Responses to Reviewer #2 Comments:

This manuscript proposes an encoder–decoder LSTM-based framework for flood nowcasting, using radar rainfall estimates and forecast products, combined with water level observations and LiDAR DEMs, to perform water level prediction and inundation mapping. The manuscript has a sufficient workload and some potential for operational application. However, it has certain shortcomings in terms of core novelty; the encoder–decoder LSTM architecture, radar rainfall forcing, and other components have already appeared in prior literature. At the methodological level, several issues require further clarification, such as the use of STEPS-NWP ensemble information, error decomposition of the HAND method, and the selection of validation events. In addition, there is partial overlap with the first author's MODSIM2025 conference paper, and the authors must explain this issue. Therefore, my recommendation is Major Revision, with specific comments as follows.

We greatly appreciate the time and effort the reviewer has devoted to evaluating our manuscript. The thoughtful comments and constructive suggestions have been invaluable in strengthening the manuscript. Below, we provide our detailed responses to the reviewer's comments.

R2C1) The novelty is unclear and needs to be explicitly defined relative to existing work. The core methodological combination in this manuscript—encoder–decoder LSTM, radar rainfall forcing, and HAND-based inundation mapping—has been extensively covered in the existing literature. For example, Kao et al. (2020) systematically validated the effectiveness of encoder–decoder LSTM for multi-step flood forecasting, and this work is also cited in the manuscript. A study in the Tamanduateí catchment in Brazil implemented LSTM coupled with HAND-based inundation mapping, also targeting a rapidly responding catchment and using a short antecedent window. Kratzert et al. (2018, 2019) and Nearing et al. (2024), among others, have demonstrated the strong capability of LSTM in rainfall–runoff modeling, even achieving global-scale flood forecasting with lead times of up to five days. The question raised in the Introduction—"whether models trained solely on real-time radar rainfall data can effectively nowcast floods"—is more a choice of experimental design than a fundamental theoretical or methodological breakthrough. Applying an existing architecture to a new region and validating its performance under three flood regimes is closer to regional application validation than to methodological innovation. It is recommended that the authors clearly state the novelty of this study in the Introduction.

Thank you for this important comment and useful suggestion. We agree that the individual methodological components of this study, including encoder–decoder LSTM, radar rainfall forcing and HAND based inundation mapping, have been investigated in previous studies. We acknowledge that the novelty of our study does not lie in developing a fundamentally new deep learning architecture or inundation mapping method, but rather in evaluating the potential of integrating these established approaches into an operationally oriented flood nowcasting framework and assessing whether it can provide locally relevant forecasting information across different flood regimes and hydrological conditions. And the aim of this study is closely aligned with the scope and objectives of the NHESS special

issue "Early Warning Systems from Research to Operations: Status, Innovations and Multi-Hazard Applications.", to which our manuscript was submitted.

In addition, this study makes several further contributions. First, we investigate the potential of LSTM models driven by operational radar rainfall products from the Australian Bureau of Meteorology, including calibrated, high temporal resolution (5–10 min) radar-based rainfall observations and short-term rainfall forecasts, to provide accurate water level predictions over prediction windows ranging from 15 min to 12 h. This addresses an important practical consideration for operational flood nowcasting, where the availability, quality and timeliness of input data are critical. Second, we integrate the predicted water levels with very high spatial resolution (1–5 m) airborne LiDAR derived DEMs through the HAND approach to generate locally relevant flood inundation forecasts. This allows the evaluation to extend beyond water level prediction accuracy to the spatial information needed for flood warning and emergency response. Third, we systematically evaluate the framework across three locations representing contrasting hydrological conditions, from relatively slow responding riverine flooding at Greta to rapidly responding and more complex flash flood environments at Wollombi and Wallis Creek. This provides an opportunity to assess the applicability and limitations of the approach across different flood regimes.

Importantly, our study focuses on the practical forecasting capabilities of these established methods rather than claiming methodological innovation. By assessing model performance under both observed radar rainfall forcing (reanalysis) and forecast radar rainfall forcing (forecast), we examine not only the predictive capability of the LSTM model but also its performance when driven by rainfall forecasts, which is particularly relevant to operational applications. Beyond that, our study provides insights into the evaluation of deep learning models for operational flood forecasting and early warning. In particular, we demonstrate the importance of assessing model performance beyond conventional accuracy metrics, by considering operationally relevant measures, including the advance indication and consistency of flood threshold exceedances, flood peak magnitude and timing errors, and the available lead time for emergency response. We also highlight the importance of benchmarking deep learning forecasts against at least a zero-skill model to establish their additional predictive value, while recognising the need for further comparisons with established operational forecasting approaches. As emphasised by Li et al. (2026a), the selection of flood thresholds, timing tolerances and evaluation benchmarks should reflect the intended forecasting application. Together, these findings provide practical insights into how deep learning approaches can be evaluated to support their transition from research to operational flood forecasting.

To address the reviewer's concern and clarify the contribution of this study relative to previous research, we will revise the Introduction to better acknowledge existing methodological developments and explicitly define the research objectives and contributions. Specifically, we will revise the last two paragraphs in the Introduction as follows:

"LSTM rainfall–runoff modelling and multistep forecasting are well established (Kratzert et al., 2018,

2019; Kao et al., 2020; Nearing et al., 2024). Subsequent studies have demonstrated the integration of water level forecasts with inundation mapping and operational alerts (e.g., Nevo et al., 2022), as well as LSTM–HAND coupling in a rapidly responding catchment (e.g., Silva et al., 2026). Beyond methodological development, increasing attention has been given to translating hydrological research into practical applications that provide reliable, contextually relevant and actionable information for decision making. In flood forecasting, this requires not only accurate model predictions but also an understanding of how forecasting approaches perform under operational data constraints and across different hydrological conditions. Li et al. (2026a) further emphasise that assessments of operational readiness depend on the flood thresholds, timing tolerances and benchmarks used in evaluation, and that these choices should reflect the intended forecasting application. Building on the methodological advances, this study aims to contribute to the translation of deep learning research into operational flood forecasting by investigating how deep learning approaches can provide locally relevant and actionable flood forecasting information in a regional application. In particular, it evaluates the potential of LSTM models supported by the highest quality, calibrated and high temporal resolution (i.e., 5-10 min) radar rainfall data to deliver accurate water level predictions, combined with very high resolution (i.e., 1-5 m) airborne LiDAR DEM data to generate flood inundation forecasts across different prediction horizons. By evaluating model performance under both observed and forecast rainfall forcing across three contrasting flood environments, this study provides insights into the practical capabilities and limitations of deep learning approaches for supporting operational flood forecasting and early warning.

