Response to reviewer #1

We thank the reviewer for the time and efforts in reviewing the manuscript. Below is our response where the reviewer comments are in black, our response is in green, and the copied text from the manuscript is in blue.

This paper presents a complete workflow for an operational flood forecasting and mapping case study.

The paper details all the critical steps in an operational system for data scarce regions, including remote sensing data integration, forecasting, inundation mapping, and communicating forecasts to the public.

We thank the reviewer for this encouraging statement.

The model performance comparison against linear regression and simple hydraulic models is not a very challenging task compared to evaluating performance against better performing ML and physical models cited in the paper.

As a start, it is important to emphasize that almost all physically-based, process-based, and even most of the ML-based, hydrological models in research and in operation are not using past water stages as their input, while for our models this is the main input data. This input makes a major difference for large gauged rivers (which is the focus of our system) and increases the prediction accuracy significantly. Even the Linear model, with past river stage input, has a median NSE >0.98. Therefore, it is not a “fair” comparison to evaluate our stage forecasting models against other hydrological models that do not use the past stage data as input. We realize this is an important point that needs to be better emphasized in the paper. Therefore, we will add the following paragraph to the revised version at the beginning of Section 2.2:

“The primary input to the stage forecast models is past river stages and the output is the future river stage. It should be noted that in this aspect they differ from many standard hydrological models where the input does not include river stages but mainly forcing data and watershed properties.”

As an ML stage forecast model we adapted the LSTM. This model is the most accurate one in the published literature to our operational use case, with some modifications that were necessary. The Linear model, which was used in earlier operational versions, was kept for

very few locations where the LSTM results were not sufficient. As already explained above, even the relatively simple Linear model shows high skills (Table 2, Figure 8) and, as we state in the paper, this is "...not surprising given that we are modeling large rivers with a typically slow response and that we use historic water stages as well as upstream water stages as model inputs". Therefore, it was important to show the added value for the transition from the Linear model into the, much more complex, LSTM model. For this reason these models are compared.

The physically-based Hydraulic model presented in the paper was used operationally in earlier versions of the warning system (it was operational in India during the 2018 and 2019 monsoon seasons). The model is based on the numerical scheme of the hydrodynamic model proposed in de Almeida et al. (2012) which is a modification of the simplified inertial formulation for the 2D shallow water model (Bates et al.,2010). Such hydraulic models constitute the vast majority of existing operational inundation models (e.g. using DHI's MIKE, Deltares' Delft-FEWS, etc.) and was a good choice for our operational framework that involves hundreds of stream gauges and river stage data (rather than discharge).

While used operationally, large efforts were made to use the Hydraulic model with the best parametrization and to improve its accuracy. But, because the comparisons described in this paper showed the new ML-based inundation models to be superior in both accuracy, scalability, and computational costs, the Hydraulic model is currently not operational. It was important to demonstrate in the paper the improvement in accuracy achieved with the ML inundation model, compared to the ubiquitously-used Hydraulic model, and this is the reason we included this part in the paper. True, there are other physically-based hydraulic models that can be tested against the ML inundation models we have used, and possibly their accuracy would be higher. But given the scalability limitation and computational cost of such models, we do not think adopting these models to our system and adding more comparisons with those models would lead to any important findings that justify such large efforts.

The performances of our models can be judged by the metrics we present for the validation data sets in terms of forecasted stages and flood inundations (Table 2, Figure 8, Figure 10). Both are showing good results. We were looking for other publications with metrics from operational systems, in order to evaluate our system, but there are only a few studies that include such metrics. We refer to this issue in the discussion.

The spatial and temporal resolution of the input data for rainfall, elevation, and other datasets is rather low to achieve comparable forecasts, but still useful in data scarce regions with limited resources.

For rainfall in data-scarce regions, satellite-based products are the most practical option. We are using the IMERG data that has an original resolution of 0.1° and 30-min, which is

among the highest available for real-time or near real-time global rainfall products. The rainfall data are further averaged in space over the watersheds area and in time to match the hourly resolution (for the implementation in India and Bangladesh). While this lowering of spatial resolution may be debatable, we emphasize two important points: 1) as many hydrological studies have shown, there is a tradeoff in spatially averaging precipitation data. Surely, floods are sensitive to the spatial patterns of precipitation but at the same time unsystematic errors in this (often inaccurate) information are reduced with averaging; 2) as opposed to standard hydrological models, where precipitation is the most important input, for the present system the main input is past river stages (from the target gauge and its upstream gauges). Precipitation input improves the prediction accuracy, on top of the accuracy achieved with past stages (Table 2, Figure 8), but the model is already quite accurate even without precipitation input. In this situation, it is better to keep the precipitation spatially averaged, since adding many precipitation features would have considerably increased the complexity of the model.

Elevation data, used in this system for inundation modeling, is actually of quite a high resolution compared to other publicly available global DEMs. We use a 1-meter resolution DEM, while SRTM is a 30-meter resolution DEM. Lidar-based DEMs can get to a much higher resolution but these are not currently available globally. The high-resolution DEM is produced by Google and is being kept up to date. Through training and validation of the inundation models, it was found out that 16 m resolution achieves similar accuracy to using higher-resolution versions of the DEM but at vastly lower computational costs.

Following the reviewer's comment, we noticed that we did not write the base resolution of the DEM. We will add these details in the revised manuscript in Section 2.3 under DEMs for target gauge AOIs:

“Consequently, higher-resolution (1x1 meter^2), up-to-date DEMs are constructed for each AOI…”

Some of the details, like how AOI pixels are defined in the flooding regions, are not clear in the paper.

Indeed this detail is missing. We will add to the revised manuscript the following text about AOIs (as a part of Section 2.5):

“AOIs for the target gauges were defined as polygons around the river at the gauge location. They were computed, by both manual and automatic procedures, by observing the flood patterns in the area around the gauge from historical flood inundations maps.”

We will also add some missing details requested by reviewer #2 about training and validation data sets and number of flood events in training and validation, which we are highlighting here as well for context, as follows:

In Section 2.2 (Stage forecast modeling) under “Training and validation” we will add: “The training and validation data sets are composed of samples where the features are past river stages at the target and upstream gauges and past spatially-averaged precipitation (for the LSTM model). The labels for training are future river stages at the target gauge for a given lead time.”

In Section 2.3 (Inundation modeling) under “Training and validation” we will modify the present text to clarify the structure of the training and validation data sets: “The training and validation data sets for the inundation models are composed of samples representing historical flood events where the features are gauge water stage measurements and the labels are the corresponding flood inundation extent maps from satellite data.”

We provided the numbers of flood events used for training the ML inundation models for the 2021 implementation in Section 2.5 (although they would be corrected in the revised manuscript). But we did not provide the numbers for the cross-validation analysis in Section 3. We will add these details in the revised manuscript: “...in these years there are a total of 4815 flood events across all the AOIs (on average 34 events were used for training and 10 for validation per AOI).”