Physically Anchored Multi-Resolution Neural Operator Framework for Flood Inundation Prediction

Abdolmehdi Behroozi[1], Kathryn Lawson[1], and Chaopeng Shen[1]

[1]Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA

Correspondence: Chaopeng Shen (cxs1024@psu.edu)

Abstract. Accurate flood inundation modeling using high-resolution hydrodynamic simulations is computationally demanding, limiting their use for large-scale analysis and rapid scenario evaluation. Although machine learning surrogates have been developed, many struggle to reproduce the full spatiotemporal evolution of flood dynamics while maintaining physically meaningful behavior across spatial scales. In particular, simultaneously capturing basin-scale wave propagation and fine-scale inundation boundaries remains challenging. This study presents a multi-resolution deep learning framework for rapid shallow water equation surrogate emulation of flood depth and extent fields. The approach combines a coarse-resolution neural operator that captures large-scale hydrodynamic behavior with a terrain-aware refinement module that reconstructs fine-scale inundation boundary structure. The framework is trained on high-fidelity two-dimensional shallow-water simulations and evaluated across riverine, dam-break, and complex floodplain systems, including tests under structured bathymetric uncertainty. Results demonstrate accurate reconstruction of continuous water depth fields, wet–dry delineation, and peak flood magnitude and timing. The model closely reproduces the evolution of domain-integrated water volume over time, providing a volume-consistency diagnostic beyond purely geometric agreement, and maintains probabilistic consistency when input topography is uncertain. The framework therefore provides high-resolution surrogate predictions of solver-generated flood depth and extent fields at substantially reduced computational cost relative to direct high-resolution simulation. These findings show that multi-resolution deep learning can efficiently emulate solver-generated hydrodynamic flood responses with strong depth and boundary accuracy and robustness to geometric uncertainty, supporting scalable flood hazard assessment and rapid scenario evaluation.

1 Introduction

Floods are among the most consequential natural hazards because they combine strong physical complexity with widespread exposure of populations, infrastructure, and supply chains. Across riverine, pluvial (urban), coastal, and dam-break contexts, inundation emerges from nonlinear coupling among rainfall–runoff generation, channel–floodplain hydraulics, tide–surge processes, terrain controls, and engineered hydraulic structures (Merwade et al., 2008; Teng et al., 2017; Bates, 2022; Zhao et al., 2025). These controls operate across spatial scales from street-level microtopography (e.g., curbs, underpasses, levees) to catchment-to-regional connectivity, and across temporal scales from minutes (convective bursts and rapid runoff responses) to days (large-basin routing and compound surge–river interactions). Because impacts depend not only on peak water levels but

also on arrival time, duration, and flow pathways, decision-relevant prediction increasingly requires spatiotemporally resolved fields of water depth and velocity over large, high-resolution domains (Bubeck et al., 2017).

Physics-based hydrodynamic models are widely used as reference tools because they directly solve the governing flow equations. However, their primary limitation is computational scalability: capturing the relevant physics requires fine spatial discretisation, small time steps for stability, and repeated solution of large numerical systems. As a result, runtime and memory costs grow rapidly with resolution, domain size, and model complexity, limiting their applicability for rapid forecast-conditioned scenario simulation and large-ensemble workflows such as Monte Carlo design (Rahman et al., 2002), uncertainty quantification (Idowu and Alfahid, 2025), and sensitivity analysis (Pappenberger et al., 2008; Alipour et al., 2022), where the cost of a single simulation is amplified across many realizations (Bates, 2022; Guo et al., 2020; Luo et al., 2022). Furthermore, when boundary forcings are supplied from external hydrologic or meteorological forecasting systems, high-fidelity solvers may still be too expensive for rapid scenario evaluation within operational time constraints. Substantial effort has been devoted to mitigating this computational bottleneck through high-performance computing and numerical acceleration. Parallelisation strategies (Neal et al., 2009; Sanders and Schubert, 2019), improved numerical schemes (Bates and De Roo, 2000; Leijnse et al., 2021; Sridharan et al., 2021; Behroozi et al., 2023), and GPU-based implementations (Buttinger-Kreuzhuber et al., 2022; Ming et al., 2020; Morales-Hernández et al., 2021; Rak et al., 2024) have significantly reduced runtimes. Nevertheless, these approaches do not eliminate the underlying computational constraints.

Data-driven and surrogate approaches aiming to approximate high-fidelity solvers at a fraction of the computational cost have emerged as an alternative paradigm, with most models trained on synthetic datasets generated from numerical solvers (Mosavi et al., 2018; Bentivoglio et al., 2022; Karim et al., 2023; Sun et al., 2024). Despite major speed gains, two distinct but coupled limitations remain. The first is a *local geometric-fidelity* problem. For flood inundation surrogates, the most consequential errors are often not basin-averaged depth errors, but local wet–dry boundary misplacement and broken channel/floodplain connectivity in narrow conveyance corridors. Because many studies are evaluated primarily with aggregate metrics (e.g., CSI, RMSE), these decision-critical hydraulic failures can be masked even when global scores are high (Wing et al., 2017; Mosavi et al., 2018; Bentivoglio et al., 2022; Karim et al., 2023; Taghizadeh et al., 2025). These failure modes arise where fine-scale topography controls dry-wet interfaces and channelized flow, so a coarse global surrogate alone is insufficient.

The second is a *cross-scale consistency* problem. Recent advances in physics guidance and operator learning improve efficiency and generalization: hybrid and reduced-order surrogates enable large-scale acceleration (Fraehr et al., 2024; Bass and Bedient, 2018; John and Nagaraj, 2023), while neural operators capture global spatiotemporal dynamics and transfer across scenarios, and convolutional refinements recover local detail (Xu et al., 2025; Rivera-Casillas et al., 2025; Li et al., 2020; Behroozi et al., 2025; Baruah et al., 2025; He et al., 2023). However, a fundamental limitation persists: models optimized for global dynamics tend to smooth terrain-controlled transitions, whereas local refinement approaches often lack explicit global hydrodynamic context, producing inconsistencies between large-scale flow evolution and fine-scale inundation geometry (Fraehr et al., 2024; Xu et al., 2025; Thomas Steven Savage et al., 2016; Cappato et al., 2022; Balachandran et al., 2025; Qiu et al., 2026; Noh et al., 2018). Thus, the key need is not only higher local resolution, but a framework that preserves both local fidelity and global–local physical consistency.