A critical factor in effective flood nowcasting for emergency services is the provision of timely early warnings to support actions that protect lives and property (Gissing et al., 2010; Ebert et al., 2024). In practical operational settings, real-time data on variables such as evaporation, soil moisture and drainage may not be readily available, while radar systems can provide frequently updated rainfall estimates. Therefore, an important practical question arises: instead of requiring a wide range of hydrological inputs, can LSTM models driven by radar rainfall and antecedent water level observations effectively nowcast floods with short lead times, from a few minutes to a few hours? Another major challenge is that forecast rainfall inputs often deviate from observed values in both intensity and location, reducing prediction accuracy. Calibrated real-time radar rainfall can help capture the spatial and temporal distribution of rainfall within a catchment, which is important for both flood nowcasting and modelling general flood processes. Integrating calibrated radar rainfall nowcasts with LSTM models therefore provides a basis for an end-to-end, operationally oriented nowcasting framework. This study develops and assesses an encoder–decoder LSTM framework using operational radar rainfall products from the Australian Bureau of Meteorology. These products provide high-frequency updates, every 5 to 15 min, and include historical rainfall estimates from RAINFIELDS, radar-based rainfall nowcasts up to 90 min from STEPS-ADV (Velasco-Forero et al., 2020), and rainfall forecasts up to 12 h from STEPS-NWP, which blends radar and numerical weather prediction information (Seed et al., 2013). To evaluate the applicability of the proposed framework across contrasting hydrological conditions, the approach is tested at three locations representing distinct flood characteristics and varying levels of hydrological complexity (Table 1). Greta represents a relatively simple riverine setting

with a slow hydrological response dominated by upstream catchment flows; Wollombi represents a moderately complex riverine flash flood setting with rapid catchment response and local inflows; and Wallis Creek represents a complex flash flood setting influenced by backwater effects and local drainage interactions."

[2] Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, Hydrology and Earth System Sciences, 26, 4013–4032, doi:10.5194/hess-26-4013-2022, 2022.

[3] Silva, S. V. da, Shiguemori, E. H., Shiguemori, A. P. A. de C., Silva, E. V. E., and Santos, L. B. L.: River-level forecasting using LSTM (Long Short-Term Memory) networks integrated with mapping of flood-risk areas, Proceedings of Computer on the Beach, 17, 242–249, doi:10.14210/cotb.v17.p242-249, 2026.

[4] Li, K., Razavi, S., Maier, H. R., Hrachowitz, M., Nabavi, E., Harvey, N., Akhtar, K., and Unduche, F.: Response to Comment on "When are AI models ready for deployment? Reassessing Google's global AI flood forecasting system through the lens of responsible modelling", Journal of Hydrology X, 31, 100230, https://doi.org/10.1016/j.hydroa.2026.100230, 2026.

R2C2) The overlap with the MODSIM2025 conference paper must be clearly clarified. The first author, Jiawei Hou, has published a conference paper entitled "Flash flood nowcasting based on deep learning and radar rainfall estimates" at MODSIM2025, and its abstract substantially overlaps with this manuscript. The authors need to explain this issue, clarifying which parts are a continuation of the conference paper and which parts are new.

Thank you for pointing this out. We would like to clarify the relationship between the MODSIM2025 contribution and the present manuscript.

MODSIM2025 accepts different types of submissions, including full conference papers published in the conference proceedings with individual DOIs, and conference abstracts intended primarily to support conference presentations. Our MODSIM2025 contribution, entitled "Flash flood nowcasting based on deep learning and radar rainfall estimates," was submitted as a conference abstract for an oral presentation, rather than as a full conference paper. It does not have a DOI and does not constitute a full publication of the study.

Our MODSIM2025 abstract provides only a brief summary of the research and therefore naturally overlaps with the present manuscript in terms of the study motivation, methodology, and key findings. However, the present manuscript provides the complete study, including detailed methodology, data processing and experimental design, comprehensive results and evaluation, discussion of model

performance and limitations, and the full scientific interpretation that could not be presented in the conference abstract. Overall, the MODSIM2025 abstract represents preliminary dissemination of the study through an oral conference presentation, whereas the present manuscript provides the full and substantially more detailed scientific account of the work.

R2C3) The use of STEPS-NWP ensemble information requires further justification. In Sections 2.3.2 and 3.3, the manuscript mentions that STEPS-NWP provides eight ensemble members, but the study used the ensemble mean as the decoder input "to maintain consistency with the modeling framework." This choice discards probabilistic forecast information, whereas ensemble information is crucial for uncertainty quantification in operational flood forecasting. The authors need to explain this issue and are encouraged to supplement experiments using ensemble members (or at least ensemble spread) to drive the LSTM, comparing forecast performance between ensemble-mean and ensemble-member inputs. In addition, it is recommended to add a qualitative or quantitative analysis in Section 3.3 or 3.4 of how STEPS-NWP forecast errors propagate to water level forecast errors, so as to enhance the credibility of the framework for operational applications.

