https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Flood nowcasting based on deep learning and radar rainfall estimates: A reliable and efficient framework for diverse flood regimes

Jiawei Hou[1], Wendy Sharples[2], Jayaram Pudashine[2], Mandi Thran[2], Carlos Velasco-Forero[2], Paul Fox-Hughes[2], Elisabetta Carrara[2] and Holger R. Maier[3]

[1]Bureau of Meteorology, Canberra, Australian Capital Territory, Australia
[2]Bureau of Meteorology, Melbourne, Victoria, Australia
[3]School of Civil Engineering and Construction Management, Adelaide University, Adelaide, South Australia, Australia

Correspondence to: Jiawei Hou (jiawei.hou@bom.gov.au)

Abstract. Floods pose a severe risk to lives, infrastructure, and ecosystems, necessitating accurate and timely forecasts to support early warning and emergency response. Flash floods in particular are among the most destructive flood hazards due to their rapid onset, short response times, and thus limited time for warning. However, physics-based or conceptual hydrological models often struggle to deliver reliable short lead-time predictions, particularly in small or fast-responding catchments where complex and rapidly evolving hydrological processes are at play. Additionally, in the case of flashy catchments, there is often insufficient time to run numerical models and issue timely warnings. This study explores the use of encode-decode Long Short-Term Memory (LSTM) networks for short-term flood forecasting using high-frequency radar rainfall and river level data across three locations representing diverse flood regimes in the Hunter Valley in Australia. Validation across these locations demonstrates strong agreement between predicted and observed flood levels, with an average RMSE of 0.09 m and MAE of 0.06 m at a 90-minute lead time, and 0.45 m (RMSE) and 0.24 m (MAE) at a 12-hour lead time. By coupling LSTM-predicted water levels with airborne LiDAR-derived digital elevation models (DEMs), inundation maps were generated to translate point-based flood level predictions into spatially distributed flood extent information. These maps showed high agreement with Sentinel-2 and Sentinel-1derived flood products, achieving over 94% overall accuracy and up to 84.5% critical success index at a12-hour lead time. We also demonstrated the workflow's operational capability, achieving accurate flood level forecasts supported by radar-based rainfall nowcasts and numerical weather predictions, albeit lower quality longer 12-hour forecasts when input precipitation forecasts have high uncertainty and error. Overall, this study presents a scalable, data-driven approach for real-time flood nowcasting, providing a practical tool to support early warning systems and inform emergency planning in vulnerable regions.

1. Introduction

Flooding is one of the most damaging natural hazards worldwide, causing substantial loss of life, disruption to communities, and damage to infrastructure and ecosystems. Effective flood forecasting plays a critical role in reducing these impacts by supporting early warning systems, emergency response, and risk-informed decision making. However, flood forecasting remains challenging because flood generation processes vary considerably across catchments, reflecting differences in climate, topography, land cover, drainage networks, and

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

hydrological response characteristics. These challenges are particularly evident for flash floods, which can develop within hours of intense rainfall and provide limited time for warnings and emergency actions (Knocke and Kolivras, 2007; Collier, 2007). In addition, urbanisation can exacerbate flood risk by increasing impervious surface cover, resulting in higher runoff rates, reduced infiltration, and shorter response times (Shuster et al., 2005). The highly localised nature of flash floods and the complex, non-linear interactions among rainfall, topography, soil moisture, and drainage conditions make them especially difficult to predict accurately (Zoccatelli et al., 2010). These challenges are expected to become more pronounced under climate change, which is projected to increase the frequency and intensity of extreme rainfall events, while ongoing development continues to expand exposure in flood-prone areas (Tellman et al., 2021; Westra et al., 2014).

Simulating floods is inherently complex, requiring the integration of multiple hydrological and hydrodynamic processes. The accuracy of flood predictions is challenged by uncertainties in input data, model structure and parameterization, and the influence of land cover, soil properties, and human infrastructure. A wide range of modelling approaches have therefore been developed and applied operationally, including lumped to distributed catchment implementations, and conceptual through to fully physically based models, each offering distinct advantages as well as limitations. Lumped models provide computationally efficient basin-scale simulations but may fail to capture localized extreme rainfall responses, while distributed models better represent spatial heterogeneity at the cost of high data requirements, parameter uncertainty, and computational burden (Haruna et al., 2022). Conceptual models calibrated to historical events can provide robust operational forecasts but may struggle to extrapolate to unprecedented extremes, whereas empirical models are constrained by limited observations of rare flood events (Voit and Heistermann, 2024). Physically based hydrological and hydrodynamic models simulate governing processes in detail and remain essential for scenario analysis and physically interpretable forecasts. However, their operational application is often restricted by high-resolution data requirements, sensitivity to initial and boundary conditions, parameter equifinality, and long runtimes (Vieux et al., 2004; Bhanja et al., 2023). Large-scale hydrodynamic models such as LISFLOOD-FP (Bates and De Roo, 2000) and CaMa-Flood (Yamazaki et al., 2011) have demonstrated robust predictions of flood timing and magnitude in major river systems, though their reliability can be affected by uncertainties in precipitation forcing, channel and floodplain geometry, and roughness parameterization (Paiva et al., 2013; Fleischmann et al., 2019). These challenges are particularly pronounced for flash floods in small, steep catchments, where rapid hydrological response, limited calibration data, and the computational cost of resolving fine-scale dynamics can hinder timely and accurate forecasts. Together, these considerations highlight that no single modelling paradigm universally outperforms others, and that model selection involves trade-offs between physical realism, data availability, computational cost, and forecast timeliness.

In recent years, machine learning (ML) and deep learning (DL) approaches have emerged as complementary tools for flood forecasting, offering the ability to learn complex non-linear relationships directly from data without explicit formulation of physical processes. Artificial neural networks (ANNs) have been widely employed for rainfall-runoff simulations, demonstrating performance comparable to physical models in streamflow prediction while significantly improving computational efficiency (Dawson and Wilby, 1998; Hsu et al., 1995; Sudheer et al., 2002; Sajikumar and Thandaveswara, 1999). Feed forward neural networks have been

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

applied to analyse time series data, but they assume independence between data points and erase their state after each step, making them unsuitable for capturing temporal dependencies (Chiang et al., 2004; Li et al., 2021), unless these are represented explicitly via lagged model inputs (Galelli et al., 2014; Maier et al., 2023). As a result, recurrent neural networks (RNNs) have gained popularity for their ability to capture sequence dynamics while retaining the functions of feedforward neural networks (Lipton et al., 2015). Long short-term memory (LSTM) networks emerged as an advancement over traditional RNNs and were designed to overcome the vanishing gradient problem, where the influence of earlier inputs fades as information is passed through many steps, and to better capture long-range dependencies in sequential data (Gers et al., 2000; Hochreiter, 1998). Kratzert et al. (2018) and Lees et al. (2021) found that LSTM networks can predict runoff with performance comparable to, or even exceeding, that of lumped hydrological models. The LSTM-based Encoder-Decoder model has also been proven to provide accurate and reliable flood forecasts across various lead times (Kao et al., 2020). It has also been applied globally to predict extreme flood events with lead times of up to five days, outperforming the current state-of-the-art global river discharge nowcasting system, the Copernicus Emergency Management Service Global Flood Awareness System (Nearing et al., 2024). Feng et al. (2020) implemented an LSTM-based data assimilation approach to integrate recent discharge measurements, enhancing forecast accuracy and reliability. Furthermore, Kratzert et al. (2019) demonstrated the effectiveness of LSTM approaches in predicting streamflow in ungauged basins, achieving skill levels comparable to those of hydrological models in gauged basins. Nevertheless, ML/DL models remain sensitive to training data quality, non-stationarity under climate change, and limited extrapolation capability during unprecedented extremes. Consequently, current research increasingly views ML/DL approaches as part of a continuum of methods (Mount et al., 2016) and therefore not as replacements for process-based models, but as complementary tools that can enhance real-time forecasting, computational efficiency, and data assimilation when used alongside physically based modelling frameworks.

The advanced application of LSTM networks in modelling the rainfall-runoff process has demonstrated significant success in streamflow prediction, as shown in the studies discussed above. More recently, research has begun to explore the use of LSTM models for flash flood prediction (Oddo et al., 2024). However, a critical factor in effective flash flood nowcasting for emergency services, which is essential for saving lives and protecting property, is the provision of timely early warnings enabled by real-time forecasting (Gissing et al., 2010; Ebert et al., 2024). In practical operational settings, real-time data on variables such as evaporation, soil moisture, and drainage are typically unavailable, while real-time rainfall data can be reliably obtained from radar systems. Therefore, an important question arises: instead of training LSTM models on a wide range of hydrological inputs, can models trained solely on real-time radar rainfall data effectively nowcast floods with short lead times (from a few minutes to a few hours)? Another major challenge is that forecasted rainfall inputs often deviate from observed values in both intensity and location, reducing prediction accuracy. Leveraging calibrated real-time radar rainfall can better capture the actual rainfall distribution within the catchment, which is critical not only for flood nowcasting but also for modelling general flood processes. To achieve this, calibrated radar rainfall nowcasting, which leverages high-resolution, real-time radar data to predict near-term precipitation dynamics, can be integrated with LSTM models to form an end-to-end, operationally oriented nowcasting system. This study aims to develop an encoder–decoder LSTM-based flood nowcasting system

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

using operational real-time radar rainfall data from the Australian Bureau of Meteorology. This operational radar rainfall data provides high-frequency updates (every 5 to 15 minutes), including historical estimates from RAINFIELDS and nowcasting products such as STEPS-ADV (radar-based nowcasting up to 90 minutes (Velasco-Forero et al., 2020)) and STEPS-NWP (which blends radar and numerical weather prediction for nowcasting up to 12 hours (Seed et al., 2013)).