A separate operational challenge is that the impacts of river bathymetry and bed topography conditions are difficult to observe and represent in surrogate models at large spatial scales, an issue acknowledged by benchmark-focused reviews (Karim et al., 2023). Because topography is a high-dimensional, distributed input, it is difficult to provide enough training data to fully constrain its impacts. Flood responses are strongly sensitive to bed geometry, especially near wet–dry transitions and connectivity bottlenecks, yet river bathymetry is rarely available. Therefore, to support probabilistic flood analysis, a surrogate should remain reliable under at least modest plausible bed-topography perturbations rather than only under a single deterministic geometry. Prior data-driven studies have often focused on predictions of fixed bed topography, including CNN-based rapid inundation mapping trained on event libraries within a given terrain (Kabir et al., 2020) and neural-operator/domain-adaptation frameworks targeting transfer across basins or forcing regimes rather than explicit bed perturbations (Xu et al., 2025; Taghizadeh et al., 2025). Such models may lose relevance even under slight changes in bed topography. This means they always need extensive training data for a given domain, which partly negates their efficiency. Consequently, systematic evaluation of surrogate sensitivity to modest bed-topography perturbations remains limited.

In this work, we introduce a physics-aligned, operator-based deep learning framework for rapid shallow water equation surrogate emulation of solver-generated flood depth and extent fields. The proposed approach separates large-scale wave propagation from local topographic refinement, enabling efficient inference over large spatial domains while preserving decision-critical inundation features. Specifically, we leverage operator learning to model global flood dynamics and incorporate a physics-anchored refinement stage to recover sharp wet–dry boundaries and connectivity governed by fine-scale terrain. By combining these components within a unified and inference-stable framework, the method addresses the trade-off between computational efficiency and geometric fidelity observed in existing surrogate models. Beyond deterministic accuracy, we explicitly evaluate whether the framework preserves predictive skill and calibration under moderate, spatially coherent bed-topography perturbations. This test targets practical robustness for uncertainty-aware scenario analysis in settings where bed data are incomplete or uncertain. After training on high-fidelity hydrodynamic-solver outputs, the resulting surrogate reduces the need for repeated numerical-solver runs and enables rapid scenario evaluation and large-ensemble analysis, while maintaining high-resolution geometric fidelity in complex terrain.

2 Methodology

This section details the multi-resolution surrogate architecture used to rapidly emulate solver-generated flood depth and extent fields. It summarizes forcing construction, the coarse-resolution Fourier Neural Operator, and the localized Magnifier refinement module, with emphasis on scale separation, terrain-aware refinement, wet–dry boundary reconstruction, and volume-consistency diagnostics.

In simple terms, the framework first learns the broad flood response over the whole domain at coarse resolution, where large-scale flood-wave timing and storage dynamics can be represented efficiently. This coarse prediction is then locally refined using high-resolution terrain information. Instead of simply interpolating the coarse depth field, the Magnifier reconstructs a terrain-aware local depth pattern and corrects it using learned residuals from the hydrodynamic solver. This design allows the model to

combine global flood evolution with fine-scale wet–dry boundary reconstruction, while keeping the computational cost much lower than a full high-resolution simulation.

2.1 Neural Operators and Fourier Neural Operators

Neural operators aim to learn mappings between function spaces rather than finite-dimensional vectors, making them well suited for systems governed by partial differential equations. In the context of flood modeling, this corresponds to approximating a nonlinear solution operator

$$\mathcal{G}:\ a(\mathbf{x},t)\ \mapsto\ u(\mathbf{x},t), \tag{1}$$

where $a(\mathbf{x},t)$ denotes inputs distributed over space $(\mathbf{x})$ and time (t), such as terrain and boundary forcing, and $u(\mathbf{x},t)$ represents the resulting spatiotemporal water-depth field. Among existing operator-learning architectures, the Fourier Neural Operator (FNO) provides an efficient mechanism for learning such mappings by parameterizing $\mathcal{G}$ in the spectral domain. The FNO applies truncated spectral convolutions in Fourier space combined with local linear transformations to capture long-range spatial correlations while maintaining computational efficiency (Figure A.1). In this work, the FNO is employed as a *coarse-resolution hydrodynamic surrogate* that captures the dominant spatiotemporal structure of flood evolution on a uniform grid, while intentionally smoothing fine-scale, terrain-controlled features addressed in a subsequent refinement stage. A detailed formulation of neural operators and the Fourier Neural Operator architecture, including spectral convolution operators and implementation specifics, is provided in Appendix A.

2.2 Coarse-Resolution Spatiotemporal Flood Modeling with a Fourier Neural Operator

The first stage of the proposed framework employs a *coarse-resolution Fourier Neural Operator (FNO)* to model spatiotemporal flood dynamics over the full computational domain. The FNO operates on a fixed, uniform grid and is trained to capture the dominant patterns of flood response in space and time, conditioned on externally prescribed forcing signals and static terrain information. In this formulation, the FNO should be interpreted as a forecast-conditioned hydrodynamic emulator: it uses a prescribed forcing trajectory over the simulation horizon and does not independently forecast future inflows from past observations. By design, this model targets large-scale flood evolution rather than fine-scale inundation geometry, yielding a physically coherent low-resolution representation that is computationally efficient and stable over large domains.

The FNO maps spatially continuous forcing fields—constructed from sparse point-scale observations—and bed topography to coarse-resolution temporal flood-depth predictions (Figure 1). These predictions encode the overall spatiotemporal structure of the flood response, including timing, broad routing behavior, and aggregate attenuation, while intentionally smoothing terrain-controlled features that cannot be reliably resolved at coarse resolution. Such fine-scale effects are instead addressed in a subsequent localized refinement stage.

In this study, the global FNO uses two input channels ($C = 2$): (i) the aggregated spatiotemporal forcing field and (ii) the static bed-topography field on the coarse grid.

