Response to reviewer #2

We thank the reviewer for the time and effort and for the detailed review report that will help us improve the paper. Below is our response where the reviewer comments are in black, the response is in green, and the copied text from the manuscript is in blue.

This paper presents ML models that respectively (i) directly predicts the river stage (rather than predicting discharge and then translating to stage); (ii) predict wet/dry of pixels depending on gauge stage; (iii) and estimate flood inundation depth. Among these, (i) was trained based on historical stream gauge data and near-real-time upstream gauges; (ii) was trained using historical satellite data and coincident stream gauge height data. (iii) was not really a model, per say, but an interpolation procedure.

I think the paper demonstrated strong performance from a completely data-driven model. It highlights the idea of directly simulating stream gauge height, which breaks many barriers. If they didn't do this, they need to simulation discharge and then resolve the highly-variable (in space) relationship between discharge and stage height. Most of time we cannot resolve it. In the authors' case, there is no discharge data to begin with. So directly tackling gaging height is a good and necessary idea (but it also leads to some issues I will discuss below). The paper also demonstrates a very efficient forecasting scheme based on upstream gauge data. The whole paper demonstrated how to stack different models together. The authors also showed a unique flood inundation component that is accurate. The work is very useful for hundreds of millions of people and it takes lots of courage to take on such a responsibility.

We thank the reviewer for highlighting this paper's contributions.

While there are many reasons why I like this paper and I encourage the publication of this paper, I also noticed a few major issues. These issues are raised here in the hope to make the manuscript more balanced and comprehensive.

Thank you for pointing out these issues. Your comments are very helpful and helped us to improve the paper. Please see below our point-by-point response to each of them.

(a1) There should be some discussion of the potential scientific limitations (even if caused by practical data availability) of the approach and the conditions under which this approach is applicable. As far as I can see, all the models were posed in a highly case-specific way. The gauge height LSTM model has weights that are shared across multiple gauges but it also needs gauge-specific weights that are tuned to local data with a particular configuration. (how much worse will it get if you don't use those gage-specific weights?) The inundation extent model is tied to the gauge and the particular river bathymetry downstream from that gauge. In other words, it seems these models can only be applied where gauge data is available for training. The trained relationship is not portable anywhere else (if so, it poses a requirement on the available data records). Don't get me wrong. I think the model is highly useful operationally. In India there are many places where the model is applicable. It just might make sense, if these limitations are true, authors can discuss where and when this model formulation is valid so it is easier for the readers to understand if these algorithms are sound for their purpose. Maybe they can come up with a more uniform model and show its accuracy.

Thank you for this comment. Indeed, as we already emphasize in the paper, and would be better clarified in the revised version, this model is aimed at gauged locations in relatively large, slow-responding rivers. Indeed, the LSTM model training has both shared and gauge-specific weights. Gauge-specific weights are required for incorporating upstream gauge data as input, which is a very informative input. Without gauge-specific weights, the LSTM performs much worse. It is also true that the inundation model is also gauge-specific. This indeed requires having some data for training (past stage data and flood inundation from satellite). We do not claim these weights can be transferred to other locations, only that adding new locations which meet the above conditions is easy, or in other words, the models are scalable. See our response to the next comment concerning scalability.

We agree the limitations need to be stated clearly and therefore we will add the following paragraph to the discussion section in the revised manuscript:

"The presented system has some limitations that need to be emphasized. First, Google's flood warning system in its present form is designed for flood forecasting in gauged river locations. Specifically, this requires, at a minimum, river gauges providing their stage data in real-time that have records of historical stage data and a few cases of flood events to be used for training and validation of inundation models (here, records of 6-7 years were used). Second, the stage forecasting and inundation models would work the best in slow-responding, large rivers. Apparently, even with these restrictions, such flood warning systems can be useful for a large population worldwide. We are currently working on extending the system to accomplish two additional goals:..."

Actually, our plans (which are already underway) include expanding the model to ungauged locations, and improving performance on flashier rivers. This is already explained in the discussion section in the current version of the paper.

We are not sure we understand what the reviewer means in a “more uniform model”. If the meaning is a model with the same configuration for all gauges then our model indeed has the same configuration across all gauges, albeit with some parameters that are gauge specific. This is similar to any hydrological model that has some watershed-specific parameters. If the reviewer meant something else, we would love to better understand and address the concern.

