Response to reviewer #3

We thank the reviewer for the time and effort. The comments are very helpful. Below the general comments, we copied the reviewer's comments from the paper's pdf. We present the reviewer comments in black, the responses in green, and the copied text from the manuscript in blue.

I found your manuscript very well organized and for sure impressive for the amount of information managed.

To be honest, the scientific advances you present are not enough described to make this paper useful to the community. In my opinion, this manuscript would be a perfect contribution to be published almost as is for NHESS, but fails to some part meeting the scientific innovation standard I expect in HESS.

We respectfully understand the reviewer's point here, but we argue that the revised paper would be a valid and valuable contribution to HESS. The detailed and very useful comments from this reviewer and from the two other reviewers indeed highlight aspects that require modifications such as a better description or additional information, which we intend to address carefully in the revised manuscript (if we will be given the opportunity). We explain below how we intend to do this.

In the revised paper we believe the scientific innovations of this study will be clear. The most important are: 1) a full chain of flood warning systems with main core machine-learning models; 2) a new machine-learning flood inundation algorithm; 3) accuracy metrics from the "real world" of river stage and of flood inundation forecast. We believe these provide a valuable and relevant scientific contribution to HESS.

L22: How many were useful? I.e. dispatched with enough anticipation to allow for measures? How many false alarms? This is the key of success. I can also send an alert each time it rains and rivers are above a certain level at start of rainfall. Sending an alert is not a scientific measure of success

We divide our answer into two. First, regarding a scientific measure of success, note that alerts are sent based on the modeled inundation maps and river stage, both are validated against SAR inundation maps and observed stages, respectively. From cross-validation analysis, we can say something about the success or failure of our model to predict whether a given pixel is inundated or not, using the F1-score, which is computed from the elements in the contingency table, and combines hits, false alarms, and missed. Its median is 69%

(Table 2, Figure 10), which is quite good. The models also predict well the water stage at the gauge locations, based on cross-validation and NSE or NSE-persist metrics (Table 2, Figure 8). We do note claim that the number of alerts sent is a scientific measure of success, this provides an indication of scale.

The second part of the answer refers to the usefulness of the alerts. This is of course a very important issue; there is no point in sending alerts that aren't useful. 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 but it was not completed yet. 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.

In response to this comment and another comment by reviewer #2, we plan to add a discussion paragraph to the revised manuscript that explains this:

"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."

L49: Well, this is what we are doing for years. Most of our studies have only few years of data just because we analyze collected forecasts in operational environment:

a selection

Bogner K, Liechti K, Bernhard L, Monhart S,Zappa M. 2018. Skill of hydrological extended range forecasts for water resources management in Switzerland. Water Resources Management, 32(3), 969-984. http://doi.org/10.1007/s11269-017-1849-5

Andres N, Lieberherr G, Sideris IV, Jordan F, Zappa M. 2016. From calibration to real-time operations: an assessment of three precipitation benchmarks for a Swiss river system. Met. Apps, 23: 448–461. HYPERLINK "http://onlinelibrary.wiley.com/doi/10.1002/met.1569/abstract"doi: 10.1002/met.1569

Liechti K, Zappa M, Fundel F, Germann U. 2013. Probabilistic evaluation of ensemble discharge nowcasts in two nested Alpine basins prone to flash floods. Hydrological processes. 27: 5-17. HYPERLINK "http://dx.doi.org/10.1002/hyp.9458"doi:/10.1002/hyp.9458

Addor N, Jaun S, Fundel F, Zappa M. 2011. An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios, Hydrol. Earth Syst. Sci., 15, 2327-2347, doi:10.5194/hess-15-2327-2011. [HYPERLINK "http://www.hydrol-earth-syst-sci.net/15/2327/2011/hess-15-2327-2011.html"Direct Link]

Zappa M, Jaun S, Germann U, Walser A, Fundel F. 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. Atmospheric Research. . Thematic Issue on COST731. Volume 100, Issues 2-3, 246-262. HYPERLINK "http://dx.doi.org/10.1016/j.atmosres.2010.12.005"doi:10.1016/j.atmosres.2010.12.005

Zappa M, Rotach MW, Arpagaus M, Dorninger M, Hegg C, Montani A, Ranzi R, Ament F, Germann U, Grossi G, Jaun S, Rossa A, Vogt S, Walser A, Wehrhan J, Wunram C. 2008. MAP D-PHASE: Real-time demonstration of hydrological ensemble prediction systems. Atmospheric Science Letters. DOI: 10.1002/asl.183

Thank you very much for those valuable references. Some of these works are relevant for the present study and we intend to refer to those papers in the revised manuscript. The most relevant for our case are those presenting the full chain of tasks and models in operational frameworks and especially where evaluation metrics are shown (e.g., Zappa et al., 2008 and Addor et al., 2011).

