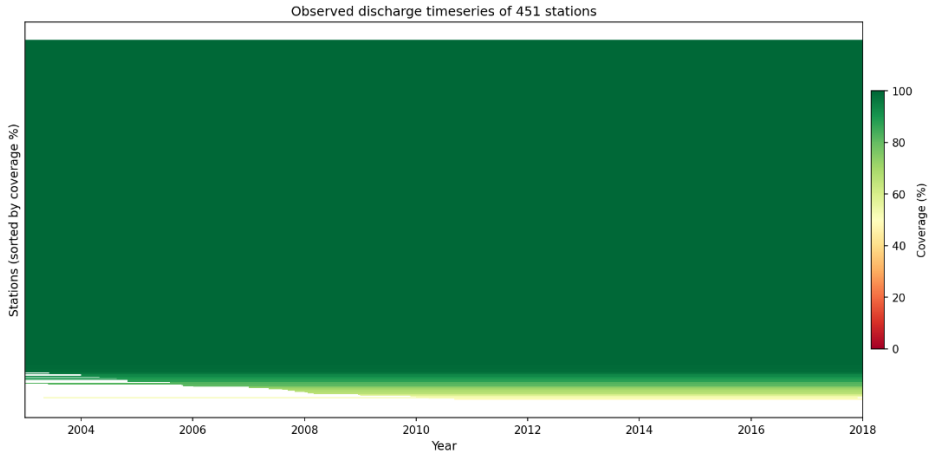
All data products mentioned in Section 2.1 (ERA5-Land, ECMWF-HRES, E-OBS, MSWEP, and GLEAM) are used as dynamic meteorological forcings in the experiments, following the approach of Kratzert et al. (2021), who demonstrated that combining multiple meteorological data products in a single LSTM significantly improves simulation accuracy compared to using individual products alone. The full set of variables from all data sources is listed in Table A1.

2.2. Appendix A: Are all the variables in appendix A used? If so, I suggest adding a column to indicate which experiment they are used in, and if not, I suggest indicating this

All variables listed in Table A1 are used in the experiments. For all experiments except the Cross Domain evaluation, the forecast driven Transfer Learning and the purely forecast driven model, all reanalysis variables from ERA5-Land, E-OBS, MSWEP, and GLEAM are used in combination as dynamic inputs, following Kratzert et al. (2021). The Cross Domain experiment is the only exception, where only the five ERA5-Land variables that have a direct equivalent in the ECMWF-HRES forecast data are used, in order to keep the input feature space identical between training and inference. We add a note to Table A1 to clarify this.

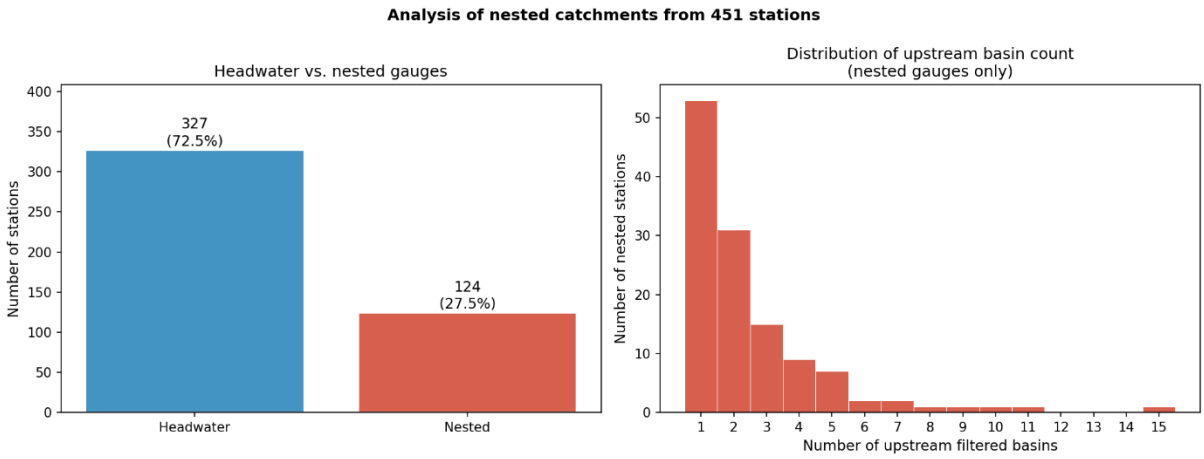
3. Section 2.1: What is the distribution of observed timeseries available at each station? Do they all cover the full experimental period from 2003-2017:

