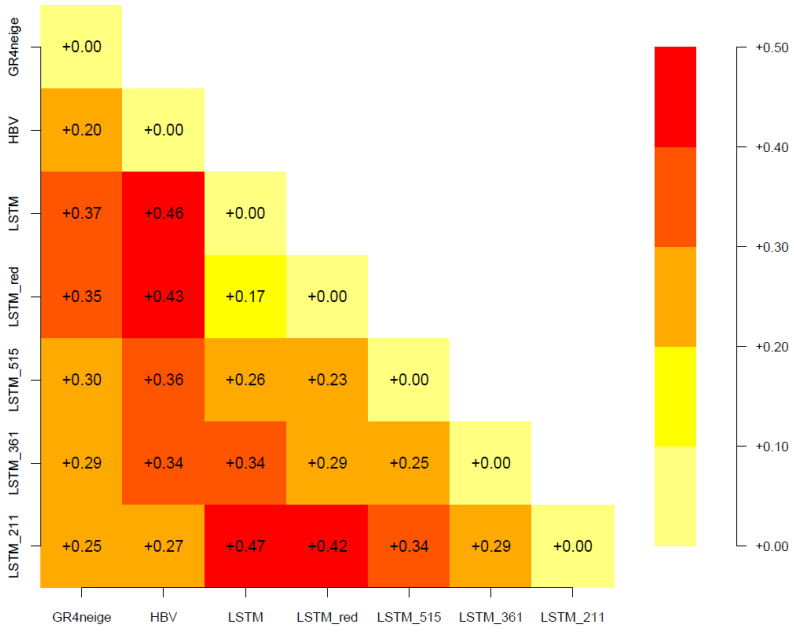
RC1:

Reichert and co-authors describe a study in which they test the qualitative response (metamorphic testing) of two conceptual hydrologic models and a deep learning model trained on CAMELS-US dataset to perturbed temperature and precipitation, aimed at mimicking the qualitative performance of these models under climate change scenarios. The deep learning model (LSTM) outperforms the conceptual models during calibration and validation, but exhibited unexpected hydrologic response in low-elevation basins when temperature was perturbed. Solely training on the low-elevation basins from CAMELS-US improved this qualitative response to the perturbed temperature, suggesting that fine-tuning or limiting datasets to prediction task may help improve out-of-bounds predictions. I provide some comments below that I think will improve the manuscript.

- Generally, I encourage the authors to attempt to summarize all the basins used for metamorphic testing rather than providing individual plots across all basins in the supplementary. There does not appear to be a reason why certain basins are displayed in the main text vs. supplementary. Summarizing across all basins used will help the reader understand better if the pattern is common without having to look through dozens of individual plots in supplementary.

  We agree that a numerical summary across the basins improves the manuscript and we suggest to add the following figure with the summary statistics. As we cannot compare with the truth or with data, we compare the sensitivities of the different approaches by calculating the root mean squared differences of all pairs of approaches and average them over the seeds (if available) and over the catchments within the same class. We suggest to include the results for the most important case, the temperature sensitivities of the low altitude catchments in the paper:



  In this figure, we clearly see that the LSTM sensitivities approach those of GR4neige (decreasing differences [in mm/d]) when moving from the basic LSTM to the LSTM_211 (first column) and they also approach those of HBV (second column). In parallel to that, we see that the sensitivities of the LSTM deviate more and more from the basic LSTM when moving from the basic LSTM to the LSTM_211. These are the changes discussed in the paper.

  Still, we believe that showing one example for each class of catchments in the paper is very important to demonstrate the shapes of the sensitivities and, for being transparent, showing the results for all chosen catchments in the supporting information is useful to demonstrate that the response is typical for the class. For this reason, we suggest to add the summary table/figure but to keep the examples.

Line 40 and elsewhere: Instead of 'modified driving forces', could you use something more generic like 'out-of-bounds predictions'?

We think that "modified driving forces" more clearly expresses what we do than "out-of-bound predictions".

- Line 114-115: the initial clause seems a bit clunky. Would it be clearer to either remove the first clause entirely or consider placing it as a separate sentence?

  We agree and will split the sentence to express the point more clearly.

- Lines 120-139: this is a flat 10% increase in precipitation for every precipitation data point in the dataset? What if there is no rain on a given day? I assume that will still be 0 precipitation increase scenario given the equation 1 but clarification would be helpful.