This study develops and assesses the utility of an LSTM-based flood nowcasting framework for predicting flood water levels at lead times from 15 minutes to 12 hours and, using high-resolution LiDAR-derived DEMs, estimates corresponding flood inundation extents. The central objective of this study is to assess whether a deep learning-based flood nowcasting system driven by radar rainfall estimates can provide reliable predictions of flood water levels, flood peaks, and inundation extents across contrasting flooding environments. To evaluate the generalisability of the approach, the framework is tested at three forecasting locations representing a gradient of hydrological conditions and flood types: Greta, a large low-gradient catchment characterised by relatively simple riverine flooding and a slow hydrological response; Wollombi, a small steep catchment with a rapid response and riverine flash-flood characteristics; and Wallis Creek, a small steep catchment with a hydrologically complex flash-flood regime influenced by backwater effects (Table 1). By evaluating model performance across these diverse settings, this study investigates whether a single deep learning framework can support operational flood nowcasting across a wide range of flood types and hydrological conditions, thereby enhancing early warning capabilities and emergency response.

2. Data and Methods

Figure 1 provides a schematic overview of the modelling framework, illustrating how the components of the encoder–decoder LSTM approach are interconnected. The figure summarises the complete workflow, including data inputs, model architecture, training and validation strategies, and resulting outputs for both historical flood prediction and real-time flood nowcasting. The aim of this study is to assess the utility of LSTM-based models for flood prediction by evaluating how well they can (i) reproduce historical flood water levels and associated inundation extents, and (ii) provide reliable short-term flood nowcasts in operational settings. This assessment is conducted across three contrasting catchments in the Hunter Valley, Australia, representing flash flood dominated, main channel, and backwater influenced flooding conditions. The modelling framework integrates multiple data sources, including radar-derived historical rainfall estimates (RAINFIELDS), radar-derived and radar-NWP blended rainfall forecasts (STEPS-ADV/STEPS-NWP), in situ flood water level observations, and high-resolution LiDAR-derived terrain data. These inputs are used within a sequence-to-sequence encoder–decoder LSTM architecture, where historical rainfall and water level observations inform the model state, and forecast rainfall drives predictions of future water levels. The predicted water levels are further combined with HAND-based terrain analysis to derive spatial flood inundation extents. As illustrated in Fig. 1, the framework is applied in two complementary modes: (1) historical flood prediction, where observed rainfall is used to evaluate model skill across multiple lead times, and (2) flood nowcasting, where radar-based rainfall forecasts are used to simulate real-time operational conditions. Model performance is evaluated using both error-based metrics (e.g., RMSE and MAE), skill relative to a zero-skill model benchmark, and key operational performance metrics (e.g., lead time for flood threshold exceedance, flood peak magnitude error, and peak timing error), as

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

well as spatial inundation metrics (e.g., overall accuracy and critical success index). Further details of the case study area and methods used are given in the subsequent sections.

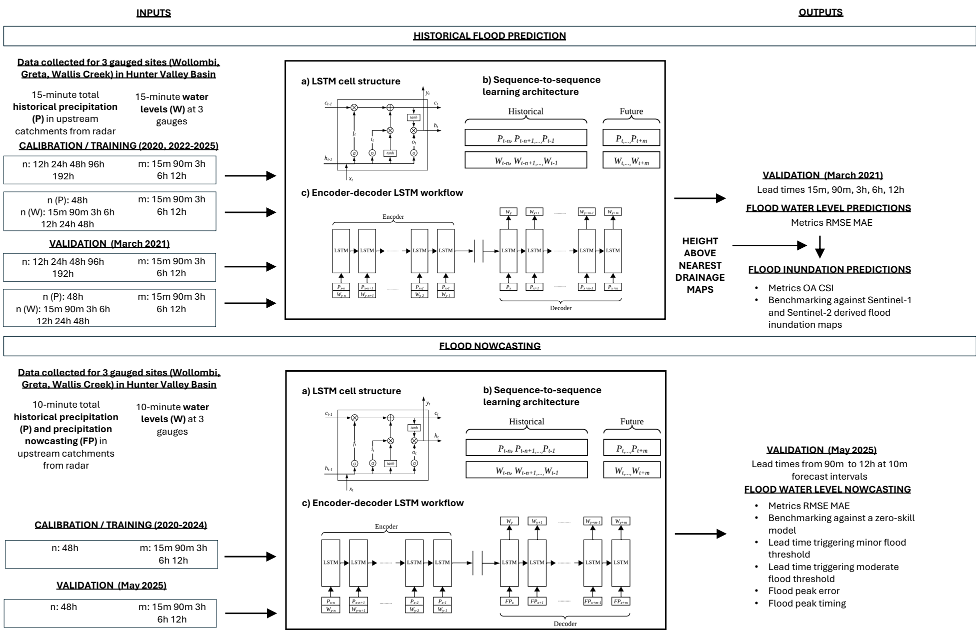

Figure 1 Schematic overview of the encoder–decoder LSTM framework for flood prediction and nowcasting in the Hunter Valley Basin. The figure illustrates the inputs and outputs of the deep learning model, as well as the model structure and workflow for historical flood prediction and flood nowcasting. (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 nowcasted upstream total radar rainfall at time *t*).

2.1 Study Area

As shown in Fig.1, the study area is located in the Hunter Valley, New South Wales, Australia, as this is a hydrologically diverse and flood-prone region and therefore ideal for testing data-driven flood forecasting models such as LSTM models. Encompassing an area of approximately 21,500 square kilometres, the valley is dominated by the Hunter River and its tributaries, with varied topography ranging from coastal lowlands to steep hills. The region experiences a mix of climatic conditions that influence rainfall-runoff dynamics, leading to frequent flood events that impact urban infrastructure, agricultural production, and mining operations. Rainfall seasonality is influenced by several major climate drivers, including the El Niño–Southern Oscillation (ENSO), the Indian Ocean Dipole (IOD), and the Southern Annular Mode (SAM), which modulate the frequency and intensity of wet and dry periods (Risbey et al., 2009). Additionally, the region is affected by East Coast Lows, which are intense low-pressure systems that can produce extreme rainfall and widespread flooding, particularly in winter and early spring (Dowdy et al., 2019). Convective thunderstorms during the summer months also contribute to short-duration, high-intensity rainfall events (Pepler et al., 2020). These diverse climatic influences, combined with the region's complex topography and land use, make the Hunter Valley an

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

ideal setting for developing and validating the encoder–decoder LSTM model to forecast water levels and inundation extents during rapidly evolving flood events. We selected three gauged locations for testing different scenarios, including Wollombi at Wollombi Brook, Greta, and Wallis Creek at Louth Park (Fig. 2 and Table 1). The Wollombi sub-catchment, located in the upper reaches of the Hunter Valley, exhibits a rapid hydrological response to intense rainfall events, often resulting in flash flooding. In contrast, the Greta region lies along the main Hunter River channel and integrates the cumulative runoff from across the basin, making it a key area for observing basin-wide flood responses. Flooding in Wallis Creek is primarily driven by backwater effects and elevated water levels in the main Hunter River system, in combination with heavy rainfall within the tributary catchment, leading to significant impacts on adjacent residential areas. The flooding of Wallis Creek can have particularly severe consequences. For instance, during the significant July 2022 flash flood event, the suburb of Gillieston Heights (a relatively new residential development) was completely isolated by rising floodwaters, effectively turning it into an island. Road access was severed, significantly hindering emergency response efforts, including rescue and evacuation operations. Together, these three locations represent the diverse hydrological and hydrodynamic settings that lead to flooding across the catchment, including fast-responding flash flood settings, slow-moving flood peaks in the lower catchment, and tributaries affected by backwater processes and local rainfall contributions (Table 1).

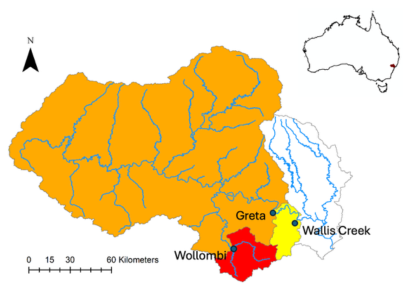

Figure 2 The three selected gauged locations in the Hunter Valley (blue line: major river network; grey line: the Hunter Valley Basin boundary; blue dot: selected gauged locations; red color: upstream catchment at Wollombi; orange and red color: upstream catchment at Greta; orange, red and yellow color: upstream catchment at Wallis Creek).

Table 1 Characteristics of the three study locations representing different flooding environments.