(a2) This point also contradicts the authors’ claim that the model is highly scalable. You cannot take the model to a new terrain and directly apply it. In addition, the learned relationships may not always stand --- what if you have heavy rainfall in the region between your upstream gauge and the gauge of interest? It seems your model cannot consider such forcings (this may not matter that much for large-scale Indian monsoons, but it could be important elsewhere). This means, while the model is fast to run, it is not scalable in the sense of expanding to new areas ---- you must spend the time and effort to collect the data and train the model in every new area of interest, and that is assuming you are lucky enough to have the data. Hence, it is uncertain how the authors intend to use the model on large areas.

Thanks for raising this point. Please note we did not use the term “highly scalable”. The term “scalable” appears two times in the paper, one in the introduction describing the findings from previous studies and the second when we explain that the ML inundation models were found more scalable than the Hydraulic model. Nevertheless, we do claim the ML models described in this paper are generally scalable in the sense explained below.

We distinguish between scalability in the sense of applicability to different regimes and scalability in the sense of how easy it is to deploy the models at enormous scales within the applicable regimes. Here we refer to the latter, while for the former we clarified above (response to comment a1) what conditions our model is applicable to. When comparing scalability of our model to other hydrological and hydraulic models the tradeoff with accuracy needs to be remembered. Many hydrological and hydraulic models were developed to work with static (e.g., DEM, soils, landuse) and meteorological forcing data (e.g., precipitation, temperature) as input. If no calibration is needed, such models in principle have high scalability. However, without any calibration the accuracy of those models is very low, as was discussed in many hydrological publications. If high accuracy is needed, calibration against observed records is required, and this is known to be a demanding process. In particular operational frameworks require high accuracy models and thus models’ calibration is an essential and important step in their deployment. We compare to those models when claiming an improvement to scalability.

As a measure for scalability, one can assess the time and effort it takes to deploy the system to a new large region (e.g., country, with hundreds of gauges), within the applicable

regime. Requirements for data collection and quality control are identical for all models using gauge data and are therefore not included in the assessment. For the presented flood warning system we conservatively estimate that the ML stage forecast and inundation models deployment would take for such a new region all together about 6-7 days of CPU time and manual work. On the other hand, the deployment and calibration of standard hydrological and hydraulic models to hundreds of new gauges would most probably take significantly longer.

It is also important to emphasize that the flood warning system is certainly intended to cover large areas. First, the system in its current version already covers 450K sq. kms and a population of 360 million, and is being extended to several new countries (in Asia, South America and elsewhere), including some that are not in the Indian Monsoon climatic regime. In parallel, we are actively developing a modified generalized global model which can be deployed both in gauged and in ungauged locations. This effort is mentioned in the paper, but we do not yet have results for this model to report.

Concerning the reviewer's point of heavy rainfall between the upstream gauge and the target gauge: although a situation as the reviewer describes can happen (though not commonly), we think the model is set up to be able to handle most of these cases as well. The main reason is that, in addition to the input from upstream gauges, the model input includes the watershed-averaged precipitation and the target gauge past stages. The situation the reviewer describes can typically happen when the upstream gauge is at a relatively large distance from the target gauge, which implies the additional watershed area between the upstream and the target gauges cannot be too small. For such a rainfall event to cause a large rise of the river stage at the target gauge location (especially in the large rivers we address) it must produce substantial rain amounts over this downstream watershed area. It is therefore probable this rainfall would produce some signal in the watershed-averaged precipitation, even though it does not cover the entire watershed. It is also probable that the target gauge will start rising due to the rainfall event. The signals in both inputs (i.e., precipitation and past target gauge stage) are likely to lead to a forecast of a general river rise. It is true such an event would have a shorter response time compared to more spread rainfall and it is possible such an event can be forecasted only with shorter lead times than the lead time selected for the gauge - yet it's worth noting that our system includes support for automatically shortening lead times in such events. The scenarios above are currently speculated and analyses are required to confirm their validity.

(a3) It also exerts some constrain on the eligibility of sites. Because you have to train a site-specific model, you can only use sites with long-enough records to train the model. The model cannot be large, and information from other sites do not help with a particular gauge of interest.

This is true but the required record length is not large. In this study, the training data set included data for 7 years for the stage forecast model and 5 years (i.e., flood events from this period) for the inundation model. The cross-validation analysis (presented in Section 3) was based on training records of 6 and 4 years for the two models, respectively. We clarify this point in the new paragraph we will add to the discussion Section (see text in the response to comment a1 above).

“The model cannot be large”: this is true, ensuring the model does not overfit to the data is a key concern. The architecture presented is built specifically to address this concern - the shared weights among all locations allow for a much larger and more complex model (since it is trained across sites), while the per-gauge weights allow for some site-specific customization but are kept small. Our validation and test results show that the models do not overfit to the data.