  This is correct, $1.1 \times 0 = 0$. We will add a sentence to clarify this.

- Line 160: can you give examples of what these precipitation or temperature related attributes would be?

  The full and reduced sets of attributes are given in Table B1. We will add this reference in line 160. Examples are "mean daily precipitation" and "fraction of precipitation falling as snow".

- Line 168: what was the validation NSE range for these catchments?

  All individual NSE values for both, calibration and validation periods, are given in the Figures in the Supporting Information. The range for calibration is 0.82-0.92, the range for validation is 0.67-0.91. We will add this to the paper.

- Line 213: I encourage the authors to include a link to the working repository at the moment, or a draft code release.

  The link for the working repository for the conceptual models is https://gitlab.com/p.reichert/hyperflex.
  The LSTM code used in this work can be accessed at http://doi.org/10.5281/zenodo.3993880.

- Figure 2: describe in the legend what the numbers next to the points represent.

  These are the identifiers of the CAMELS basin. We will add this statement to the legend.

- Line 253: clarify that the sensitivities are in relation to the outlet discharge and not the overall model performance – 'sensitivities of the models are essentially negative' makes it sound like the models had a poor/unexpected outcome, but this is quite the opposite. I suggest changing to something like, "The predicted outlet discharge for the GR4neige and HBV models was lower with increased temperature, which is expected … "

  ΔQ_P and ΔQ_T are defined by the equations (1) and (2), respectively. They are negative if the predicted output is lower than for the base simulation and positive if it is higher. We will add an explanation about the meaning of the sensitivities and what is expected.

- Figure 3: Please change the colors of the third panel to be different than the colors used to indicate the different types of models in the other panels? This is confusing to switch the meaning of the colors in the same figure.

  We agree and will do that.

- Figure 3: for the top panel, why is there a y-axis that extends to -4 when there are no negative values? Also seems like the max differences are cutoff at the top of the y-axis, please extend higher.

  We think that it makes sense to have the same scale for the sensitivities across all catchments to make it easier to compare them. The chosen scale if a compromise of seeing sufficient detail and not cutting some of the sensitivities too much.

- Figure 3: for the fourth panel, indicate what the black circles are – I assume they are observations

  Yes, they are. We will add this statement to the legend.

- Figure 4: please include the full legend here so the reader doesn't have to refer to a separate figure

  We agree and will do that.

- Line 276: change "rainfall" to "precipitation" as a lot of this precipitation is falling as snow in this basin and other basins.

  We agree. Thank you for the hint.

- Figure 4: it is hard to distinguish the different model traces on here and I cannot tell what I'm supposed to take away from the bottom panel. I suggest splitting out the LSTM traces into a separate panel and/or show the deviation of the LSTM_x compared to the base LSTM results.

  We agree that this is difficult to see and will improve the signatures.

- Lines 395-402: I'm curious if the authors tried pre-training on the entire dataset with early stopping criteria as to not overfit, and then fine tune on the reduced dataset with only low-elevation basins. That seems like it might be the best of both worlds – providing better fits by using more data but also passing the metamorphic test.

  We agree that strategies like pre-training can be highly effective. In a separate study (Ma et al. 2021, cited in the paper), we have leveraged pre-training and transfer learning technique to enhance the performance on smaller datasets. However, for the purpose of this study, our focus is on employing a standard LSTM model. This approach is intended to provide insights into the typical methodologies used in training ML models and to highlight the discernible differences that emerge from such conventional training practices.

- Lines 398-399: by how much did the quality of fit deteriorate? By 1%, 30%? I suggest adding in a quantitative measure so the readers can evaluate how much the tradeoff is for passing the metamorphic test with less data.

  We agree. Taking the basin 11468500 as an example, when the number of data reduces from 211 to 99, the validation NSE drops from 0.86 to 0.30 which is far below the range of validation NSE values for all approaches used in the paper (0.67-0.91).

- Lines 412-414: See Topp et al. 2023 https://doi.org/10.1029/2022WR033880 for comparison of ML architectures to prediction in unseen conditions. They suggest ML process/physics guidance helps improve predictions in unseen conditions. Likewise, see Read et al. 2019 https://doi.org/10.1029/2019WR024922 for out-of-bounds predictions using process-guidance for ML models.

  Thank you very much for these hints. These papers are about stream water temperature which is not a topic of this paper. We can still add these references to the discussion that hybrid approaches are also relevant in other areas of hydrology and beyond.