Forecasting locations	**Greta**	**Wollombi**	**Wallis Creek**
Geographic setting	Large, low-gradient catchment with substantial flow buffering and floodplain storage	Small catchment with steeper terrain and confined valleys	Small, steep catchment with a complex drainage network
Hydrological response	Slow riverine response dominated by upstream catchment flows	Rapid response with flashy runoff generation and short flood rise times	Highly responsive flash-flood behaviour driven by local runoff processes
Hydrological complexity	Relatively simple, dominated by riverine processes	Moderate complexity due to rapid catchment response and local inflows	Complex, influenced by backwater effects and local drainage interactions
Flood classification or regime	Riverine flooding (Cessnock City Council, 2019)	Riverine flash flooding (Singleton Shire Council, 2016)	Flash flooding (Hudson, 2026)

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

2.2 Data

The majority of the modelling data used (Fig. 1) are obtained from the Australian Bureau of Meteorology (the Bureau), which manages a comprehensive network of more than 70 weather watch radars across the country and is recognized globally for its advancements in radar rainfall research. Each radar can monitor weather within approximately 250 km, though coverage depends on terrain, buildings and various sources of interference including transient sources such as smoke, birds, and insects. RAINFIELDS is the Bureau's operational system that processes real-time radar reflectivity data into high-quality, gridded rainfall estimates called Quantitative Precipitation Estimates (QPEs). These QPEs are enhanced through calibration with rain gauge measurements from areas within radar coverage and undergo comprehensive quality control measures to remove ground and sea clutter, signal interference, anomalous propagation, second-trip echoes, bright band contamination, and partial beam blockages. The QPEs generated by RAINFIELDS are then used to produce Quantitative Precipitation Forecasts (QPFs) using a Short-Term Ensemble Prediction System (STEPS; (Bowler et al., 2006)). This forecasting system creates two types of predictions: a 2-hour forecast based solely on radar-derived QPFs, and a 12-hour forecast that integrates the QPEs with downscaled output from the Bureau's ACCESS-CE and ACCESS-G numerical weather prediction (NWP) models. To support reproducibility, radar-based rainfall nowcasting products similar to those employed in this study can be produced using the open-source *pySTEPS* library implementing the STEPS framework (Pulkkinen et al., 2019; Imhoff et al., 2023).

The Bureau maintains and quality-controls a database of gauged water level data, which underpins the Bureau's Water Data Online site (http://www.bom.gov.au/waterdata/). Water level data are sourced from a variety of partner providers across the country (in addition to Bureau owned flood sites). For the Wollombi Brook, Greta, and Wallis Creek gauges, the raw data were provided by Cessnock City Council, WaterNSW, and the New South Wales Department of Climate Change, Energy, the Environment and Water (NSW DCCEEW), respectively. Raw river level data are collected and transmitted via Event-Reporting Radio Telemetry (ERTS) systems, where water level readings are transmitted at 15-min intervals whenever a change in water level is detected at the gauge, and readings are transmitted less frequently when water level changes are not detected. Once in Bureau systems, the raw water level data are processed to remove spikes and upsampled to a regular 15-min temporal frequency during intervals of steady water levels. The river network and sub-catchment boundaries, derived from the Australian Hydrological Geospatial Fabric (Geofabric: http://www.bom.gov.au/water/geofabric/) and developed by the Australian Bureau of Meteorology, are used to delineate the upstream areas of the selected gauging stations. High-spatial-resolution airborne LiDAR DEM data (1 m or 2 m) are obtained from ELVIS (Elevation Information System) Elevation and Depth https://elevation.fsdf.org.au/) to support accurate flood inundation mapping and enhance flood impact analysis in practical application.

2.3 Methods

2.3.1 Flood Level and Inundation Prediction Models

LSTM networks have been widely applied in diverse time series modelling tasks, as they are inherently structured to preserve long- and short-term dependencies in sequential data, a property attributed to their architectural design (Hochreiter, 1998). At the core of this capability are two key internal states (Fig. 1): the cell state (c_t) for the timestep t, which captures long-term dependencies and gradual patterns, and the hidden state

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

(h_t), which encodes short-term contextual changes. The model regulates information flow through gating mechanisms, sigmoid-activated neural layers, that perform three essential functions: (1) the forget gate (f_t) filters out irrelevant information from the cell state, (2) the input gate (i_t) controls the incorporation of new inputs, and (3) the output gate (o_t) determines how much of the updated cell state is exposed to subsequent layers (Fig. 1a). Together, these gates allow LSTMs to retain crucial long-term information while dynamically adapting to new inputs, addressing challenges such as the vanishing gradient problem that affects simpler recurrent architectures. The cell state evolves smoothly as a stable memory reservoir, while the hidden state remains flexible to cater to immediate temporal variations, making LSTMs particularly effective for complex time series modelling.

We implemented the LSTM model within a sequence-to-sequence learning framework. Each sequence was divided into two components within a moving time window: a historical segment representing antecedent conditions and a future segment representing prediction conditions (Fig. 1b). Rainfall and water level time series were temporally aligned, and the sequence window advanced at each time step (e.g., 10- or 15-minute intervals) to generate calibration/training and validation samples. The historical segment was provided to the encoder, while the future segment was used by the decoder in the LSTM architecture (Fig. 1c). For example, each sequence spanned 60 hours, comprising 48 hours of historical data used as input to the encoder and a subsequent 12-hour prediction window used by the decoder. This design enabled the model to learn temporal dependencies and dynamic rainfall–runoff relationships across multiple time steps. During training, the encoder processed historical rainfall and water level inputs to generate latent representations in the form of hidden and cell states. These states were then passed to the decoder, which took radar-based rainfall nowcasts as input to produce forecasts of flood water levels. This encoder–decoder architecture facilitated short-term forecasting while effectively capturing nonlinear hydrological responses governed by antecedent rainfall and water level conditions.

We applied the Height Above Nearest Drainage (HAND) method (Nobre et al., 2011) to process high-resolution LiDAR-derived DEMs, generating relative elevation maps that represent the vertical distance and horizontal connectivity between land surface points and their nearest drainage channels, to enable water level predictions to be converted to flood inundation extents (Fig. 1). The LiDAR DEMs were first pre-processed through hydrological conditioning steps, including sink filling to remove spurious depressions and enforcing hydrologically correct flow paths. Flow direction and flow accumulation grids were then derived using standard D8 algorithms to identify stream networks and channel initiation points. The HAND values were calculated by determining, for each grid cell, the elevation difference to its nearest downstream drainage cell along the flow path, resulting in a spatially continuous surface of relative height above drainage. To ensure that DEM-derived stream networks or drainages grids accurately represent the true river channels, the extracted networks were corrected using the Geofabric surface river network (https://www.bom.gov.au/water/geofabric/) and the DEA Landsat-derived Water Observations (WOs) surface water summary dataset (Mueller et al., 2016). The Geofabric database comprises more than 2.5 million river and stream segments across Australia, while the DEA WOs summary characterises historical inundation within river–floodplain systems, including the minimum and maximum observed extents of surface water. By integrating predicted flood water levels from the LSTM model with the corrected HAND maps, we generated spatiotemporal flood inundation maps that reflect the dynamic

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

progression of floodwaters over time, offering valuable insights for flood risk assessment and emergency planning.

2.3.2 Model Training and Evaluation Approach

In the historical flood prediction experiments (Fig. 1), we utilised 5-minute RAINFIELDS radar rainfall data and derived 15-minute total precipitation (P) by aggregating observations over upstream catchment areas delineated using the Geofabric for each selected gauging station during the period 2020–2025. Concurrently, corresponding 15-minute flood water level (W) time series were compiled for the same period. The LSTM model was trained using rainfall and water level data from 2020 and 2022–2025, encompassing two major flood events and several high-flow episodes, with particularly intense activity in 2022. The trained model was then independently evaluated on the March 2021 flood event as a validation case study, focusing on three gauging locations: Wollombi, Greta, and Wallis Creek. Model predictions at multiple forecast lead times (15 minutes, 1.5 hours, 3 hours, 6 hours, and 12 hours) were compared against observed in situ water level measurements, using observed radar rainfall as a proxy for forecast inputs. Flood responses to extreme rainfall are governed by catchment-specific response times. To assess the model's ability to capture these dynamics, we evaluated a range of historical input sequence lengths for rainfall and water level (12, 24, 48, 96, and 192 hours) during the training period (2020 and 2022–2025). The influence of input sequence length on predictive performance was then examined across multiple forecast lead times during the March 2021 validation event. To further distinguish whether the LSTM primarily learns rainfall–runoff relationships or relies on the autocorrelation structure of water level time series, we conducted an additional sensitivity experiment. In this setup, the rainfall input sequence length was fixed at 48 hours, while the historical water level input length was varied (15 minutes, 1.5 hours, 3 hours, 6 hours, 12 hours, 24 hours, and 48 hours). Model performance was subsequently evaluated across the same forecast lead times during the March 2021 flood event.

In the flood nowcasting experiments (Fig. 1), we utilised 5-minute RAINFIELDS radar rainfall data and derived 10-minute total historical precipitation (P) by aggregating observations over upstream catchment areas for the period 2020–2025 as encoder inputs. In addition, we incorporated 10-minute STEPS-NWP radar rainfall forecasts to derive total precipitation nowcasting (FP) over the same upstream catchment areas for the May 2025 flood event (15–30 May 2025) as decoder inputs. Concurrently, corresponding 10-minute flood water level (W) time series were compiled for the same period as both encoder inputs and validation data. The LSTM model was trained using all available historical radar rainfall and water level observations at 10-minute resolution during periods when both datasets were concurrently available (2020–2024). Following training, the model was applied to forecast flood water levels for the May 2025 event, driven by radar-based rainfall nowcasts. The STEPS-NWP product was selected because it blends radar observations with numerical weather prediction (NWP) forecasts and provides lead times of up to 12 hours, which are critical for emergency response applications. Although STEPS-NWP provides eight ensemble members, the ensemble mean was used to maintain consistency with the modelling framework. Each sequence was constructed using a 60-hour sliding window, with the first 48 hours of historical rainfall and water level serving as antecedent inputs to the encoder, and the subsequent 12 hours of rainfall nowcasts provided to the decoder to generate flood level forecasts. This experimental design enables the model to learn from a wide range of historical hydrometeorological conditions and to evaluate its

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

ability to predict future flood levels across lead times ranging from 90 minutes to 12 hours. Forecast performance was assessed at 10-minute intervals over lead times from 90 minutes to 12 hours, using observed water levels as the reference. In addition, the evaluation focused on key operational performance metrics, including lead time for exceeding minor and moderate flood thresholds, flood peak magnitude error, and flood peak timing error, all of which are critical for emergency response during flood events.

