The comments of the reviewer are written black, our answers in purple.

Anonymous Referee #1

The paper describes the use of deep learning rainfall-runoff models based on LongShort Term Memory networks for combining multiple forcing products and improve the model accuracy relative to using only individual input datasets. The approach is demonstrated over 531 basins in the CAMELS dataset. Overall, the approach is technically sound, the manuscript is very well written, and the general topic is interesting for HESS readership. However, there are a few of points that I would recommend to clarify before the paper is accepted for publications.

We want to thank Reviewer #1 for their sincere comments and suggestions. We made two major changes based on this review:

- The first was to add a new set of benchmarks related to how multiple forcings inputs are used in a traditional hydrological modeling situation.
- The second was to shorten the manuscript by moving a lot of the existing analysis to supplementary material and reorganizing the introduction to speak more clearly to the main point of the paper, which is leveraging multiple forcing products to help address challenges in existing methods.

1. The main contribution of the paper should be better contextualised with respect to the existing (and fast growing) literature on the topic. The current manuscript introduction is indeed relatively short (i.e. 30 lines) and only introduces the purpose of this study without illustrating other existing methods. while I found the idea of the proposed approach interesting, neither the use of deep learning hydrologic models or the idea of data fusion is completely new and, therefore, the paper will benefit from a critical analysis of existing methods and how the proposed model is advancing the state of the art. Moreover, I would recommend to better clarify the novel contribution of this paper wrt the sequence of previous publications by the same authors using LSTMs for rainfall-runoff models (I'm not saying this paper is not advancing the previous ones, but considering also the concerns related to the benchmarking discussed at point 2 I believe the authors should clearly demonstrate that the contribution of this paper is beyond the "minimum publication unit").

We agree with this assessment. The introduction in our original submission was short and missing a clear statement about why this manuscript is clearly advancing over previous publications. In the revised manuscript we include new introductory material that outlines challenges related to leveraging multiple inputs in traditional hydrology models as well as related literature. We expect that this, along with the added benchmarks related to these traditional methods (see answer to remark 2), will help clarify the new contribution presented in this manuscript.

2. The set up of the benchmarking analysis is not fully convincing as the authors are comparing their model accuracy against (A) models calibrated using a single product and (B) traditional hydrologic models from Kratzert et al. (2019b). While the first analysis is the core of the paper, I don't understand the reason for the second one for two main reasons: in Kratzert et al. (2019b) the authors have already demonstrated the superiority of LSTMs wrt standard hydrologic model; if the new models that combines multiple inputs outperform the LSTMs using a single forcing as shown in (A), it comes straight that the new models also perform better than standard hydrologic models. In addition, this second benchmarking might confuse some readers who may attribute the reported improvements to the combination of inputs, whereas they are mostly due to the model structure. Rather than the comparison with traditional hydrologic models (which cannot use multiple meteo forcing data as the LSTMs), I would suggest the paper will benefit much more from a benchmarking against other state-of-the-art data driven models.

We do understand why one would come to these conclusions. Nevertheless, we believe that the model comparison in Figure 4 is important, since it contextualizes and highlights the improvement we see due to using multiple inputs in a single LSTM. It gives a sense of how much this improvement really is (the multi-forcing LSTM almost - not quite - doubles the performance gap between LSTM-based models and traditional hydrological models).

We added this analysis to contextualize the results of our current manuscript to our previous studies, where we trained LSTMs just on a single forcing product. The purpose of the hydrological benchmark models is to highlight the improvement of the model performance over single-forcing LSTMs .

However, we agree with the reviewer that including a different set of benchmarks improves the manuscript. In the revised manuscript (uploaded on invitation by the editor) we benchmarked against arguably the most common method of using multiple forcing products in the context of traditional hydrological models, which is to train separate hydrological models for each forcing product, and to combine their outputs using ensembling techniques. We used the SAC-SMA + Snow-17 model, which is used for operational forecasting in the US and was also the model originally included in the CAMELS data set. To account for stochasticity in the optimization process, we calibrated multiple models per basin and forcing (similar to what was done in the original CAMELS paper by Addor et al.). The code and simulation outputs will be made available.

Regarding adding different state-of-the-art data driven models: We are not aware of any other data-driven modeling approach (something that is not based on LSTMs) that yields similar performance for regional/continental modeling tasks (i.e. one model that predicts discharge everywhere) and can thus also be applied to forecasting (e.g., PUB,as was shown for the LSTMs in one of our previous publications).

3. Lastly, the paper is in my opinion a bit lengthy with 14 figures that make the narrative a bit scattered. I would then suggest to explore the option of selecting the main findings-figures worth to be discussed in the main paper (e.g. Fig. 6 and 7) and move some content to a supplementary material.

Thanks for the suggestion. We agree with this assessment and thus moved a lot of material from the original manuscript to supplementary sections.