Review

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Cohen, D., and Gilon, O.:
How to deal w___ missing input data
EGUsphere [preprint]
https://doi.org/10.5194/egusphere-2025-1224, 2025.

Dear Martin and others,

It was a pleasure to review your manuscript on "*How to deal w___ missing input data*", submitted to the Hydrology and Earth System Science journal. I found your study to be well-structured, informative, and a very pleasant read. The experiments are clearly motivated, performed, and analyzed. I think this is a valuable contribution to the field.

I have provided a list of minor comments for your consideration below. Most of these do not require urgent revisions. Most comments are regarding clarity especially for readers that may not have much experience regarding methods like embedding and attention. The authors have already provided additional material which is much appreciated.

I have no doubts that the authors will be able to respond to all my comments without any problem. Hence, I am recommending minor revisions. I appreciate the effort you have put into this work and would be happy to have another look at the revised version.

Best regards,
Julie Mai

Detailed comments:

Introduction

- Title & Abstract: I like the title and the clue that this manuscript deals with missing information. Maybe it would be good to make that connection at the end (or elsewhere) in the abstract. Something like: "[...] or not arrive at all; like the missing characters in 'to deal w___ missing input data'."
- Figure 1: Mention in the caption that "gray" indicates missing data and shades of "blue" mean data available (?!). What I don't understand is why the yellow box indicates where models are <u>not</u> robust. Isn't that where all data across both basins are available and complete and hence this is where the model <u>is</u> robust (given that only forcing group 1 and 2 are used and 3 is discarded)? Some more detailed explanation may be required.

- Lines 37-42: Would it be possible to make a connection of these three cases to Figure 1? E.g., first setting is top case in Fig 1? Maybe there is a way to adjust Figure 1 to be used not only for illustrating various kinds of missing data but also used for these three cases? Optional to address. May just be helpful for readers to understand the three experiments better.
- Lines 52-53: Use of term "downstream model". I am assuming that it is "downstream" in the modeling process and not downstream in a hydrologic sense. It may be good to clarify this.

Data and methods

- Footnote on page 3: "Unlike what is mentioned in the paper, …". I am assuming that the Kratzert et al. (2021) is referred to. I haven't consulted both tables of static attributes yet and am confused if "p_seasonality" is or is not a static attribute in the study presented here. Maybe this footnote would be better placed with the list of actual static attributes to be less confusing as I assume that using/not using this static attribute does not make a huge difference?!
- Figure 2:
  - "NaNs in the input data for a given time step are replaced by zeros, …" —> I would mention that this is what happened to the three example entries for forcing group 2 (grey means zero).
  - I would recommend having four instead of three example entries for each forcing group. This way it becomes clearer that the binary flags are per forcing group and not per basin.
  - Maybe mention that this example shows forcing available for three/four basins. Initially I thought time steps… I know it is all in the caption, but it took me a second to wrap my head properly around this.
  - Why is the third entry in forcing group 3 not set to zero even though it seems it is not available (NaN) for one of the basins? I would recommend explaining this a bit more in detail. This seems crucial as the binary flag seems to be only set to 0 if a forcing is not available at any basin.
- Lines 82-83: "each of them yielding an embedding vector of the same size" —> Not an expert of embedding networks. What is the size of the resulting embedding vector. In your example (Fig 3) it is 4. Is that a hyper-parameter of the embedding network? Some information on that may be helpful.
- Line 83: "average the non-NaN embeddings" —> Again, my limited knowledge of embeddings here: Under which circumstances can embeddings be NaN? It may be helpful to have such an example in your Figure 3 (vectors shown as list in "avg()").
- Lines 92-93: "Appendix D provides a brief introduction to the concept of attention for readers who are not familiar with the topic." —> Much appreciated! Great job explaining this with intuitive examples.
- Figure 4:
  - There is some overlap of text in top left corner.
  - Is it a coincidence that there are 4 static attributes and the length of your embeddings is 4 as well?
  - What are the binary flags used for? Aren't they already implicit by you only creating two embeddings and none for forcing group 2? If the binary flags are important to inform the

number of embeddings created, why are the binary flags not part of the masked mean figure (Fig 3)? Guessing that this is a "requirement" of the attention framework?!
  - Make explicit in caption that "k/v" means "keys and values".
- Lines 96-98: In the masked mean you decided not to use static attributes as they only deteriorated the performance. For attention networks you decided to use them. I am assuming that the performance increased by using them. If so, I would mention this like you reporting on this already for the masked mean (lines 84-86).
- Figure 5: The caption does not make a connection to the legend. I know the text explains that the (random) subset of basins was assumed to have three data products while the rest is assumed to have only 2 products available, but this should be put in the caption to explain the figure without the need to look up in the text what it actually means.
- Experiment 3: Lines 129-137: What do you think is the impact of you picking the worst performing of the three products, i.e., NLDAS (see lines 115-116) as the additional one for the subset of 51 basins? Your motivation for this experiment was that "higher quality" forcing may be available locally. Shouldn't you have picked the best performing forcing as the additional one?

Results

- Figure 6:
  - "NaN probability" equals "p_time", right?
  - I would probably sort the various approaches in the legend in the order they are introduced (replacing, masked mean, and attention). Same for figures 7 and 8.
  - I'd potentially cite Kratzert et al. (2021) for the two reference results (dashed and dotted line) to emphasize that this was done previously. Again, this is only to make the figure content somewhat independent from the rest of the text.
- Line 145: "while masked mean is slightly better in KGE" —> it looks like input replacing is also better in KGE at p_time=0.0.
- Lines 149-150: "except for p_time = 0.2, the masked mean results are significantly better than those of input replacing" —> I can't really see that in the plot. The median value for blue (masked mean) and yellow (input replacing) seem both to be around 0.775. Are they only significantly different in terms of the one-sides Wilcoxon signed-rank test? Is that the only result where a significant difference was detected?
- Line 161: "exact forcings that are available at inference time" —> "exact forcings that are available at inference time (dashed line)"
- Line 161: "performs significantly better" —> based on a statistical test or just from looking at plots? I can't really see a difference between the dashed CDF lines and the colored CDF lines; especially in the right column of Figure 7. It may be that you are talking in this paragraph about the single forcing results (left column) but it's not clear from the text. Only when you start the next paragraph it is suggesting that you were only looking at single-set results (probably)?
- Line 160: Is the highlighted result for NLDAS-only likely happening because NLDAS was the worst performing dataset in general?

- Line 172: "The three-forcing model trained only on the 51 basins …" —> Maybe call this "regional model" here explicitly or say that this refers to the dashed line in the plot to make it easier to connect these things
- Figure 8: Can "regional model" in the legend include that it is using all three forcings? Like the legend entry for the "global model". Easier to connect to the text (line 172 etc.).
- Lines 174-175: "However, from a practical hydrological perspective, all approaches perform quite similar, despite the statistical significance." —> I like this sentence. I think this may also be something that could be added to the results for experiment 2.

Discussion and conclusions

- Line 183: "unable to outperform the baseline trained on all three forcings but only 51 local basins (experiment 3)" —> may this also be caused by the additional third forcing not being "better" than the others but indeed shown previously to be the least performing?
- General: Is there any notable difference in computational expense or difficulty in implementation between the three methods? This may also be factors for practitioners to select one method over another.

Appendix

- Figure C1/C2: It would be great if you could add a reference for the definition of all the additional metrics.