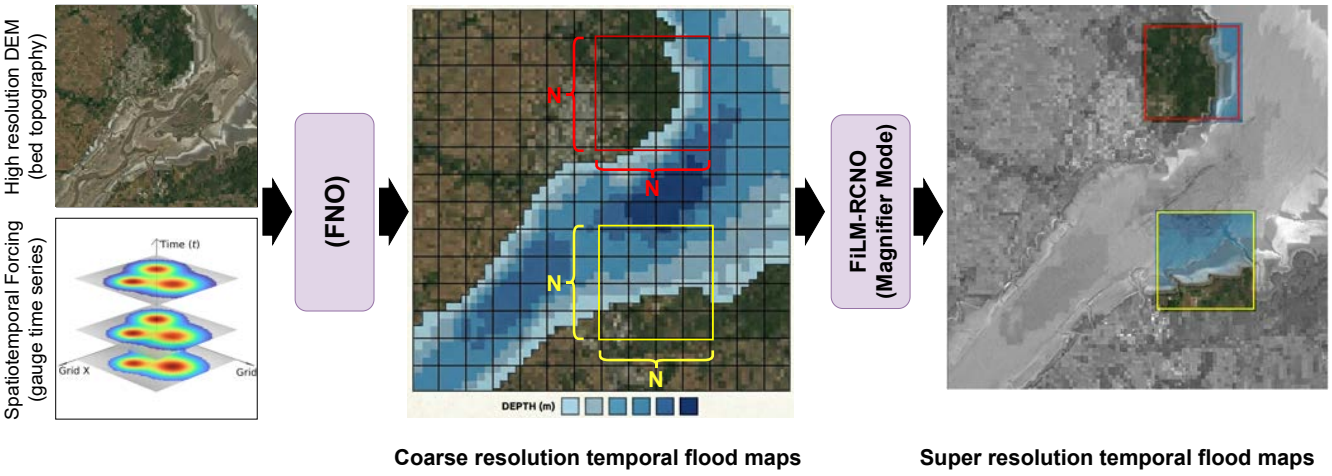


Figure 1. Overview of the Multi-Resolution Flood Mapping Framework. The pipeline consists of two primary stages: (1) A **coarse-resolution FNO** that predicts coarse-resolution temporal flood maps across the entire domain based on spatiotemporal gauge forcing and bed topography; and (2) a localized **Feature-wise Linear Modulation based Residual Convolutional Neural Operator (FiLM-RCNO)** in Magnifier Mode. The Magnifier operates on specific $N \times N$ windows extracted from the coarse resolution prediction, utilizing high-resolution topographic data and a hydrostatic baseline to produce final **Fine-Resolution Temporal Flood Maps** via hydraulic residual correction.

The construction of the spatiotemporal forcing fields, including the spatial projection and temporal aggregation procedures, is described in detail in Appendix B. Within this framework, the FNO provides the globally consistent coarse-scale context used by the localized Magnifier for high-resolution terrain-aligned correction, without the computational cost of end-to-end high-resolution operator learning.

2.3 Residual Convolutional Neural Operator Magnifier

2.3.1 Motivation and scale separation:

A natural question when employing a coarse-resolution flood model is whether fine-resolution inundation maps can be reconstructed through straightforward spatial upscaling. The most direct strategy is to apply standard interpolation schemes—such as nearest-neighbor or bilinear interpolation—to project coarse water depth predictions onto a finer computational grid. Although computationally efficient and simple to implement, these geometric interpolation approaches implicitly assume that inundation behaves as a smoothly varying field. In doing so, they neglect the governing hydraulic processes that control flow routing, wet–dry transitions, and topographically constrained connectivity. A coarse-resolution solution represents a spatially averaged flow field in which subgrid hydraulic variability has already been filtered out. Because spatial averaging is inherently a many-to-one operation, simple interpolation cannot reconstruct the information lost during coarsening. Consequently, interpolation-based upscaling fails to reproduce fine-scale features such as channelized flow structures, sharp wet–dry fronts, and elevation-controlled flow connectivity (Fraehr et al., 2024), particularly in shallow and topographically complex regions

that are most critical for flood hazard assessment. To address these limitations, we introduce a localized refinement model, referred to as the *Magnifier*, which replaces purely geometric upscaling with a physically-anchored, data-driven correction mechanism. The Magnifier is built on an explicit separation of spatial scales: coarse-resolution flood evolution—such as overall timing, propagation patterns, and large-scale attenuation—is assumed to be sufficiently represented by the Fourier Neural Operator, while discrepancies relative to high-resolution reference solutions arise primarily from localized effects driven by fine-scale terrain variability, bathymetric gradients, and hydraulic connectivity. Accordingly, the Magnifier does not seek to re-learn coarse flood dynamics, but instead refines a locally plausible representation to recover high-resolution, terrain-aligned inundation structure.

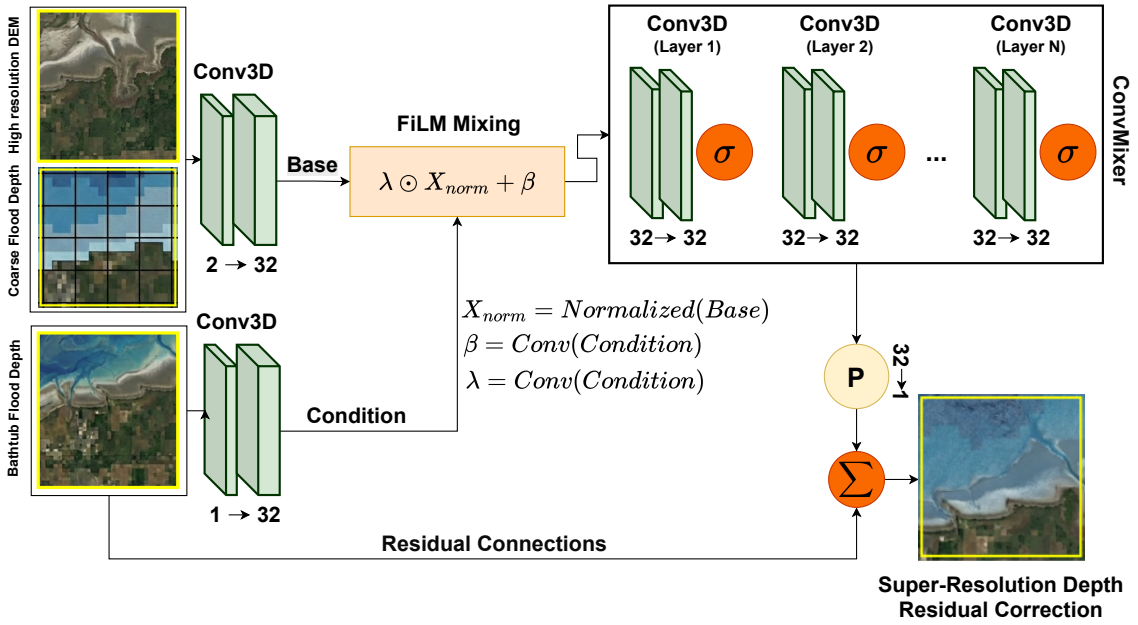


Figure 2. Architecture of the FiLM-RCNO Magnifier Model. The Magnifier performs localized refinement on an $N \times N$ window of coarse-resolution flood predictions produced by the Fourier Neural Operator. The model ingests the high-resolution digital elevation model (DEM), the coarse flood depth, and a physically anchored hydrostatic bathtub baseline. A Feature-wise Linear Modulation (FiLM) conditioned residual convolutional neural operator predicts a hydraulic residual correction, which is added to the bathtub baseline to produce the final super-resolution flood depth for each local window.

