Responses to Reviewer #1 Comments:

The manuscript presents an encoder–decoder LSTM framework using radar rainfall and gauge observations to forecast water levels across three contrasting flood regimes in the Hunter Valley, Australia, and further converts predicted water levels to inundation extents using high resolution LiDAR-derived HAND. The study evaluates historical predictions and additionally tests operational nowcasting using STEPS-NWP rainfall forecasts. This combination of high-frequency operational radar products, water level forecasting, operationally relevant metrics, and inundation mapping gives the study practical relevance.

The manuscript addresses an important problem, and it contains a useful framework and encouraging results. I particularly appreciate the authors' effort to move beyond conventional error statistics by considering operationally relevant measures such as flood threshold warning lead time, peak magnitude and timing, as well as comparison against a benchmark from a zero-skill model.

Overall, I find the study promising and potentially suitable for publication in the special issue "Early Warning Systems from Research to Operations: Status, Innovations and Multi-Hazard Applications". However, several aspects of the methodology require further clarification and justification, particularly regarding model reproducibility, the validation strategy, and the distinction between historical prediction and operational forecasting. I've provided my comments below to help strengthen the manuscript.

We are grateful to the reviewer for the positive assessment of our manuscript and for the detailed and constructive comments and suggestions, which have helped us improve the quality of the manuscript. Below please find our responses to the reviewer's comments in detail.

R1C1) More detail is needed on the model architecture, training procedure and reproducibility. The methods provide a useful overview of the encoder–decoder LSTM framework, but the description of the actual model implementation and training procedure is currently rather limited. More information is needed on the model architecture and configuration, training and optimisation strategy, and how key hyperparameters were determined. Providing these details would help readers better understand how the framework was implemented and assessed and would improve the reproducibility of the study. It should also be clarified whether separate models were developed for each site, lead time, and input sequence configuration, or whether a common model setup was applied across the different experiments.

Thank you for pointing out this issue. We acknowledge that the original description did not provide sufficient information to reproduce the model implementation and training procedure. We will expand Section 2.3.1 to provide further details on the model architecture and configuration, the training and optimisation strategy, and how the key hyperparameters were determined. We will also clarify the model training configurations across the different experiments, including the use of separately trained models for the reanalysis and sensitivity analyses and a single trained model for the nowcasting analysis, as follows:

"The model implementation uses separate rainfall and water level encoders, each comprising a single LSTM layer with 32 hidden units. Their final hidden and cell states are concatenated and transformed through separate fully connected layers with hyperbolic tangent activation to initialise a 32-unit LSTM decoder. A linear output layer predicts the change in water level relative to the latest observed level. Training uses the AdamW optimiser with a learning rate of 0.001, weight decay of 10^{-5}, a batch size of 128, and a gradient-norm limit of 1. The model is trained by minimising the mean squared error (MSE) on the training data, while MSE calculated over the validation period is used to monitor model performance, apply early stopping, and select the best model checkpoint. Training is limited to 24 epochs, with at least eight epochs executed and early stopping after five epochs without sufficient improvement in the inner-validation loss. For the reanalysis and sensitivity analyses, a separate model is trained for each combination of site, held-out event, and input configuration. The model architecture and optimisation settings are kept consistent across sites and held-out events, while the fitted parameters are estimated independently for each model. For the nowcasting analysis, a single model is trained for the Wollombi site to more closely reflect an operational forecasting setting, in which the same trained model is used to generate forecasts as new observations and rainfall forecasts become available."

R1C2) The validation strategy requires further justification. For the historical experiment, the model is trained using data from 2020 and 2022–2025 and evaluated on the March 2021 flood event. Although holding out an entire flood event is preferable to randomly dividing a highly autocorrelated time series, validation against only one event makes it difficult to assess robustness across different event magnitudes and hydrometeorological conditions. The manuscript also acknowledges the limited radar record and number of available flood events. I recommend either conducting event-based cross-validation/leave-one-event-out testing where possible, or providing a stronger justification for the current split and moderating claims of generalisability accordingly.

Thank you for your suggestion. We acknowledge that evaluation against the March 2021 flood alone provides insufficient evidence to assess the model's generalisability across different flood events. To address this limitation, we have conducted an additional leave-one-event-out reanalysis using eight flood events (i.e., eight high flow episodes shared across the three gauges were identified using site-specific thresholds defined as the 95th percentile of historical hourly maximum water levels) at Greta, Wollombi and Wallis Creek. In the revised manuscript, we will include the event periods and observed peak water levels in Table 1, together with the corresponding event level prediction errors in Figure 2.