We agree that ensemble information is important for quantifying uncertainty in operational flood forecasting. The original analysis presented a deterministic forecasting experiment using the STEPS–NWP ensemble mean rainfall forecasts. To address the reviewer's concern, we have extended the analysis to explicitly examine the variability in water level predictions arising from individual rainfall ensemble members.

In addition to evaluating the RMSE of water level forecasts driven by ensemble mean rainfall, we have extended the analysis to examine the variability in forecast performance across all eight rainfall ensemble members. For both the 90 min and 12 h lead time configurations, we calculated the RMSE over the complete forecast trajectory at each forecast issue time, with forecasts issued at 10 min intervals, for the water level predictions generated by all eight ensemble members. The new Figures 2 and 3 present the minimum and maximum RMSE among the eight member driven water level forecasts, relative to the RMSE of the ensemble mean water level forecast.

The results show that forecast uncertainty varies throughout the flood event and becomes more pronounced during periods of intense rainfall and at longer forecast lead times. In particular, the largest spread in RMSE among ensemble members occurs around 18–23 May 2025, during a period of heavy rainfall. For example, the largest difference between the minimum and maximum member RMSE occurred on 22 May. For the 90 min forecasts issued at 12:40 UTC, member RMSE ranged from 0.24 to 0.44 m, compared with 0.31 m for the ensemble-mean forecast. For the 12 h forecasts issued at 07:50 UTC, member RMSE ranged from 0.16 to 1.46 m, compared with 0.45 m for the ensemble-mean forecast. In contrast, outside periods of heavy rainfall, the differences among ensemble members are relatively small, with their RMSE values remaining close to that of the ensemble mean forecast. Overall, although uncertainties in ensemble rainfall forecasts introduce variability in predicted water levels, the individual

ensemble members generally maintain good predictive performance, even during periods of intense rainfall and at longer forecast lead times.

We have also conducted an additional diagnostic experiment to investigate how uncertainties in STEPS–NWP radar rainfall forecasts propagate into water level prediction errors. In this experiment, the LSTM was trained using all observed rainfall and water level records from 2020–2025, including the May 2025 flood event, and subsequently used to generate water level forecasts driven separately by the eight STEPS–NWP rainfall ensemble members. The training was driven by historical radar rainfall, while the evaluation was driven by forecast radar rainfall. The underlying assumption is that, if the forecast rainfall were accurate, the resulting water level prediction errors would primarily reflect residual LSTM model errors and observational uncertainties, which are expected to be relatively small. Therefore, when the model is driven by forecast rainfall, the resulting water level prediction errors reflect the combined effects of rainfall forecast uncertainties propagated through the LSTM model, residual model errors and observational uncertainties, with rainfall forecast uncertainties expected to be a major contributor. Although these contributions cannot be completely separated in this experiment, the analysis provides useful insights into the influence of rainfall forecast uncertainty on water level predictions. Across the issued member forecasts, mean RMSE increases from 0.15 m at 90 min to 0.19, 0.40, 0.58 and 0.79 m at 3, 6, 9 and 12 h, respectively. Maximum RMSE increases from 0.42 m at 90 min to 1.50 m at 12 h. Mean MAE similarly increases from 0.12 to 0.54 m, while maximum MAE increases from 0.41 to 1.34 m. These results, reported in Table 2, indicate that uncertainties in forecast rainfall contribute to water level prediction errors, with their potential influence becoming more pronounced at longer forecast lead times, although residual model errors and observational uncertainties cannot be fully excluded.

We will incorporate the additional ensemble-based analysis methods and results into Sections 2.3.2 and 3.3, respectively, and add an analysis of forecast error propagation and expand the discussion of forecast uncertainty in Section 3.4. These additions provide a more comprehensive assessment of the influence of rainfall forecast uncertainty on water level predictions and highlight the importance of considering individual ensemble members alongside the ensemble mean forecast, particularly during major flood events.

We have also incorporated ensemble forecasts into the comparison with the zero-skill model benchmark, evaluating absolute and relative mean skill gains across forecast lead times, variability among the eight ensemble members, and the distribution of skill gains across all ensemble members and all forecast issues. Further details are provided in our response to R2C6.

