Responses to Reviews
How to deal w___ missing input data

https://doi.org/10.5194/egusphere-2025-1224

RC2

General comments

The paper “How to deal w___ missing input data” provides a thorough implementation and analysis of three strategies for how to deal with missing input data for operational AI-based models that depend on the real-time availability of meteorological forcings. The strategies are: input replacing, masked mean, and attention. Through three sets of experiments that train models with different permutations of missing forcings, they show that the masked mean strategy tends to perform the best, if only marginally. They show that attention appears to be unnecessary, as it is complicated and tends to collapse to the simpler weighted mean approach. This is a very nice result.

While the paper is well structured, the text requires some refinement. The description of the key experiments lack some details and make it difficult for the reader to follow. For instance, experiment 2 can only be understood after reading the figure caption, not from the text body. Furthermore, the reviewer has some concerns regarding the reproducibility, since the training code is not made available. Combined with some technical errors, a major revision is recommended.

We’d like to thank the reviewer for their detailed comments. We have responded to each comment directly below.

Specific comments

Major concerns

- Code is missing. While the authors provide detailed code for the replication of the figures as well as the trained models, the code for the model training and inference is not available. In my opinion, and in line with GMD’s practices, the code should be made available before this manuscript can be published.
  This is not correct. The code to reproduce our experiments is publicly available as part of the NeuralHydrology Python library as it is already described in the code & data availability section of the paper.

- The description of the LSTM model is missing in this manuscript. While the authors provide a thorough explanation of the model architectures for the input-data-processing

methods, a description of the architecture for the core LSTM model is missing. The description of the target output of the LSTM model (i.e., what it is actually predicting) is also missing. i.e., Section 2.1 describes the forcing variables, but it is unclear what the target is.
You are right that we missed to explicitly state the target variable (streamflow); we have corrected that in the updated manuscript. We have further added a brief statement to explain that, apart from the input layers, our model architecture is identical to that of (e.g.) Kratzert et al. (2020). We believe that it is not necessary to reiterate the LSTM architecture in detail, as there are no changes to what has been described in much of the cited prior work.

- Section 2.1: Is there a mistake in the testing period? The testing period should not overlap with the training dataset; otherwise, this is a major error, as the models could be overfitting.
  This was a typo in the text, thank you for catching it. The correct periods are: 1 October 1999 to 30 September 2008 for training, 1 October 1980 to 30 September 1989 for validation, and 1 October 1989 to 30 September 1999 for testing. Hence, there is no data leakage between training and validation/test periods.

- Starting from section 3.2, following and understanding the experiment setup is difficult and needs the be clarified. For instance, in section 3.2, you mention that the results show the cases of missing forcings at inference. For a given line/plot, it is not clear what parts of each model are trained (just the LSTM, or also the data-preprocessing components?). It is not also clear what the difference is between the dashed lines and the masked mean/attention/input replacement lines are. I understood the dashed line as a LSTM model trained with 1 (or 2) forcings, and the solid lines are the same LSTM, but with an additional trained encoder (with the appropriate data-processing techniques). However, I could be misunderstanding this. For the dashed (reference) lines, how are the forcings passed to the LSTM (concat, mean, attention?).
  We trained the entire model from scratch in all experiments. We are not sure where the idea that only parts of the model might be trained originates from.
  The dashed and dotted lines stem from Kratzert et al., 2021, i.e., from models that are not robust to missing input data. For these models, the forcings are simply concatenated, following the standard training approach for LSTMs. We've updated the description of experiment 2 and rephrased the caption of Figure 7 to better guide readers and help to avoid confusion.

Minor concerns and questions

- Figure 6: When you describe the attention mechanism for high probability p, it is an interesting result that the weights fluctuate around 1/3. Is this also true for the lower missing data probabilities, where the attention method performs worse? If not, what would your analysis be around this?
  Our finding about the equal attention weights was actually run on a NaN probability of zero. We can see that the previous sentence about high NaN probability suggested otherwise, so we have updated the text to state this clearly. Further, in preparation for this response, we re-ran the analysis with p_time = 0.6 and found the effect to be similar there.

- From reading the manuscript, it is unclear to the reader why the metric becomes CDF for experiments 2 and 3. Please highlight the connection between these two metrics in the text, and specify that the ideal result is a delta function at NSE=1, and thus, curves that are closer to the right are better.
  The main metrics (in the main paper) is NSE and/or KGE in all experiments (the appendix shows additional metrics). We chose cumulative density functions (CDFs) to visualize the distribution of metric values in two of the experiments. As experiment 1 further shows the uncertainty across multiple seeds, we felt that a CDF would be too cluttered and hard to read.
  CDFs are extremely common in hydrologic literature, so we do not see a need to explicitly describe how to read them. Nevertheless, we added the explanation that curves further to the right are better.

Specific, minor comments

- The abstract is missing key results. It would be beneficial to briefly mention the different solutions and hint at which provide the most robust results (seemingly, masked mean)
  We agree and have expanded the abstract in the revised manuscript to include short descriptions of the methods and a peek at the results.

- Minor comment: for readability, it would be a smoother flow if the literature review section on other fields and models with missing input data occurs earlier in the introduction, before you present the three strategies to accomplish this goal.
  We would like to push back on this, as we believe that it makes sense to explain early in the manuscript how we intend to achieve the goals outlined at the beginning of the paper. This gives a high-level overview to the reader, before we dive into the details of related approaches. It also allows us to point out differences between those approaches and our proposed methods that would be hard to understand if we hadn't introduced our methods beforehand.