Not all discharge observations cover the full experimental period from 2003 to 2017. The temporal coverage of all 451 stations is shown in the figure below. Basins with insufficient discharge data in a given period are automatically excluded from training by the NeuralHydrology framework, which silently drops any basin for which no valid target samples can be constructed. Training proceeds normally on the remaining basins.



4. Line 122: Are nested catchments included within the 451 basins?

Yes, nested catchments are included in the 451 basins. 72.5% of the basins are headwater catchments, while 27.5% are nested, meaning they have at least one other gauged basin from the selection upstream. Nesting was determined by walking the river network downstream via the NEXTDOWNID attribute in the catchment attributes of the LamaH-CE dataset (Klingler et al., 2021). If any other basin from the filtered set was encountered along the downstream path, the downstream basin was classified as nested. A figure illustrating the distribution of headwater versus nested catchments, as well as the number of upstream basins per nested gauge is shown hereafter. We will add this information as text to the Data section.



5. Figure 2: It would be very helpful if this figure (or an additional figure) included a diagrammatic comparison of the model architectures, and, if possible, of the experiments performed.

Figure 2 provides an overview of the experimental setup in the "Experimental Split" block, but we agree that a more explicit architectural diagram would aid the reader. We add a diagram of the Sequential Forecast LSTM architecture to the appendix and refer the reader to Nearing et al. (2024) for a detailed illustration of the Encoder-Decoder LSTM architecture.

6. Line 141-142: For clarity, do the temporal splits contain the lookback window of 365-days such that the validation and testing is done over 2-years' of forecasts (2011-2013 and 2015-2017, respectively) rather than 3 years?

We thank the referee for this question, which prompted us to look more carefully into the NeuralHydrology source code. The temporal splits do not absorb the 365-day lookback window. The NeuralHydrology framework explicitly loads a warmup period prior to each split's start date, drawn from the preceding period, so that the model state is properly initialized before the first evaluation timestep. Predictions and evaluation therefore span the full defined periods: four years of validation (2010–2013) and four years of testing (2014–2017). This approach is consistent with earlier studies using the same framework, such as Gauch et al. (2021), Gauch et al. (2025) and Acuña Espinoza et al. (2025), who applied the same warmup strategy, with the only difference that their periods started in October rather than January, but equally with only one day difference between the end of one period and the start of the next.

7. Line 160: "exclusively driven by reanalysis (e.g., ERA5) or spatially interpolated observational (e.g., E-OBS) data sources". Does the RA baseline use ERA5 or E-OBS?

The RA baseline uses all available reanalysis and observational data products simultaneously (ERA5-Land, E-OBS, MSWEP, and GLEAM) following the approach of Kratzert et al. (2021), who demonstrated that combining multiple meteorological data sources improves simulation performance. We revise line 160 to make this explicit and avoid any ambiguity introduced by the use of "e.g." in the original text.

8. Section 2.2.1/2.2.2: Which LSTM architecture is used for the baselines and the cross-domain experiments? Are they all multi-basin models (i.e., trained on all data)? Please clarify and include in the experimental figure (see comment 3).

As stated in Section 2.2.1, all baseline experiments use a standard LSTM runoff simulation framework (CUDA LSTM in NeuralHydrology), with the core tensor equations presented in Appendix E and a detailed description of the LSTM in the context of hydrological modelling available in Kratzert et al. (2018). All experiments, including the baselines and cross-domain evaluation, are trained in a multi-basin setting across all 451 catchments simultaneously, following the regional training approach of Kratzert et al. (2024). Single-basin training was not considered, as multi-basin training has been shown to yield substantially better performance by leveraging information across catchments.

9. Line 230: I appreciate that the authors have kept the manuscript concise; however, since the premise of the paper is that differences between the reanalysis and forecast domains lead to errors, I feel that more discussion of these differences is needed. This could be addressed in a short additional section, or at minimum by adding text to Appendix B to guide readers on the key differences between the two domains.