2.3.2 Physically anchored bathtub baseline:

Within each local window, refinement is anchored to a physically consistent hydrostatic *bathtub baseline*, which provides a terrain-aligned, mass-consistent initial estimate of inundation. Consider a coarse grid cell $\Omega_{ij} \subset \mathbb{R}^2$, indexed by (i,j), with area $|\Omega_{ij}|$. Let $\bar{d}_{ij}(t)$ denote the coarse-scale mean water depth in that cell at time t, where $t \in [0,T]$. The bathtub baseline assumes a locally horizontal water surface within Ω_{ij} and defines the fine-scale water depth at spatial location $(x,y) \in \Omega_{ij}$ and

time t as

$$h_{\mathrm{bt}}(x,y,t) = \big(H_{ij}(t) - B_{\mathrm{fine}}(x,y)\big)_+, \tag{2}$$

where $h_{\mathrm{bt}}(x,y,t)$ is the reconstructed fine-scale bathtub water depth, $H_{ij}(t)$ is the water surface elevation associated with coarse cell Ω_{ij} at time t, $B_{\mathrm{fine}}(x,y)$ is the fine-resolution bed elevation at location (x,y), and $(\cdot)_+ := \max(\cdot, 0)$ enforces non-negativity so that dry locations have zero depth.

The water surface elevation $H_{ij}(t)$ is determined by enforcing conservation of water volume within the coarse cell:

$$\int_{\Omega_{ij}} \big(H_{ij}(t) - B_{\mathrm{fine}}(x,y)\big)_+ \, dA = \bar{d}_{ij}(t)\,|\Omega_{ij}|, \tag{3}$$

where dA denotes the differential area element. The left-hand side represents the total reconstructed water volume over the fine-scale topography inside Ω_{ij}, while the right-hand side is the water volume implied by the coarse-scale mean depth $\bar{d}_{ij}(t)$ over the same coarse-cell area.

This construction provides a terrain-consistent hydrostatic baseline that aligns the coarse predicted water level with fine-scale topography and supports volume-aware depth reconstruction. Because the baseline is hydrostatic, it should not be interpreted as enforcing full shallow-water-equation dynamics or momentum consistency.

2.3.3 Residual refinement with FiLM-RCNO:

To correct residual depth differences between the hydrostatic baseline and the reference fields, including effects associated with backwater, spillover, and channelized inundation patterns, the Magnifier is formulated as a residual learner. The final refined depth field is defined as

$$h_{\mathrm{final}}(x,y,t) = \big(h_{\mathrm{bt}}(x,y,t) + \Delta h(x,y,t)\big)_+, \tag{4}$$

where $\Delta h(x,y,t)$ is the hydraulic residual predicted by the FiLM-RCNO. To encourage local volume consistency with the coarse FNO prediction, the residual correction Δh is constrained to have zero net contribution over the $N \times N$ local refinement window:

$$\int_{\text{window}} \Delta h(x,y,t)\, dA = 0. \tag{5}$$

This constraint encourages the refinement stage to redistribute water locally according to fine-scale terrain gradients and learned inundation structure, rather than introducing large artificial volume changes. In this architecture, multiscale behavior is achieved through the explicit coarse-to-fine decomposition between the global FNO and the local Magnifier, while Feature-wise Linear Modulation (FiLM) (Turkoglu et al., 2022) serves as an internal conditioning mechanism within the Magnifier. FiLM modulates intermediate feature channels using coarse-scale flood state and local terrain descriptors, allowing the network to adapt residual corrections to local hydraulic conditions (Figure 2).

2.4 Training, inference, and integration:

During training, local windows are sampled using a spatial random-walk strategy to efficiently cover large spatial domains. To address the strong imbalance between wet and dry regions, a wet-patch quota is enforced by probabilistically downsampling windows that are entirely dry in the reference solution. Here, wet and dry states are defined using a depth threshold of $0.01\,\mathrm{m}$, consistent with the criterion used in ANUGA (Roberts et al., 2015) to distinguish inundated and dry cells. For additional efficiency, the Magnifier is applied only to windows that are not dry at the coarse resolution, as identified from the Fourier Neural Operator prediction. At inference time, the Magnifier is applied in a sliding-window manner with overlapping windows, retaining only the central region of each refined patch to ensure smooth global reconstruction. Together, the coarse-resolution Fourier Neural Operator and the Magnifier form a unified multi-resolution surrogate that separates coarse spatiotemporal flood evolution from local geometric refinement. The coarse model provides scalable solver-consistent depth and inundation-boundary predictions within the tested configurations, while the Magnifier recovers decision-critical fine-scale inundation structure in a terrain-aware and computationally-efficient manner.

3 Model Training and Experimental Design

This section outlines the development of the multi-resolution flood modeling framework, detailing the hydrodynamic simulation setup, neural operator implementation, training configuration, and experimental design used for systematic evaluation.

3.1 Framework overview and workflow

Figure 3 illustrates the overall workflow adopted in this study for developing and evaluating the proposed flood inundation learning framework. The process begins with the generation of a diverse ensemble of upstream hydrographs representing a broad range of flood scenarios, spanning variations in peak discharge, timing, and duration. These boundary forcings are simulated with ANUGA (Roberts et al., 2015) to produce physically-based, time-evolving reference water depth fields over the computational domain; detailed solver formulation and configuration information is provided in Appendix D. The reported evaluation is therefore a surrogate-to-hydrodynamic-solver assessment: model predictions are compared against ANUGA-generated hydrodynamic reference fields under the tested configurations.

