AC Comment to RC2

We are thankful for the referee's valuable time in helping to improve the manuscript. This document will address all reviewer comments (gray boxes). We believe that the proposed changes will majorly improve the manuscript.

Considering the referee's comments, our planned improvements for the revised submission can be summarized as follows:

(1) We will conduct and present a more in-depth statistical analyses of the residuals, as also shown in the Literature suggested by the referee. As the referee pointed out, many studies report a high degree of heteroscedasticity in the residuals. Our analysis indicate the same (see evaluations below). To address this, we will apply a Box-Cox transformation to the data to decrease the degree of heteroscedasticity. Furthermore, we will investigate and report the ACF and PACF and also use this information for redefining the search space of our ARIMA model (see also the comment from referee 1).

(2) We agree with the referee that the HLSTM-PBHM does not perform a direct error correction as not the residuals are forecasted but the discharge directly. We will address this accordingly in the relevant sections of the manuscript (especially the title).

In our opinion, the above presented modifications address all major comments of referee 2 and will majorly benefit the manuscript.

> This study proposes an application of Long Short-Term Memory Networks (LSTM) complemented by the results of a hydrological model (PBHM) for operational flood forecasting in a smaller mountainous catchment. The performance of the resulting HLSTM-PBHM is compared with an ARIMA error correction model and a standalone application of the LSTM (HLSTM).
>
> The results of this study are particularly significant, as they reveal performance improvements for the HLSTM-PBHM, especially for larger lead times. These findings have practical implications for flood forecasting in similar catchments.
>
> The paper is within the scope and very interesting for the readers of HESS. The authors address a topic of high relevance for flood forecasting since studies focusing on small catchments and requiring sub-daily time steps are limited.
>
> The authors have done a commendable job of presenting the scientific results concisely and well-structured. However, I have some fundamental comments on the interpretation of the proposed method and the concept of the experimental design to compare the different approaches:

> From my perspective, the proposed HLSTM-PBHM is an informed approach that uses precalculated results of the hydrological model (PBHM) combined with observations for the hindcast rather than applying an explicit error correction as the ARIMA error correction model does. Therefore, the title of the paper should reflect this, and I suggest revising it.

We agree with the referee's comment that, in principle, the ARIMA and LSTM models follow a different mythology of how the final forecasts are obtained. ARIMA forecasts the residuals, which are used to correct the simulations of the hydrologic model, while the LSTM uses the hydrologic modelling results as a feature to directly forecast the runoff. This implies that the LSTM does not apply an error correction, which should be revised in the relevant sections of the manuscript, especially the title. We will thus change the title and also make this fact more clear in other sections of the manuscript.

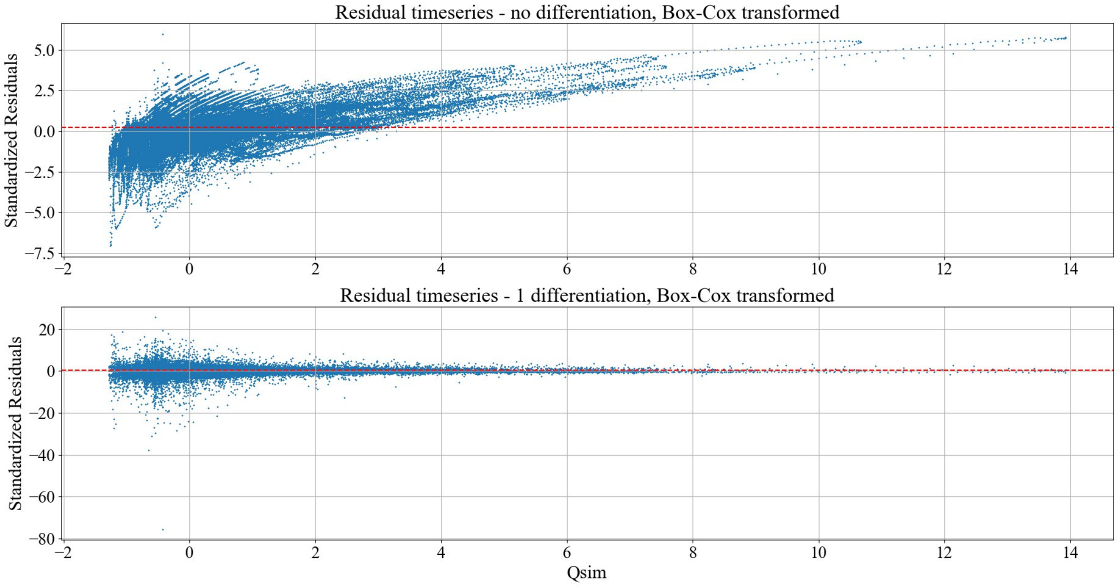
> This approach's consequence is that the input data used and the internal corrections of HLSTM-PBHM cannot be compared with the residual errors of the hydrological model and the corrections calculated by the explicit error correction models.