“information from other sites do not help with a particular gauge of interest”: we do not agree with this statement. The LSTMs are shared models, and as such allow the model for each target gauge to benefit from the training data of all gauges. This point relates to the previous ones - each specific gauge does not have enough historical training data to train a model as complex as LSTM - these models aggressively overfit when trained on individual gauges. However, the shared LSTM model allows the architecture to model complex rainfall-runoff patterns without overfitting.

(a4) If my understanding is incorrect, I stand corrected and the authors can show a test case where the model is applied to an “ungauged” location.

The reviewer is correct. This flood warning system indeed focuses on gauged locations. We wrote this in the current version of the paper but will better emphasize this in the revised version with the new discussion paragraph presented above.

(b) The training dataset for the models were not clearly described. For the inundation extent model, there should be descriptions of how many events were included as training and test images.

Thank you, these details will be added to the revisions 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).”

(c) It is not clear if the model accuracy drops as we go further downstream from the gauge. Some exploration here will be useful.

It is not clear if the reviewer refers here to the stage forecast model or to the inundation model. The stage forecast is only for the gauge location so we assume that the reviewer’s comment refers to inundation modeling. We did not explore this issue, but we visually examined numerous flood inundation examples and did not identify such signal within the scale of the current AOIs. Nevertheless, we will do a small test to decide whether such analysis is worthwhile to pursue and if a trend will be recognized we will present it in the revised manuscript.

(d) regarding authors’ criticism on the hydraulic model --- are we sure you feed it the best parameters and inputs? There is no description about calibration. Back to point (a), in a region without past observations, the hydraulic model may still function but the ML inundation model may not --- which means these models have their own use cases. If I’m wrong please correct.

Indeed many efforts were made to calibrate and benchmark the Hydraulic model, especially during the first years of the project when it was a part of the operational system (until 2019). We have tested our calibrated hydraulic models against standardized benchmarks, and paid well established third-party hydraulic engineering companies to set up and manually calibrate hydraulic models using standard platforms (e.g. Delft-FEWS) on our areas of interest - achieving similar metrics and forecasts to our own hydraulic models. Our analysis has also shown that the inundation pattern from the Hydraulic model is most sensitive to the upstream discharge and the downstream normal slope, while the effect of roughness coefficients on the inundation was much weaker and interacted with the other factors. Therefore, we ended up fixing the roughness coefficient per pixel (one of two values,

representing either a river-bed or a non-river-bed pixel) and optimized the upstream discharge and, if needed, also the downstream slope. We ran about 1000 simulations per AOI to optimize those parameters. One cannot guarantee that those are the best parameters, as the optimization of such models is a complex process. But given all simulations, we assume the optimization is quite good and it is fair to compare the optimized Hydraulic model with the ML models. The reviewer is correct that we do not elaborate on the Hydraulic model optimization. We do so for the sake of conciseness since we are not using it in the operational system.

The reviewer raises a fair point that the Hydraulic model could be useful for locations without data on past flood inundations, while the ML models are not applicable in such cases. This would be correct in cases where the gauge input is discharge, rather than the stage. However, with stage input data the Hydraulic model must be optimized for the relation between stage and inundation pattern and therefore requires records of historical flood inundation data and their respective gauge stage measurements, similarly to the ML models. As the reviewer points out in this same comment, the Hydraulic model often requires some level of calibration, unless some regional parametrization can be used. Ideally, if high quality, high resolution DEMs (including bathymetry) and real-time discharge data are available and if regional parametrization is applicable, Hydraulic models can indeed perform well. But in reality, this is a rare situation, and in most gauged sites the reliable application of the Hydraulic model, same as ML models, would require historical flood data.

(e) there seemed to be no description of network configurations such as hyperparameters, hidden size, minibatch (maybe there is not a minibatch), training epochs, etc.

Thank you. We will add a new table with all hyperparameters to the revised manuscript.

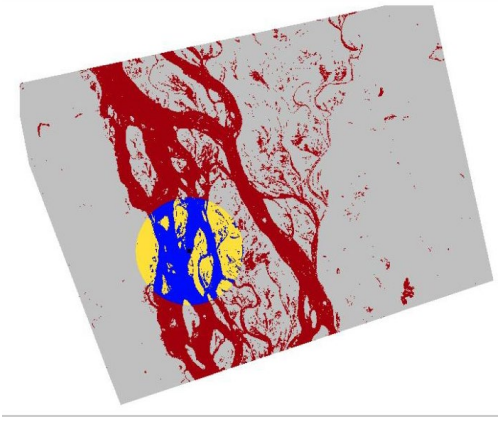
(g) does it make sense to average precipitation for a drainage area > 100,000 km^2?

