Response to Reviewers' Comments for 'Brief Communication: Rise of the Guadalupe River- A Multifaceted Post Event Analysis of the July 4, 2025, Flood in Central Texas'

We thank the reviewers for their comments and feedback. Our responses are marked in blue.

Reviewer #2:

The paper presents a comprehensive comparison of precipitation, discharge and flood inundation forecasts for a specific flash flood event of the Guadalupe River basin in Central Texas. The study evaluates forecasts against gauge observations, high-water-mark-derived flood maps, and building exposure datasets, enabling the identification of uncertainty propagation throughout the forecasting chain.

The manuscript addresses an important topic, namely understanding the sources of forecast failure during extreme flood events. The multifaceted evaluation framework is valuable, particularly the attempt to propagate forecast uncertainty all the way to flood impacts rather than limiting the analysis to precipitation or discharge verification alone.

However, the manuscript would benefit from clearer articulation of its novelty and stronger emphasis of learned lessons from this multifaceted framework compared to only precipitation or discharge evaluation framework. Figure consistency and presentation also need substantial improvement to make comparisons between lead times and locations clearer.

Overall, I believe the paper has potential, but considerable revisions are needed before publication.

Response:

Main points to improve are:

1. The motivation for diagnosing failed forecasts is clear, but the manuscript does not sufficiently explain the novelty of the study. First of all, in the abstract it is stated that the objective of the study is to diagnose the drivers of this catastrophic event. While this is important, this does not seem like the main objective nor does it highlight novelty. Currently, one of the main conclusions appears to be that shorter lead-time forecasts produce better precipitation, discharge, and flood predictions. While expected and useful to quantify, this is not in itself novel. The more novel aspects seem to be propagating uncertainty through the full forecasting chain toward impacts, and the use of post-event USGS high-water marks to generate a spatially continuous benchmark flood inundation map. However, the manuscript should more explicitly explain what additional insights are obtained from evaluating flood impacts

rather than only discharge forecasts, and why the inundation analysis materially improves understanding of forecast failure. Currently, it is unclear whether the impact-based analysis provides substantially more information than a direct comparison of forecasted versus observed discharge. One important advantage that should be emphasized more clearly is that high-water-mark-derived flood extents may enable evaluation in locations where gauge observations are unavailable or failed during the event. Relatedly, propagating uncertainty to impacts can help place forecast errors into societal context, which is currently only implicitly addressed. The novelty statement in the Introduction and Conclusion should therefore be strengthened considerably.

Response: We sincerely thank the reviewer for this constructive and incisive comment. Based on your suggestion, we modified the introduction and concluding remarks.

Three specific contributions are now highlighted in the revised manuscript:

1. An end-to-end assessment of the NOAA operational forecasting chain for a single high-consequence event, propagating from HRRR-driven precipitation forecasts through NWM short-range streamflow forecasts to OWP HAND-derived flood inundation maps and, finally, to building-level impact estimates. To our knowledge, no prior study has evaluated the full operational pipeline for a flash-flood event in this end-to-end manner.

2. A methodology for generating a spatially continuous benchmark flood inundation map from post-event USGS high-water marks, combining quality filtering, IQR-based outlier removal, Local Moran's I spatial filtering, semivariogram-informed neighborhood definition, and hydrologically conditioned interpolation. This is a transferable methodology that can be applied to future events and, critically, enables forecast evaluation in locations where in-situ streamflow observations are unavailable, either because no gauge exists (Camp Mystic in this study) or because the gauge failed during the event (Hunt). We have made this point explicit in both the Introduction and the Closing Remarks.

3. An impact-based evaluation that translates forecast error into societally meaningful information (number of buildings flooded), enabling forecast failure to communicate the social context rather than reported only as pixel-level skill scores. We now state explicitly that two locations may have similar CSI but very different building-impact counts depending on whether the spatial errors fall in populated areas, and that impact-based metrics provide complementary information to extent-based skill scores.

We have revised (highlighted) the relevant passages in the Introduction (second paragraph), and Closing Remarks. The specific changes are summarized below:

Revised Introduction: The last paragraph of the introduction (from line number 37) is modified as follows:

"*The main motivation of this study was to utilize operational forecast pipeline and conduct an extensive impact-based and categorical assessment of the event across different forecast hours against a high-water-mark-derived benchmark FIM. We employed the National Weather Service (NWS) National Water Model (NWM) short-range forecast and the FIMserv tool (Baruah et al., 2025) to generate 306 FIMs at the Hydrologic Unit Code (HUC)-8 scale. Additionally, we developed a methodology to use post-event USGS observed high-water marks to create a spatially continuous FIM benchmark for evaluating the forecast FIMs. This benchmark is particularly valuable because it enables spatial evaluation of forecast FIMs at locations where in-situ streamflow observations are unavailable either because no USGS gauge exists, as is the case near Camp Mystic, or because the gauge failed during the event, as occurred at Hunt. Our objective is to investigate the trade-off between NWM forecast range and the reliability of flood impact predictions across the operational forecasting chain, and to translate forecast error into meaningful information by quantifying building-level impacts in addition to standard pixel-level skill metrics. Specifically, we examined how FIM forecast skill, including building-level impacts, varied with respect to different forecast hours at five different locations (four streamflow gauges and near Camp Mystic).*"