The additional analysis covers four events in 2022 and four events in 2025, with each event held out in turn for evaluation. Predictions were assessed at lead times of 15 min, 90 min, 3 h, 6 h and 12 h. These events span a range of flood magnitudes, with observed peak water levels ranging from 5.59 to 14.23 m at Greta, 2.99 to 14.19 m at Wollombi, and 3.40 to 7.17 m at Wallis Creek. This extends the evaluation beyond the original March 2021 case and allows differences in model performance among events to be examined explicitly. The results show consistently small absolute errors at short lead times, followed by

increasing errors and greater variability among events as lead time increases. At 90 min, RMSE ranges from 0.09 to 0.15 m at Greta, 0.07 to 0.16 m at Wollombi, and 0.03 to 0.07 m at Wallis Creek. At 6 h, the corresponding ranges are 0.29–0.60 m, 0.27–0.58 m and 0.10–0.27 m, respectively. At 12 h, RMSE increases to 0.50–1.06 m at Greta, 0.43–1.13 m at Wollombi, and 0.17–0.54 m at Wallis Creek. The event producing the largest 12-h error also differs among sites: E03 at Greta, E01 at Wollombi, and E07 at Wallis Creek. Overall, these results demonstrate that the LSTM model maintains consistently high predictive accuracy at shorter lead times across flood events of varying magnitudes. Although prediction errors increase with lead time, the model continues to show promising performance at longer lead times, including up to 12 h.

These results provide broader evidence of model performance on independently withheld events within the available observational record and therefore provide stronger evidence of the generalisability of the proposed approach across different event magnitudes and hydrometeorological conditions. We will describe the additional validation procedure in the Methods, present the event level results alongside the reanalysis evaluation, and revise the Results and Discussion and Conclusions accordingly.

In addition to the leave-one-event-out reanalysis, we have expanded the forecast analysis for the May 2025 flood event at Wollombi using all eight STEPS–NWP rainfall ensemble members. These additions examine variability in water level forecast errors among ensemble members, investigate the influence of rainfall forecast uncertainty, and compare the ensemble forecasts with the zero-skill benchmark across forecast lead times and individual forecast issues. We will also expand the discussion of operational evaluation metrics, including the advance indication of flood warning thresholds and peak water levels, forecast timing errors, warning consistency, and the balance between missed warnings and false alarms. The discussion also emphasises the importance of co-designing operational guidance with emergency management agencies. Details of these additional analyses, results and proposed manuscript revisions are provided in our responses to Reviewer 2, particularly R2C3 and R2C6.

Table 1. High flow event periods and observed peak water levels (m) at Greta, Wollombi and Wallis Creek used in the leave-one-event-out reanalysis.

Event	Period	Greta	Wollombi	Wallis Creek
E01	2–16 Mar 2022	12.71	11.23	4.18
E02	24 Mar–13 Apr 2022	6.35	4.05	3.86
E03	3–25 Jul 2022	14.23	14.19	7.17
E04	7–14 Oct 2022	8.46	4.70	3.43
E05	27 Apr–2 May 2025	5.59	2.99	3.40
E06	19 May–8 Jun 2025	9.74	9.14	6.35
E07	2–13 Aug 2025	10.05	6.90	5.09
E08	21–26 Aug 2025	6.88	3.31	3.48



Figure 2. Event level prediction errors under leave-one-event-out reanalysis at Greta, Wollombi and Wallis Creek.

R1C3) The distinction between historical prediction and operational forecasting needs to be made more explicit. In the historical evaluation, observed radar rainfall is used as a proxy for future forecast rainfall. Therefore, the very good 6 h and 12 h results in this experiment do not include rainfall forecast uncertainty and are not directly comparable with operational forecasts at the same lead times. This is particularly important because the nowcasting experiment shows larger and more variable errors at 12 h as the rainfall forcing transitions towards NWP. I suggest making this distinction clear throughout the Abstract, Results, Discussion and Conclusions, so that potential predictability using observed rainfall is not confused with forecast skill using forecast rainfall.