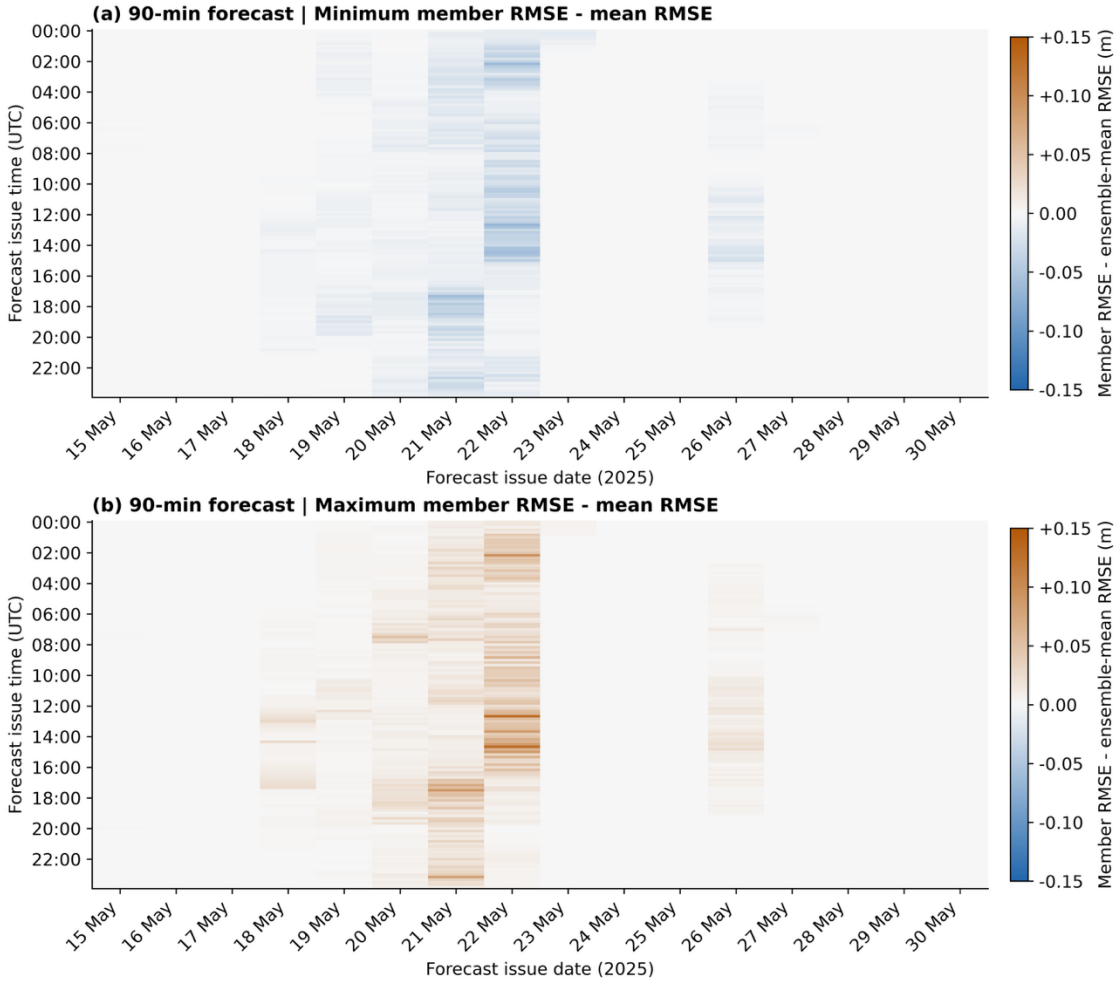


Figure 2. Differences in root-mean-square error (RMSE) relative to the ensemble mean forecast for the 90-min forecast during the May 2025 flood event at the Wollombi site. Panels (a) and (b) show the minimum and maximum RMSE across the eight ensemble members minus the ensemble-mean forecast RMSE, respectively, evaluated at forecast issue times at 10-min intervals. RMSE is calculated over the complete 90-min forecast results at each issue time.