Revised Closing Remarks: From L No-198

"*In this brief communication, we have presented an end-to-end evaluation of the NOAA operational forecasting chain for the Texas flash flood event on July 4, 2025,using precipitation forecasts, streamflow predictions, and flood inundation mapping. Using USGS gauge observations and a benchmark flood inundation extent generated from post-event ground-sampled high-water marks, we analyzed the NWM short-range streamflow forecasts and the resulting OWP HAND-derived flood inundation maps. Across all four gauged locations (North Fork, Hunt, Kerrville, and Comfort), the NWM short-range forecasts considerably underpredicted the peak flow, with a pronounced lag between observed and forecasted flows at downstream sites. We found that underprediction and lag in short-range forecasts are mainly due to errors in rainfall estimation and the failure of data assimilation ("nudging") caused by malfunctioning USGS gauges during the peak flow. These streamflow underpredictions seem to have propagated, resulting in substantially underestimated flood extents in forecasted FIMs at different lead times. Using the gauge-recorded peak time as a reference, we evaluated forecast FIMs at different hours against the HWM-derived*

benchmark. At upstream gages, forecast skill improved as the forecast time approached the reference peak, whereas at downstream locations (Kerr and Comfort), evaluation scores were very poor at the reference time but improved after the observed peak flow. The HWM-derived benchmark also enabled forecast evaluation at the ungauged Camp Mystic location. It is also clear from the impact assessment that forecast FIM was unable to capture the majority of flooded buildings compared to the benchmark, with this decoupling between pixel-level skill and impact-level skill underscoring the operational value of evaluating forecasts in the societally relevant units in which response decisions are actually made. This case study points to the need for communicating the uncertainty and error bounds in operational forecast FIMs for early response and decision-making. One promising approach is probabilistic FIM which can effectively represent and leverage uncertainty in streamflow forecasts, providing more informative and actionable guidance for emergency responders."

The manuscript should clarify whether forecast deficiencies were actually a primary driver of the impacts during this event. Currently, it remains unclear why this specific event was selected, whether losses were mainly caused by forecast shortcomings, or whether other factors such as exposure, warning dissemination, evacuation timing, infrastructure vulnerability, and emergency response played larger roles. Providing this context is important to justify the relevance of diagnosing forecast uncertainty for this case study. In addition, the Conclusions should acknowledge that reducing or quantifying forecast uncertainty does not necessarily translate directly into reduced impacts or losses. Other uncertainties remain present throughout the forecasting chain, including those associated with the simplified inundation modelling approach, exposure estimation, warning interpretation, and early-action decision making. In many cases, deficiencies in communication and response may contribute more strongly to disaster impacts than forecast uncertainty itself. Explicitly discussing these broader limitations would provide a more balanced perspective on the practical implications of the study and on where improvements in risk reduction efforts may be most effective.

Response:

This is an important framing point. We thank the reviewer for this question. The catastrophic loss of life was driven by multiple compounding factors: the overnight onset of the rising limb, the extreme convective rainfall rates over rugged Hill Country terrain with steep, clay-rich soils, the limited time to peak (less than 1.5 h to exceed the 500-yr return period at Hunt), the failure of in-situ instrumentation, and gaps in warning communication and evacuation. Our motivation is not to argue that forecast deficiencies were the dominant cause of impacts; rather, we use this event to diagnose whether the operational forecast chain provided actionable insights. We have deliberately avoided making specific claims about the warning or response system for the event. We have rewritten the closing remarks to (a) explicitly acknowledge that improved forecast skill alone does not necessarily translate into reduced losses, (b) list the additional sources of uncertainty, including exposure estimation,

the simplified HAND-FIM approximations, warning interpretation, and early-action decision making, and (c) emphasize that improving the forecast chain is one element of a broader risk-reduction strategy.

"*While this study focuses on the technical components of the operational forecasting pipeline, we acknowledge that forecast quality is only one of the components, while warning dissemination, warning interpretations by at-risk populations, evacuation timing, and the operational decisions of emergency responders all contribute substantially to the alert dissemination process*"

2) There are several inconsistencies between the figures that make cross-comparison unnecessarily difficult. In addition, the overall figure design and readability should be improved throughout the manuscript. Specific suggestions are provided below.

Suggestions to improve figures:

Figure 1

- Fig 1c: I suggest using different colours for rain gauge stations and streamflow gauges (currently

  Response: Thank you for your suggestion. We have updated the plot (Fig 1c)