2.3.3 Flood Prediction Evaluation Metrics

Model predictions of flood water levels were validated against observed values using standard error metrics, including Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) (Fig. 1), as defined below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

In the assessment of nowcasting capacities in an operational context (Fig. 1), we evaluated the key performance metrics of the LSTM model for water level forecasting, focusing on its ability to accurately trigger warnings for different flood thresholds (i.e., service level specification for minor, moderate and major flooding (Bureau of Meteorology, 2025)), and predict flood peak magnitude and timing. These metrics capture both the critical water level magnitudes and the timing of the most important characteristics of flood behaviour. Early warnings at different flood thresholds enable emergency agencies to prepare and respond effectively. Accurate prediction of peak level and timing is particularly critical, as it defines when the most severe impacts are likely to occur. All lead-time characteristics were evaluated relative to the lead times associated with threshold triggering, including the nominal forecast horizons (i.e., 90 min and 12 h). When the effective lead time exceeds the nominal forecast horizon, the model may overestimate water levels during the rising limb, and vice versa. Nevertheless, a modestly longer effective lead time is operationally preferable to a shorter or delayed one, because earlier, still accurate, warnings are favoured over missed alarms in flood risk management.

To provide a baseline for model nowcasting performance, we also benchmarked the LSTM forecasts against those produced by a zero-skill model (Fig. 1), which assumes that the most recent observed water level remains unchanged over the forecast horizon. Formally, the zero-skill prediction at lead time τ is defined as

$$\widehat{W_{t+\tau}^{\mathrm{ZS}}} = W_t \tag{3}$$

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

where W_t is the observed water level at time t, and $\widehat{W_{t+\tau}^{\mathrm{ZS}}}$ denotes the zero-skill forecast at time t+ τ. Model performance was evaluated by comparing the LSTM predictions $\widehat{W_{t+\tau}^{\mathrm{LSTM}}}$ against this benchmark using RMSE. To quantify the added predictive value of the LSTM model, we further defined a skill score relative to the zero-skill model as

$$Skill(\tau) = RMSE_{ZS}(\tau) - RMSE_{LSTM}(\tau) \quad (4)$$

where positive values indicate improved performance over persistence. This comparison enables a more robust assessment of the model's added value beyond persistence, especially during periods of rapid water level change associated with flood events.

To evaluate the accuracy of predicted inundation maps, we compared them against flood inundation maps from the Copernicus Global Flood Monitoring (GFM) product, which are derived from Sentinel-1 Synthetic Aperture Radar (SAR) imagery (Fig. 1). However, SAR-derived flood extents may underestimate true inundation, particularly in areas affected by double-bounce scattering from vegetation and urban structures. To provide a complementary assessment, we also employed optical remote sensing data from Sentinel-2. We calculated the Normalized Difference Water Index (NDWI) and applied Otsu's automatic thresholding method to delineate flood extents. As such, our flood inundation products were evaluated against both Sentinel-1 and Sentinel-2-derived flood maps. Validation was conducted using standard classification metrics, including Overall Accuracy (OA) and the Critical Success Index (CSI), defined as follows:

$$\mathrm{OA} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\mathrm{CSI} = \frac{TP}{TP + FP + FN} \quad (6)$$

where TP represents true positives, or the number of correctly identified flood pixels; TN represents true negatives, or the number of correctly identified non-flood pixels; FP represents false positives, or the number of non-flood pixels incorrectly identified as flood pixels, and FN represents false negatives, or the number of flood pixels incorrectly identified as non-flood pixels. Due to the relatively low spatial resolution and long revisit time of Sentinel-1 and Sentinel-2 in the Greta and Wollombi sites, flood detection in these areas was limited. Consequently, the evaluation of flood inundation maps was conducted only for the Wallis Creek site, where data quality was sufficient for reliable analysis.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

3. Results and Discussion

3.1 Evaluation of Flood Water Level Predictions Across the Three Flooding Regimes

A comparison of the LSTM model predictions and observed water levels during the March 2021 flood event demonstrates strong forecasting performance across three locations representing distinct flood regimes: Greta (riverine flooding), Wollombi (riverine-to-flash flooding), and Wallis Creek (flash flooding) (Fig. 3 and Table 2). Despite the contrasting hydrological characteristics of these catchments, the model successfully reproduced flood dynamics across lead times ranging from 15 minutes to 12 hours, highlighting the robustness of the encoder–decoder LSTM framework under diverse flood conditions.

Forecast accuracy varied according to flood regime. The best overall performance was achieved at Wallis Creek, the flash-flood-dominated catchment, where RMSE values remained below 0.30 m even at the 12-hour lead time and showed little sensitivity to the length of the historical input sequence (Table 2). For example, RMSE at the 12-hour lead time ranged only from 0.26 to 0.30 m across all historical sequence lengths. This suggests that flood behaviour at Wallis Creek is primarily controlled by recent rainfall and runoff conditions, allowing the model to effectively capture rapid hydrological responses without requiring long antecedent records.

Greta, which is characterised by a slower riverine flood regime and substantial floodplain storage, showed a moderate sensitivity to historical input length. While differences between historical sequence lengths were relatively small at shorter lead times, the best overall performance was generally achieved using 24–48 hours of antecedent information. For example, at the 12-hour lead time, RMSE ranged from 0.43 to 0.56 m across the different historical input lengths, with the 24-hour sequence producing the lowest error. This suggests that incorporating approximately one to two days of antecedent conditions is sufficient to capture the catchment memory and flow-storage processes that influence flood evolution in riverine systems, while longer historical sequences provide limited additional benefit.

Wollombi, representing a transitional riverine-to-flash flood regime, showed intermediate behaviour between the two end-member systems. Forecast accuracy benefited from longer historical sequences, with 24–48 hour inputs generally producing lower errors than shorter 12-hour inputs. At the 12-hour lead time, RMSE decreased from 0.56 m for the 12-hour historical sequence to 0.38–0.41 m for the 24–48 hour sequences. This indicates that both antecedent catchment conditions and rapid runoff generation contribute to flood dynamics at Wollombi, reflecting its mixed hydrological character.

Across all sites, forecast skill declined progressively with increasing lead time, as expected due to the accumulation of uncertainty in rainfall nowcasts and model predictions. Nevertheless, the model maintained useful predictive skill at lead times up to 12 hours, with particularly strong performance during the rising limb of flood events. While minor peak underestimation and timing delays were evident at longer lead times, the model consistently reproduced flood recession behaviour across all three flood regimes. This result is noteworthy because accurately simulating recession processes remains challenging for many hydrological and hydrodynamic models (Mishra et al., 2022). Overall, the results demonstrate that the proposed LSTM framework can successfully forecast water levels across a spectrum of flood regimes, ranging from slow riverine

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

systems influenced by catchment storage, through mixed riverine–flash systems, to highly responsive flash-flood catchments dominated by local runoff processes.

The results of the additional analysis conducted in this study eliminates the possibility that the LSTM model's strong performance arises solely from autocorrelation in water levels. This is demonstrated by the RMSE and MAE values for each configuration based on varying input sequence lengths of historical water level data (15 minutes, 90 minutes, 3 hours, 6 hours, 12 hours, 24 hours, and 48 hours), with the rainfall input sequence length 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 stable, demonstrating strong skill across different water level sequence lengths. Specifically, RMSE values ranged from 0.09–0.19 m in Wollombi, 0.05–0.15 m in Greta, and 0.04–0.10 m in Wallis Creek. Corresponding MAE values ranged from 0.06–0.10 m, 0.03–0.12 m, and 0.02–0.07 m, respectively. In contrast, for longer lead times (e.g., 12 hours), model accuracy generally improved with increasing historical water level sequence length. The RMSE decreased from 0.69 m to 0.39 m in Wollombi, from 0.69 m to 0.45 m in Greta, and from 0.39 m to 0.27 m in Wallis Creek, indicating that longer historical water level sequences enhance predictive capability. These results suggest that incorporating past water level dynamics, in addition to rainfall data, plays a significant role in forecasting future water levels, especially at longer lead times. Importantly, the model's ability to maintain robustness even with minimal water level input (e.g., 15-minute sequences) indicates that the LSTM model does not solely depend on the autocorrelation of water level time series. Instead, it effectively learns hydrological response patterns, supporting its robustness in capturing rainfall–runoff interactions.

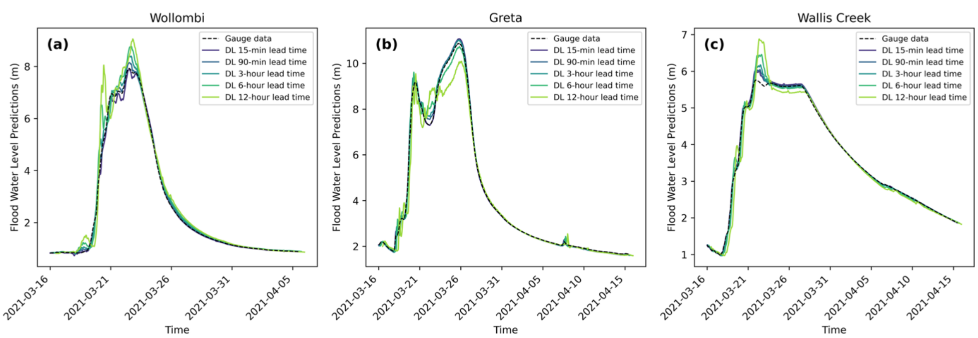

Figure 3 Comparison between LSTM model-predicted flood water levels (light to dark green lines) and in-situ observations (dashed black line) at Wollombi, Greta, and Wallis Creek for the March 2021 flood event, shown at different lead times (i.e., 15 minutes, 1.5 hours, 3 hours, 6 hours, and 12 hours).