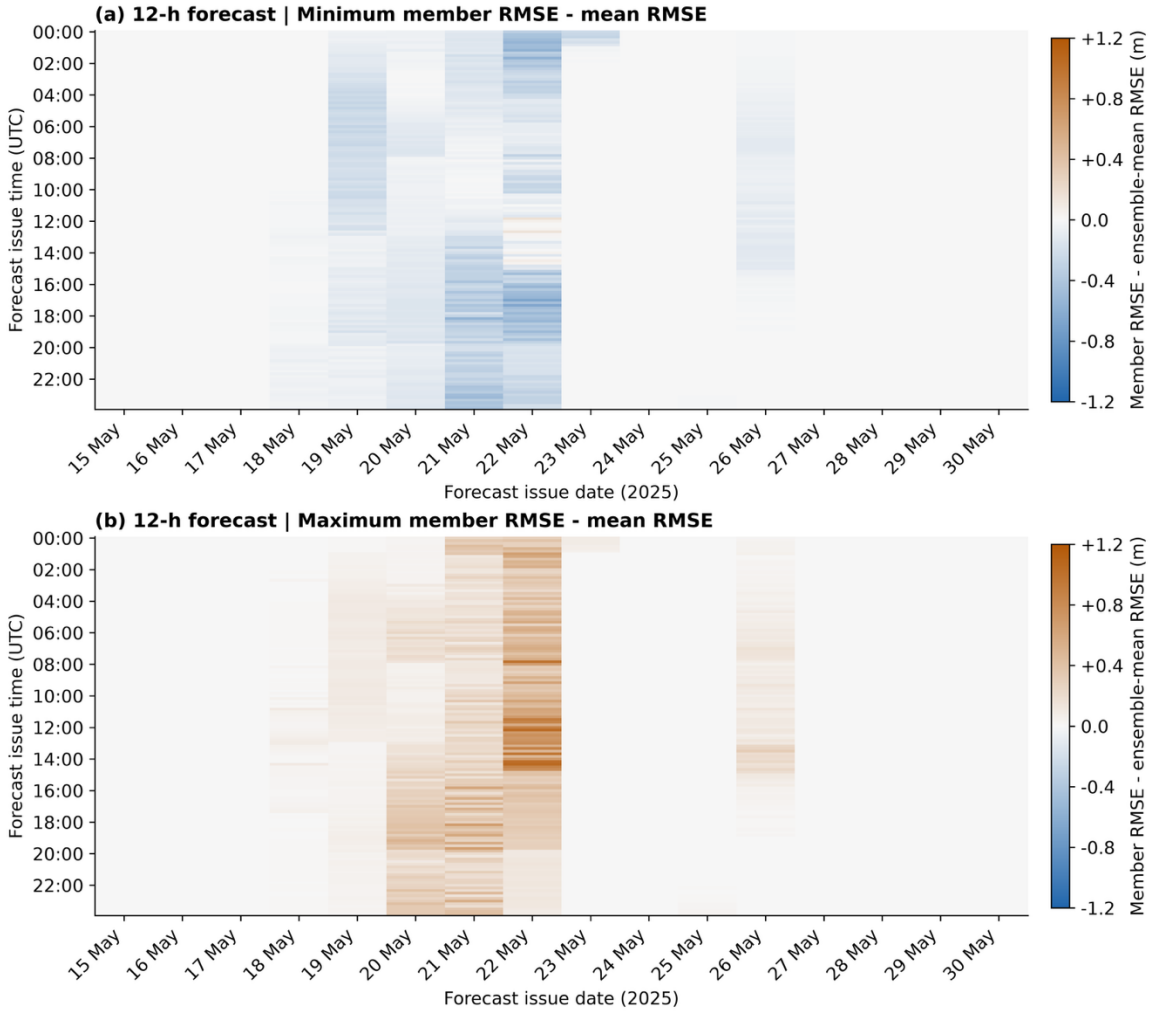


Figure 3. Differences in root-mean-square error (RMSE) relative to the ensemble mean forecast for the 12-h forecast during the May 2025 flood event at the Wollombi site. Panels (a) and (b) show the minimum and maximum RMSE across the eight ensemble members minus the ensemble-mean forecast RMSE, respectively, evaluated at forecast issue times at 10-min intervals. RMSE is calculated over the complete 12-h forecast results at each issue time.

Table 2. Mean and maximum root-mean-square error (RMSE) and mean absolute error (MAE) of Wollombi water level forecasts at lead times of 90 min, 3 h, 6 h, 9 h and 12 h during the May 2025 flood. The LSTM was trained using observed historical rainfall and water levels from all usable 2020–2025 records, including the target event, and subsequently driven by STEPS–NWP ensemble radar rainfall forecasts to examine rainfall error propagation into water level predictions. Ranges cover all issued forecasts across the eight ensemble members.

Lead time	Mean RMSE (m)	Maximum RMSE (m)	Mean MAE (m)	Maximum MAE (m)
90 min	0.15	0.42	0.12	0.41
3 h	0.19	0.58	0.12	0.52
6 h	0.40	1.03	0.27	0.90
9 h	0.58	1.29	0.41	1.16
12 h	0.79	1.50	0.54	1.34

R2C4) The sources of error in the HAND method have not been decomposed. Section 3.2 presents comparisons of inundation extents with Sentinel-1/2 (OA 94–96%, CSI 63–85%), and Section 3.4 discusses several limitations of the HAND method, such as the uniform hydraulic gradient assumption, neglect of temporary storage, and the inability of LiDAR to penetrate water surfaces. However, the manuscript does not evaluate the specific contribution of these simplifications to inundation mapping errors. In the current results, errors may arise simultaneously from LSTM water level prediction errors, HAND terrain simplification errors, and errors in the satellite products themselves (e.g., omission within river channels in Sentinel-1 and vegetation occlusion in Sentinel-2). The authors are advised to add an error decomposition analysis, for example, by driving HAND mapping with observed water levels rather than LSTM-predicted water levels and comparing the results with those driven by LSTM predictions, so as to separate the contributions of water level errors and terrain simplification errors. If full decomposition is not possible, at least clearly state the limitations of the current validation in the discussion and provide a qualitative judgment of the error sources. In addition, consideration could be given to using a high-accuracy hydrodynamic model for comparison over selected events to evaluate the impact of HAND simplifications.

Thank you for this helpful suggestion. We agree that the original comparisons with satellite derived inundation maps combine several sources of error and do not separately quantify the effects of water level prediction errors. Following the reviewer's recommendation, we have applied the HAND mapping approach using observed gauge water levels for the nine inundation validation cases and compared the resulting maps with those generated using LSTM predicted water levels at 90 min, 6 h and 12 h lead times.

In this comparison, the HAND terrain representation and mapping procedure were held fixed. The HAND flood inundation maps generated using observed water levels therefore provide a reference for assessing the additional changes in inundated area associated with water level prediction errors. The relative area differences are reported in Table 3, with negative values indicating a smaller inundated area and positive values indicating a larger area relative to the HAND inundation maps generated using observed water levels. At the 90 min lead time, relative area differences range from −2.13% to −0.01% across the nine cases. All differences are negative, indicating slightly smaller mapped areas when predicted water levels are used. At 6 h, the differences range from −4.81% to +4.13%, while at 12 h they range from −8.80% to +4.57%. The mean absolute relative area differences across the nine cases are approximately 0.85%, 2.83% and 5.12% for the 90 min, 6 h and 12 h predictions, respectively. These results indicate that water level prediction errors introduce relatively small changes in total mapped inundation area across the evaluated lead times, although the differences generally increase with lead time.

This experiment provides a partial separation of error sources by quantifying the sensitivity of mapped inundation area to water level prediction inputs within the existing HAND framework. However, it does not constitute a complete decomposition of errors relative to satellite observations. In particular, neither the HAND-derived inundation maps generated using observed water levels nor the satellite-derived

inundation products represent an independent, error-free observation of the actual flood extent, as both are subject to uncertainties and limitations. We will therefore clarify that the reported OA and CSI values measure the agreement between the complete inundation mapping workflow and the satellite derived inundation products. These metrics cannot be attributed exclusively to either HAND mapping errors, LSTM prediction errors or errors in the satellite products themselves. The relatively small differences in total mapped inundation area across different lead times suggest that water level prediction errors have a limited influence on the total inundated area in the evaluated cases. However, these results do not establish that terrain simplification is the dominant remaining source of error. Other potential sources include the omissions or classification errors in satellite-derived inundation products. The individual contributions of these error sources cannot be quantified from the present comparison. We also agree that an independently calibrated and validated hydrodynamic model would provide a useful additional benchmark for examining the effects of HAND simplifications. Such a comparison has not been included in the present analysis. We will acknowledge this limitation and identify it as a direction for future evaluation.

This additional analysis will be presented in the revised manuscript, with its interpretation and the remaining sources of uncertainty discussed explicitly in Section 3.2.

Table 3. Relative errors (%) in inundated area between HAND-based flood maps generated from predicted water levels at 90-min, 6-h and 12-h lead times and those generated from observed gauge water levels for nine flood inundation validation cases.

