Overview

This is a summary of the work as I have understood it. It does not need an explicit response but if there are any misunderstandings perhaps it would be worth clarifying them.

The paper seeks to answer two related questions. Firstly, whether using multiple input (forcing) datasets improves LSTM performance in rainfall-runoff modelling. Secondly, to extract insights about how the LSTM uses information in different times and places ("... in spatiotemporally dynamic ways.").

In order to answer these questions the authors completed two experiments:

1) tested the accuracy of the LSTM with different combinations of meteorological input datasets
2) tested the sensitivity of the LSTM to different precipitation inputs (as a demonstration of one meteorological variable)

The major finding was that LSTM model performances were improved with multiple meteorological inputs. Indeed, results are a further improvement on the previous benchmarks set by the authors in their 2019 study (Kratzert et al 2019). Not only were results improved, but the authors were able to show:

a) the DayMet dataset had the most information for rainfall-runoff modelling
b) the LSTM does learn to use different products in different locations, and different times. A simple example is that the LSTM learns a time-lag in the Maurer precipitation product, reproducing findings from elsewhere.

The novel contributions are as follows:

1) Showing that LSTMs can effectively be used with an ensemble of inputs, suggesting a simple method for combining different datasets without prior information of those datasets reliability.
2) Interpreting how an LSTM is able to use input information.
3) Demonstrating new state of the art results for rainfall-runoff modelling in a large sample study in the USA

The techniques used by this paper present an opportunity for the hydrological community to better understand LSTM based models. This fits neatly within the context of recent calls for studies to interpret machine learning methods (Nearing et al 2020, Beven 2020). While the techniques themselves (triple collocation analysis, integrated gradients, ablation studies) are not novel, the ability to use deep learning models to better understand hydrological datasets (or processes) is certainly a growing and important direction for this subfield of hydrology. This work demonstrates one use of these methods, and clearly meets the objectives set out in the introduction.

Overall, the research manuscript meets the aims of HESS and advances hydrological modelling in three ways:

1) by demonstrating the ability to utilize information from multiple input data sources, without a priori information about the reliability of this data.
2) by demonstrating techniques for interpreting LSTM based models.
3) by further improving rainfall-runoff model accuracy and providing competitive benchmarks for future rainfall-runoff modelling studies.

References:
Beven, Keith. "Deep learning, hydrological processes and the uniqueness of place." Hydrological Processes 34.16 (2020): 3608-3613.
Kratzert, Frederik, et al. "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets." Hydrology and Earth System Sciences 23.12 (2019): 5089-5110.
Nearing, Grey S., et al. "What role does hydrological science play in the age of machine learning?." Water Resources Research (2020): e2020WR028091.

Specific comments

I was grateful for the following:

1) The paper is focused and the structure is clear. The subsection titles signpost exactly what the reader is expecting to find.
2) The thoroughness of the description and analysis in Appendix E - Analysis of Precipitation uncertainty is great. The experiment using triple collocation analysis is very well described and overall, this section is an exemplar of a valuable appendix.
3) The comparison with both the SAC-SMA and LSTMs trained with single meteorological products was useful to demonstrate that the traditional hydrological models are not able to utilise this information as effectively as the deep learning methods. While this is not surprising it would have been easy to exclude it.
4) I was very keen to download and play with the code available on github. Indeed, I was able to download and run the models as described and am always impressed when results are made available and reproducible like this. Thank you very much!

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 data products 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.”

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.

Formatting

These are a list of small formatting / spelling errors.

P5 L103: Space between “(1)all” -> “(1) all”

P8 L178: Spelling “calbrated” -> “calibrated”

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

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. If Appendix E becomes Appendix C, update Figure E1. to Figure C1)

Suggestions

Feel free to incorporate these or to ignore them. I have tried to offer my best suggestion for how a 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"

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.

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 than the 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.

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."

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)”

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.

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 data products” (or something more correct along these lines!)

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“