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Table 2 Root mean square error (RMSE) and mean absolute error (MAE) errors (in meters) at different lead times (i.e., 15 minutes, 1.5 hours, 3 hours, 6 hours, and 12 hours) for different historical sequence lengths.

Locations	Lead time	t+15minutes		t+90minutes		t+3hours		t+6hours		t+12hours	
	Historical input (hours)	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Wollombi (riverine to flash regime)	192	0.07	0.04	0.09	0.05	0.13	0.07	0.24	0.13	0.42	0.23
	96	0.10	0.06	0.18	0.13	0.16	0.10	0.30	0.15	0.57	0.26
	48	0.11	0.06	0.13	0.08	0.16	0.09	0.24	0.11	0.38	0.19
	24	0.09	0.06	0.10	0.07	0.14	0.08	0.24	0.12	0.41	0.22
	12	0.15	0.09	0.12	0.08	0.20	0.11	0.39	0.20	0.56	0.32
Greta (riverine regime)	192	0.09	0.06	0.13	0.09	0.21	0.13	0.37	0.21	0.78	0.41
	96	0.09	0.07	0.12	0.06	0.26	0.15	0.43	0.24	0.73	0.38
	48	0.04	0.03	0.06	0.05	0.11	0.09	0.22	0.16	0.56	0.33
	24	0.05	0.04	0.05	0.02	0.10	0.05	0.21	0.12	0.43	0.25
	12	0.12	0.08	0.13	0.07	0.13	0.08	0.21	0.14	0.46	0.28
Wallis Creek (flash regime)	192	0.03	0.02	0.06	0.03	0.10	0.05	0.17	0.08	0.29	0.15
	96	0.03	0.02	0.04	0.02	0.07	0.03	0.14	0.07	0.29	0.18
	48	0.04	0.03	0.08	0.06	0.12	0.08	0.20	0.12	0.30	0.17
	24	0.03	0.03	0.03	0.02	0.06	0.03	0.14	0.07	0.27	0.14
	12	0.02	0.01	0.05	0.03	0.07	0.04	0.13	0.08	0.26	0.14

3.2 Evaluation of Flood Inundation Predictions

The evaluation results indicate that the LSTM approach can rapidly produce highly accurate real-time flood extent maps (Fig. 4, Fig. 5, and Fig. S1). The performance of the LSTM-based flood inundation predictions did not vary substantially across different lead times (i.e., 1.5 hours, 6 hours, and 12 hours), as the range of errors in water level predictions was not large enough to cause significant differences in the corresponding inundation extents. For the March 2021 flood event, comparisons with Sentinel-2 NDWI-derived maps indicated a slight overestimation of inundation in the northern floodplain, achieving an OA of 94.62% and a CSI of 67.24% at a 12-hour lead time. The comparison with the GFM Sentinel-1 flood product yielded comparable metrics, with an average OA of 95.65% and CSI of 63.15% at the same lead time. However, it is worth noting that the Sentinel-1 flood product does not effectively capture water extents within river channels, which limits its utility for validation purposes. The results for the July 2022 flood event demonstrated even stronger performance. Comparisons against Sentinel-2 flood maps yielded an average OA of 96.50% and CSI of 84.54%, while the Sentinel-1 comparison resulted in an average OA of 94.76% and CSI of 75.89%. It is important to note that satellite-derived flood products, while valuable, are inherently limited by factors such as vegetation canopy interference in optical imagery, signal distortions from double-bounce effects in SAR data (e.g., from trees and buildings), and the relatively coarse spatial resolution of Sentinel sensors. As such, these comparisons are treated as benchmarking exercises rather than absolute validation, given the absence of ground truth data. Overall, these findings underscore the robust capability of combining LSTM-based flood forecasting with LiDAR-derived DEMs for accurate and timely flood inundation forecasting. Such predictive flood maps can be further integrated with spatial infrastructure data, such as building footprints, road networks, and bridges, to

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

support impact assessments and inform emergency response strategies including evacuation planning and resource allocation.

While LiDAR-derived DEMs offer high spatial resolution and vertical precision, their application in flood inundation mapping is subject to several limitations. As inherently static datasets, LiDAR DEMs do not capture temporal changes in river morphology or human-induced alterations such as sediment deposition, bank erosion, infrastructure development, or land use changes, factors that can significantly change flood-inundated areas and their extent (Cook and Merwade, 2009; Wedajo, 2017). Furthermore, LiDAR sensors cannot penetrate deep water surfaces, resulting in the absence of riverbed bathymetry in the DEM. This omission is particularly problematic when combining DEMs with gauge-based water level measurements, as the gauge's zero reference level should ideally align with the actual riverbed elevation to ensure accurate inundation estimation. In addition, the HAND method relies on the assumption of a uniform hydraulic gradient and continuous connectivity to the drainage network. It also does not incorporate key hydrodynamic processes such as temporary floodplain storage (Afshari et al., 2018). These simplifications can lead to systematic under- or over-estimation of inundation extents, especially in low-gradient terrains or areas with complex hydraulic behaviour (Sanders et al., 2024). As such, while LiDAR-based approaches enhance spatial realism, they should be used with an awareness of these limitations and ideally be complemented by traditional modelling methods, in-situ observations, and/or satellite-based products where feasible.

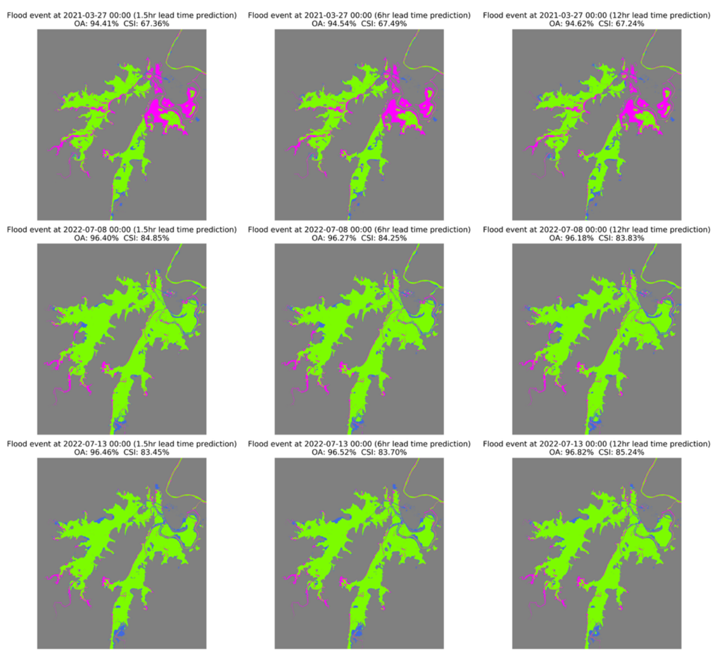

Figure 4 Comparison of LSTM model predicted flood inundation maps against Sentinel-2 derived flood map for the March 2021 and July 2022 flood events at lead times of 1.5 hours, 6 hours, and 12 hours (green: wet agreement; grey: dry agreement; pink: overestimation; blue: underestimation).

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

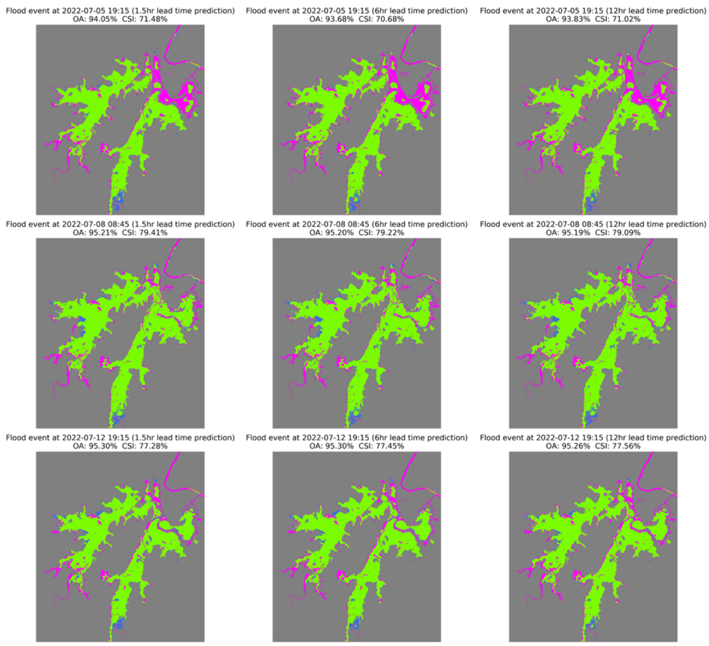

Figure 5 Comparison of LSTM model-predicted flood inundation maps with the Copernicus GFM Sentinel-1 flood map for the July 2022 flood event at lead times of 1.5 hours, 6 hours, and 12 hours (green: wet agreement; grey: dry agreement; pink: overestimation; blue: underestimation).