Time (UTC)	90-min lead time water level prediction driven flood area error (%)	6-h lead time water level prediction driven flood area error (%)	12-h lead time water level prediction driven flood area error (%)
2021-03-27 00:00:00	−1.39	−4.72	−8.80
2022-07-08 00:00:00	−0.02	−1.37	−2.29
2022-07-13 00:00:00	−0.14	+0.37	+3.51
2022-07-05 19:15:00	−1.04	+4.13	+2.22
2022-07-08 08:45:00	−0.01	−1.46	−2.36
2022-07-12 19:00:00	−0.12	+1.73	+4.57
2021-03-24 19:15:00	−1.37	−4.46	−8.06
2021-03-27 08:45:00	−1.46	−4.81	−8.58
2021-03-31 19:00:00	−2.13	−2.38	−5.68

R2C5) The selection of validation events and the temporal split of training data require further justification. Section 2.3.2 states that the historical prediction experiments used data from 2020 and 2022–2025 for training, skipping the entire year 2021, and then used the March 2021 flood event as an independent validation case. Although this approach avoids data leakage, it also means that the validation event is not within the temporal

neighborhood of the training distribution. For data-driven models such as LSTM, the temporal coverage of training data directly affects their ability to generalize to extreme events. In addition, the manuscript notes that radar rainfall data currently cover only approximately 2020–2026, so the amount of training data is relatively limited. The authors are advised to explain why 2021 was skipped rather than using other splitting strategies (e.g., proportional random splitting or event-based splitting); if 2021 was excluded because of data quality or event-specific characteristics, this should be clearly stated.

We understand the reviewer's concern regarding the training and evaluation strategy. In the original analysis, the March 2021 flood was selected as an independent evaluation event because it represented an impactful flood event within the limited period of radar rainfall data coverage and provided a representative case for evaluating model performance and conducting sensitivity analyses across different input configurations. The selection was not motivated by data quality concerns or other event specific characteristics. The entire year of 2021 was excluded from training to ensure temporal separation between the training and evaluation datasets.

To further address the reviewer's concerns regarding the robustness of the training and evaluation strategy, we have conducted an additional leave-one-event-out reanalysis involving eight flood events from 2022 and 2025 across all three study sites. This additional evaluation provides a more comprehensive assessment of model performance across different flood magnitudes and hydrometeorological conditions, while maintaining independence between training and evaluation events. We will clarify the rationale for the original temporal split in Section 2.3.2 and incorporate the additional validation method and results into the revised manuscript.

Please refer to our detailed response to R1C2 for the training and evaluation procedures and quantitative results of the additional leave-one-event-out analysis.

R2C6) The interpretation of operational evaluation metrics needs to be more cautious. Table 3 presents operational metrics for the 90-minute and 12-hour lead times. For the 90-minute configuration, the threshold exceedance lead times are "+2 h 40 min" (minor) and "+20 min" (moderate), both earlier than the observed exceedance times; for the 12-hour configuration, the minor threshold lead time is as long as "+1 d 1 h 50 min." The manuscript considers that "these detections remain true positives" but that "excessively early warnings may lead to premature mobilization of emergency resources." This interpretation is reasonable, but a more in-depth discussion is needed: the cost of excessively early warnings in operational flood warning can be very significant, such as reduced public trust and wasted emergency resources, and the manuscript needs to discuss how to balance "not missing alarms" and "not issuing false alarms." For the 12-hour configuration, the flood peak timing error is "+4 h 50 min" (early bias) and the flood peak magnitude error is "+0.14 m"; for operational applications, is this timing error acceptable? This should be discussed in the context of the response time of the specific catchment. In addition, the comparison with the zero-skill benchmark (Figure 7) shows that the LSTM outperforms persistence forecasts at all lead times, but the absolute skill gain at 12 hours is about 0.29 m; the significance of this gain for actual warning decisions needs to be discussed.

We agree that the operational value of flood forecasts should be assessed not only by prediction accuracy but also by the timeliness, consistency and reliability of flood warnings (e.g., Li et al., 2026a, b). We will therefore revise the interpretation of the operational evaluation metrics and expand the discussion to address the balance between providing sufficient advance warning and minimising unnecessary false alarms. In particular, we will add the assessment of warning consistency across successive forecast issues, which shows that the early indication of minor flooding was followed by consistent predictions of threshold exceedance, rather than intermittent warnings that were subsequently withdrawn. This provides further evidence of the potential usefulness of early flood indications, even when they occur considerably earlier than the observed threshold crossing. Nevertheless, we acknowledge that excessively early warnings may lead to premature mobilisation of emergency resources and reduced public confidence, while delayed or missed warnings may have more serious consequences for communities at risk. We will therefore expand the discussion to consider the trade-off between missed and false alarms, taking into account forecast uncertainty, warning consistency and the consequences of different warning decisions. We also emphasise the importance of co-designing operational flood forecast guidance with emergency management agencies to ensure that acceptable uncertainties in forecast timing and magnitude reflect operational needs and support effective warning decisions.

Regarding flood peak timing error, we acknowledge its relevance to emergency response. However, from an operational perspective, we consider how early the model can provide reliable indications of consequential flood levels, particularly minor and moderate flood thresholds, and the approaching flood peak, together with the accuracy of predicted flood severity, to be more directly relevant to warning decisions (Li et al., 2026a). This is particularly important in rapidly responding catchments such as Wollombi, where the time available between the onset of flood impacts and the flood peak can be critical for emergency preparedness and response. Although the 12 h forecasts exhibited an early peak timing bias of 4 h 50 min, they provided an indication of the observed flood peak 18 h 50 min in advance, with a relatively small peak magnitude error of +0.14 m. This substantial advance indication provides potentially valuable preparation time, even though the predicted peak timing is biased. Nevertheless, the operational acceptability of the peak timing error cannot be established from this evaluation alone and requires consideration of the catchment's flood response timescale and the time needed for protective actions. We have therefore revised Table 3 to emphasise the advance indication of flood warning thresholds and flood peaks, together with peak magnitude errors, as key operational evaluation metrics, while acknowledging the importance and limitations of peak timing accuracy.

