Reviewer: Juliane Mai

Dear Daniel and others,
It was a pleasure to review your manuscript on "EARLS: A runoff reconstruction dataset for Europe", submitted to the Earth System Science Data journal. I found your study to be well-structured, informative, and a very valuable contribution to the field.
I have provided a list of minor comments for your consideration. Most of these do not require urgent revisions, but I have highlighted in blue those that may benefit from some additional attention.
My main comments focus on the following aspects:

- Slight inconsistencies between datasets – While these are not major issues and likely do not require changes, I would appreciate your expert opinion on what potential impacts these inconsistencies might have on your findings.
- Figure clarity – Some suggestions are provided to improve comparisons and ease interpretation.
- Simulation data and uncertainties – Comments regarding the data shared to ensure transparency and clarity.

Overall, 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

Dear Juliane,

Thank you for your extremely productive review. Your feedback was very detail oriented and wherever possible we implement the proposed changes for our revisions (see detailed comments). We think this will greatly improve the paper!

Detailed comments:

Section 1:

- None

Section 2:

- Line 74: “not able to produce a simulation for 227” —> Why? Are these 227 part of the 14,161 basins without data gaps, or the 2,655 with gaps, or additional ones?

For these basins we have so much missing data in the inputs that we were not able to model to make simulations at all. The 227 are not part of the 2,655. We will clarify this indeed.

- Line 83: “The basin shapes are either derived from EStreams (do Nascimento et al., 2024) or HydroATLAS (Linke et al., 2019)” —> For a given basin, what made you decide which one to use? If you specify this later please refer to that section here.

We used both whenever possible. We will make sure to mention this.

- Line 84: “We use 4 dynamic inputs” —> You state later that EOBS has more forcing variables. Are the other ones not contributing to a better performance? Did you test other variables? I think it would be great to have some additional information on this here.

The reviewer's intuition is completely right. The additional inputs can be used to improve model performance (albeit not by much). We chose a small amount of commonly available inputs to make it easy to apply the model for different use cases/scenarios. We will discuss this more in the revised manuscript.

- Line 88: “[EOBS is] spanning 1 January 1950 to present.” —> Your dataset starts in 1953 though. Why are the first 3 years skipped. I think that’s fine. It would just be good to have some information why you skipped 3 years.

This is indeed confusing for readers. We use the first three years as buffer years. We will mention this explicitly here.

- Line 90: "we use 13 attributes that we aggregate from HydroATLAS, namely: basin area, the average elevation, the average slopes, the average stream gradient, the average long-term air temperature, the minimum long-term air temperature, the maximum long- term air temperature, a global aridity index" —> I am wondering what happens if you are using a basin shape from EStreams (not HydroATLAS) to aggregate forcing etc but then you use basin characteristics that are (technically) for a (potentially) different basin shape. For example, basin area may be different but also meteorologically based basin characteristics would not be consistent. Further, why are the meteorologically based attributes to directly derived from the forcing dataset? I'm sure this is all just minor differences but some kind of note for the reader may be helpful.

This is an interesting question. If the introduced differences are systematic the LSTM could learn to compensate for them and there might not be a difference at all. If they would, however, have random components it would inject additional noise to the forcings. One can only guess how heavy the degradation in performance would then be. We will mention the possibility to create new EARLS version by recombining different datasets in the conclusions of the revised manuscript, together with the discussion of using different dataset all together for the statics (see reviewer comment regarding Table 1).

As for the meteorologically based attributes: For the static attributes we follow the convention of the Caravan dataset. The goal here was to make it so that, in theory, the model can be used for a wide range of situations. We will mention this explicitly in the revised manuscript.

- Figure 2: The positioning of the three-parameter Laplace distribution of simulation results just at the end of the LSTM time line looks bit like there is only one set of the three parameters estimated. It maybe clearer if each of the orange LSTM "boxes" would have an arrow (or connection) to the gamma, b,

and tau. Or maybe place an index "t" at each of the three parameters to show they are time dependent?

We agree that this depiction might be misleading. To remedy this we will add a time index to the figure (as proposed by the reviewer), and mention this explicitly in the caption.

- Line 146: "seperately delineated gauged basins" —> What does that mean? Basins have another shape than you used for EARLS? Doesn't that mean that the forcings are (slightly) different? What impact may it have that EARLS estimates are for a potentially slightly larger or smaller basin than mHM given that it is another basin shape? Also, typo: "seperately" —> "separately"

The formulation is indeed unfortunate. What we wanted to express is that they are not part of EARLS but are specifically used for the experiment. We will make sure to say this explicitly in the revised manuscript.

- Table 1: You list the datasets used to setup mHM. Technically these could be used to derive the static basin attributes for the LSTM. How different are estimates from those datasets compared to what you use from HydroATLAS? Would these differences matter?