3.3 Nowcasting Capabilities Assessment – Focus on the Riverine to Flash Regime

Singling out the Wollombi (riverine to flash regime) we examine key performance metrics of the LSTM model forecasts for the 90-minute and 12-hour lead-time configurations during the May 2025 flood event at Wollombi as defined by the Bureau Service Level Agreements (Bureau of Meteorology, 2025), which are shown in Table 3. The 90-minute configuration demonstrates strong short-term forecasting skill, providing operationally meaningful early warnings of 2 h 40 min and 20 min for the minor and moderate flood thresholds, respectively. Although these warnings are slightly earlier than the observed exceedance times, this behaviour remains advantageous in practice, as the thresholds are correctly identified without missed alarms. Peak prediction performance is also robust, with a small peak-level error (+0.03 m) and a late bias in peak timing (-1 h 30 min). In contrast, the 12-hour configuration substantially enhances operational preparedness by providing much earlier threshold warnings, with lead times of 1 day 1 h 50 min for minor flooding and 14 h 40 min for moderate flooding. The very early detection of the minor threshold is primarily due to a slight overestimation of water levels near the flood threshold. While these detections remain true positives, excessively early warnings may lead to premature mobilisation of emergency resources. The longer lead-time forecast exhibits a low peak-level error (+0.14 m) but an early bias in peak timing (4 h 50 min), reflecting the increased uncertainty associated with extended forecast horizons.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Table 3 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	+2h40min	+1d1h50min
Lead time triggering moderate flood threshold	+20min	+14h40min
Flood peak error	+0.03m	+0.14m
Flood peak timing	-1h30min	+4h50min

The 90-min lead-time nowcast is particularly critical for flash flood conditions, where water levels can rise within minutes to hours. The results demonstrate that this configuration provides operationally useful lead times for exceedance of key flood thresholds, while also accurately predicting peak water level and timing. Rapid flood development during the most critical moments of emergency response is often difficult for conventional modelling approaches to represent; however, the 90-min LSTM model configuration is able to closely track and forecast the rapid rising limb, making it especially valuable for short-range tactical decision-making and real-time warning. In contrast, the 12-h lead-time forecast provides greater value for riverine flooding, which is more common across many flood-prone regions and typically evolves over longer timescales. The results indicate that this configuration can offer approximately more than 6–12 h of anticipatory information regarding exceedance of key flood thresholds, meeting the operational requirement of providing emergency agencies with sufficient time for planning and response prior to peak flooding. However, forecast errors are larger than those of the 90-min configuration. This difference arises because the first 90 min of prediction is driven primarily by radar observations, which provide the most reliable representation of near-term rainfall, whereas forecasts beyond 90 min increasingly rely on numerical weather prediction (NWP), the uncertainty of which grows with lead time. Consequently, although the 12-h forecast captures the overall flood evolution well, its accuracy in peak magnitude and timing remains limited compared with the short-lead nowcast.

Figure 6 presents the temporal distribution of RMSE at each forecast time step (every 10 min) for the 90-min and 12-h lead-time forecasts during the May 2025 Wollombi flood event. The 90-min configuration maintains consistently low RMSE values (< 0.35 m) throughout most of the evaluation period, indicating stable short-term predictive accuracy across different flood phases. Relatively high errors (RMSE > 0.2 m) are visible around the main rising and peak stages of the event (approximately 19–24 May), but the magnitude remains comparatively small, confirming the reliability of short-range forecasts for operational warning purposes. In contrast, the 12-h configuration exhibits substantially higher and more variable RMSE, with a mean of 0.3 m and a standard deviation of 0.34 m. This behaviour reflects the transition from radar-driven nowcasting within the first 90 min to NWP forcing at longer lead times, where forecast uncertainty increases progressively toward the 12-h horizon. Elevated error bands (RMSE > 0.75 m) are observed during the flood development and peak phases, reflecting increased forecast uncertainty under rapidly changing conditions. Outside the main flood window, RMSE values decrease but remain generally higher than those of the 90-min forecast, indicating persistent long-lead uncertainty. Overall, the RMSE patterns reinforce the performance trade-off identified from the threshold- and peak-based evaluation metrics. The short lead times provide greater accuracy and temporal stability, whereas long lead times enhance early situational awareness but introduce larger magnitude and timing errors, particularly during rapidly evolving flood conditions.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

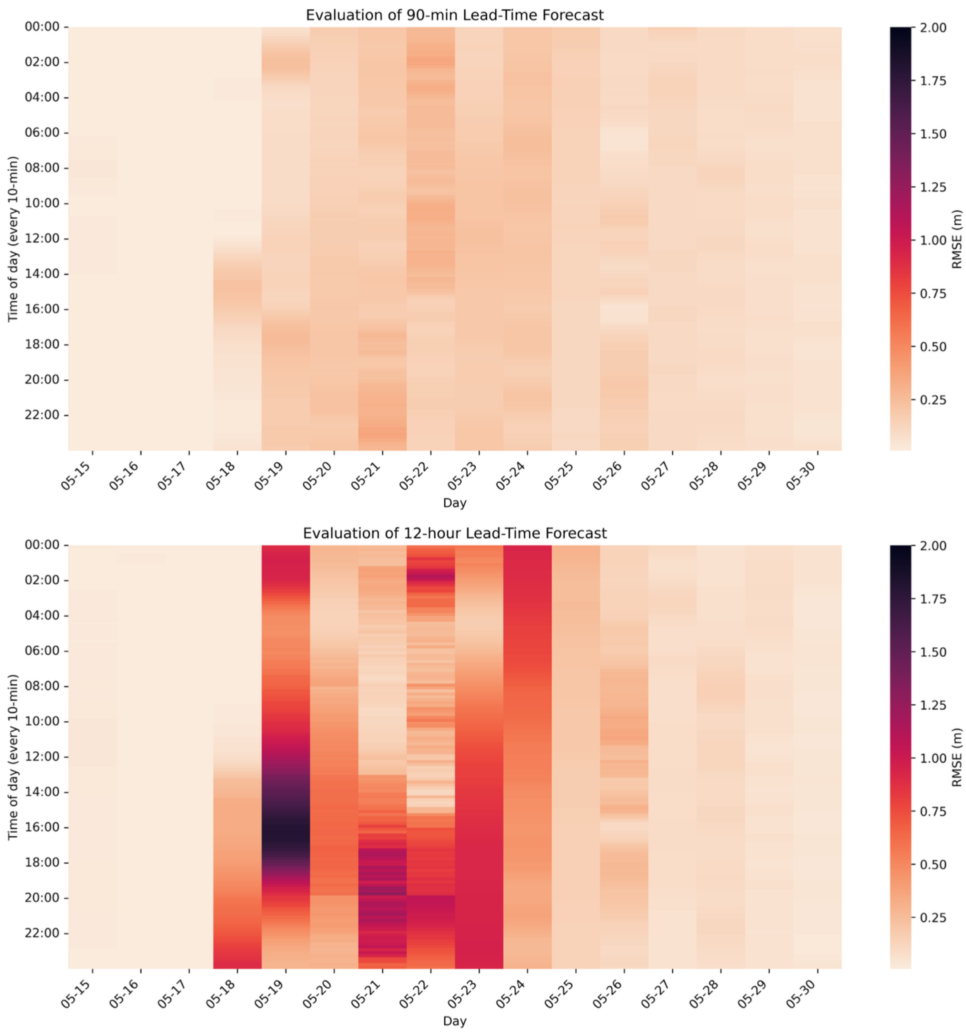

Figure 6 Root-mean-square error (RMSE) evaluated at each forecast time step (10-min intervals) for the 90-min lead-time forecast (top panel) and the 12-h lead-time forecast (bottom panel) during the May 2025 flood event at the Wollombi site.

The utility of the LSTM model is further evidenced by the increased skill it offers compared with the zero-skill benchmark model in terms of reduction in RMSE at each 10-minute forecast issue time (i.e., water level forecasts issued every 10 minutes) over the period from 2025-05-15 00:00:00 to 2025-05-30 00:00:00 (Fig. 7). As can be seen, the LSTM model outperforms the benchmark model for all forecasting lead times considered, with its relative skill increasing with lead time. This is likely because in contrast to the zero-skill benchmark, which assumes a constant future water level equal to the most recent observation, the LSTM model is able to capture the temporal evolution of flood dynamics beyond simple persistence.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

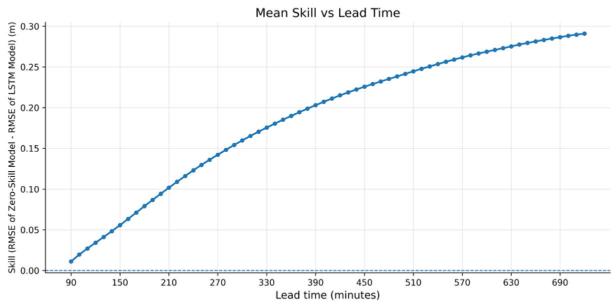

Figure 7 Mean forecast skill of the LSTM model relative to a zero-skill benchmark as a function of lead time, ranging from 90 minutes to 12 hours at 10-minute intervals. Skill is defined as the difference in Root-mean-square error (RMSE) between the zero-skill model and the LSTM model, with positive values indicating improved performance of the LSTM model.

3.4 Model Strengths and Limitations

A key strength of the proposed framework is its simplicity, computational efficiency, and suitability for operational deployment. The workflow requires only routinely available rainfall and water-level observations as inputs and does not depend on complex hydrological or hydrodynamic model calibration. Once trained, the encoder–decoder LSTM model can generate flood level forecasts within seconds, making it well suited for real-time forecasting applications where rapid updates are essential. The integration of radar-based rainfall nowcasts with water-level observations allows the framework to exploit both catchment memory and future rainfall information while maintaining a relatively low computational burden. Compared with physically based flood forecasting systems that often require extensive catchment parameterisation, model calibration, and substantial computational resources, the proposed approach offers a practical and scalable alternative for operational agencies.