The simulated depth sequences are then organized into structured datasets, where hydrographs serve as inputs, and the corresponding spatiotemporal depth fields form the learning targets. After preprocessing and resolution alignment, the neural framework is trained in three stages. First, the coarse-resolution Fourier Neural Operator (FNO) is trained independently to learn global flood-wave propagation. Second, the coarse operator is frozen, and the high-resolution magnifier model is trained to reconstruct terrain-aligned inundation patterns using physics-anchored residual learning. Finally, the full framework is fine-tuned end-to-end to ensure consistency between global dynamics and local topographic refinement. The trained system is then evaluated on independent test scenarios, with predictions assessed against ANUGA-generated hydrodynamic reference

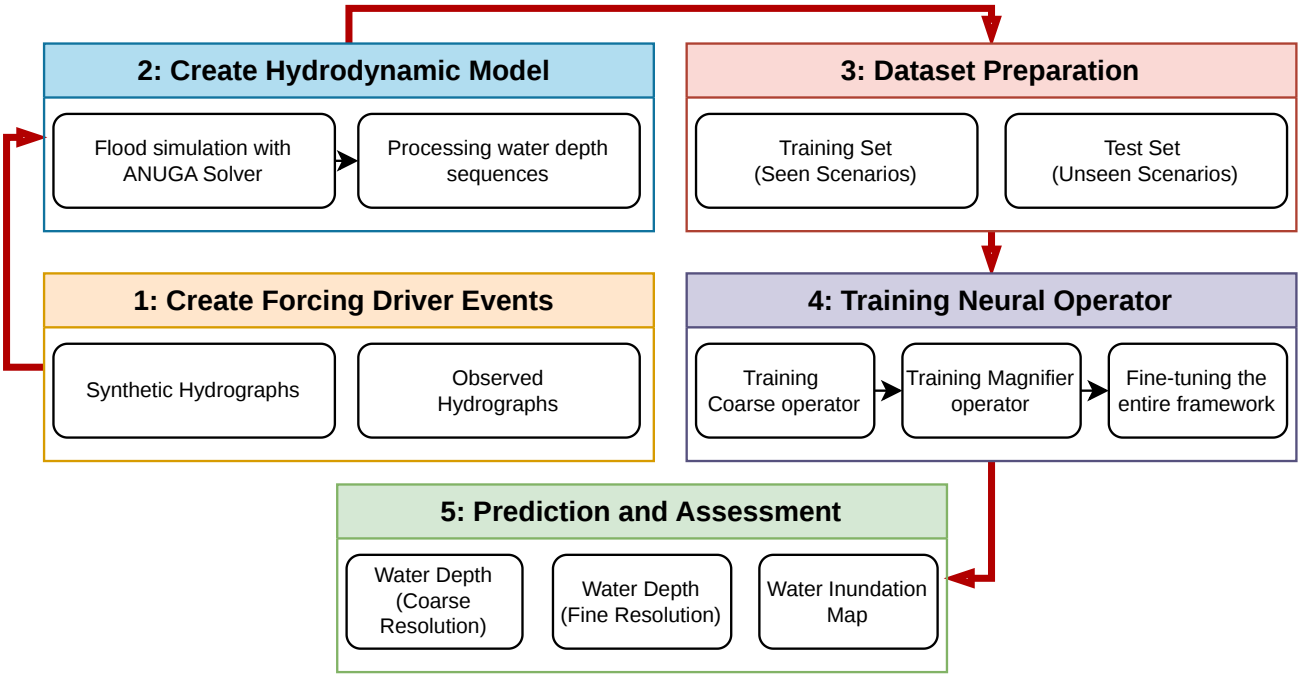


Figure 3. Implementation phases of the multi-resolution flood mapping framework.

simulations using both depth-based and extent-based metrics. In this context, the proposed framework should be interpreted as a fast emulator for repeated ANUGA-like scenario simulations.

3.2 Study sites

We evaluated the proposed framework across three hydrologically distinct flood-prone environments: (i) a hurricane-driven fluvial flooding event along the Neuse River upstream of Quaker Neck Lake (North Carolina, USA), (ii) a dam-break overtopping failure scenario at Fall River Lake Dam (Kansas, USA), and (iii) a long-duration, regulated floodplain system in the Chowilla floodplain within the Murray–Darling Basin (South Australia, Australia). Table 1 and Figure 4 summarize their geographic context, domain configuration, and elevation characteristics. Detailed hydrodynamic setup and boundary condition specifications are provided in the subsequent subsections.

The first case study focuses on hurricane-driven fluvial flooding along the Neuse River upstream of Quaker Neck Lake near Goldsboro, North Carolina, USA (Figure 4a,c). This low-gradient plain reach contains extensive floodplains and exhibits strong sensitivity to extreme precipitation and upstream runoff forcing. During Hurricane Matthew in October 2016, sustained basin-wide rainfall produced severe riverine flooding and widespread inundation throughout the corridor (Musser et al., 2017). The computational domain is defined by eight upstream streamflow gauges that provide the boundary forcing configuration (Figure 5a). Hydrologic forcing data is derived from the observed Hurricane Matthew hydrographs spanning 5–16 October

2016 (Watson and Musser, 2017), which define the inflow conditions for the simulations. Spatial variability in hydraulic resistance is represented through Manning's roughness coefficients ranging from 0.015 to $0.06\ \mathrm{s \cdot m^{-1/3}}$, obtained by converting land use and land cover (LULC) classes from the National Land Cover Database (Dewitz, 2021) into roughness values using a standard lookup table following Arcement and Schneider (1989), reflecting contrasts between confined channel sections and adjacent floodplain areas. The inundation extent associated with this reference Neuse River event was obtained from the FIMbench product, which reconstructed it using remote-sensing imagery, DEM-based gap filling, and the hydrologically guided region-growing (HGRG) algorithm (Cohen et al., 2024). This inundation information was used to support examination of the base hydrodynamic setup for the Neuse case. However, FIMbench represents a single observed event, whereas the main objective of this study is to evaluate surrogate performance across many controlled synthetic forcing scenarios. Therefore, the reported surrogate metrics focus on agreement between the surrogate and ANUGA-generated reference simulations over the independent test scenario ensemble, rather than direct surrogate-versus-FIMbench validation for one event.

The second case study considers an overtopping-driven dam-break scenario at Fall River Lake Dam in Greenwood County, Kansas (Figure 4a,d). The structure is an earthen embankment dam with a height of 28.6 m and a contributing drainage area of approximately 1,515 $\mathrm{km^2}$. According to the National Inventory of Dams, the reservoir has a normal storage capacity of 316 million $\mathrm{m^3}$, increasing to approximately 501 million $\mathrm{m^3}$ at crest elevation. The dam is designated as a High Hazard Potential facility, indicating that failure could result in substantial downstream consequences, including possible loss of life (Jakubauskas et al., 2011; Balachandran et al., 2025). Overtopping is selected as the governing failure mechanism due to its documented prevalence and destructive impact (Balachandran et al., 2025). This process occurs when reservoir levels exceed the crest elevation, initiating progressive embankment erosion and producing a rapid, high-intensity outflow. For this scenario, the upstream boundary condition is represented by a sharp triangular breach-release pulse characteristic of overtopping-induced discharge evolution. Representative hydrographs are shown in Figure 6c, and the mathematical formulation of the synthetic hydrograph generation is provided in Appendix C. A spatially-uniform Manning's roughness coefficient of $n = 0.06\,\mathrm{s \cdot m^{-1/3}}$ was applied throughout the downstream domain to represent effective hydraulic resistance across the floodplain.