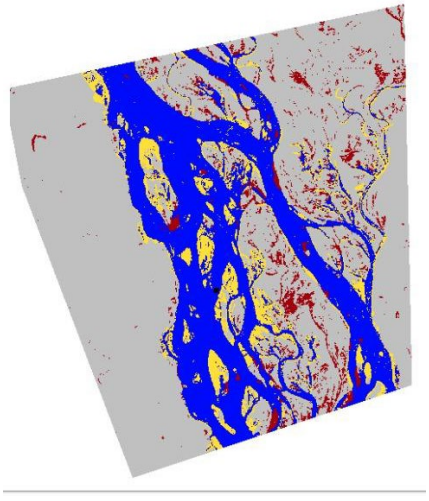
This is a good point. As many hydrological studies have shown, there is a tradeoff in spatially averaging precipitation data. Surely, floods are sensitive to the spatial patterns of precipitation but at the same time errors in this often inaccurate information are dumped out with averaging. However, as opposed to standard hydrological models where precipitation is the most important input, for the present system the main input signal is from 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, as shown in section 3.1, but the model is already quite accurate even without precipitation input. In this situation, it is better to keep the precipitation spatially averaged rather than adding many precipitation features and increasing the model complexity. Even when averaged over such large areas

of hundreds of thousands square km, the mean-areal precipitation time series has value as it is correlated with the main precipitation systems that lead to an increase of these large river stages. It might be that in a very different climatic regime, where precipitation systems are substantially smaller (e.g., in the arid regions), or if past river stage data are not used as input, such an average would not make sense and another strategy would be taken. Indeed, in our efforts towards extending the models to more climatic regimes and to ungauged locations we are currently exploring averaging precipitation over sub-basins, but cannot report the results yet.

(h) We have no intuitive understanding of what F metrics mean. Do you mind showing some observed vs simulated maps for different values of the F metric?

Below are two maps for the same flood event with two different models. On the left, a trivial "model" simply assumes an inundated circular area, while the model on the right is a result of one of our inundation models. The colors represent true-positive (hits) in blue, false-positive (false-alarm) in yellow, false-negative (miss) in red, and true-negative in gray. The map on the left has: precision=53.8%, recall=16.4%, F1=25.1%. The map on the right has: precision=76.3%, recall=85.4%, F1=80.6%. Note that the map on the right has some gray area that was cut out, but true-negatives are not included in the computation of the F1-score so it does not affect this metric. We hope this is helpful. We will provide these examples in the supplementaries of the revised manuscript.





(i) the flooding depth model was never tested and we do not know its accuracy. Can you talk about its value in the real world? Also, low-resolution could also give you discontinuity.

We agree with this point. In fact, we are putting considerable effort to collect ground truth depth data and get feedback from people in the affected regions via surveys in collaboration with local organizations. Unfortunately, the data we collected so far is too limited and too scattered - due to COVID limitations and other hurdles. Therefore we cannot yet present its accuracy, as we emphasize in the paper. We are continually investing in collecting more data to validate flood inundation and depth. To better clarify this point we will add the following discussion paragraph to the revised manuscript:

"A few studies examined the effectiveness of flood alerts in operational frameworks. For example, Rotach et al. (2009), as a part of an end-to-end operational flood warning system in the Alpine region, collected feedback from end-users through questionnaires, interviews, and workshops, and some initial insights are given on the utility of the system for the decision-makers and how well the information was perceived. It should be noted, however, that the current knowledge about effective flood warnings in countries like India and Bangladesh is very limited. For example, a literature review by Keller et al. (2021) shows that the large majority of the published literature about this topic focused on industrial countries while less than 6% focused on Asia and none on South America or Africa; they emphasize that little is known about the transferability of findings from industrial to non-industrial countries. An important input to these investigations is feedback from the population and from local aid organizations on whether alerts were received, how accurate they were (in terms of flood inundation and flood depth), how useful they were and what actions have been taken. Our research efforts are ongoing in this direction and the analysis indicates flood alerts being effective. The full details of this research would be reported separately and are expected to help both in validating and in improving the flood warning system."