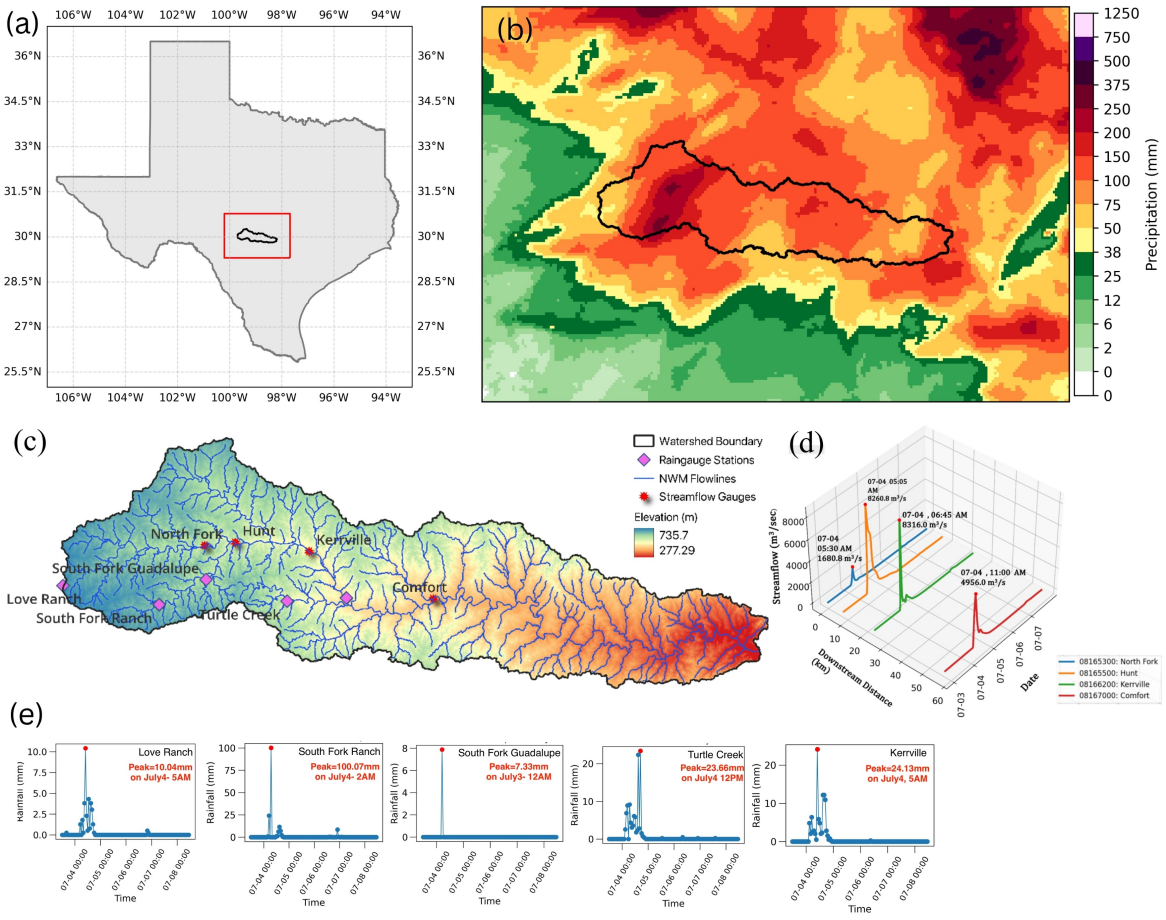


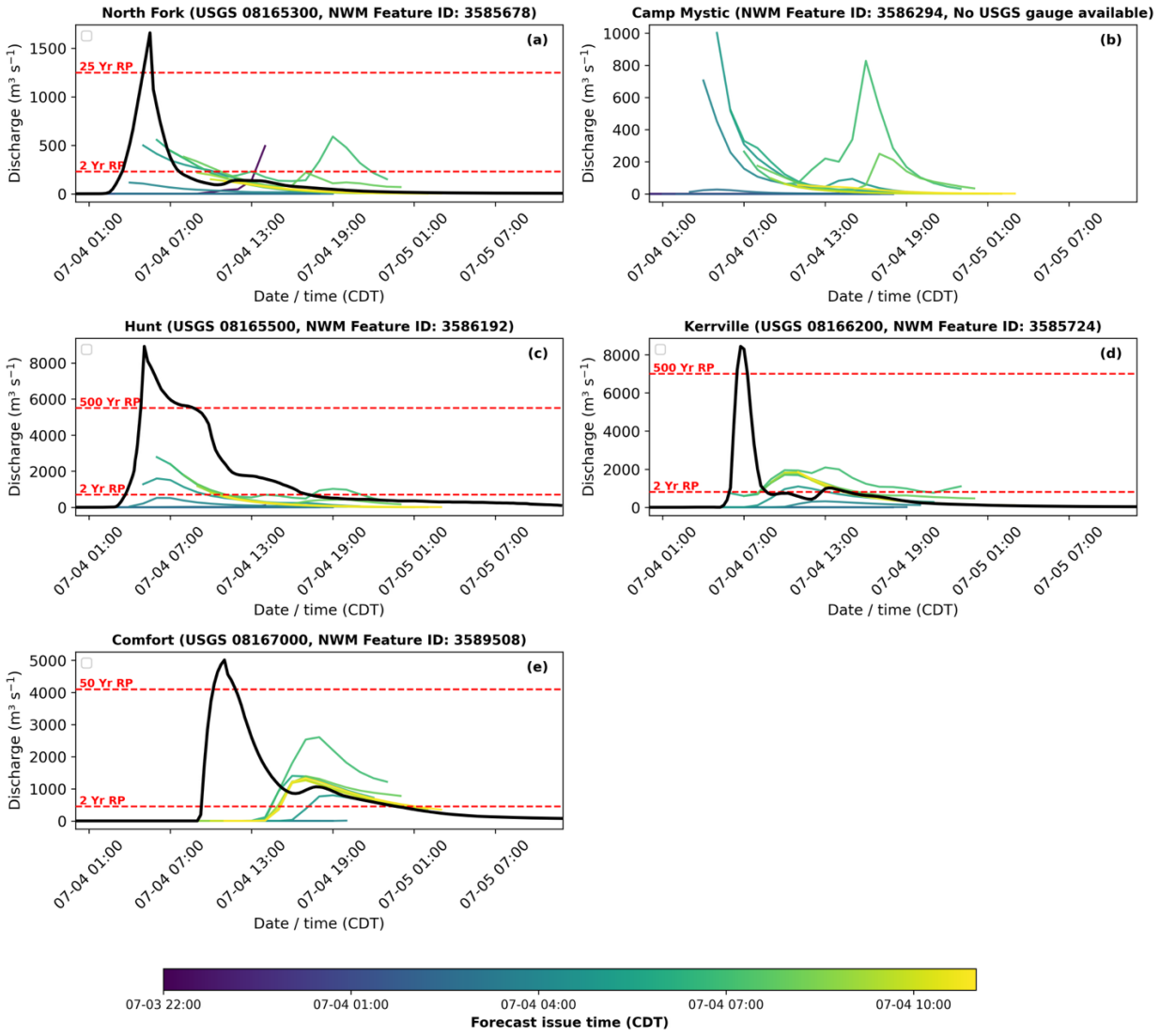
Figure 2

Regarding panel e, it is confusing that it does not have any USGS gauge readings and no return periods. This “station” (?) is also not shown in Figure 1 panel c and therefore leads to confusion. Also the subtitle is not informative and the time interval and coverage is different. Generally, it could be an idea to use a colour scale that intuitively changes colour with lead time (for example increasing colour intensity), as now certain lead times have the same colour and cannot be distinguished. Also the time range can be shortened to only show relevant data. This might allow the panel to be a big larger and the data to be more visible.

Response:

Thank you for your question. Panel (e) in Figure 2 indicates the National Water Model short-range forecasts near Camp Mystic, where the maximum casualties occurred. There was no USGS gauge available at that location. We have updated the caption in Figure 2 accordingly.

Based on your suggestion we have also updated the color scheme, and time range in Figure-2.



It is confusing that L109-113 states USGS gauge data near Hunt failed during the fast-rising streamflow (July 4th to July 5th), whereas panel b of Figure 2 does show USGS gauge data for Hunt for the period from July 4th to 5th. The only failed gauge from Figure 2 seems to be for Camp Mystic "panel e". It is not clear how the accuracy of the forecast for a specific gauge location can be tested when observational data is missing (for which the HWM estimates seem to be the solution, if I understand correctly)

Response: We thank the reviewer for raising this point and we will revise the manuscript accordingly

The USGS gauge at Hunt (USGS 08165500) was indeed disabled by the rapidly rising flood waters on July 4, 2025. The streamflow values shown in Figure 2(c) between approximately 04:50 on July 4 and 15:20 on July 5 are not direct gauge measurements, those are post-event

reconstructions developed by USGS from surveyed high-water marks using indirect measurement techniques (Benson & Dalrymple, 1967). This is confirmed on the USGS National Water Information System page for the site, where the peak of 315,000 $ft^3\ s^{-1}$ (≈ 8,910 $m^3\ s^{-1}$) is documented as estimated value rather than an observed one (USGS, 2025; https://waterdata.usgs.gov/monitoring-location/USGS-08165500). We have clarified this in the revised manuscript.

Regarding panel (b) of Figure 2, there is no USGS gauge near Camp Mystic, and consequently no observation against which to evaluate the NWM forecast at that location. We have retained this panel because it illustrates the NWM short-range forecast behavior at an ungauged reach where maximum casualties have happened. In place of direct discharge verification, we used surveyed high-water marks at this location to reconstruct the observed flood extent and used that to evaluate the corresponding forecast-derived flood maps.