We agree that a more explicit discussion of the differences between the reanalysis and forecast domains would strengthen the motivation of the paper. We add descriptive text to Appendix B to guide the reader through the key distributional differences between the data products.

10. Line 243: Are the weights for the embedding of both the dynamic and static features updated or just the dynamic?

Only the dynamic embedding network weights are updated during the selective transfer learning, as the static catchment attributes remain unchanged across all experimental setups. The static embedding network weights are therefore kept frozen. We clarify this in the revised text.

11. Section 3: It would be useful to have a table that compares the statistics of each experiment. For example, in some sections the standard deviation is mentioned and in others it's not. The interquartile range would also be an interesting statistic as the standard deviation is influenced a lot by the very worst performing stations.

We thank the referee for this suggestion. We add a table at the beginning of Section 3, reporting the median, mean, standard deviation, and interquartile range of NSE values for all experimental configurations, allowing for a more consistent and complete comparison across setups.

12. Figure 3: I'm struggling to understand this figure. From the experiment description I would expect there to be three lines: reanalysis baseline, forecast baseline, and the Cross domain where the model used for the reanalysis baseline is driven by forecast data in the test period. What is the difference between the "Cross Domain (Reanalysis,

Pre train)" and the "Baseline Reanalysis"? It would be nice to see a discussion on the cause between the difference between the two, particularly as it seems to have a big impact on the worst performing basins. It would also be useful if the caption labels corresponded to the experiment description more directly.

The Baseline Reanalysis and CrossDomain (Reanalysis, Pretrain) differ in their input feature space: the Baseline Reanalysis ($\mathcal{D}_{RA,Baseline}$) uses all 31 available reanalysis variables from ERA5-Land, E-OBS, MSWEP, and GLEAM, while the CrossDomain (Reanalysis, Pretrain; we call it here now $\mathcal{D}_{RA,CrossDomain}$) is restricted to the five variables (ERA5L_2m_temp_mean, ERA5L_2m_dp_temp_mean, ERA5L_surf_net_solar_rad_mean, MSWEP_RR,GLEAM_ETA) that have a direct equivalent in the ECMWF-HRES forecast data. The model pretrained on $\mathcal{D}_{RA,CrossDomain}$was then used for inference with the corresponding five ECMWF-HRES forecast variables from $\mathcal{D}_{FC}$. So, formally:

$\mathcal{D}_{RA,Baseline} \neq \mathcal{D}_{RA,CrossDomain}$

The key message of the cross-domain experiment is precisely this: training a model on reanalysis data and directly applying it with forecast inputs from the same variables leads to substantial performance degradation, and should be avoided in operational settings. The performance gap between the two reanalysis driven experiments in Figure 3 therefore proves the benefit of combining multiple meteorological data sources, consistent with the findings of Kratzert et al. (2021). Formally:

$$\mathcal{D}_{RA,Baseline} = \{(x_t^{RA}, q_{max,t})\}_{t=1}^{T} \; with \; x_t^{RA} = x_t^{RA,Baseline} = \sum_{i=1}^{n=31} x_i$$

$$\mathcal{D}_{RA,CrossDomain} = \{(x_t^{RA}, q_{max,t})\}_{t=1}^{T} \; with \; x_t^{RA} = x_t^{RA,CrossDomain} = \sum_{i=1}^{n=5} x_i$$

such that $\mathcal{D}_{RA,Baseline} - \mathcal{D}_{RA,CrossDomain} = \Delta NSE$, since the median NSE increases with an increasing number of meteorological reanalysis inputs.

We will include a clear formulation such as here above in the description of the experiments.

13. Section 3.1: It would strengthen the manuscript to explore the relationship between differences in meteorological reanalysis and forecast datasets and the resulting streamflow predictions. For example, do locations with the largest input differences show the greatest performance degradation in the cross-domain experiments?