Thank you for this helpful comment. We will revise the Abstract, Methods, Results, Discussion and Conclusions to ensure consistent terminology and clearly distinguish the rainfall inputs and evaluation procedures used in each experiment. Specifically, experiments using observed radar rainfall over the prediction window will be described as "reanalysis", while those using forecast rainfall will be described as "forecast". This clarification will distinguish the assessment of model performance under observed rainfall forcing from that under forecast rainfall forcing. We will use "reanalysis lead time" or "prediction horizons" when reporting reanalysis results and reserve "forecast lead time" for forecast evaluations. The term "prediction window" will consistently refer to the entire period over which the model generates water level predictions.

We have also revised Figure 1 (this figure number refers to the original manuscript) below and its caption to distinguish the reanalysis and forecast pathways, including their respective rainfall inputs and outputs, and ensure that the relevant figure and table captions use the same terminology consistently.

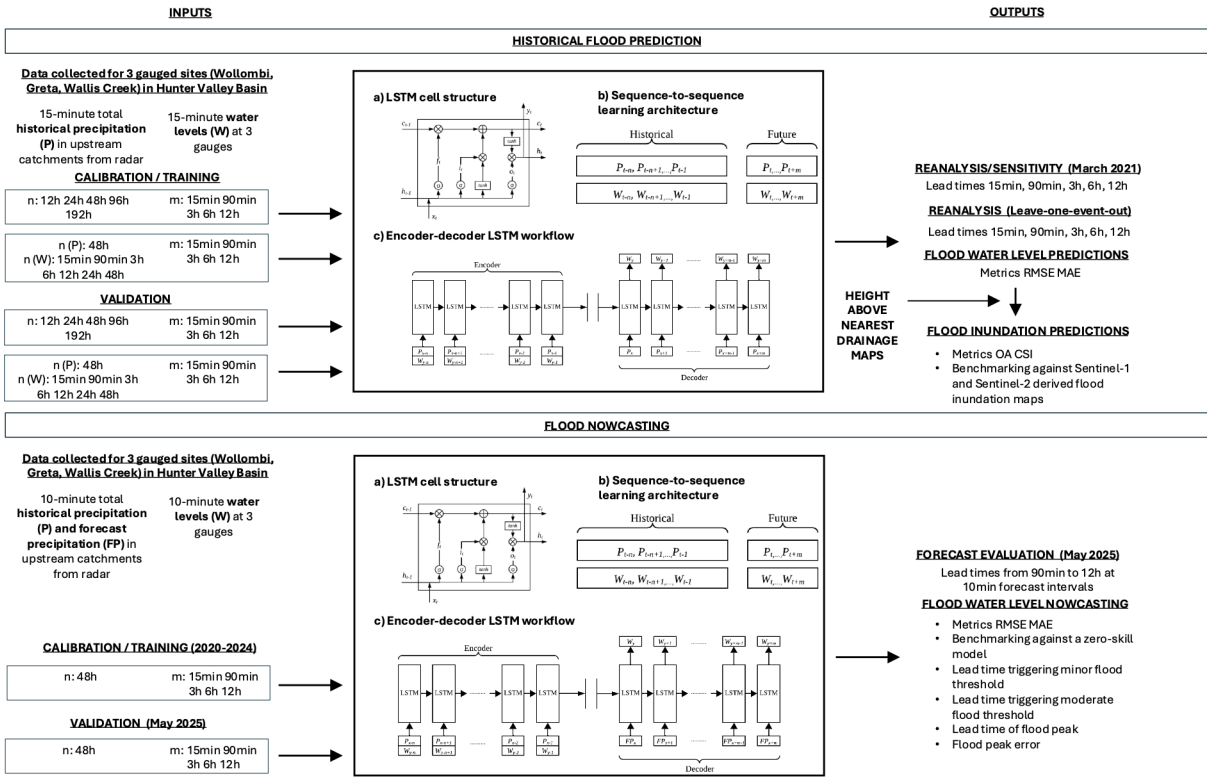


Figure 1 *Schematic overview of the encoder–decoder LSTM framework for reanalysis and forecast evaluation in the Hunter Valley Basin. The figure illustrates the inputs and outputs of the deep learning model, as well as the model structure and workflow for reanalysis using observed rainfall and forecast evaluation using rainfall forecasts. (a) LSTM cell structure (c represents the cell state; h, the hidden state; x, the input variables; y, the output variables; f, the forget gate; i, the input gate; o, the output gate; and t, time.); (b) sequence-to-sequence learning architecture (n denotes the length of the antecedent time series in each sequence, m denotes the length of the future time series and l denotes the full length of time series); (c) encoder–decoder LSTM workflow (P_t denotes the historical upstream total radar rainfall at time t, and W_t denotes the flood water level at time t, FP_t refers to the forecast of upstream total radar rainfall at time t). OA, overall accuracy; CSI, critical success index; RMSE, root mean square error; and MAE, mean absolute error.*