L84-84: Nice!

Thank you.

Section 2 title: About end-to-end systems

Rotach MW, Ambrosetti P, Ament F, Appenzeller C, Arpagaus M, Bauer HS, Behrendt A, Bouttier F, Buzzi A, Corrazza M, Davolio S, Denhard M, Dorninger M, Fontannaz L, Frick J, Fundel F, Germann U, Gorgas T, Hegg C, Hering A, Keil C, Liniger MA, Marsigli C, McTaggart-Cowan R, Montani A, Mylne K, Ranzi R, Richard E, Rossa A, Santos-Muñoz D, Schär C, Seity Y, Staudinger M, Stoll M, Volkert H, Walser A, Wang Y, Wulfmeyer V, Zappa M. 2009. MAP D-PHASE: Real-time Demonstration of Weather Forecast Quality in the Alpine Region. Bulletin of the American Meteorological Society. 90. Pages 1321-1336. HYPERLINK "http://dx.doi.org/10.1175/2009BAMS2776.1"doi:10.1175/2009BAMS2776.1

Thank you. This is indeed a relevant paper as an end-to-end operational flood warning system; it also contains an interesting part about end-user feedback. We will refer to this work in the revised manuscript.

L117-121: Automatically or by operators using some kind of tools?

Automatically, as a part of the data management module. To clarify we will rephrase this sentence to be:

“Therefore, all near-real-time stage data go through a series of automatic validation and correction procedures”

L136: And with respect to more physically-oriented models?

Both conceptual and physically-based models were compared to ML hydrological models. The sentence will be modified in the revised version to:

“whereas the LSTM has been shown in recent years to improve hydrological simulations relative to conceptual and physically-based models (e.g., Kratzert et al., 2018; Kratzert et al., 2019a,b; Hu et al., 2019; Feng et al., 2020; Xiang et al., 2020).”

L143: per gauge

Correct. The sentence starts with “for example, a target gauge with…” so we think this is clear, but we will modify the sentence to be:

“(for example, a target gauge with a selected maximal lead time of 24 hours and hourly resolution implies 24 trained Linear models for this gauge)”

L152: Arbitrary number of steps or kind of based on analysis?

The number of steps is not arbitrary but is also not gauge-specific. We selected a number of steps to look back which we found to work well. For the system implemented in India and Bangladesh, we used 72 hours, as indicated in Table 1.

L159-161: Nice

Thank you.

L166: Described in Klotz et al., I suppose

Indeed. We will add this citation again where CMAL is mentioned.

L174-176: Reference on your validation efforts would be desirable

We are not sure what the reviewer means in this comment. The validation results are presented later on in Section 3 of the paper. We would be happy to provide more information once we understand this comment better.

L198: Well, if this is novel, it might need more detail to be reproducible. If this is a previously developed method, it needs to be referenced.

The reviewer is right, thanks for noting this. This is not a new model, but a modified version of the model presented in Ben-Haim et al., 2019. The main improvement from the original is the optimization algorithm which is described in the paper. However, we recognize that the description of the Thresholding algorithm needs more details to be better understood. The description of its categorized output into three levels of certainty was also missing. We will therefore modify the text of Section 2.2a to the text presented below. Also, note our plans for adding Python codes or pseudo-codes, as explained below (response to comment L221). That should help in understanding the algorithm in a more complete way and make it reproducible.

"(a) Thresholding model (modified from Ben-Haim et al., 2019) (Figure 3a): The model includes pixel-specific thresholds for each pixel in the AOI. Pixels that the water stage in the target gauge exceeds their thresholds are assumed to be inundated (i.e., wet) while the others are dry. These thresholds are learned from the series of historic stage data at the

target gauge and the corresponding state of the pixel (dry/wet) during these events (Figure 3a). Each pixel in the inundation map is treated as a separate classification task, predicting whether the pixel will be inundated or not. We refer to the "wet" class as the positive class.