It has to be mentioned that we previously tried a residual LSTM model. However, the prediction accuracy was inferior compared to the presented LSTM variant that directly predicted the streamflow. The reason for this might be that the predictions of the hydrologic model (in this study) are very poor when for instance compared to the study cited by the referee, where the original model's performance was already quite good. This leads to the fact that the LSTM does not rely that much on the simulated stream flow in some conditions (i.e., at peak runoff). This was demonstrated in our reply to RC1 and will be included in the final manuscript.

A general shortcoming of comparability between ARIMA and the HLSTM-PBHM was also pointed out by referee 1. We will address this issue by introducing an intermediate model that uses the same data as ARIMA and has the same architecture as the HLSTM-PBHM. This, however, still does not resolve the fact that residual errors across the models cannot be compared like it was done in the literature suggested by the referee. One option would be to compute the LSTM's residuals **after** the streamflow was forecasted. However, we do believe that a more objective way (for this specific study!) is still comparing the final forecasts of each model. However, we will make the fact that one model forecasts residuals and one runoff directly more clear throughout the manuscript.

> In general, a comprehensive analysis of the residual errors of the PBMH model, e.g., the underlying statistical distribution, would be helpful and give the reader more insight to interpret the results. It would also prove the assumption of whether the errors are normally distributed. Many studies (among them [1]) found a high heteroscedasticity variance of residuals, which should be checked and considered for the residuals in the study.

We will add the residual statistics of the PBHM to address the comment of referee 2. Similar to other studies, the literature pointed out by the referee amongst them, we also found a high degree of heteroscedasticity in the residuals. We will reduce the degree of heteroscedasticity by applying a Box-Cox transformation as suggested in the literature presented by the referee (see figure below).



We will also report on the autocorrelation functions (PACF and ACF) and use this information for redefining the search space for ARIMA but also to discuss the implications on the LSTM as shown in the literature suggested by the referee.

> Please also briefly introduce the PBHM model in the Methods Section as it is used in the study.

We will include more information on the PBHM in the revised manuscript such that the reader gets a general picture of the model without having to go to the cited literature.

> Multiple errors exist in flood forecasting due to meteorological uncertainties and those rising from the structure and parametrization of the hydrological model. Please elaborate on how the different contributions could be considered in future developments of HLSTM-PBHM in the discussion.

We agree with the referee that a variety of uncertainties exist in operational flood forecasting (e.g., meteorology, streamflow observations and all uncertainties concerning the PBHM). As for meteorologic uncertainties, we specifically chose the hindcast-forecast architecture to address some uncertainties and characteristics of the meteorologic input data. First, the chosen architecture provides a more or less straightforward way of including meteorologic ensemble predictions in the future. This can be achieved by simply modifying the forecast LSTM, whilst no modification of the hindcast LSTM is required. Second, the chosen architecture allows for a clear distinction between meteorologic hindcast and forecast data. This is especially relevant, since often quite large differences between meteorologic hindcast data and forecast data are present. The reason for this being that the hindcast data incorporates observations (ground stations and radar), whilst he forecasts heavily rely on numeric weather predictions. This leads to the fact that the forecast LSTM also has the potential to learn from uncertainties in the meteorologic forecasts. Unfortunately, for this study we did not have meteorologic forecasts available but addressing this topic is planned in the future. As for the hydrologic model, it is evident, that some underlying problems exist, given its poor performance. Noteworthy, this specific catchment was part of a broader study, where multiple catchments were modelled. Interestingly, most neighbouring catchments achieved quite a high model accuracy (NSE values slightly below and above 0.8) – calibration was done equally. In our opinion the poor performance of this catchment is a result of various uncertainties, one definitely being the inability of the employed HBV model to capture some important process in the catchment.

Furthermore, changing environmental conditions can lead to a change of the outputs of the hydrologic model (which the ML model has not yet seen), i.e., the calibrated parameters are not well suited anymore to depict the catchment processes. In our opinion an effective, yet simple, way to reduce these uncertainties is regular retraining of the LSTM model. This may be done automatically, or manually.