A detailed investigation of the causal link between meteorological input bias and catchment-scale performance degradation goes beyond the scope of this study, which focuses on methods to reduce forecast-induced bias rather than explaining its origins. We agree with the referee that this is a highly interesting research direction, and the explanation of forecast-induced bias patterns including their spatial controls and

catchment-scale drivers is the dedicated focus of a planned follow-up study. As a first step in this direction, we add a new Section 3.1.1 "Linking meteorological domain shift to topographical features" to the revised manuscript. For each of the 451 basins, we compute the 1-Wasserstein distance between the reanalysis and forecast distributions for each of the five shared meteorological variables (from the CrossDomain Experiment), and correlate these input distributional differences with static catchment attributes to identify which basin characteristics are associated with larger meteorological domain shifts. The Wasserstein distance was chosen over simpler error measures such as MSE or PBIAS because it captures differences in the full shape of the distribution, rather than penalizing differences point-wise, making it particularly suited for characterizing domain shift in meteorological variables where extreme value differences are of special hydrological relevance.

14. Figure 4:
    14.1. Are the arrows showing the median change in NSE or the change in median NSE (the caption suggests the former, but the arrows suggest the latter)? Please also check other figure captions for consistency.

    The arrows show the change in median NSE, not the median change in NSE. We correct the figure caption accordingly and check all other figure captions for consistency.

    14.2. Also please check the 0.3 value associated with the orange line (Sequential Forecast LSTM) as it doesn't seem consistent with the other values.

    The 0.3 value associated with the Sequential Forecast LSTM is a typo introduced when inserting the arrows. The median NSE value of this architecture is 0.63, which corresponds to a change in median NSE of 0.24 from the forecast baseline (0.33). We correct this in the revised figure.

15. Line 312: Interestingly, the difference between the two TL methods disappears at the 90th percentile. Can the authors explain this behaviour?

The convergence of the two transfer learning strategies at the 90th percentile is not fully understood, but we hypothesize that it may reflect an upper performance ceiling imposed by the quality of the forecast data itself. Regardless of the transfer learning strategy applied, the inherent uncertainty and bias in the ECMWF-HRES forecast inputs may limit the achievable skill for all models equally, such that neither strategy can outperform the other beyond a certain threshold.

16. Line 329: How many and what size layers does the handover network have, and are they tuned?

The handoff network consists of a single fully connected layer with 128 hidden units, consistent with the original implementation of Nearing et al. (2024), and was not subject to hyperparameter tuning. Its architecture was kept fixed to ensure comparability with their results and to reduce the complexity of the hyperparameter search space, as the

primary focus of this study was on evaluating and comparing different LSTM-based training strategies rather than optimizing individual architectural components.

17. Lines 335-336: I really like these summary sentences (e.g. “In short…”) at the start of each results section - very clear and helpful!

Thank you!

18. Line 346: If pretrained with both reanalysis and forecast data as input, what is the model fine-tuned on? Is it the loss function that is changed rather than the inputs?

The loss function remains unchanged across all experiments, always using NSE* as the optimization objective. In this experiment, the model is pre-trained on the combined forecast and reanalysis domain $\mathcal{D}_{FCRA}$, which includes all reanalysis variables alongside the five ECMWF-HRES forecast variables as additional inputs. The embedding network is then fine-tuned on the forecast-only domain $\mathcal{D}_{FC}$, using only the five ECMWF-HRES variables, while all LSTM weights remain frozen. The intuition behind this approach was that pre-training on both data sources simultaneously might allow the model to learn representations that bridge the two domains, potentially facilitating a more effective adaptation of the embedding network to forecast-only inputs during fine-tuning. However, as shown in Section 3.3, this strategy did not yield consistent improvements over pre-training on reanalysis data alone. As the experimental setup is clearly described in Section 2.2.5 and the results are thoroughly discussed in Section 3.3, we believe the manuscript already provides sufficient clarity on this point and do not see a need for further changes.

19. Section 3.2/Section 3.4: The study focuses on training strategies and LSTM architectures to mitigate forecast-induced biases. However, the inclusion of near real-time streamflow observations may also be compensating for other uncertainties. A brief discussion of this effect would be helpful. Additionally, please justify the use of forecast- and reanalysis-based baselines as upper and lower benchmarks within this section (Seibert et al., 2018).