The results demonstrate the operational value of LSTM-based flood nowcasting for flood warning and emergency management. Accurate water-level forecasts can support timely emergency response, targeted evacuations, and reduced flood impacts on communities (Pappenberger et al., 2015). Importantly, the predicted water levels can be directly translated into high-resolution flood inundation nowcasts, providing spatially explicit information on flood extent and depth in near real time (Hou et al., 2026). Such information can improve situational awareness, support evacuation planning, and guide the deployment of emergency resources, particularly in flash-flood-prone regions where warning lead times are limited (Schubert et al., 2022). The strong forecasting skill achieved at lead times ranging from 90 minutes to 12 hours is especially valuable operationally, as it bridges the gap between immediate response actions and longer-term planning. The strong predictive skill achieved at lead times of around one hour is particularly important for flash-flood forecasting, as even short warning periods can substantially enhance emergency response, public warning dissemination, and evacuation preparedness in rapidly responding catchments (Nayak et al., 2005). The ability to further extend forecasts through the incorporation of NWP rainfall forecasts enhances these planning opportunities. These findings highlight the importance of maintaining robust radar rainfall, river gauge, and real-time data infrastructure to support operational flood forecasting services.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

While the performance of the model is encouraging based on the experiments conducted, it is vital that relevant practical limitations are considered before operational deployment (Li et al., 2026). One of these limitations is the spatial coverage of radar stations, which can result in incomplete rainfall input. In Australia, more than 70 weather watch radars provide extensive coverage over populated areas, with data gaps often supplemented by NWP forecasts. In contrast, radar coverage in many other countries is sparse or inconsistent, limiting the feasibility of radar-driven LSTM models for flood prediction. Another operational challenge is the real-time acquisition of river level data, which can be affected by communication lags, sensor malfunctions, and poor integration among monitoring networks. Despite these constraints, studies such as Nearing et al. (2024) and Kratzert et al. (2019) have shown that LSTM models can predict flood levels up to five days in advance, even in ungauged basins. While our study specifically targets riverine flash floods, typically driven by cumulative upstream rainfall and runoff, localized flash flood events may benefit more from models like Convolutional LSTM (ConvLSTM) models, which can learn both temporal and spatial patterns and identify the most influential rainfall cells associated with rapid inundation (Oddo et al., 2024). Finally, a commonly cited limitation of deep learning models in hydrology is their lack of physical interpretability, as they do not explicitly represent catchment-scale hydrological processes. To address this, hybrid modelling approaches that combine data-driven and physically based methods are gaining traction. Such frameworks offer the dual benefit of enhanced predictive performance and improved process-based understanding, making them attractive for both research and operational forecasting (Slater et al., 2023).

Although the model demonstrated strong performance across all three regimes investigated in this study, the transferability of the framework to other flood regimes remains uncertain. In particular, two important flood regimes have not yet been evaluated. The first is highly urbanised catchments, where flooding is often dominated by overland flow, drainage network capacity, impervious surfaces, and highly localised convective rainfall (Xing et al., 2019). In these environments, flood response times can be measured in minutes rather than hours, and accurate forecasting may require higher spatial-resolution rainfall information and modelling approaches capable of explicitly representing spatial rainfall patterns, such as ConvLSTM models. The second is highly regulated or modified catchments containing dams, reservoirs, and retarding basins (Volpi et al., 2018). In such systems, downstream flood behaviour may be strongly influenced by reservoir storage levels, operational release decisions, and infrastructure management practices rather than rainfall-runoff processes alone. Because these controls are not represented in the current model inputs, additional predictors would likely be required to achieve reliable forecasts.

Several additional limitations warrant future investigation. First, the study evaluates a limited number of catchments and flood events, and broader testing across different hydroclimatic regions would provide greater confidence in model generalisability (Kratzert et al., 2024). The current framework is also site-specific and requires local training data, which may limit its transferability to ungauged or poorly monitored basins. Furthermore, the analysis is constrained by the availability of historical radar rainfall data, which currently spans only a relatively short period (approximately 2020–2026). As longer radar archives become available through ongoing efforts to extend the Bureau’s historical radar datasets, a substantially larger number of flood events could be incorporated into model training and validation. This expanded dataset would likely improve model

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

robustness, particularly for rare and extreme flood events that are currently underrepresented in the training data (Frame et al., 2022). Another limitation is that the flood events examined in this study, while representing different flood regimes, do not include some of the most rapidly responding flash-flood systems observed in Australia. In certain catchments, water levels can rise dramatically within less than an hour following intense localised rainfall. Such events provide very limited warning time and may challenge the ability of LSTM models to accurately predict both the timing and magnitude of rapid flood rises. Future work should therefore investigate model performance across a wider range of flash-flood severities and response times to better understand the practical limits of data-driven forecasting approaches under highly dynamic hydrological conditions.

4. Conclusions

This study demonstrates the potential of an encoder–decoder Long Short-Term Memory (LSTM) framework for short-term flood forecasting and high-resolution inundation mapping using radar rainfall estimates, river-level observations, and airborne LiDAR-derived DEMs. Validation against observed flood events showed that the framework can accurately predict flood water levels and generate inundation maps that agree closely with satellite-derived flood products.

We show that the proposed framework is suitable, reliable, and computationally efficient across riverine, riverine-to-flash, and flash-flood regimes. At a 90-minute lead time, RMSE values ranged from 0.05–0.13 m for the riverine regime (Greta), 0.09–0.18 m for the riverine-to-flash regime (Wollombi), and 0.03–0.08 m for the flash-flood regime (Wallis Creek). At a 12-hour lead time, RMSE values ranged from 0.43–0.78 m, 0.38–0.57 m, and 0.26–0.30 m, respectively. The model performed particularly well in the riverine-to-flash and flash-flood regimes, maintaining strong predictive skill in both reanalysis and forecasting modes despite the rapid hydrological response of these systems. Operational testing demonstrated that the framework can provide reliable flood nowcasts at lead times from 90 minutes to 12 hours. Short lead-time forecasts showed high accuracy and were particularly valuable for rapidly evolving flash-flood conditions, while longer lead-time forecasts provided additional warning time for preparedness and response despite increased rainfall forecast uncertainty. Importantly, the LSTM model consistently outperformed a zero-skill persistence benchmark, demonstrating that forecast skill arises from learned rainfall–runoff relationships and hydrological response dynamics rather than persistence in the water-level record.

Overall, the results suggest that a single deep-learning framework can support flood forecasting across diverse hydrological settings without the need for flood-regime-specific model structures. A further operational advantage is that the framework continuously incorporates observed radar rainfall data, allowing forecasts to automatically adapt to evolving rainfall patterns and reducing the need for manual intervention when rainfall intensity or location differs from earlier forecasts. This enables operational staff to focus on warning interpretation, communication, and emergency response activities rather than updating model configurations or correcting forecast inputs. By combining real-time rainfall observations with high-resolution terrain data, the approach enables the rapid generation of operationally relevant inundation forecasts that can strengthen early warning systems, improve situational awareness, and support more effective emergency management. Future work should focus on extending the framework to additional flood regimes, including highly urbanised and regulated catchments, incorporating longer historical radar archives and more extreme flood events, and developing hybrid machine-learning and physically based modelling approaches to improve forecast robustness, transferability, and interpretability.

Data availability

Flood water level observations used in this study were obtained from the Bureau of Meteorology's Water Data Online service (https://www.bom.gov.au/waterdata/). Operational radar rainfall nowcasts can be viewed through the Bureau of Meteorology Rain Radar and Weather Maps portal (https://www.bom.gov.au/weather-and-climate/rain-radar-and-weather-maps). Historical radar rainfall data were accessed through the Australian Unified Radar Archive hosted by the National Computational Infrastructure (https://opus.nci.org.au/spaces/NDP/pages/399803502/Australian+Unified+Radar+Archive). Airborne LiDAR elevation data were retrieved from the Elevation and Depth – Foundation Spatial Data maintained by Geoscience Australia (https://elevation.fsdf.org.au/).

Author contribution

JH conceived the idea, developed the methodology and software, carried out the investigation and analysis, curated the data, and wrote the first manuscript draft. JH, JP and MT curated the data. HRM developed the methodology figure. WS, JP, CVF, MT, PFH, EC and HRM reviewed and edited the manuscript.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This research is supported by Natural Hazards Research Australia. We express our sincere gratitude to Christopher Leahy and Andrew Frost for their internal reviews at the Australian Bureau of Meteorology. This research was conducted with the support of the high-performance computing system, Gadi, provided by the National Computational Infrastructure (NCI) and funded by the Australian Government.

References

Afshari, S., Tavakoly, A. A., Rajib, M. A., Zheng, X., Follum, M. L., Omranian, E., and Fekete, B. M.: Comparison of new generation low-complexity flood inundation mapping tools with a hydrodynamic model, Journal of Hydrology, 556, 539-556, 2018.

Bates, P. D. and De Roo, A.: A simple raster-based model for flood inundation simulation, Journal of hydrology, 236, 54-77, 2000.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Bhanja, S. N., Coon, E. T., Lu, D., and Painter, S. L.: Evaluation of distributed process-based hydrologic model performance using only a priori information to define model inputs, Journal of Hydrology, 618, 129176, 2023.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 132, 2127-2155, 2006.

Bureau of Meteorology: Service Level Specification for Flood Forecasting and Warning Services for New South Wales and the Australian Capital Territory, 2025.

Cessnock City Council: Greta Flood Study: Final Report, Cessnock City Council, 2019.

Chiang, Y.-M., Chang, L.-C., and Chang, F.-J.: Comparison of static-feedforward and dynamic-feedback neural networks for rainfall–runoff modeling, Journal of hydrology, 290, 297-311, 2004.