Specific comments:

R1C4) Line 85: This sentence needs a bit more context, e.g., under what kind of hydrological/catchment conditions

We acknowledge that the original statement was too general and did not provide sufficient context regarding the hydrological and catchment conditions under which LSTM models have demonstrated strong performance in the studies by Kratzert et al. (2018) and Lees et al. (2021). We will revise this sentence as follows:

> *"Across a range of hydroclimatic conditions, LSTM networks have achieved performance comparable to or better than lumped hydrological models, particularly in catchments where runoff dynamics are well represented by the available hydro-meteorological inputs, although performance can be lower in arid, intermittent, or groundwater-dominated catchments (Kratzert et al., 2018; Lees et al., 2021)."*

R1C5) Line 130: Perhaps categorise those three catchments based on their hydrological complexities, for example, from relatively simple to complex settings as described in Table 1

We agree that categorising the three study locations according to their hydrological complexity provides a clearer distinction among the sites and is more consistent with Table 1. We will revise the sentence as follows:

> *"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."*

R1C6) Line 160 Fig.1: replace 15m, 90m with 15min and 90min, and briefly describe OA, CSA, RMSE and MAE in the caption

We will make this change to Fig.1 in the revised manuscript. Please see the revised figure in our response to R1C3.

R1C7) Line 169: Typo - Fig.2

We will correct it in the revised manuscript.

R1C8) Line 170 Section 2.1 study area: this section would benefit from a general description of rainfall conditions for your study region, e.g., mean annual rainfall and/or typical rainfall intensity

In the revised manuscript, we will add a general description of the Hunter Valley climate and rainfall, including its mean annual rainfall, spatial variability, and seasonal rainfall pattern, to Section 2.1, as follows:

> *"The region has a subtropical to temperate climate, with an average annual rainfall of approximately 870 mm. Rainfall varies considerably across the region, from more than 1,100 mm yr^{-1} near the coast and Barrington Tops to less than 600 mm yr^{-1} in parts of the Upper Hunter, and is generally greatest during summer and autumn (NSW Office of Environment and Heritage, 2014)."*

[1] Office of Environment and Heritage (OEH) (2014). Hunter Climate Change Snapshot. NSW Office of Environment and Heritage, Sydney, Australia

R1C9) Line 210 Data: A general summary of datasets used would be helpful, such as radar derived rainfall, water level, DEM, etc. Also it would be helpful to briefly describe the resolution of the datasets.

We agree that a concise overview of the datasets and their spatial and temporal resolutions will improve the clarity of the data section. We will add a summary at the beginning of Section 2.2 describing the four main datasets used in the study as follows:

> *"The modelling framework uses four main datasets: radar-derived rainfall estimates (i.e., RAINFIELDS) and forecasts (i.e., STEPS), gauged water level observations, river network and sub-catchment vector data, and high-resolution LiDAR DEMs. Radar rainfall estimates and forecasts are provided as gridded datasets with temporal resolutions of 5–10 min and spatial resolutions of 1–2 km, depending on the product. Gauged water level observations are processed to a regular 15-min temporal resolution. Geofabric data provide the river network and upstream sub-catchment boundaries, while airborne LiDAR-derived DEMs provide high-resolution topographic information at spatial resolutions of 1–5 m."*

R1C10) Line 220: Please clarify the abbreviation ACCESS-CE

We will add the full name of ACCESS-CE and ACCESS-G (i.e., Australian Community Climate and Earth-System Simulator City Ensemble (ACCESS-CE) and Global (ACCESS-G)) in the revised manuscript.

R1C11) Line 225: It is unclear to me whether your water level datasets were directly sourced from the Bureau online portal or from other partners.

In the revised manuscript, we will clarify the provenance of the water level data as follows:

> *"Gauged water level data were sourced from the Australian Bureau of Meteorology database, which underpins Water Data Online (http://www.bom.gov.au/waterdata/). The underlying observations are collected by a range of partner providers across Australia (and by Bureau-owned flood sites), and are quality-controlled by the Bureau prior to publication."*

R1C12) Line 255-260: The description states that the "future segment was used by the decoder" and that the decoder takes "radar-based rainfall nowcasts" as input. It is unclear whether this description applies to both model training and prediction. During historical training, are the decoder inputs observed future rainfall, while during operational forecasting they are STEPS-NWP rainfall forecasts? Please clarify this distinction.

