We thank the reviewer for his comments as well as text suggestions. We adapted most of them as is and for the others, we left an explanation below.

I have some comments:

My main comment was that it was only clear to me on second reading that the model received multiple meteorological inputs rather than just multiple precipitation products. Having looked through some of the previous reviews it seems that I wasn't the only one. I understand that it has been clearly stated here: P4 L78: "We used all five meteorological variables of all three dataproducts as inputs for our model". The confusion I think, stems from two places. 1) the abstract 2) the analysis with only precipitation products.

1) in the abstract you have written P1 L3-4 "Using multiple precipitation products (NLDAS, Maurer, DayMet) in a single LSTM significantly improved simulation accuracy relative to using only individual precipitation products." Unless I have misunderstood, would it not be better to write that "Using meteorological inputs from different data products (NLDAS, Maurer,DayMet) in a single LSTM significantly improved simulation accuracy relative to using only individual meteorological products."

We adapted the abstract as proposed by the reviewer.

2) The analysis focuses on precipitation products (rightly given that it is the most important input variable). I think that you should explicitly write this somewhere, perhaps using the response that you used in your comments to reviewer 2: "We only looked at the influence of the three precipitation products because a) precipitation is arguably the most important variable in the process of rainfall-runoff modeling b) according to Behnke et al. (2016) there is little difference in all other meteorological variables in between these data products, c) we know from other research projects that precipitation has by far the most influence in LSTM-based rainfall-runoff models (see e.g. Frame et al., 2020), and d) nothing that we show or conclude implies that other variables are not important, but the point is to show that the LSTM learns to mix forcings in dynamically heterogeneous ways. We show this using the

precipitation input. It is trivial to perform similar analysis on all other variables (and we invite everyone to do this with the code we provide with our paper),". Perhaps something like: "For the analysis that follows we only consider the sensitivity of the LSTM to precipitation inputs. This is for two reasons. 1) Precipitation is consistently found to be the most important variable for rainfall-runoff modelling (Frame et al., 2020) 2) according to Behnke et al. (2016) there is little difference in all other meteorological variables in between these data product"

This can go in BOTH or EITHER of P5 L129 Section 2.5 & P6 L243 Section 2.5.2. Just to make clear why you are only looking at sensitivity to precipitation. This should also clear up any confusion about the inputs to the LSTM for future readers.

We updated the description of the experiment in Sect. 2.5.

Formatting

These are a list of small formatting / spelling errors.

P5 L103: Space between "(1)all" -> "(1) all"

Thanks, corrected.

P8 L178: Spelling "calbrated" -> "calibrated"

Thanks, corrected.

P12 L228: Double word "... in the left-subplot, and the the overall sensitivity" -> "in the left-subplot, and the overall sensitivity"

Thanks, corrected.

Appendix (need to be considered together)

- P16 L270 Re-label Table C1 -> A2 (or B1 depending on whether it needs its own subsection)
- P17 Appendix C is missing (or is Table C1 meant to be Table B2 ?)
- Appendices are shifted by 1 (B, D, E); Replace with ABC or ABCD (depending on whether Appendix C has it's own
- Figure captions updated depending upon the chosen appendix structure (e.g. IfAppendix E becomes Appendix C, update Figure E1. to Figure C1)

Thanks, we corrected the labeling in the appendix.

Suggestions

Feel free to incorporate these or to ignore them. I have tried to offer my best suggestion for howa suggestion could be addressed in red.
P2 41-42: Feel free to ignore, but it is perhaps useful to use the same units for the two resolutions. You write: "the former has 1 km x 1 km spatial resolution and the latter two have one-eighth degree spatial resolution". Perhaps replace with: "the former has 1 km x 1 km spatial resolution and the latter two have 12.5 km x 12.5 km"

We used two different units, because we preferred to report the spatial resolution as reported by the original dataset producer. However, we adapted the sentence slightly to:

"...the former has 1 km x 1 km spatial resolution and the latter two have one-eighth degree (approximately 12.5 km x 12.5 km) spatial resolution."