We have also expanded the comparison with the zero-skill model benchmark to better demonstrate the additional predictive value of the LSTM forecasts. The revised analysis evaluates both absolute and relative skill gains across different forecast lead times, as well as the distribution of improvements across individual forecast issues. The revised analysis incorporates ensemble results and examines the distribution of skill gains across individual forecast issues at different lead times. While the original analysis presented average skill gains across all forecast issues, the revised analysis further examines the

distribution of skill gains for individual forecasts. The additional results show that the majority of forecasts achieve positive skill gains, with substantial improvements exceeding 0.5 m in many cases, although some forecasts exhibit limited or no improvement over the zero-skill model benchmark.

We will incorporate these revisions into Results and the Discussion to provide a more comprehensive and balanced assessment of the potential operational value and limitations of the proposed approach. The detailed revisions are provided below.

> *"Table 4 summarises the performance of the ensemble mean LSTM forecasts for the May 2025 flood event at Wollombi, focusing on the advance indication of flood warning thresholds and peak water level prediction errors. First alert advance is defined as the interval between the first forecast issue indicating exceedance of a specified threshold and the observed exceedance. The 90 min forecasts first indicated minor flooding (6.7 m) 2 h 40 min before the observed threshold crossing and moderate flooding (8.6 m) 20 min before the observed crossing. The corresponding advances for the 12 h forecasts were 25 h 50 min and 14 h 40 min, respectively. The observed peak water level of 9.15 m was first indicated 1 h 50 min and 18 h 50 min in advance by the 90 min and 12 h forecasts, respectively, with corresponding peak magnitude errors of +0.03 and +0.14 m. These results demonstrate the potential of the LSTM forecasts to provide valuable advance indications of flood warning thresholds and accurate predictions of peak water levels. Early indications of moderate flooding, together with advance predictions of the flood peak, can provide valuable preparation time for emergency management agencies to plan road closures, pre-position specialist flood rescue personnel and equipment, and arrange adequate staffing. Although the nominal forecast horizon was 12 h, the forecasts first indicated minor flooding 25 h 50 min before the observed threshold crossing, indicating that the initial forecasts predicted threshold exceedance too early. However, this early indication of minor flooding was followed by consistent predictions of threshold exceedance across successive forecast issues. This consistency supports the potential usefulness of the early indication for preparedness, despite the initial error in the predicted timing of threshold exceedance.*
>
> *Nevertheless, excessively early warnings can present operational challenges, including premature mobilisation of emergency resources and potential reductions in public confidence. Similarly, forecasts that repeatedly trigger and withdraw warnings, or provide insufficient warning time before the observed threshold crossing, may reduce their operational value. These consequences also depend on how forecast information is interpreted and used. Demeritt et al. (2013) found that European Flood Alert System alerts encouraged greater vigilance among forecasters but generally did not, by themselves, lead to additional preparatory action or earlier warnings to emergency services or the public. This suggests that early model indications can support closer monitoring while forecasters assess whether external warnings are warranted. Furthermore, in a study of forecast-based humanitarian financing, Coughlan de Perez et al. (2016) incorporated disaster managers' willingness to undertake specific preparedness actions when flooding might not subsequently occur into the selection of forecast triggers. This approach*

recognises that some costs associated with false alarms may be acceptable when weighed against the expected benefits of early action.

A key challenge for operational flood warning systems is therefore balancing the need to minimise missed flood warnings while avoiding unnecessary false alarms. More sensitive warning criteria may provide earlier indications of potential flooding and reduce the risk of missed events, but can also increase false alarms and premature warnings. Conversely, more conservative criteria may improve warning reliability but reduce the available time for emergency response. Achieving an appropriate balance requires consideration of forecast uncertainty, the consistency of threshold exceedance predictions across successive forecast issues, and the consequences of missed and false alarms. Therefore, while the results highlight the potential of the LSTM model to provide substantial advance warning, they also emphasise the importance of considering warning timing, consistency and reliability alongside prediction accuracy when evaluating its suitability for operational flood forecasting (Li et al., 2026a, b). Operational flood forecast guidance should be co-designed with emergency management agencies to ensure that thresholds for acceptable uncertainty in forecast timing and magnitude reflect operational needs and are incorporated into forecast and warning preparation."

"Figure 4 evaluates the additional predictive skill of the LSTM ensemble forecasts relative to the zero-skill model. For each forecast issue and lead time, skill gain is calculated as the absolute error of the zero-skill model minus that of the LSTM over the forecast window, with positive values indicating improved prediction accuracy. At the 12-h lead time, the ensemble mean forecast achieved a skill gain of approximately 0.29 m, averaged across all forecast issues. The shaded range shows the variation in these issue averaged gains across the eight individual ensemble members, spanning approximately 0.13–0.32 m. The relative skill gain, calculated as the mean gain divided by the mean zero-skill error and expressed as a percentage, increased from 6.3% at 90 min to a maximum of 27.4% at 320 min and remained substantial at 21.5% at 12 h. These results demonstrate that the LSTM model provides additional predictive skill beyond the zero-skill baseline, with particularly notable improvements at longer forecast lead times. The moderate decline in relative gain beyond 320 min may reflect increasing uncertainties in the blended rainfall forecasts, which rely progressively more on NWP predictions beyond 90 min as rainfall forecast skill decreases. Figure 5 further illustrates the distribution of skill gains across individual forecast issues at five selected lead times. Positive skill gains were achieved in 50.5%, 78.5%, 79.2%, 76.5% and 69.1% of forecast issues at 90 min, 3 h, 6 h, 9 h and 12 h, respectively, demonstrating consistent improvements for the majority of forecast issues from 3 h onwards. Notably, positive gains were maintained in approximately 70% of forecast issues even at the longest evaluated lead time of 12 h. Although some forecasts showed limited or no improvement, potentially due to uncertainties in rainfall forecasts, substantial gains were achieved in individual cases, reaching approximately up to 0.67 m at 3 h, 1.23 m at 6 h, 1.83 m at 9 h and 2.25 m at 12 h. Overall, these findings demonstrate the added value of the LSTM approach over the zero-skill baseline, particularly its

ability to improve water level predictions at extended forecast lead times, highlighting its potential for operational flood forecasting and early warning."