The algorithm described below identifies pixel-specific thresholds and is aimed at maximizing an F_β-score using an optimized global parameter called *minimal ratio*. F_β-score is defined as $(1+\beta^2)\cdot\frac{precision\cdot recall}{\beta^2\cdot precision+recall}$ (Sokolova et al., 2006), where precision represents here the fraction of all wet pixels that are predicted as being wet and recall is the fraction of all pixels that are predicted to be wet and are really wet. An iterative process is applied to each pixel. In each iteration, we find the threshold that maximizes the ratio of true wet events (where the water stage at the gauge is above the threshold and the pixel was wet) to false wet events (where the water stage at the gauge is above the threshold and the pixel is dry). The threshold that maximizes this ratio is the most cost-effective threshold in the sense that it provides the most true wet pixels per false wet instance. At the first iteration all training events are considered; then, after each selection of a threshold and its respective true-false ratio, events with stage measurements above the threshold are discarded and a new iteration starts with the remaining events. If the new true-false ratio calculated is lower than the *minimal ratio* parameter value, the process stops and the final threshold for the pixel is the one found in the previous iteration. It can be shown that for every *minimal ratio* parameter value, no other set of pixel-specific thresholds achieves simultaneously better precision and recall; implying it is Pareto optimal. Therefore, for any β there exists some value of the *minimal ratio* parameter which finds the thresholds that optimize this respective F_β-score. After repeating the algorithms for different values of the *minimal ratio* parameter, the one that maximizes a specific target β is selected and the respective pixel thresholds for this parameter value are used.

The Thresholding algorithm is applied to compute categories of certainty that a given pixel is wet. This is done by computing the respective thresholds for two β values of 0.3 and 3, where lower β values imply higher thresholds. When a given water stage in the target gauge is considered, the two thresholds are examined for each pixel. If the threshold of β=0.3 is exceeded (implying the threshold of β=3 is also exceeded) it is classified as wet with a high probability. If only the thresholds of β=3 is exceeded, it is classified as wet with a low probability. The pixel is classified as dry if no threshold is exceeded.

In cases where the river stage input is higher than all past stage data, the Thresholding model's output inundation map is initialized from the most severe inundation extent seen in the historical events and expanded in all directions. The expansion distance is a linear function of the difference between the forecasted stage and the stage of the highest historical event. This 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."

L203: bit "sloppy" formulation

The reviewer is right. The text in this section will be rephrased in the revised manuscript, as presented in response to the comment above.

L221: All nice, well explained in plain text, but possibly to few information for something you declared being presented for the first time

We were debating whether to provide full details at the cost of making the paper very long and somewhat technical, vs. keeping a clear and concise explanation, albeit probably not enough to make it reproducible. Following the reviewers comments (reviewer #3 and reviewer #2) we would do our best to provide, with the revised manuscript, either a standalone python code or accurate and reproducible pseudo-code for the two ML inundation models which would be practically equivalent to the model runs in the operational system. Given this, we think it is better to leave the explanations in the main text of the paper at the current level of detail, where from the code one can understand the exact algorithm.

L271: Numbers?

We assume the reviewer is asking about the number of flood events used for training and validation. The information is given in Section 2.5 which describes the implementation of the system in India and Bangladesh. The Thresholding and Manifold models were applied to 228 gauges total; they were trained based on historical flood events for 2016-2020, where the mean number of events per AOI is 29 and the median is 15 (these numbers were modified from the those presented in the original manuscript which included the AOIs with no inundation model). We will put the modified numbers in the revised version of the manuscript.

L305: Are there also channels that do not need smartphones

Yes. The alerts are sent to relevant agencies, such as CWC, NDMA, IFRC and the Indian Red Cross, that further distribute them in their own channels. Alerts are also shown in Google Search and on Google Maps when viewing an area with an active flood alert; both can be reached not only from smartphones but through any computer or other devices with an internet connection (see paper's Sections 2.4 and 2.5).

L326-327: Reference? Link? Documentation?

The reference is World Bank (2015). It was mentioned in the sentence that followed this text but for clarity, we will add this reference also here.

L335-337: Are these separately displayed in Figure ?

No. We decided not to make different marks for different types of gauges since the map is already quite busy. But the Supplementary Table (S1) includes all the required information.

L356-366: Are you aware of action taken to mitigate loss of lives and infrastructure during this event?

This question is well justified and we are very much aware of its importance, but, unfortunately, we cannot answer it yet. We are putting a lot of effort into getting feedback from people in the area trying to get surveys either directly or through collaborations with local organizations. Unfortunately, the data we collected so far is too little and too scattered, probably, among other reasons, due to COVID limitations. So we cannot write yet about the actions that were taken. As explained in response to the comment above (L22) we have some preliminary results from our collaboration with the Yale Economic Growth Center. They 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. We do hope to publish these results soon.

L358-359: Missing information on training

We are not sure what the reviewer means. The line for which this comment was written provides the mean and the median number of flood events used to train the inundation models of 2021 in India and Bangladesh. What information is missing?