P5 L103-105: Experimental Design. It might be useful if you label this experiment "Feature Ablation", since you describe it once here and then again in Section 2.5 L125-128. That would make it clearer that Analysis 1 - Feature Ablation is the same as the experiments that you describe at the start of Section 2.4. Or else, include the following sentence in L128. "The feature ablation study describes the seven input configurations with different input datasets, discussed above", or words to that effect.

We think that Section 2.4 (Experimental design) and Section 2.5 (Analysis) are two different things. Section 2.4 describes the experiments that we run (i.e. train LSTMs for all different combinations of meteorological forcings) and Section 2.5 describes how we analyzed the results of these experiments. Both types of analysis (Ablation and Sensitivity) were performed on the same set of experiments (from the experimental design section).

P9 L202-206 I am confused about what the difference is between two different benchmarkings.You write: “The three-forcing LSTM outperformed the single forcing LSTMs almost everywhere. Individual exceptions where “less is more" do, however, exist (e.g., Southern California). Concretely, the three-forcing model was better than the best single forcing model in 66% of the basins (351 of 531) and had a higher NSE than the individual single-forcing LSTMs in over 80% of the basins”. What is the difference here? That the 3-forcing LSTM does better thanthe best single forcing LSTM in 66% of basins (n = 531), but better than 80% of all basin-feature combinations (n=531 * 3)? Perhaps the confusion comes from the non-specificity of better. I was thinking initially that “better” meant something other than improved NSE, since you explicitly write “higher NSE” in that sentence but leave it vague before.

In both cases, better is referring to higher NSE. In the first comparison, we took the highest NSE of the three single forcing LSTMs per basin and compared those values to the result of the 3 forcing LSTM. Here, the 3 forcing LSTM is better in 66 % of the cases. In the second case, we compared the 3 forcing LSTM to each single forcing model in all basins separately. Here, the 3 forcing model is better in 441 basins than the LSTM trained with DayMet forcings (83%), better in 456 basins than the LSTM trained with Maurer forcings (86%), and better in 472 basins than the LSTM trained with NLDAS (89%). We had the detailed numbers in the caption of Fig. 5, and thus shorten this information in the text to “in over 80% of the basins”. We reformulated this passage to make the comparisons clearer.

P11 Figure 5: Is it possible to have more information in the caption. Perhaps including a description of what +ve and -ve values mean. “Positive (blue)

values represent basins where the improvement of the 3-forcing LSTM over the comparison single-forcing model is larger. Negative values (brown) values reflect basins where the comparison single-forcing model outperforms the 3-forcing LSTM."

Thanks, we adapted the figure caption to include qualitative information on the color scale.

P12 Figure 6: Could you include information about which model you are using this information from? I am assuming it is the 3-forcing LSTM, but it could possibly be the learned contribution for each single-forcing LSTM for each respective product (DayMet, Maurer, NLDAS). "The integrated gradients were calculated for the 3-forcing model (the model with all of the precipitation products used as input)"

Thanks, we adapted the figure caption.

P13 Figure 7: Would a key be useful? Or at least a description of the colours in the figure caption (probably easier). "The integrated gradient of Daymet is shown in blue, Maurer in orange and NLDAS in green." I know it's the same throughout the paper but I think it would help people to navigate the figure.

Thanks, we adapted the figure caption.

P19 Figure E1: Is it worthwhile including two more pieces of information in the caption? 1) That each point is a basin 2) Define what rho and sigma represent. 1) is definitely implicit and easy to understand in the context of the other figures in the appendix. However, this reviewer feels that it would be useful to be explicit about this, at least for this first plot in the appendix. 2) is defined in the text (Equation E4,E5), however, it might be useful to have a caption that fully describes the axes labels. Perhaps you can also describe what the values describe E.g. "\rho describes how much correlation there is between the given data product and the estimated truth"; "\sigma describes the estimated disagreement between the given data product and the other dataproducts" (or something more correct along these lines!)

Thanks, we adopted these suggestions.

P22 Figure E5: Similar to above, is it possible to describe what the log-determinant of the covariance matrix describes? E.g. “|\Sigma| increases when there is a larger disagreement between the three datasets, approximating the joint entropy of the three products“

Thanks, we adopted these suggestions.