[5] Demeritt, D., Nobert, S., Cloke, H. L., and Pappenberger, F. (2013). The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. Hydrological Processes, 27(1), 147–157.

[6] Coughlan de Perez, E., van den Hurk, B., van Aalst, M. K., Amuron, I., Bamanya, D., Hauser, T., Jongman, B., Lopez, A., Mason, S., Mendler de Suarez, J., Pappenberger, F., Rueth, A., Stephens, E., Suarez, P., Wagemaker, J., and Zsoter, E. (2016). Action-based flood forecasting for triggering humanitarian action. Hydrology and Earth System Sciences, 20, 3549–3560.

Table 4 Key performance metrics for the 90-min and 12-h lead-time forecasts during the May 2025 flood event at the Wollombi site.

Key Performance	90-min lead time mode	12-hour lead time mode
Lead time triggering minor flood threshold (6.7 m)	+2h40min	+1d1h50min
Lead time triggering moderate flood threshold (8.6 m)	+20min	+14h40min
Lead time of flood peak (9.15 m)	+1 h 50 min	+18 h 50 min
Flood peak error	+0.03m	+0.14m

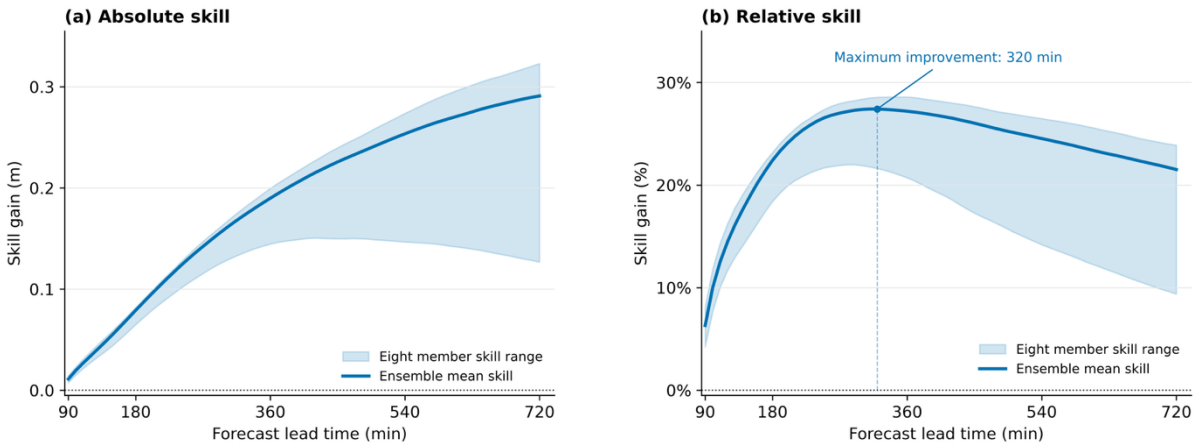


Figure 4. Average ensemble water level forecast skill for Wollombi during the May 2025 flood event. (a) Absolute skill, defined as the zero-skill model error minus the LSTM model error, in metres. (b) Skill improvement expressed as the ratio of LSTM skill gain to the zero-skill model error, multiplied by 100 and shown as a percentage. Blue lines represent the skill of the ensemble mean forecast; shaded bands show the minimum–maximum skill across eight ensemble members.

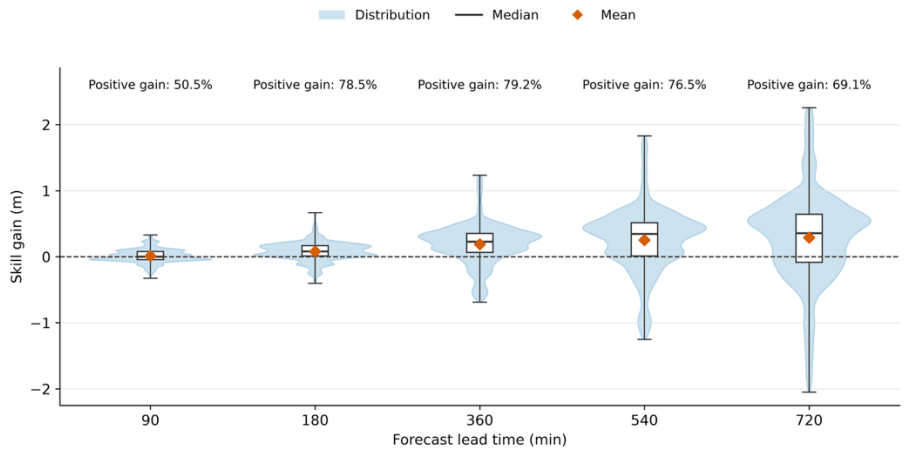


Figure 5. Distributions of skill gains across forecast issues for Wollombi during the May 2025 flood. Skill gain is the zero-skill model absolute error minus the ensemble mean LSTM absolute error at the selected lead time; positive values indicate improved predictions. Blue violins show distributions, boxes span the 25th–75th percentiles, black lines indicate medians, whiskers show the full range, and orange diamonds mark means. Labels indicate the percentage of issues with positive gain.