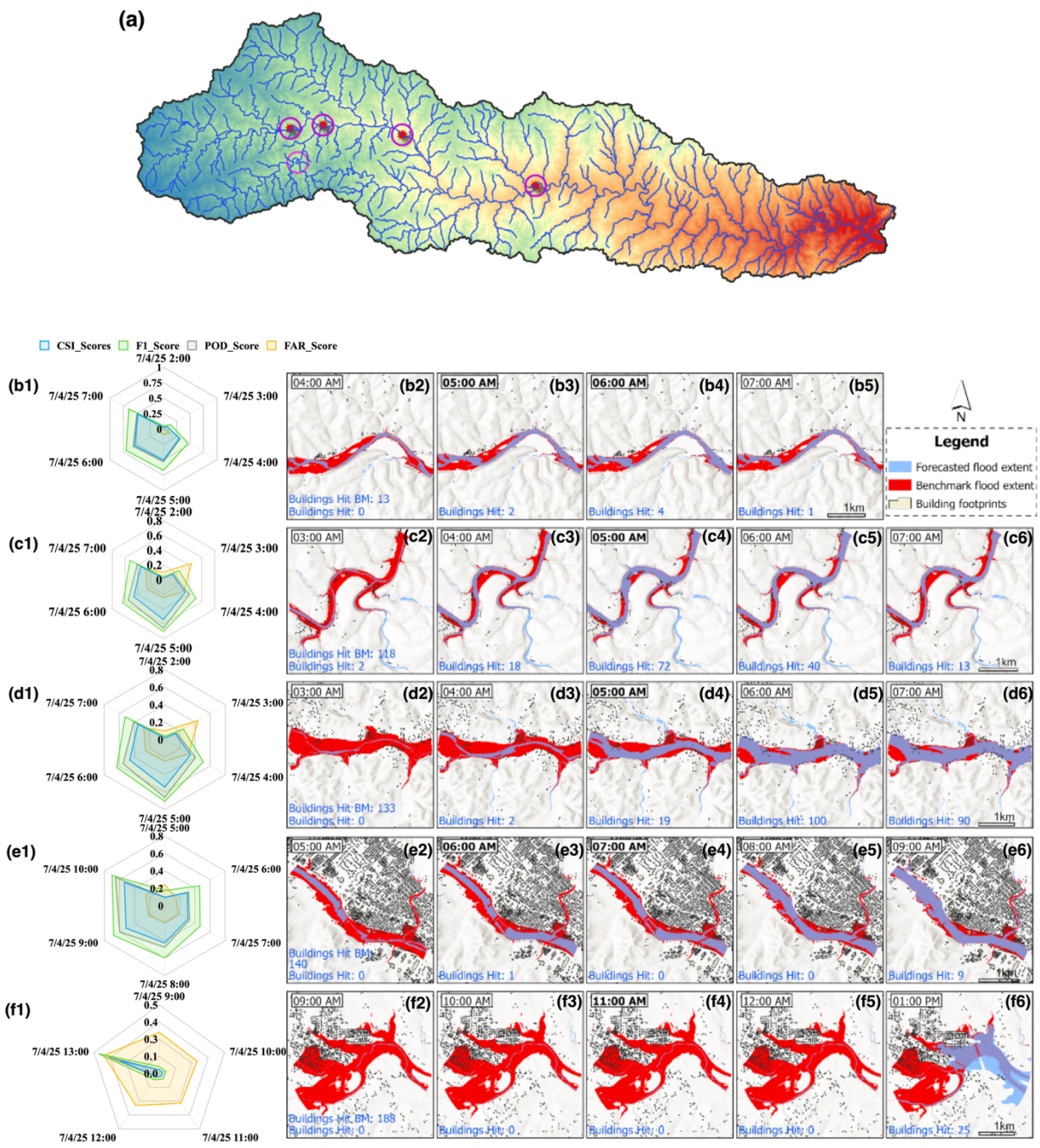
Reference

Benson, M. A., & Dalrymple, T. (1967). General field and office procedures for indirect discharge measurements (U.S. Geological Survey Techniques of Water-Resources Investigations, Book 3, Chapter A1). U.S. Government Printing Office. https://doi.org/10.3133/twri03A1

Figure 3

Panel numbering is missing for upper map. In this panel, please add the name of the fifth purple circle and explain this is Camp Mystic (I assume). Also, please center the purple circle for the gauge at Hunt in top map.

Response: Thank You for your comment. We have added the panel number, named the fifth purple circle and put the purple center at the center of the Hunt gauge.



Please use the name order of gauge names as in Figure 2, e.g. in Fig 2 panel e is Camp Mystic and here (Fig 3) panel b. I prefer the station order used in Figure 3, as it follows the downstream progression of the catchment flow and therefore improves interpretability. I suggest using this same ordering consistently across all figures and throughout the text. It is

good how the order of in text results on p. 8 matches the order of those presented in figure 3.

Response: Thank you for your suggestion. We have updated the plot

The text in the a-e1 panels is very small and I suggest to increase the size where possible. I do like the way these panels visualise the different metrics per FIM.

Response: We have updated the plot

For the legend of panels a-e2-6, is flood extent the forecast flood extent? If so, suggestion to change to "Forecasted flood extent".

Thank you. The legend has been updated.

The colour of the building footprints makes is very hard to see them, especially at this scale. I recommend to fill them and choose a colour combination that makes them easy distinguishable.

Thank you. The building footprint markers have been filled in light grayish yellow (#F3EDD3) and outlined with 80% gray to make them more seeable. Also, the text in the lower left of each panel has been changed to light blue to make them more visible after changing the symbol of building footprints.

A question: Is the catchment of Camp Mystic a larger determinant of the peak discharge for the Hunt station than the North Fork catchment, as the Hunt peak discharge happens at 5 AM and that of North Fork only later at 6 AM? If so, this might be worth mentioning in the text.

Thank you for this question. At the Hunt station, the peak was recorded at 5 AM, and the upstream Gauge at North Fork was at 6 AM. From Figure 1, it is observed that the Hunt gauge station is situated at the downstream confluence of North Fork Guadalupe and South Fork Guadalupe River. The Upstream gauge in the Northfork River, doesnot shows a very high streamflow value (crossing the 25 Year Return period), but at Hunt the flow crosses the 500-year RP, which is only possible if there is a large contribution from the South Fork River, which passes through the Camp Mystic catchment and has a confluence near Hunt (with North Fork River). Due to the lack of a USGS gauge, there is no observation to monitor the flow, but it is very reasonable to conclude that the Major contribution was from the South-Fork River. We will add this section in the revised manuscript:

"*At the Hunt Station (USGS 08165500), the peak discharge was recorded at approximately 05:00 CDT on July 4, 2025, while the upstream gauge on the North Fork Guadalupe River (USGS 08165300) peaked roughly one hour later, at 06:00 CDT. As shown in Figure 1, the Hunt gauge is located immediately downstream of the confluence of the North Fork and South Fork branches of the Guadalupe River. The observed flow at the North Fork gauge, did not exceed the 25-year return-period, whereas the peak at Hunt exceeded the 500-year*

return-period level. This large difference cannot be explained by inflow from the North Fork alone and is consistent only with a substantial contribution from the South Fork Guadalupe River, which drains the Camp Mystic catchment and joins the North Fork just upstream of the Hunt gauge.

Because no operational USGS stream gauge exists on the South Fork in this reach, the streamflow contribution from the South Fork River cannot be directly verified from in-situ observations. However, the timing of the peak and the spatial distribution of rainfall during the event strongly indicate that the South Fork was the dominant contributor to the very high flashy streamflow peak at Hunt."

Another thought is that it seems that the importance of forecast uncertainty increases for downstream locations. If so, this is an important finding to mention explicitly.