The third case study focuses on the Chowilla floodplain, a low-relief multi-anabranch system within the Murray–Darling Basin at the border of Victoria, New South Wales, and South Australia (Figure 4a,e). The domain consists of flat terrain interlaced with a dense network of distributary channels that facilitate strong hydraulic connectivity between the main river corridor and the surrounding floodplain. Regulation through 22 weirs and control structures produces spatially variable inundation patterns governed by both managed releases and natural channel interactions (Fraehr et al., 2022, 2023, 2024). Inflow enters the domain through three primary upstream rivers (G1–G3; Figure 4e), forming a multi-boundary forcing configuration. The dataset was derived from a prolonged hydrologic period extending from 5 August 2010 to 1 June 2011, characterized by a gradual water-level rise followed by a sustained recession phase under regulated conditions rather than a short-lived extreme flood pulse. Representative reference and synthetic hydrographs are displayed in Figure 5(b). Spatially-varying Manning's coefficients were prescribed with $n = 0.026\,\mathrm{s \cdot m^{-1/3}}$ within river channels and $n = 0.083\,\mathrm{s \cdot m^{-1/3}}$ across the floodplain to account for differences in momentum dissipation between confined flow paths and vegetated overbank areas.

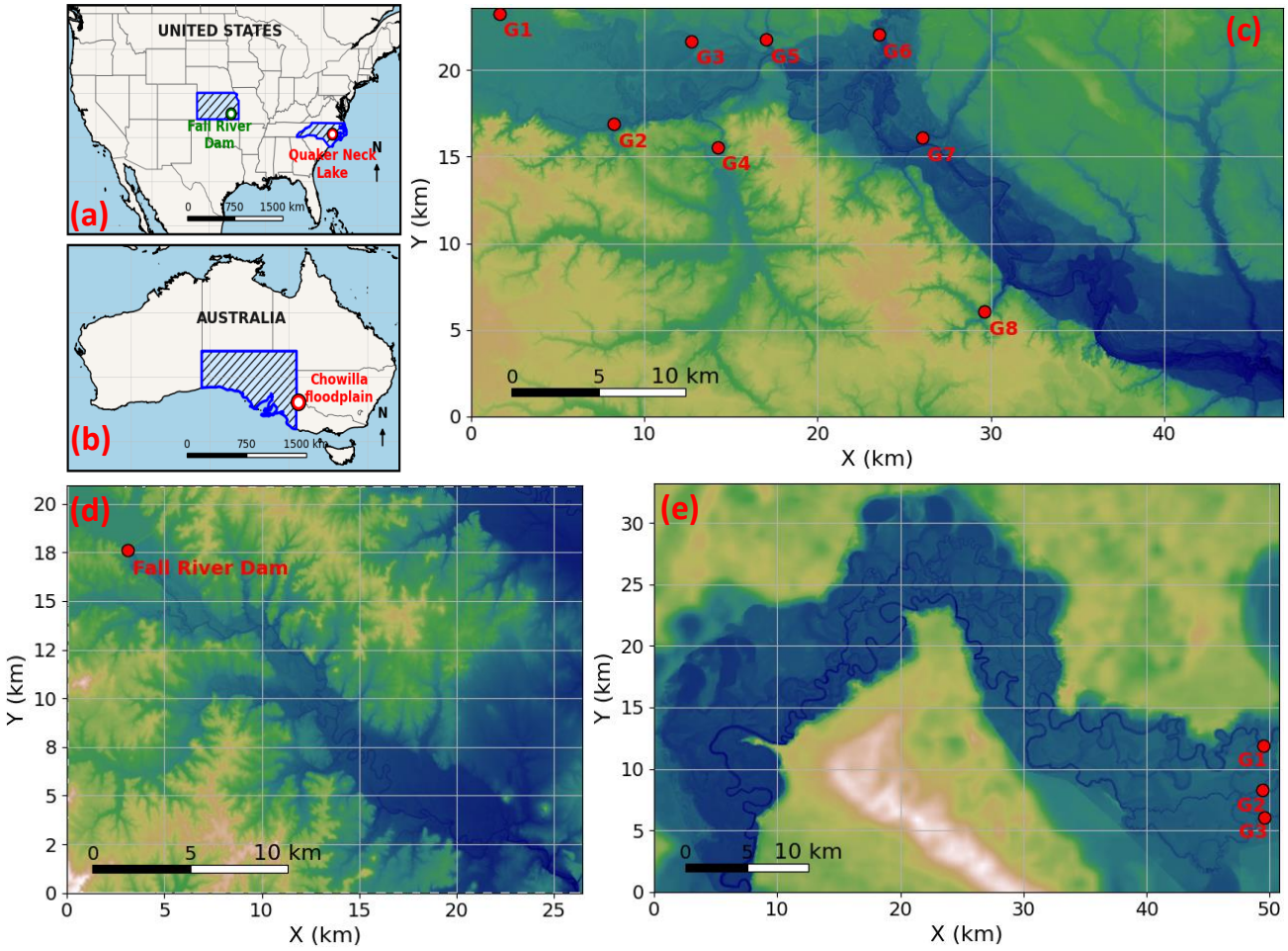
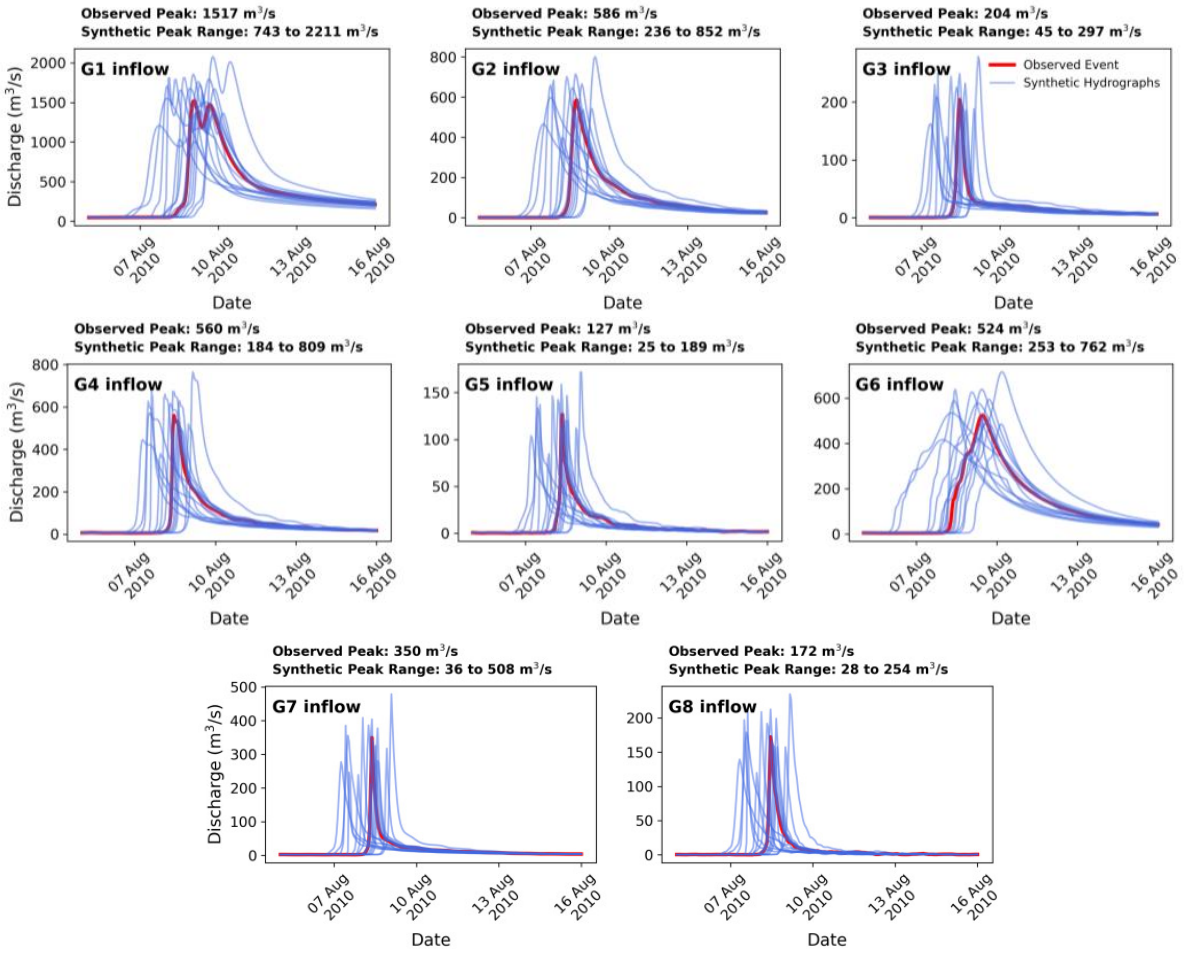