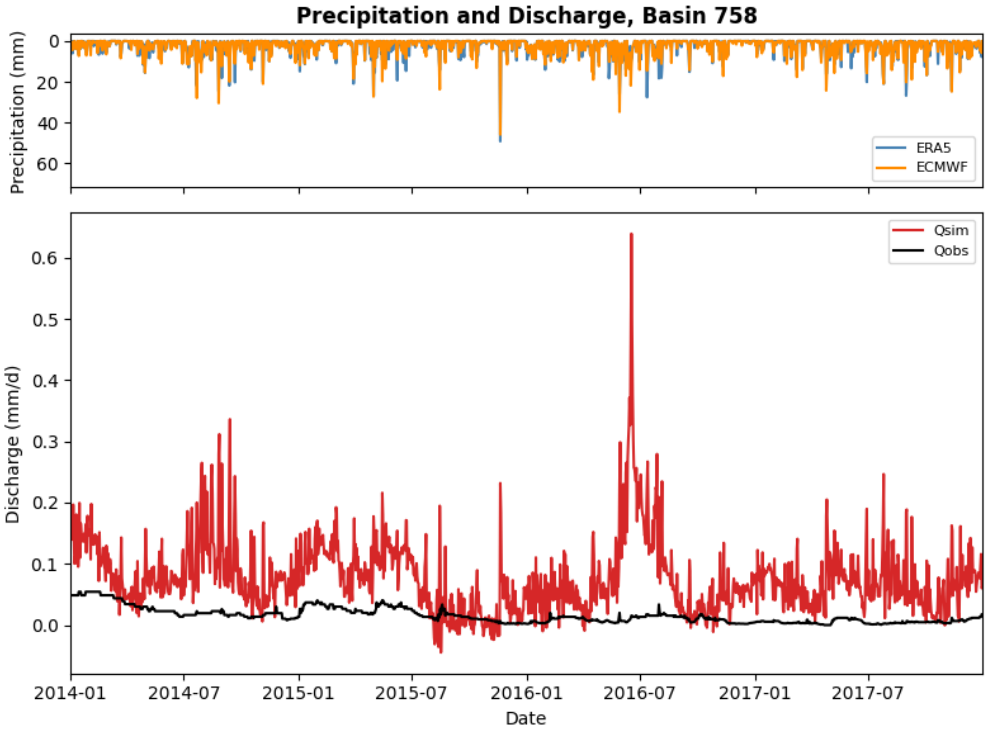
We agree with the referee's observation. Integrating near real-time discharge observations in the hindcast phase reduces uncertainties beyond forecast-induced bias alone - most notably by providing a more accurate representation of the initial hydrological state, from which the forecast phase is initialized. However, an improved initial state directly translates into better forecast skill, as the model enters the forecast phase with a more accurate representation of catchment conditions. The performance gains from discharge integration therefore serve the main objective of this paper: reducing forecast-induced bias in runoff predictions. We add a brief discussion of this in the revised manuscript.

Regarding the use of upper and lower benchmarks, we follow the framework of Seibert et al. (2018), who argue that model performance can only be meaningfully interpreted when

related to benchmarks representing what could and should be expected. The forecast-only baseline ($\mathcal{D}_{FC}$) serves as the lower benchmark, representing the minimum performance achievable when a model is trained and tested exclusively on weather forecast data, while the reanalysis-only baseline ($\mathcal{D}_{RA}$) serves as the upper benchmark, representing the maximum performance achievable under ideal hindcast conditions, consistent with the strong LSTM simulation skill demonstrated in reanalysis-driven settings by Kratzert et al. (2018, 2019a, 2021). All experimental configurations are evaluated relative to these two bounds. We add an explicit reference to Seibert et al. (2018) the revised text.

20. Line 358: Where are the outliers? What is causing these outliers? For example, are they at high elevations? Do they have short observation records?

A systematic explanation for the outliers has not been identified, but individual inspection of specific basins suggests that residual anthropogenic influences may play a role. For example, the overall worst performing Basin 758 shows a persistent and strong overestimation of simulated discharge relative to observations throughout the test period (see figure afterwards), which is consistent with the presence of water abstractions, retention structures, or other human interventions not represented in the model. Despite filtering the 451 basins for low anthropogenic influence based on the LamaH-CE catchment classification, some residual human influence cannot be excluded.