Thank you for your comment. We added the following section in the revised manuscript

"*In the National Water Model (NWM), forecast errors are partially constrained by streamflow nudging from upstream USGS gauges, where simulated discharge is adjusted toward observed values, with the correction subsequently propagating downstream through the routing scheme (Cosgrove et al., 2024; McCreight et al., 2024). In this case, when the gauge at Hunt fails during the rising limb, this correction mechanism did not work, and forecast error at downstream locations is no longer bounded by upstream observational constraints. The rainfall forecast errors that drove the initial overestimation or underestimation of inflows are therefore amplified through the routing network rather than corrected at intermediate points, leading to the progressive degradation of forecast skill at downstream gauges, systematically affecting the FIM forecast."*

Supplementary figures:

Figure S3 states "six-hour rainfall accumulation at 8:00 AM derived from MRMS (1 km) Pass 2 observations and HRRR (3-km) forecasts issued at successive lead times, from 0 hour through 7-hour lead time", however the plots shown (panel b-h) are from 2:00 AM – 20:00 PM, which do not seem to match the mentioned lead times. Also, a difference of plots could be more informative to see the magnitude of differences in specific locations than leaving this to the estimation skills of the reader.

We agree with the reviewer on both points and thank them for flagging the caption/lead-time inconsistency and for suggesting a direct difference plot. Figure S3 has been replaced. The new figure shows the HRRR − MRMS error for the full 18-hour HRRR cycle across twelve successive initializations on 2025-07-04 (00:00–12:00 UTC). Red indicates more rainfall in HRRR, blue more in MRMS, and each panel title gives the init time and the 18-hour valid window. The updated caption is below.

Figure S3. Spatial error (HRRR − MRMS) in 18-hour accumulated precipitation over the Guadalupe (HUC8) basin and a 1° buffer, for twelve successive HRRR initializations from 2025-07-04 00:00 UTC through 11:00 UTC. Each panel (a–l) shows the difference field for one initialization: the 18-hour HRRR forecast accumulation (regridded to the MRMS grid) minus the matching 18-hour MRMS Pass-2 hourly QPE sum over the same valid window. All values are in mm. A common symmetric scale (±200 mm) is used across panels; red denotes more rainfall in HRRR, blue more rainfall in MRMS. The black outline marks the HUC8 boundary.