This is an interesting question. We did some tests with the static attributes of EStreams. They are, generally speaking, of higher quality than the ones from HydroATLAS, but secluded to Eruope. Roughly speaking, what we saw there is that the ungauged basin performance increased a little. In order to be able to use the first EARLS LSTM to a wider area we did, however, decide to use the static attributes that are globally available. That said, the broader idea here is perhaps to build different EARLS versions with other static attributes in the future. We will make sure to discuss this in the conclusions of the revised manuscript

Section 3:

- "potentially difficult to predict ungauged basins" —> It is stated that all basins have "natural" streamflow (line 53). I was hence expecting that these are mostly pristine basins with low human impact. Is that correct? What

would make these basins then "potentially difficult to predict"? It would be great if some reasons why a basin may be "potentially difficult to predict" would be added to the manuscript. Also, out of curiosity, how do you know that basins have "natural" streamflow. Is that a HydroATLAS attribute for basins? If so, maybe mention that?

The statement of the reviewer is correct. We use the pre-filtering described in appendix A1 (However, this still leaves a lot of ungauged basins that can be difficult to handle, e.g., Karst-affected catchments, anthropogenic influences, etc.). This specific sentence, however, specifically refers to the data splitting approach that we describe in appendix A3. That is, we do not split the data completely randomly, but make it so that the train-, validation-, test-data exhibit distributional shifts. The way we formulated this does indeed make this unclear. In the revised manuscript, we will make this clearer and refer explicitly to appendix A3.

- Line 169: I think it would be helpful for the reader if you could start the section with repeating the time period you evaluate here.

We will do so in the revised manuscript.

- Line 171: suggestion: "opposed to Kratzert et al., 2021" —> "opposed to Kratzert et al., 2021 using multiple facing datasets"

We will change this as proposed.

- Line 172: "These results are similar to the ones in Kratzert et al. (2019a), but worse than the ones from Mai et al. (2022)." —> Would it be possible to state how many where below 0.0 in those two studies? It would help for comparison.

If this is helpful we will include the specific information in the revised manuscript. Specifically: TK

- Figure 3: The validation data line is barely visible. Maybe use thinner lines for all three and remove (some) gridlines? It would also be nice to state clearly

how many basins are in each of the three sets. Maybe add to the legend (e.g., "test data (N=500)")?

We agree with the analysis. Of course the fact that all lines are closely alone is part of the argument. Still more information is good for readers here. Hence, we will adapt this as proposed by the reviewer.

- Figure 4:
  - The caption states that "The colors mirror Fig. 3". But figure 3 has colours for test, training and validation data. Figure 4 shows NSE performances of the validation (?) data.
  - I think figure 3 and 4 should be combined since they are somewhat connected. The continuous colourbar in figure 4 is not really needed I think. I can only distinguish 3 colours- blue, red and purple. I would just pick 3-4 distinct colours (or the ones you use in for example Figure 8) and then use those to get the point across. Also, I think the colourbar needs to be open ended at the lower end as I am assuming that there is values smaller than -0.5 that are also coloured in red?!
  - It would be great to exactly state which basins are in figure 4. I am guessing validation basins. I would use the same exact wording as used in figure 3.

We did indeed mess this up. We will adapt this as proposed by the reviewer.

- Line 175-190: I really like this discussion. It states a lot about the methodology (data pre-processing and filtering of basins) that I was expecting beforehand (see comments above). I think it maybe advisable to have a separate section called "data pre-processing" where all this is placed before we dive into model performance etc.

If the editor allows for it we will adapt this as proposed by the reviewer.

- Figure 5: caption suggestion:
  - "Bertola distances" —> "Bertola distances d"
  - Please add to the caption what panel a and b depicts.

- Typo: “D_{<0.8}” —> “D_{>0.8}” (in caption)

We will adapt this as proposed by the reviewer.

- Line 199: “Tn” —> “In”

Thank you.

- Line 201: “roughly starting at the 60th percentile” —> I would actually say much earlier; like around 37.5th percentile. Unless the authors look at another feature or criterion to determine when models start looking similar in performance… It maybe helpful to have those lines added as horizontal lines in Figure 6 for the reader to better be directed towards what they are supposed to see.

We will adapt this as proposed by the reviewer.

- Line 201: “For the remaining 15%” —> I am not sure which remaining 15%.
  - 0-15th = EARLS and default are similar
  - 60-100th (I suggest 35-100th)= EARLS and local are similar
  - 15-60th (I suggest 15-35th) = in between -> but this is 45% (unless the authors agree that it should be around 20% remaining)

Exactly. As with the comment before, we will adapt the wording of the reviewer in the revised manuscript.