21. Lines 380-381: "The Sequential Forecast LSTM exhibits even more impressive improvements when discharge is integrated, achieving a median NSE of 0.71 compared to 0.63 without discharge." may be misleading because the improvement

of the median NSE value is smaller than for the Encoder-decoder model despite it being for higher NSE values.

We agree that the original phrasing is misleading. The absolute improvement in median NSE when integrating discharge is marginally smaller for the Sequential Forecast LSTM (+0.08, from 0.63 to 0.71) than for the Encoder-Decoder LSTM (+0.09, from 0.57 to 0.66). What is noteworthy about the Sequential Forecast LSTM is not the magnitude of the improvement but rather the higher absolute performance level reached, which surpasses even the reanalysis baseline (0.69), demonstrating that discharge integration enables the model to exceed the upper benchmark defined by ideal hindcast conditions. We revise lines 380–381 to reflect this distinction more accurately.

22. Section 3.4: It would be useful to explore the relationship between catchment attributes and model performance, particularly since persistence methods can sometimes achieve high skill at a 1-day lead time. This would strengthen the results of the manuscript.

The influence of static catchment attributes on LSTM-based runoff predictions has been extensively studied by Kratzert et al. (2019a, 2021), and we refer the reader to these studies for a detailed treatment of how catchment attributes shape model skill in a reanalysis-driven setting. Building on this foundation, we add a new Section 3.6 "Physiographic Controls on Model Skill", in which we compute the Pearson correlation between per-basin NSE and all 33 static catchment attributes across all experimental configurations, visualized as a heatmap with experiments sorted by median NSE. This provides a direct and comprehensive view of how catchment characteristics modulate model skill across all tested architectures and training strategies, extending the perspective of Kratzert et al. to the forecast domain investigated in this study.

Regarding persistence methods, we argue that they are most applicable as benchmarks for mean daily discharge predictions, where high temporal autocorrelation makes today's observed discharge a reasonable predictor of tomorrow's value. In this study, the target variable is daily maximum discharge based on hourly data, which exhibits substantially lower day-to-day autocorrelation, as peak flows are predominantly driven by the timing and intensity of individual precipitation events. Persistence-based skill is therefore expected to be considerably lower for daily maximum discharge, making it a less meaningful reference in our setting.

23. Section 3.5: Please check throughout that the analysis of the graph is correct. For example, “In the baseline, encoder-decoder and sequential LSTM experiments, simple linear embedding networks produced slightly higher simulation performance compared to more complex embedding architectures.” (lines 409-411) but the dashed line (simple) is primarily to the left of the solid line (complex). I think there may also be an editing error in this section as the last paragraph seems to contradict the paragraph before.

We thank the referee for catching this error. The legend in Figure 7 is incorrect — the dashed line represents the complex embedding network and the solid line represents the simple embedding network, which is the opposite of what the legend states. The text in Section 3.5 is correct and consistent with the actual figure, but the figure legend and caption contain a typo that created the apparent contradiction. We correct the figure legend and caption in the revised manuscript.

24. Section 3.6: Some relevant topics could be discussed further. For example,

24.1. The study appears to use only ECMWF-HRES as the forecast dataset. While different resolutions are mentioned, how transferable are the results to other NWP systems?

This is an important limitation that we acknowledge in the revised manuscript. From a practical standpoint, the use of ECMWF-HRES as the only forecast dataset was driven by data availability — to our knowledge, ECMWF-HRES is the only NWP system for which sufficiently long archives of forecast data are openly available to train an LSTM over the experimental period used in this study.

From a theoretical standpoint, however, the approach is not inherently tied to any specific NWP system. A key advantage of using discharge as the target variable is that the model trains on an integrated catchment signal that implicitly captures the combined effect of all meteorological inputs, regardless of their source. This means that the training strategy can in principle be applied to any NWP system by replacing or supplementing the forecast inputs accordingly. Furthermore, drawing on the findings of Kratzert et al. (2021), who demonstrated that combining multiple meteorological data sources improves LSTM simulation accuracy in a reanalysis setting, we hypothesize that combining forecasts from multiple NWP systems could yield similar benefits in a forecasting context, enabling the model to learn source-specific bias patterns and potentially improving robustness across different forecast systems. We add this discussion to the limitations section of the revised manuscript.