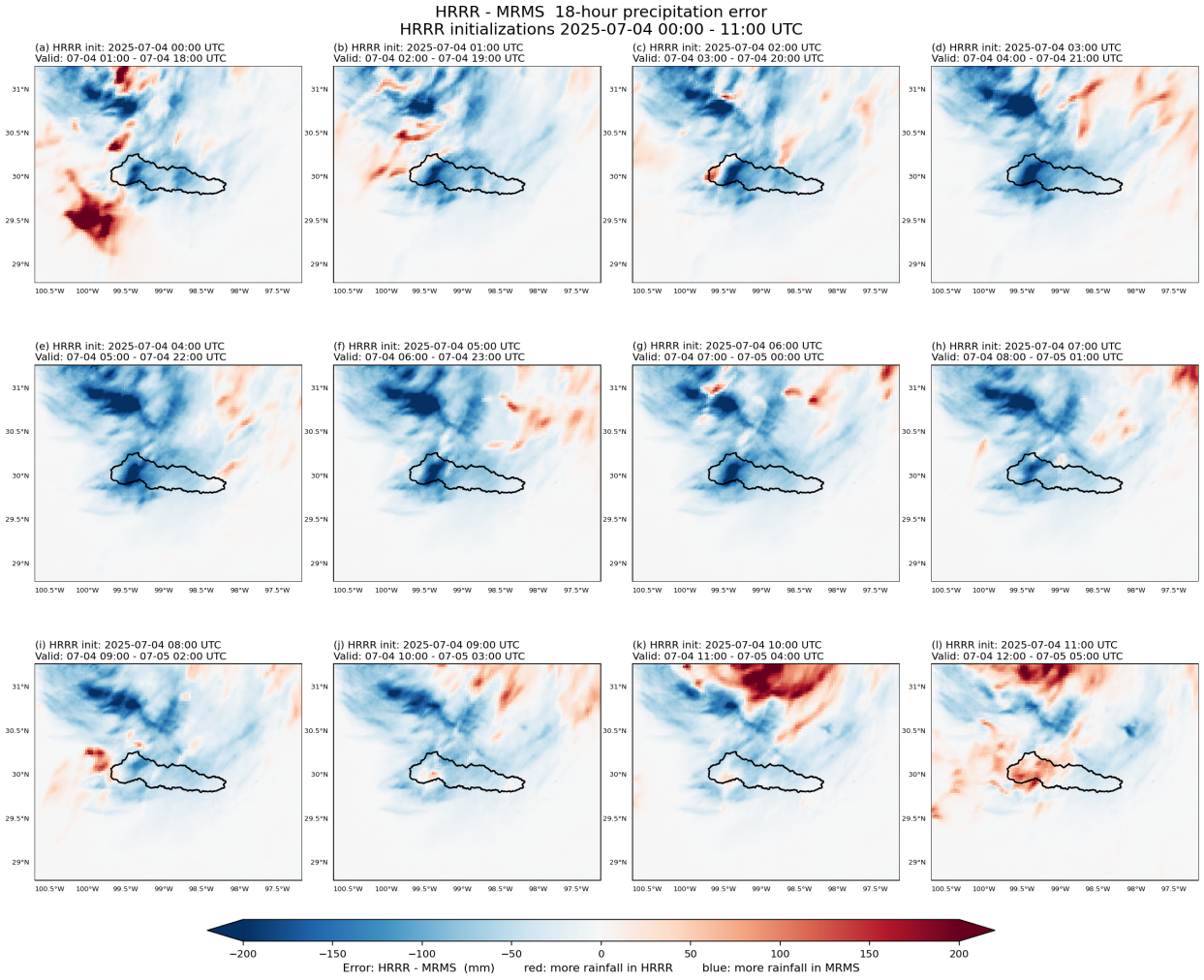


Figure S3: Spatial error (HRRR − MRMS) in 18-hour accumulated precipitation over the Guadalupe (HUC8) basin and a 1° buffer, for twelve successive HRRR initializations from 2025-07-04 00:00 UTC through 11:00 UTC. Each panel (a–l) shows the difference field for one initialization: the 18-hour HRRR forecast accumulation (regridded to the MRMS grid) minus the matching 18-hour MRMS Pass-2 hourly QPE sum over the same valid window. All values are in mm. A common symmetric scale (±200 mm) is used across panels; red denotes

more rainfall in HRRR, blue more rainfall in MRMS. The black outline marks the HUC8 boundary.

Minor comments:

- In the abstract, please clarify which type of drivers are diagnosed, L11 (if this is indeed the main objective of the study, which I question in point 1 of the main points to improve). From L49-55 I understand you are referring to large-scale climatic drivers.

  Response: We thank the reviewer for this important observation. Our main objective is to perform an evaluation of the NOAA operational forecasting chain, from HRRR-driven precipitation forecasts through NWM short-range streamflow forecasts to OWP HAND-derived flood inundation maps, benchmarked against a high-water-mark-derived inundation extent.

  Based on your suggestion, we have revised the abstract. The phrase "diagnose the drivers of this catastrophic event" has been replaced with the following:

  "*The objective of this study is to evaluate the performance of NOAA operational flood forecasting pipeline, including Quantitative Precipitation Forecasts, National Water Model short-range streamflow forecasts, and Office of Water Prediction flood inundation mapping during this catastrophic event, and to characterize how forecast skill and impact-based predictions varied with forecast lead time at gauged and ungauged locations*."

- Please clarify what "impact-based and pixel-based assessments" means in L16, as impact-based can also be pixel-based, right?

  Response: We thank the reviewer for this question, and we agree with your view. The impact assessment is in fact also pixel-based in its computation, since the building counts are derived from intersecting building footprints with flooded pixels in the forecast and benchmark FIMs. Based on your suggestion, we have revised the terminology throughout the manuscript. The two evaluations are now referred to as:

  Skill-based assessment, a pixel-level comparison of forecast and benchmark FIMs to compute the Critical Success Index (CSI), Probability of Detection (POD), False Alarm Rate (FAR), and F1-Score. This quantifies the spatial accuracy of the forecast flood extent. While Impact-based assessment is also derived from the pixel-level comparison, it is summarized as the number of buildings intersecting flooded pixels in the forecast versus the benchmark FIM.

The abstract (L16) and Section 3.3 have been updated accordingly. The terms "skill-based" and "impact-based" are now used consistently throughout the manuscript.

- L22, use of not common scientific units: “3-4 inches per hour”, please use the format as done in L57 “500 mm (20 inch)” or ignore inches all together as this is not consistently used, e.g. L62 & L63. Please choose one format and use this consistently.

  Response: Thank you for your comment. We have used same units consistently in the revised manuscript.

- Typo in L75, “gages” should be “gauges”, also L28, L86, L92, L157, 207

  Response: Thank you for your comment. We have corrected this spelling error.

- L109 “changing signal location” is not easy to understand what is meant here. Please be explicit.

  Response: We thank the reviewer for this question. By using the phrase "changing signal location" we mean that there is a spatial mismatch in successive HRRR forecast cycles in the hours leading up to the event over different geographic locations in Upper Guadalupe Basin from one cycle to the next. Consequently, the NWM short-range forecasts driven by these HRRR cycles produced spatially inconsistent runoff signals.

- USGS gauge number in L111 “08165500” is not that important and can be put in parenthesis.

  Response: Thank You. We made the changes as suggested.

- L100, it is not clear whether “MRMS Quantitative Precipitation Estimate (QPE)” is a forecast or reanalysis product.

  Response: Thank you for the comment. MRMS Quantitative Precipitation Estimate (QPE) is not a forecast product; it is a near-real-time precipitation analysis product derived primarily from weather radar observations and rain gauge data. We have revised this sentence in the manuscript

- L114: Unclear how this works “streamflow “nudging” to correct model states toward observed discharge.” Is there only nudged for streamflow at downstream gauges of Hunt where there is data? The whole part about the failed USGS gauge for Hunt is not clear.

Response: We thank the reviewer for this question. We have revised the relevant passage in the manuscript and provide a more detailed explanation below:

"The National Water Model applies streamflow nudging within its Analysis and Assimilation (AnA) cycle, which produces the initial conditions for each short-range forecast. At each gauged reach, the nudging scheme adds a time-weighted correction term to the channel-routing equation that adjusts simulated discharge toward the observed value (Seo et al., 2021). The corrected discharge state then propagates downstream through the routing method.