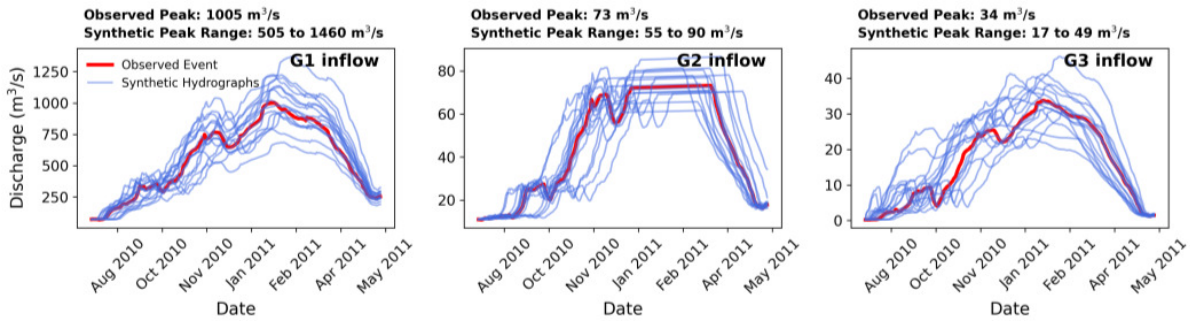
Figure 4. Overview of the case studies. Panels (a) and (b) illustrate the geographic locations of the selected study areas in the United States and Australia. Panels (c), (d), and (e) present the corresponding model domains and elevation fields for the Neuse River reach; the Fall River Lake watershed; and the Chowilla floodplain, respectively. The letter G followed by a number indicates a gage location.

3.3 Hydrodynamic Modeling, Data Preparation, and Operator Learning Configuration

Training datasets for all case studies were produced using high-resolution hydrodynamic simulations of the depth-averaged Shallow Water Equations solved with the ANUGA framework (Roberts et al., 2015). ANUGA applies a finite-volume scheme on unstructured triangular meshes, allowing stable resolution of wetting–drying processes, irregular bathymetry, and dynamically evolving flood boundaries. For each synthetic inflow realization, the solver generates the full spatiotemporal water depth evolution across the event duration. After each ANUGA simulation, the water-depth outputs defined on the unstructured triangular mesh were interpolated onto structured Cartesian raster grids at the target spatial resolution of each case study. This conversion was performed to create consistent tensor representations for neural-operator training and evaluation. The same



(a) The Neuse River case



(b) The Chowilla floodplain case

Figure 5. Upstream inflow hydrographs for (a) the Neuse River case and (b) the Chowilla floodplain case. Black curves denote the observed reference events, while blue curves illustrate representative synthetic realizations generated through controlled perturbation of peak magnitude, timing, and limb steepness (see Appendix C).

Table 1. Overview of case studies and general characteristics.