- Figure 7: I love that figure!

Thank you!

- Line 207: “We encourage readers to compare our version with these depictions” —> Please be aware that the publication is not open access. So, it may limit the ability of readers to actually do this. :(I am assuming it is not reasonable to recreate the figure with the Blöschl data and have them for comparison in the manuscript?

Unfortunately this is exactly the case.

- Figure 8: The caption is not consistent with the colorbar. “Positive trends are depicted in red” but they are blue and vice versa. Unless the legend shows Blöschl minus EARLS estimates and then you talk about EARLS minus Blöschl in the figure caption. In any case it’s confusing.

We will correct this.

- Line 219: “the original version shows slightly positive trends” —> Isn’t figure 9a all “red” in Scandinavia which is negative values which means negative trends (see caption figure 9). There is some sort of mix-up what positive and negative means I think. There maybe more in this paragraph but I leave it to the authors to revise them without pointing each one out here.

Yes. This is indeed a mixup leaching from figure 8. We will correct this together with your comment on Figure 9.

- Line 219: “positive trends” and “negative trends” —> in general, I am not sure if I would refer to them as “positive” and “negative”. I am understanding that positive values indicate more floods and I am not sure if that’s “positive”. Maybe refer to them as “trends of increasing number of floods” or “increasing trend of floods” or something…

We will adapt the description as proposed.

- Line 224: “Some differences can be explained by data availability” —> Do you think it may be helpful to actually plot the gauge stations used for the two datasets in figures 9a and 9b? It may underline your point that differences appear where more data are available while the other dataset lacks observations.

Yes. This comment is related to comment TK by Wouther Berghuis. We do indeed agree that contextualizing the nature of the differences makes the manuscript clearer and more readable. Hence, we will include the proposed solution in the revised manuscript.

- Figure 9: Would it be possible to use the same colorbar as Figure 8?

Yes.

- Figure A1 and A2: I highly recommend to merge these two figures into one figure with two panels. This would make it much easier to compare the workflows; especially when the box-diagrams (which are beautiful) are arranged in the same way.

We will adapt the figure as proposed in the comment.

> I think the "selection" step in either figure needs to be part of the methodology. Currently this is a bit vague and distributed. It's such a curial step that it should be easily findable. I later found some of that in the Appendix but maybe move it to the methods or at least refer to this section of the appendix early on in the methods.

Albeit this is an attractive proposal we don't think that we can do that. As a matter of fact, the figures were part of the method section for the first version of the manuscript. However, the editor asked us to move them to the appendix to align the manuscript more with ESSD. Hence, if the editor does not ask us explicitly to move them back up, we will keep them as part of the appendix.

> Also, there should be a comparable section like this for the section of "engaged" basins, right?

We will include such a section.

- Table B1: Wow!

All accolades go to Thiago for this one!

- Figure C1 to C3: I think these three can be merged into one figure with three panels. It would be easier to compare them if they are next to each other. Also, the KGE could just run from -2 or -1 to 1. This way one would see more of the actual interesting part of the rising limb.

We will adapt these changes for the revised manuscript.

Section 4:

- Line 256: You may want to link to DOI "10.5281/zenodo.13864842" which would always point to the latest version of the dataset.

We will do that.

- Line 263: The constructions folder is stated to include CSV files with at least date and simulation in mm/day. That is great. I am however wondering if it would be possible to include some information about the uncertainty estimates. I think these are a major selling point of this dataset and it is stated that it is included later (line 289-291: "each time step EARLS provides a conditional uncertainty estimate — which can, for example, be used to compute the likelihood of a given model" (which I really like). I was however not able to download the full 33GB dataset and check if there may be a file that contains the uncertainty information. The estimated download time was 26 hours which seemed too much to wait…
    - If the uncertainty data are contained, please make more clear where one would find these data.

The uncertainty data are indeed part of the dataset. We will describe this more thoroughly in the revised version.

    - If it's not included maybe make more clear that a user would need to setup an LSTM themselves and train it and then get those estimates themselves.

The LSTM is not provided as part of the dataset, but we will provide the model structure and weights it as part of the code of the paper. The revised manuscript will emphasize this in the code section.

    - Is the download always taking so long or is it just me? An idea would be to have a mini-example with 3-5 basins in a separate (much smaller) zip such that people could download that to see if it contains what they would expect, and setup workflows while they wait for the entire pack to download?! Up to the authors, of course.

We will do exactly that! When we tried it it did not take that long. But as the reviewer points out there can always be circumstances that lead to slower download times. Providing a small sample of the dataset solves this. Great!

Section 5:

- Line 288: “11 thousand” —> “11,000”

Acknowledgements:

- Line 433: “mHm” —> “mHM”

Thank you.