Response to RC2: 'Comment on egusphere-2025-4978', Anonymous Referee #2

We thank the anonymous referee for the careful reading of the manuscript and the constructive 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.

1. Scope

The scope of the article is inside the scope of HESS.

2. Summary

The authors proposed methods to improve model performance under forecast-induced bias (performance drop when models trained on reanalysis data are used with forecast data). They test different strategies including encoder-decoder lstms, sequential-lstms and transfer learning. Moreover, they investigate how linear embeddings, and the inclusion of past-observed discharge influence performance.

We thank the referee for summarizing our paper.

3. General comments

I think the article is really well written, with a good introduction, clear objective and well-posed experiments. The results are presented in a clean way. Even with many model variations, it was easy to follow, which is not always the case. The only major limitation of the study is that the authors are limiting themselves to one-day-ahead prediction, when the LSTMs architectures and the available data already allow them to increase the lead times. This limits the conclusions that can be drawn, especially with the effect of decaying quality of the forecast as the lead time increases. From lines 95 and 489, I understand that this article is a stepping stone for multi-day prediction, so I understand that they want to leave that for a future study, but then they should clearly state this in the limitations section.

We thank the referee for this comment. As noted, this work is intended as a foundation for a forthcoming study focused on multi-day-ahead forecasting. We will add a clear statement in the limitations section acknowledging that the one-day-ahead scope restricts the conclusions that can be drawn, particularly regarding forecast quality decay with increasing lead time.

4. Specific comment

Line 139-142: It is good that you used embeddings, but this paragraph makes it sound as if an LSTM without embeddings cannot capture non-linear relationships nor learned how to combine the features, which they actually can. The main advantage of embeddings is that (1) you can reduce large input dimensions into smaller latten spaces, (2) the embeddings can learn to compensate for systematic bias in your data, for example if you use a different embedding for hindcast and forecast, and (3) if you use different type/groups of inputs with different number of variables, you can map them to a shared dimension for further processing (e.g Acuna2025 for multiple frequencies or Gauch2025 for missing data).

We thank the referee for this correction. We agree that the original phrasing is misleading, as LSTMs are themselves capable of capturing non-linear relationships and learning feature combinations. In our specific setup, a primary motivation for using input embeddings is to enable transfer learning and consistent processing across experiments with different input dimensions. By projecting varying numbers of input variables from different data sources into a shared latent space, the model can be pre-trained on reanalysis and fine-tuned on forecast data regardless of differences in the feature space between domains. We revise lines 139–142 accordingly.

Line 146-148: I do not understand what you are trying to say, can you please rephrase or further explain?

We thank the referee for this comment. As "domain" in hydrology typically refers to a geographical region, we explicitly define its use in the machine learning sense in lines 146–148 to avoid this ambiguity. We rephrase these lines to make this distinction clearer.

Line 259-262: The problem with using tanh as the activation of the embeddings is that tanh saturates. Saturation is a known problem for lstms (Kratzert2024, Acuna2025, Baste2025), and I think it can be further increased if you also saturate the input before it goes into the LSTM. Was there a specific reason you used tanh? Have you tried if ReLu gives you better results, especially considering that in section 3.5 you indicate that more complex embeddings gave you worse performance for the enc-dec and seq-lstm. Are you using dropout in the more complex embeddings to avoid overfitting?

We thank the referee for this insightful comment! The tanh activation was used as it is the default in the NeuralHydrology framework and was not explicitly varied in our experimental design. We acknowledge that the compounding saturation effect of tanh in both the embedding and the LSTM is a valid concern and may indeed contribute to the performance degradation observed with more complex embeddings in Section 3.5. Regarding dropout, we can confirm that it was set to 0.0 within the embedding networks across all experiments, meaning overfitting in the complex embeddings was not explicitly regularized at that level. Testing alternative activations such as ReLU and introducing

dropout within the embedding networks would be a valuable direction for future work. We add these points to the discussion.

Section 3.1: Can you further explain the difference between BaseLine Reanalysis and CrossDomain (Reanalysis, Pretrain)?

We thank the referee for this comment and acknowledge that the distinction between these two configurations is not sufficiently clear in the text. The Baseline Reanalysis model is trained on all available reanalysis variables (31 variables from ERA5-Land, E-OBS, MSWEP, and GLEAM), following the best practices of Kratzert et al. (2021) by combining multiple meteorological forcings for increased model accuracy. The CrossDomain (Reanalysis, Pretrain) experiment, in contrast, uses only 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. This constraint is necessary to ensure that the input feature space remains identical between the training (reanalysis) and inference (forecast) phases, replicating the classical hydrological modelling workflow where a model is calibrated on reanalysis data and subsequently applied with forecast inputs. We clarify this distinction in the revised text.

Line 324-326: The sequential data processing is not only done in the sequential lstm, is it? I agree with what you said at the end of the paragraph, that sequential-lstm is better than encoder-decoder, because the hindcast-forecast transition is done on the same lstm instead of having to initialize a new one, especially in your case, where the forecast part is only run for one day. But this makes it sound like only the sequential-lstm process data sequentially and in temporal order, which is not true.

We thank the referee for this comment. We agree that the original phrasing is misleading, as sequential processing in temporal order is a general property of all LSTMs. With "sequential data processing" we refer to the two-phase structure of the architecture, where the hindcast and forecast phases are processed consecutively within the same LSTM. The key advantage over the Encoder-Decoder architecture, as the referee has noted, is that this two-phase processing happens within a single continuous state evolution, avoiding the disruption introduced by transferring states through a fixed handoff network to initialize a separate forecast LSTM. This is particularly impactful in our setting, where the forecast phase spans only a single day. We revise lines 324–326 to make this distinction explicit.

Line 435: The limitation of integrating reanalysis data depends on the test case. Multiple meteorological services have real-time observed data (from stations or radar), which is the data that can be included in the hindcast period, and then the forecast data comes from the meteorological models. I understand that if you are thinking on a global or continental scale, you might need reanalysis data, but in national-scale applications, you can directly use observed data (if the country has this available).

We thank the referee for this valuable point. We agree that the limitation related to data latency depends strongly on the application scale. At national scales, real-time analysis or nowcast data can indeed serve as the hindcast forcing, bypassing reanalysis latency entirely. Products such as INCA from GeoSphere Austria (see Haiden et al., 2011) are a good example of operationally available high-resolution analysis data that could fill this role. However, in large-sample or continental-scale settings such as ours, station-based data could introduce further practical challenges, as the number of stations per basin varies greatly — a large basin may contain dozens of stations while a small one may have only one, complicating consistent spatial aggregation and resulting in varying input dimensions across basins. While input embeddings as used in this study could potentially help map these varying input dimensions into a shared latent space, the implications for model performance and stability in such a setting remain an open question that warrants future investigation. We add a note to the limitations section acknowledging that at national scales, real-time observed or nowcast data may represent a more operationally viable alternative to reanalysis in the hindcast phase.

Final remarks

We thank the referee for the valuable comments and the overall positive assessment of our work. We hope that our responses and the corresponding revisions adequately address the points raised.

References

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, Weather and Forecasting, 26, 166–183, https://doi.org/10.1175/2010WAF2222451.1, 2011.

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.