When the USGS gauge at Hunt (08165500) failed, no observational constraint was available to correct the simulated discharge at that reach. As a result, errors from upstream were no longer removed by the assimilation step and continued to propagate downstream through the routing network. Downstream gauges at Kerrville (08166200) and Comfort (08167000) did continue to nudge their own respective reaches; however, by the time these corrections were applied, the flood peak had already advected through the ungauged channel segments between Hunt and these downstream gauges. The short-range forecast is launched in open-loop mode from the AnA initial condition; any uncorrected error in the initial state propagates into the forecast and is not removed during the forecast horizon. This is why the loss of nudging at Hunt is reflected in degraded forecast skill at downstream locations (Figures 2 and 3)."

We have revised L114 to read as follows:

" *This failure disrupted the NWM Analysis and Assimilation step at the Hunt reach, where simulated discharge would normally be nudged toward observed values via a time-weighted correction added to the channel-routing equation (Seo et al., 2021). With no observations available, the rising flood wave entered the routing network without observational constraint at Hunt, and the resulting error propagated downstream through the Muskingum-Cunge routing."*

- L157-158, why create an additional temporal uncertainty when you can also take the forecasted peak, as is assumed for the gauge record?

  Response: We thank the reviewer for this suggestion. The reason we used the gauge-recorded peak time with a $\pm$2-hour evaluation window is that the benchmark FIM itself does not have a well-defined time stamp.

  The benchmark FIM is derived from USGS High Water Marks (HWMs), which are post event surveyed points and record the maximum water surface elevation reached at each location. The HWMs do not carry information about when each mark was attained during the event, and there is no guarantee that all HWMs across the basin correspond to the same point in time, different locations have reached their maximum at different hours. The HWM-derived benchmark FIM is therefore inherently a maximum-extent product rather than a snapshot at a specific time.

We used the gauge-recorded peak time as a physically meaningful temporal anchor (the closest available proxy for when peak conditions occurred at each location), and the ±2-hour window allows the evaluation to accommodate the fact that the HWM recorded maximum may have been attained slightly before or after the gauged peak. Aligning to the forecasted peak would not remove this ambiguity, because the source of the temporal uncertainty lies in the benchmark itself, not in the choice of reference time.

We have revised Section 3.3 to clarify that the ±2-hour window is a direct consequence of the HWM benchmark's lack of time stamp information, rather than an arbitrary evaluation choice.

- L164, the word "although" does not make sense here, as the following point about a higher FAR at larger lead times is the same conclusion as is given in the first part of the sentence that shorter lead time result in better CSI, F1 and POD. So, I would say that there are also improvements in FAR from 4 AM to 6 AM, if the 4AM forecast exhibits a higher FAR than the 6 AM (which is how I interpreted from the way the sentence is written now)?

  Thank you for the observation. We agree with your point.

  From the analysis, the evaluation metrics indicate that there is a gradual increase (improvement) in CSI, F1 and POD from 4 AM to 6 AM, while FAR decreases from 0.125 (4 AM) to 0.088 (5 AM) and 0.084 (6 AM). We have revised the sentence in the revised manuscript.

  "*At North Fork, the flood peak occurred at 5:30 AM. Compared to earlier forecasts, the FIM forecasts for 4 AM to 6 AM show gradual improvements in CSI, F1, and POD scores, while FAR decreases over the same period, indicating an overall improvement in performance*"

- Please use same order of presenting results in Sect 3.3, in the paragraph starting from L163. First 4 AM results and then 5 AM results, e.g. L172 vs. L174-175

  Thank You for your observation. We have corrected this in the revised manuscript.

  "The impact assessment (Figures 3c2–c6) indicates that the benchmark FIM estimates 133 buildings impacted, compared to only 2 for the 4 AM forecast and 19 for the 5 AM forecast"

- L179, should FIMs be singular here, i.e. FIM? What about the buildings predicted to be flooded for the 10AM FIM, since the CSI is reported to be better?

Response: Thank you for your comment. We have corrected this in the revised manuscript.

- L185, at 1PM for Comfort, how many buildings are predicted to be flooded?

  Response: Thank you for your comment. At 1 PM for Comfort, there are 25 buildings predicted to be flooded. We have added this sentence in the revised manuscript.

- In the closing remarks, it is confusing that L199 mentions four-gauge stations, whereas in Figure 3 five are discussed (for which one data is lacking).

  Response: Thank you for your comment. We have revised the sentence

  "Across all four gauged locations (North Fork, Hunt, Kerrville, and Comfort), the NWM short-range forecasts considerably underpredicted the peak flow, with a pronounced lag between observed and forecasted flows at downstream sites. No USGS gauge is available on the South Fork Guadalupe River near Camp Mystic, so a direct streamflow evaluation was not possible at that location."

- If satellite-derived flood extent observations are available for this event, it would be valuable to compare these against both the forecasted FIMs and the HWM-derived flood extent maps. Such a comparison could provide an additional independent validation of the inundation results and help assess the robustness of the HWM-derived benchmark.

  Response: Thank you for the suggestion. Understanding the importance of complementary ground-truth that satellite observations can provide, we indeed investigated the availability of satellite imagery before proceeding with the use of field-collected High-Water Marks (HWMs). Specifically, we examined both radar (Sentinel-1) and optical imagery (Planet, Sentinel-2, and Landsat); however, no suitable imagery was available due to either (a) the absence of satellite overpasses during the flooding event or (b) extensive cloud cover in the available acquisitions. We also evaluated high-resolution Maxar imagery. Although this imagery contained limited cloud cover, no floodwater was detectable because of the temporal gap between the flood event and image acquisition. The imagery was collected approximately seven days after the flooding, by which time floodwaters had largely receded.