We will add sentences to clarify this distinction in the revised manuscript as follows:

> *"Historical rainfall inputs were obtained from the observed RAINFIELDS dataset, which was also used to provide rainfall over the future prediction window during training and historical flood reanalysis, representing a perfect rainfall forecast scenario. In contrast, during forecast evaluation, the future rainfall inputs were obtained from STEPS-NWP forecasts available at each forecast issue time. The encoder's hidden and cell states were then passed to the decoder, which used these future rainfall inputs*

to predict water levels over the prediction window."

R1C13) Line 281: This should move to the data section

We will move this sentence to the data section in the revised manuscript.

R1C14) Lines 289–290: Please clarify "15-minute total precipitation … by aggregating observations over upstream catchment areas." Does the total refer to temporal accumulation of the original 5-min rainfall data, spatial aggregation over the catchment, or both? If spatially aggregated, please specify whether the catchment mean, sum, or another statistic was used and provide the corresponding units.

We will added sentences in the method section to clarify this point in the revised manuscript as follows:

> *"Rainfall inputs were derived by spatially aggregating gridded radar rainfall data over the upstream catchment of each gauge, as delineated using Geofabric data, and temporally accumulating the 5-min radar rainfall estimates to match the model time steps (15 min for historical flood reanalysis and 10 min for forecast evaluation)."*

R1C15) Line 315-320: The description of the STEPS NWP product should move under the data section

We will remove the description of the STEPS-NWP product here, as it is already provided in the Data section of the manuscript.

R1C16) Line 340: A table showing the different thresholds for minor, moderate, and major flooding would be helpful here.

We will add thresholds information in Table 3 (this table number refers to the original manuscript) in the revised manuscript.

R1C17) Lines 428–445: The statement that the sensitivity analysis "eliminates the possibility" that model performance arises solely from water level autocorrelation seems too strong. The results show that increasing the length of the historical water level sequence substantially improves performance at longer lead times. I suggest revising this to say that the analysis provides evidence that the model does not rely solely on water-level autocorrelation. Similarly, the claim that the model "effectively learns hydrological response patterns" should be phrased more cautiously unless this is demonstrated independently.

Thank you for this insightful comment. We agree that the original statements were overly strong. We will revise the text to state that the sensitivity analysis provides evidence that model performance does not rely solely on water level autocorrelation, and we will moderate the interpretation regarding the model's ability to capture hydrological response patterns accordingly. We will revise this paragraph as follows:

"The additional sensitivity analysis conducted in this study provides evidence that the LSTM model's predictive performance does not rely solely on autocorrelation in the water level time series. This is demonstrated by the RMSE and MAE values obtained for configurations using different historical water level input sequence lengths (15 minutes, 90 minutes, 3 hours, 6 hours, 12 hours, 24 hours, and 48 hours), while the rainfall input sequence length was fixed at 48 hours during training, across different lead times at the three study locations (Table S1). At shorter lead times (e.g., 90 minutes), model performance remained relatively stable across different water level sequence lengths. Specifically, RMSE values ranged from 0.09–0.19 m at Wollombi, 0.05–0.15 m at Greta, and 0.04–0.10 m at Wallis Creek, with corresponding MAE values of 0.06–0.10 m, 0.03–0.12 m, and 0.02–0.07 m, respectively. In contrast, at longer lead times (e.g., 12 hours), model accuracy generally improved as the length of the historical water level sequence increased. The RMSE decreased from 0.69 m to 0.39 m at Wollombi, from 0.69 m to 0.45 m at Greta, and from 0.39 m to 0.27 m at Wallis Creek. These results indicate that longer records of antecedent water level conditions provide useful information for longer lead forecasts. However, the model retained predictive skill even when only a short history of water level was provided, suggesting that its performance cannot be attributed solely to persistence or autocorrelation in the water level series. Together with the inclusion of rainfall as a predictor, these findings are consistent with the model making use of information beyond water level autocorrelation, although this sensitivity analysis alone does not allow the contribution of specific hydrological response processes to be isolated."

R1C18) Line 474: Was the "significant difference" tested statistically? If not, please consider rewording.

We will change this to "relatively large differences" in the revised manuscript.