In this response document, all line numbers refer to the newest draft of the manuscript (without track-changes).

Reviewer Report #1: Paul Santi

The model provides a rigorous and honest effort to predict inundation areas using existing models. As the authors state, this is an initial step to automate a process that will certainly become more important in the future.

My primary suggestion in this second review is in the order of presentation of findings in the abstract and conclusions. As written, the authors start with the fully predictive model which, unfortunately, has unconvincing results ("Approximately 94% of the observed inundated area was forecast to have an inundation probability greater than 1%"). Next, they indicate that the model improves when using actual volumes rather than predicted ones.

I think they paper would be stronger if the results are presented in the following order instead:

1. Predicted inundations using measured debris flow volume are strong

2. Predicted inundations using prediction models for volume are not strong

3. Part of under-forecasting can be attributed to I15 prediction

So, rather than starting with a "failed" model and basically digging yourself out of a hole to justify its use, you start showing that the concept is valid, but then start adding elements to show where the weaknesses are found.

> Response:
>
> We appreciate the reviewer's suggested reorganization of the Abstract and Conclusions and have made edits to implement this recommended structure. The relevant section of the Abstract (lines 14-24 in the newest manuscript draft) have been edited to read (edits in italics):
>
> "We applied this framework to simulate debris-flow inundation associated with the 9 January 2018 debris-flow event in Montecito, California, USA. *When the observed debris-flow volumes were used to drive the probabilistic forecast model, analysis of the simulated inundation probabilities demonstrates that the model is both reliable and sharp. In the fully predictive model, however, in which debris-flow likelihood and volume were computed from the atmospheric model ensemble's predictions of peak fifteen-minute rainfall intensity,* I_{15}*, the model generally under-forecasted the inundation area. The observed peak* I_{15} *lies in the upper tail of the atmospheric model ensemble spread, thus a large fraction of ensemble members forecast lower* I_{15} *than observed. Using these* I_{15} *values as input to the inundation model resulted in lower than observed flow volumes which translated into under-forecasting of the inundation area. Even so, approximately* 94% of the observed inundated area was forecast to have an inundation probability greater than 1%, demonstrating that the observed extent of inundation was generally captured within the range of outcomes predicted by the model."
>
> The relevant section of the Conclusions (lines 380-386) was rearranged and now reads:
>
> "We applied this methodology by using a 24-hour, 100-member atmospheric model ensemble forecast of rainfall intensity associated with a destructive debris-flow event that followed the 2017 Thomas Fire. *When debris-flow volumes were well-constrained, the probabilistic model predictions were sharp and well-calibrated to the observed area inundated. In the fully predictive model*, approximately 99% of the observed inundation area was contained within a region where the

simulated probability of inundation was greater than zero. In general, however, we found that the model under-forecasted the area inundated. We attribute the under-forecasting of inundation extent to the fact that the observed peak 15-minute rainfall rates were in the upper tail of the atmospheric model ensemble distribution of forecast rainfall rates."

Also, in my reading, the influence of the volume prediction model was much more important than the accuracy of the I15 model, but the abstract seems to indicate the reverse. I think it would help to either change the emphasis, or if indeed the I15 model is more important, perhaps indicate this better in the body of the paper.

Response:

While the input debris-flow volume did exert the greatest control on the uncertainty associated with inundation outcomes, we believe that the model tendency to under- or over-forecast inundation was primarily controlled by the predicted value of peak I_{15}. This is clearest to see through comparison of the calibration curves of the fully predictive model, Scenario A, and Scenario B. We have added text to the Discussion section to make this point more explicit. Lines 336-345 in the newest manuscript draft now read (additions in italics):

"Our interpretation that the ensemble distribution of predicted I_{15} led to under-forecasting is supported by *comparison with* the reliability diagrams associated with forecast scenarios A and B, which were run using observed debris-flow volumes and observed *peak* I_{15}, respectively (Fig. 5). The calibration curves from these two scenarios indicate high sensitivity of the calibration to the input debris-flow volumes*, which are influenced by peak I_{15}. When the volumes predicted from I_{15} were too low, as in the fully predictive model, the calibration curve lies above the one-to-one line, indicating under-forecasting (Fig. 3b). The calibration curve passes through the one-to-one line when the observed volumes, which are greater than those computed using the ensemble predictions of I_{15}, were used (Scenario A; Fig. 5b). Finally, the calibration curve drops below the one-to-one line, indicating over-forecasting, when volumes are computed based on the observed I_{15} (Scenario B; Fig. 5d and Fig. S1). Volumes computed from the observed I_{15} were greater than both the observed volumes and those computed from the ensemble predictions of I_{15}. As a result, the model would have over-predicted inundation area if the atmospheric model yielded a perfect prediction of peak I_{15}.*"

Lines 16-21 - I appreciate the new information incorporated into the abstract, but there is some repetition and it could benefit from a rewrite to streamline the text.

Response:

We have removed repetitive information from the rewritten Abstract.

Lines 16-17 - Having a probability of inundation greater than 1% does not sound like a strong validation of the model to me.

Response:

We have clarified the intention of this statement in the rewritten Abstract by placing it at the end of the new sentence structure suggested by the reviewer (lines 22-24). The purpose of this statement is to convey that even though the observed I_{15} was on the extreme end of the atmospheric model ensemble, the inundation forecast was still able to capture the observed inundation patterns in the spread of model outcomes. This is a general objective of forecasting with model ensembles.

Key point in line 273 - when using actual volume, the model accuracy was much better

The authors have made a strong effort to address the reviewers comments, and in my opinion, they have sufficiently responded to justify publication, provided the modifications I suggest herein are incorporated as best they can.

Response:

We appreciate the reviewer's kind words and thoughtful critiques that have improved the quality of the manuscript.

Reviewer Report #2: anonymous

I am sorry I do not think the manuscript has been completed. In the manuscript, only three pages are occupied by the parts Results and Discussion except related figures and tables. In addition, what are the innovative results obtained from this study? If the most important result is "a first step toward a near-real time hazard assessment product (line 25)", you have to give an in-depth discussion to verify the innovativeness of result.

Response:

To further support the line quoted in the reviewer's final sentence, we have added the following sentences (in italics) to the Discussion section (lines 356-365):

"The methods presented here take a step toward near-real time assessments of postfire debris-flow hazards associated with an incoming rainstorm. *Our work builds on that of Oakley et al. (2023), who used the same atmospheric model ensemble to produce probabilistic predictions of debris-flow likelihood and volume in watersheds burned by the 2017 Thomas Fire. They did not include predictions of postfire debris-flow inundation, but they identified that a product linking together postfire debris-flow volume ensembles with runout models was an important area of focus for future research to support impact-based decision making (Oakley et al., 2023). Further, recent surveys demonstrate a need for hazard assessment products that connect debris-flow inundation models with forecasts of rainfall in the short period of time between fire containment and the first precipitation event (Barnhart et al., 2023; Gourley et al., 2020).* Considering that decision quality improves when probabilistic information is presented appropriately in weather forecasts (Ripberger et al., 2022), the types of maps generated by the model framework presented here could be used to support decisions regarding evacuations, staging of equipment and emergency personnel, and debris-flow mitigation efforts."

Additional miscellaneous edits

Line 25: we combined sentences and replaced “However” with “but” for sentence fluidity.

Lines 366 and 372: we combined existing sentences with related content into one paragraph.