Collier, C.: Flash flood forecasting: What are the limits of predictability?, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 133, 3-23, 2007.

Cook, A. and Merwade, V.: Effect of topographic data, geometric configuration and modeling approach on flood inundation mapping, Journal of hydrology, 377, 131-142, 2009.

Dawson, C. W. and Wilby, R.: An artificial neural network approach to rainfall-runoff modelling, Hydrological Sciences Journal, 43, 47-66, 1998.

Dowdy, A. J., Pepler, A., Di Luca, A., Cavicchia, L., Mills, G., Evans, J. P., Louis, S., McInnes, K. L., and Walsh, K.: Review of Australian east coast low pressure systems and associated extremes, Climate Dynamics, 53, 4887-4910, 2019.

Ebert, B., Perrels, A., Mooney, C., Hoffmann, D., Tupper, A., Mills, B., Pástor-Paz, J., Liang, X., Msemo, H., and Da Costa, J.: Value chain approaches to describe, improve, value and co-design early warning systems, 2024.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, Water Resources Research, 56, e2019WR026793, 2020.

Fleischmann, A., Paiva, R., and Collischonn, W.: Can regional to continental river hydrodynamic models be locally relevant? A cross-scale comparison, Journal of Hydrology X, 3, 100027, 2019.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377-3392, 2022.

Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., and Gibbs, M. S.: An evaluation framework for input variable selection algorithms for environmental data-driven models, Environmental Modelling & Software, 62, 33-51, 2014.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Gers, F. A., Schmidhuber, J., and Cummins, F.: Learning to forget: Continual prediction with LSTM, Neural computation, 12, 2451-2471, 2000.

Gissing, A., Keys, C., and Opper, S.: Towards resilience against flood risks, Australian Journal of Emergency Management, The, 25, 39-45, 2010.

Haruna, A., Garambois, P.-A., Roux, H., Javelle, P., and Jay-Allemand, M.: Does flash flood model performance increase with complexity? Signature and sensitivity-based comparison of conceptual and process-oriented models on french Mediterranean cases, Hydrology, 9, 141, 2022.

Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6, 107-116, 1998.

Hou, J., Sharples, W., Tarpanelli, A., Renzullo, L., Woldemeskel, F., and Carrara, E.: Advancing near-real-time flood inundation mapping in Australia, Water Resources Research, 62, e2025WR040640, 2026.

Hsu, K. l., Gupta, H. V., and Sorooshian, S.: Artificial neural network modeling of the rainfall-runoff process, Water resources research, 31, 2517-2530, 1995.

Hudson, K., Wilke, D., Mooney, C., Heinrich, V.: Flash flooding case studies to improve predictions and the communication of uncertainty: Case study reports, Natural Hazards Research Australia, 2026.

Imhoff, R. O., De Cruz, L., Dewettinck, W., Brauer, C. C., Uijlenhoet, R., van Heeringen, K. J., Velasco-Forero, C., Nerini, D., Van Ginderachter, M., and Weerts, A. H.: Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library, Quarterly Journal of the Royal Meteorological Society, 149, 1335-1364, 2023.

Kao, I.-F., Zhou, Y., Chang, L.-C., and Chang, F.-J.: Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting, Journal of Hydrology, 583, 124631, 2020.

Knocke, E. T. and Kolivras, K. N.: Flash flood awareness in southwest Virginia, Risk Analysis: An International Journal, 27, 155-169, 2007.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187-4201, 2024.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005-6022, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11344-11354, 2019.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517-5534, 2021.

Li, K., Razavi, S., Maier, H. R., Hrachowitz, M., Nabavi, E., Harvey, N., Akhtar, K., and Unduche, F.: 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, 100215, 2026.

Li, W., Kiaghadi, A., and Dawson, C.: High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks, Neural Computing and Applications, 33, 1261-1278, 2021.

Lipton, Z. C., Berkowitz, J., and Elkan, C.: A critical review of recurrent neural networks for sequence learning, arXiv preprint arXiv:1506.00019, 2015.

Maier, H. R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I. N., Sànchez-Marrè, M., Acutis, M., Wu, W., and Humphrey, G. B.: Exploding the myths: An introduction to artificial neural networks for prediction and forecasting, Environmental modelling & software, 167, 105776, 2023.

Mishra, A., Mukherjee, S., Merz, B., Singh, V. P., Wright, D. B., Villarini, G., Paul, S., Kumar, D. N., Khedun, C. P., and Niyogi, D.: An overview of flood concepts, challenges, and future directions, Journal of hydrologic engineering, 27, 03122001, 2022.

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., and Abrahart, R.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, Hydrological Sciences Journal, 61, 1192-1208, 2016.

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., Lymburner, L., McIntyre, A., Tan, P., and Curnow, S.: Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia, Remote Sensing of Environment, 174, 341-352, doi:10.1016/j.rse.2015.11.003, 2016.

Nayak, P., Sudheer, K., Rangan, D., and Ramasastri, K.: Short-term flood forecasting with a neurofuzzy model, Water Resources Research, 41, 2005.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., and Metzger, A.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559-563, 2024.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., and Saleska, S.: Height Above the Nearest Drainage–a hydrologically relevant new terrain model, Journal of Hydrology, 404, 13-29, 2011.

Oddo, P. C., Bolten, J. D., Kumar, S. V., and Cleary, B.: Deep Convolutional LSTM for improved flash flood prediction, Frontiers in Water, 6, 1346104, 2024.

Paiva, R. C. D., Buarque, D. C., Collischonn, W., Bonnet, M. P., Frappart, F., Calmant, S., and Mendes, C. A. B.: Large-scale hydrologic and hydrodynamic modeling of the Amazon River basin, Water Resources Research, 49, 1226-1243, doi:10.1002/wrcr.20067, 2013.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J.: The monetary benefit of early flood warnings in Europe, Environmental Science & Policy, 51, 278-291, 2015.

Pepler, A. S., Dowdy, A. J., Van Rensch, P., Rudeva, I., Catto, J. L., and Hope, P.: The contributions of fronts, lows and thunderstorms to southern Australian rainfall, Climate Dynamics, 55, 1489-1505, 2020.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L.: Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1. 0), Geoscientific Model Development, 12, 4185-4219, 2019.

Risbey, J. S., Pook, M. J., McIntosh, P. C., Wheeler, M. C., and Hendon, H. H.: On the remote drivers of rainfall variability in Australia, Monthly Weather Review, 137, 3233-3253, 2009.

Sajikumar, N. and Thandaveswara, B.: A non-linear rainfall–runoff model using an artificial neural network, Journal of hydrology, 216, 32-55, 1999.

Sanders, B. F., Wing, O. E., and Bates, P. D.: Flooding is not like filling a bath, Earth's Future, 12, e2024EF005164, 2024.

Schubert, J. E., Luke, A., AghaKouchak, A., and Sanders, B. F.: A framework for mechanistic flood inundation forecasting at the metropolitan scale, Water Resources Research, 58, e2021WR031279, 2022.

Seed, A. W., Pierce, C. E., and Norman, K.: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme, Water Resources Research, 49, 6624-6641, 2013.

Shuster, W. D., Bonta, J., Thurston, H., Warnemuende, E., and Smith, D. R.: Impacts of impervious surface on watershed hydrology: A review, Urban Water Journal, 2, 263-275, 2005.

Singleton Shire Council: Wollombi Brook Flood Study - Report, NSW Office of Environment and Heritage, 2016.

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., and Speight, L.: Hybrid forecasting: blending climate predictions with AI models, Hydrology and earth system sciences, 27, 1865-1889, 2023.

Sudheer, K., Gosain, A., and Ramasastri, K.: A data-driven algorithm for constructing artificial neural network rainfall-runoff models, Hydrological processes, 16, 1325-1330, 2002.

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., and Slayback, D. A.: Satellite imaging reveals increased proportion of population exposed to floods, Nature, 596, 80-86, 2021.

Velasco-Forero, C., Pudashine, J., Curtis, M., and Seed, A.: STEPS3–ADV–VERIFICATION REPORT, 6202020.

Vieux, B. E., Cui, Z., and Gaur, A.: Evaluation of a physics-based distributed hydrologic model for flood forecasting, Journal of hydrology, 298, 155-177, 2004.

https://doi.org/10.5194/egusphere-2026-3401
Preprint. Discussion started: 4 August 2026

Voit, P. and Heistermann, M.: A downward-counterfactual analysis of flash floods in Germany, Natural Hazards and Earth System Sciences, 24, 2147-2164, 2024.

Volpi, E., Di Lazzaro, M., Bertola, M., Viglione, A., and Fiori, A.: Reservoir effects on flood peak discharge at the catchment scale, Water Resources Research, 54, 9623-9636, 2018.

Wedajo, G.: LiDAR DEM Data for flood mapping and assessment; opportunities and challenges: A Review, J. Remote Sens. Gis, 6, 2015-2018, 2017.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N.: Future changes to the intensity and frequency of short-duration extreme rainfall, Reviews of Geophysics, 52, 522-555, 2014.

Xing, Y., Liang, Q., Wang, G., Ming, X., and Xia, X.: City-scale hydrodynamic modelling of urban flash floods: the issues of scale and resolution, Natural Hazards, 96, 473-496, 2019.

Yamazaki, D., Kanae, S., Kim, H., and Oki, T.: A physically based description of floodplain inundation dynamics in a global river routing model, Water Resources Research, 47, 2011.

Zoccatelli, D., Borga, M., Zanon, F., Antonescu, B., and Stancalie, G.: Which rainfall spatial information for flash flood response modelling? A numerical investigation based on data from the Carpathian range, Romania, Journal of Hydrology, 394, 148-161, 2010.