24.2. The analysis focuses on a 1-day lead time—how does this choice affect the interpretation and generalizability of the results?

The focus on a 1-day lead time was a deliberate choice to establish a proof-of-concept baseline, as forecast uncertainty generally increases with lead time, making shorter horizons the most appropriate starting point for validating bias correction capabilities. While the results demonstrate that the proposed strategies can effectively mitigate forecast-induced biases at this lead time, conclusions regarding forecast quality decay with increasing lead time cannot be drawn from this study. This work is intended as a foundation for a forthcoming study explicitly focused on multi-day-ahead forecasting, and we add a clear statement to the limitations section acknowledging this scope restriction.

24.3. Static attributes are not discussed. How might they influence the results, particularly if static embedding networks are not fine-tuned in the TL experiments?

The role of static catchment attributes in LSTM-based runoff modelling has been extensively discussed in Kratzert et al. (2019a, 2021), and we refer the reader to these studies for a detailed treatment. In the context of our transfer learning experiments, only the dynamic embedding network weights were updated during fine-tuning in both $TL_{EmbeddingNet_1}$and $TL_{EmbeddingNet\_2}$, while the static embedding network weights remained frozen. The rationale is that static catchment attributes do not differ between the reanalysis and forecast domains, and the representations learned from them during pre-training therefore require no adaptation. We acknowledge that this was not stated clearly enough in the methods section and add an explicit clarification in the revised manuscript.

Technical corrections

1. Line 11: Suggest using “meteorological-forecast-induced” for consistency with other section
2. Line 32: “magnitudinal correct” should be “magnitude-correct” or “magnitudinally correct”
3. Lines 146-148: Something is not quite right in the grammar of the sentence “Hereafter, we use …”. Please rewrite for clarity.
4. Line 284: “with equal five” doesn’t make sense.
5. Lines 583-584: This reference has no way of finding the dataset doi/url etc.
6. Lines 454-456: Something is not quite right in the sentence beginning “The substantial improvements”. Please rephrase.
7. Line 478: The Hunt et al method is not GloFAS-specific it is just applied to GloFAS but it does require an initial streamflow forecast.
8. Appendix B: The orange colour is not defined in the caption

We thank the referee for the technical corrections. We address all technical corrections in the revised manuscript: (1) "meteorological-forecast-induced" is adopted for consistency, (2) "magnitudinal correct" is corrected to "magnitude-correct", (3) lines 146–148 are rephrased for clarity, (4) "with equal five" is corrected, (5) regarding the missing dataset reference at lines 583–584, the ECMWF-HRES data does not have a formal DOI as it is accessed through the ECMWF MARS archive system, which does not issue persistent identifiers for operational forecast data. We add the URL of the ECMWF operational archive to the reference to allow readers to locate the dataset: https://www.ecmwf.int/en/forecasts/dataset/operational-archive, (6) the sentence at lines 454–456 is rephrased, (7) the description of the Hunt et al. (2022) method is revised to clarify that it is not GloFAS-specific but requires an initial streamflow forecast, consistent with our earlier response to this point, and (8) the orange colour is defined in the caption of Appendix B.

Final Remarks

We thank the anonymous referee for the careful and constructive review. The comments have substantially improved the manuscript and have been a strong motivation for future work. In particular, the referee's emphasis on understanding how domain shift varies across basins and meteorological variables has sparked the idea for a dedicated follow-up study, in which we plan to investigate the temporal and spatial patterns of forecast-induced domain shift and their controls on hydrological model performance in greater depth.

References

Bárdossy, A., Kilsby, C., Birkinshaw, S., Wang, N., and Anwar, F.: Is Precipitation Responsible for the Most Hydrological Model Uncertainty?, Front. Water, 4, 836554, https://doi.org/10.3389/frwa.2022.836554, 2022.

Hunt, K. M. R., Matthews, G. R., Pappenberger, F., and Prudhomme, C.: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States, Hydrol. Earth Syst. Sci., 26, 5449–5472, https://doi.org/10.5194/hess-26-5449-2022, 2022.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth Syst. Sci. Data, 13, 4529–4565, https://doi.org/10.5194/essd-13-4529-2021, 2021.

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

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrol. Earth Syst. Sci., 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.