The reviewer is correct about the possibility of discontinuity in inundation depth due to the discretization to 16 meter pixels. This resolution was however found appropriate in providing reliable inundation prediction; the pixelated pattern better represents actually the information these maps provide. Furthermore, the feedback we received so far from governments and NGOs doesn't indicate a need for a higher resolution than that.

(j) can this study be reproduced at all? It seems not much of the study can be reproduced or even compared to in terms of data. All the code and data are either proprietary or unavailable. We were just told they could do this and do that and there is no possible path to trying most of the steps here.

We agree this is an unfavorable situation. We have no solution for the data, which we are not allowed to distribute, and the system code which is proprietary Google code. But we do want to make the models reproducible as much as we can, and in particular, the new ML inundation model presented in this paper for the first time. Therefore, we plan to do our best

to write either a practically equivalent, standalone python code or accurate and reproducible pseudo-code for the Thresholding and Manifold models, which will be provided in supplementary with the revised manuscript. If one has samples of gauge stage and flood inundation maps (and DEM data for the Manifold model), the code can be used for training the models and computing flood inundation and potentially their depth.

Some minor points:

Line 158. What does “State handoff” mean?

State handoff means transferring the final hidden and cell states of one LSTM, through a fully connected layer, to the initial states of the next LSTM. These are represented by the chain: h(t), c(t) -> fully connected layer - > $h_0(t)$,$c_0(t)$ in the middle of Figure 2b. We can add this clarification in the paper if the reviewer recommends this. We can add this clarification in the revised paper if the reviewer recommends this.

Line 190. Should be “Quasi steady state” to be more exact

Thank you. We will modify this term in the revised paper.

Line 196. “Discarded” – see my point above, can you use a more gentle word?

We agree. The sentence would be changed in the revised paper to:

“...it was used in the operational framework in previous seasons, but it is currently not in use since in the present conditions it was found to be both less accurate and less scalable than the ML models…”

Line 198-199. “when the target gauge exceeds a (pixel-specific) threshold water stage. ” A bit confused. A gauge is just at one location, then why do you have a pixel-specific threshold linked to a gage? If it is pixel-specific, then you end up getting a map of different thresholds? Should it be image-specific thresholding?

The model results in a map of thresholds on the gauge stage. When the gauge data is higher than the threshold in a given pixel, this pixel is considered as wet. We do not think the term “image-specific” sufficiently clarifies this. To improve clarity, we will modify this text to:

"The model includes pixel-specific thresholds for each pixel in the AOI. Each pixel for which the water stage in the target gauge exceeds its threshold is assumed to be inundated (i.e., wet) while the pixels for which the target gauge is below their threshold are dry."

Line 219. Maybe I'm missing sth, although the thresholding model does not need DEM, it is tied to a particular gauge and the particular terrain/floodplain characteristics. It needs to be trained for each domain of interest using historical inundation extent and gauge height data, so it is not clear to me you can deploy to a new region without effort.

The sentence says: "Thresholding model requires almost no site-specific data like DEMs, and no manual work, making it appealing for large scale deployment across many AOIs in a short amount of time.". We do not claim that no data at all is needed, but the data required are satellite data (i.e., SAR images for a few historical floods), which are available globally (thanks to Sentinel-1 and others). We indeed also require stage data for the target gauge, and have revised to further emphasize our focus on gauged rivers in response to previous comments. Nevertheless, we agree that it is inaccurate writing "requires almost no site-specific data" and we will modify this sentence in the revised manuscript to:

"Thresholding model requires no DEM data, but only historical flood inundation maps and gauge stages for training. This makes it appealing for large scale deployment across many AOIs in a short amount of time."

Line 375 what happened to the flood and the effectiveness of the alert? You get us concerned but didn't say any outcome.

What we can currently say is that the alerts were sent to the public and to organizations, including CWC, which published it on Twitter (shown in Figure 7f). We were in touch with several NGOs, including the Yuganter organization, www.yuganter.org.in, that confirmed the flood events' occurrence (and sent us the photos presented in Figure 7d,e). However, we do not have yet information, which we agree is important, about actions that were taken in response to this specific flood alert. Evaluating actions on the ground requires completely distinct tools and methodologies (e.g. randomized controlled trials, field surveys, etc.). We are currently pursuing such research and hope to publish its results soon. We can share informally that preliminary results from a research collaboration with the Yale Economic Growth Center show that in places where our forecasts were distributed (through Yuganter volunteers) they have led to a statistically significant increase in protective actions compared to control cases. A discussion paragraph, presented in our response to comment i above, will be added to the revised manuscript to better clarify this point.