L363: Wo decides?

Alerts notifications to smartphones were sent if the forecasted water stage at a gauge exceeded a pre-defined threshold, which was set by the relevant agency (CWC or BWDB). To clarify we modified this text to be:

“In cases where flooding was sufficiently severe (i.e., the forecasted water stage at the gauge exceeded a pre-defined threshold provided by CWC and BWDB), alerts were also sent as push notifications to smartphones…”

Figure 6: Really cool

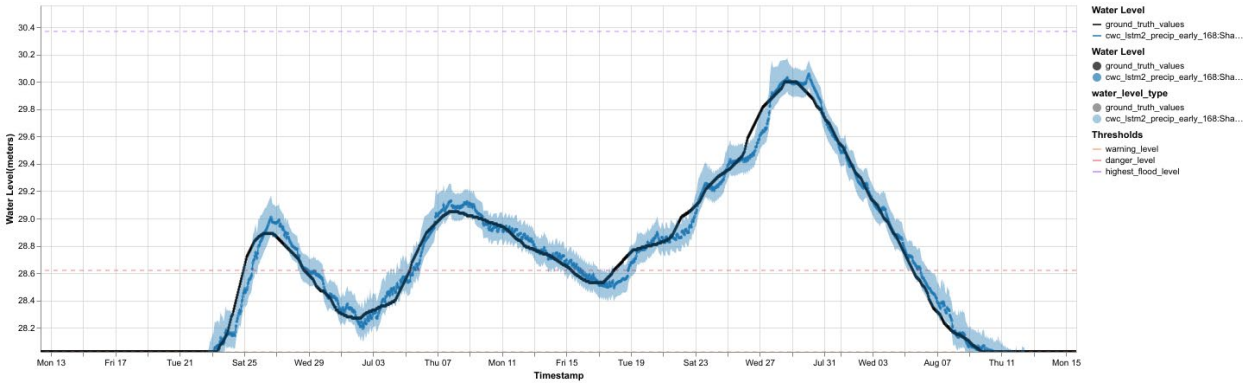
Thank you.

L390-396: This is deterministic evaluation. State of the art models are probabilistic. How you plan to go for such approach?

Thanks for raising this point. First, we need to clarify that our system does include probabilistic prediction of the forecasted river stage. It is different from the standard approach to probabilistic forecasts. We first explain why it is different, and then describe the form of probabilistic forecast in our system.

Our flood warning system relies mostly on past river stage data, as opposed to typical flood warning systems where precipitation and other meteorological forcing data are the main input. Probabilistic forecast in the latter situation is very important due to: 1) high uncertainty in observed or forecasted weather and mostly in precipitation data, and, 2) high sensitivity of floods to the precipitation input. Therefore, ensembles of weather forecasts are often utilized to produce the probabilistic flood forecast. Although precipitation improves prediction accuracy to some level, in our system, at least in its current focus (i.e., gauged, large rivers), the effect of its uncertainty on the predicted flood is substantially lower (see Table 2 and Figure 8; the median NSE is increased from 0.986 to 0.987 and the median persistent-NSE from 0.66 to 0.67).

While the contribution of precipitation to uncertainty is negligible in our system, other uncertainty sources exist, with the main one being the past river stage data. The estimation of uncertainty in the forecasted river stages, resulting from past stage data and other sources, is done by estimating the time-dependent parameters of the CMAL distribution (described in Section 2.2b of the paper). The figure below demonstrates the estimated uncertainty for one gauge:



The real-time estimated river stage probabilities are then used in the distributed flood alerts in the form of a range of river level change (taken between the 20th and 80th quantiles). An example of such a range can be seen in Figures 4a and 4d in the paper. Furthermore, when the estimated uncertainty range exceeds a given threshold (typically 50 cm) a shorter lead time is selected for the gauge. We are currently not using the uncertainty of the stage forecasting model to inform our inundation models.

We realize that this is an important point that has to be better explained. Therefore, in the revised manuscript we will add information on the river stage uncertainty, the Thresholding model's support for uncertainty, and their utilization in the flood alerts.

L419-420: Thank you

Thanks.

L441-442: Reference

We added a reference to:

Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006, December. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

L494: Very nice paper and nice that you address it!

We agree this is a nice and very interesting paper.

Conclusions: Could you give some advice to teams that are not able to access to the Google infrastructure?

This is a really good suggestion. In the revised manuscript we will add a paragraph with such advice, according to our understanding, and we hope it would be helpful.