	Neuse River	Fall River Lake Dam	Chowilla Floodplain
Domain size	$1104\,\mathrm{km}^2$	$554\,\mathrm{km}^2$	$740\,\mathrm{km}^2$
Case type	Hurricane-driven fluvial flood	Overtopping dam-break	Regulated multi-boundary floodplain
Number of inflows	8	1 (Dam operation)	3
Range of inflows (Min–Max)	0–$2{,}211\,\mathrm{m}^3/\mathrm{s}$	0–$20{,}000\,\mathrm{m}^3/\mathrm{s}$	0–$1{,}460\,\mathrm{m}^3/\mathrm{s}$
Duration of flood events	11 days	24 hours	300 days
Hydrodynamic solver resolution	379,487 triangle cells	199,216 triangle cells	154,504 triangle cells
	189,872 nodes	98,946 nodes	77,504 nodes
Total simulated flood events	150 total realizations: 100 training scenarios and 50 unseen test scenarios		
Dataset spatial resolution	$10\,\mathrm{m} \times 10\,\mathrm{m}$	$25\,\mathrm{m} \times 25\,\mathrm{m}$	$30\,\mathrm{m} \times 30\,\mathrm{m}$
Dataset temporal intervals	3-hour intervals	15-minute intervals	3-day intervals
Coarse operator resolution	$300\,\mathrm{m} \times 300\,\mathrm{m}$	$200\,\mathrm{m} \times 200\,\mathrm{m}$	$300\,\mathrm{m} \times 300\,\mathrm{m}$
Resolution scaling parameter (f)	30	8	10
Magnifying window (N)	5	5	5
Magnifier operator resolution	$10\,\mathrm{m} \times 10\,\mathrm{m}$	$25\,\mathrm{m} \times 25\,\mathrm{m}$	$30\,\mathrm{m} \times 30\,\mathrm{m}$

ANUGA-to-raster interpolation procedure was used for the reference depth fields, the coarse FNO training targets, the fine-resolution Magnifier targets, and the interpolation baseline, ensuring that all model outputs were compared on consistent spatial grids. This preprocessing produced depth tensors of size $[n_x, n_y, n_t]$ for each realization, where n_x and n_y denote the raster-grid dimensions and n_t denotes the number of archived temporal snapshots. To assess sensitivity to training-set size, we conducted an additional learning-curve analysis using expanded training libraries of 50, 100, 200, and 300 training realizations, while evaluating all models against the same 50 independent test scenarios. The results, reported in Appendix E, show that performance improves substantially from 50 to 100 training realizations but only modestly beyond 100, indicating saturation within the sampled forcing envelope. Based on this analysis, the main experiments use 150 total realizations for each case study, with 100 scenarios used for training and 50 independent test scenarios reserved only for final evaluation. Each realization yields a three-dimensional depth tensor of size $[n_x, n_y, n_t]$, where n_x and n_y correspond to spatial resolution and n_t denotes the number of archived temporal snapshots. These tensors form the supervised dataset for learning the mapping from boundary inflow forcing to flood depth evolution. For computational context, generating the 150 ANUGA realizations for each case study required approximately 450 wall-clock hours in total (about 3 hours per realization), whereas one surrogate forward pass was trivial (approximately $1\,\mathrm{s}$). This efficiency contrast was observed across scenarios with very different physical event durations, including a 24-hour dam-break case, an 11-day riverine case, and an approximately 300-day floodplain case.

3.3.1 Case (i): Neuse River – Hurricane Matthew (2016)

For the Neuse River domain, simulations were conducted on an unstructured triangular mesh with element areas ranging from 97 to $324{,}000\,\mathrm{m}^2$. This mesh configuration provides concentrated refinement along the active river channel while maintaining

computational efficiency over the surrounding coastal floodplain. Computed depth fields were interpolated onto a structured grid with $10\,\mathrm{m} \times 10\,\mathrm{m}$ spatial resolution. The upstream boundary forcing was specified using eight discharge hydrographs (Figure 4c) associated with the Hurricane Matthew event spanning 5–16 October 2016, illustrated in Figure 5a. Based on this observed event, 150 synthetic inflow realizations were generated through systematic modification of peak magnitude, timing characteristics, and hydrograph morphology following the formulation detailed in Appendix C. Simulations covered the entire 11-day interval using adaptive time steps between 310 and 2600 seconds. Water depth fields were recorded at 3-hour intervals, yielding 89 temporal snapshots per realization. Within the operator learning framework, the coarse FNO model operated on a $300\,\mathrm{m} \times 300\,\mathrm{m}$ grid to capture basin-scale flood dynamics. The Magnifier component enhanced these coarse predictions to the native $10\,\mathrm{m} \times 10\,\mathrm{m}$ resolution using a refinement ratio $f = 30$ and local window size $N = 5$, enabling recovery of terrain-controlled inundation features while preserving large-scale hydraulic behavior.

3.3.2 Case (ii): Fall River Lake – Overtopping Dam-Break Scenario

For the Fall River Lake overtopping case, hydrodynamic computations were performed on a triangular mesh with cell areas between 345 and $10{,}000\,\mathrm{m}^2$. This discretization increased spatial resolution in zones of strong hydraulic gradients immediately downstream of the breach while maintaining tractability across the broader floodplain. Simulation outputs were projected onto a structured grid with $25\,\mathrm{m} \times 25\,\mathrm{m}$ spacing. The upstream inflow boundary condition was defined as a triangular breach-release hydrograph consistent with the overtopping configuration reported in Balachandran et al. (2025) and shown in Figure 6. Synthetic breach realizations were generated by sampling peak discharge, time to peak, and release duration within physically admissible ranges, as specified in Appendix C. Simulations were carried out over a 24-hour window using adaptive time increments between 3.2 and 15 seconds. Water depths were archived at 15-minute intervals, producing 97 temporal snapshots per realization. For surrogate prediction, the coarse FNO model operated on a $200\,\mathrm{m} \times 200\,\mathrm{m}$ grid to represent domain-scale breach wave propagation. Fine-scale inundation patterns at $25\,\mathrm{m} \times 25\,\mathrm{m}$ resolution were reconstructed using a refinement ratio $f = 8$ and window size $N = 5$, allowing accurate recovery of rapidly varying downstream hydraulic structures.

3.3.3 Case (iii): Chowilla Floodplain

For the Chowilla floodplain system, simulations were executed on an unstructured mesh with element areas ranging from 400 to $256{,}000\,\mathrm{m}^2$. This mesh design resolves distributary channels and floodplain connectivity while maintaining efficiency over the low-relief basin. The simulated water depth fields were then mapped to a structured grid with spatial resolution $30\,\mathrm{m} \times 30\,\mathrm{m}$. The boundary forcing consisted of three upstream hydrographs associated with the regulated Murray–Darling Basin inflow system, depicted in Figure 5b. A total of 150 synthetic realizations were produced by perturbing the prolonged historical hydrograph spanning 5 August 2010 to 1 June 2011 using the procedure outlined in Appendix C. In particular, peak discharge was varied within $\pm 50\%$ of the reference event maximum while preserving the large-scale temporal morphology of the regulated hydrograph. Simulations were performed over the full hydrologic period with adaptive time steps between 310 and 2600 seconds. Water depths were stored at 3-day intervals, resulting in 100 temporal snapshots per realization. Within the operator learning configuration, the coarse FNO model operated on a $300\,\mathrm{m} \times 300\,\mathrm{m}$ grid to represent large-scale regulated

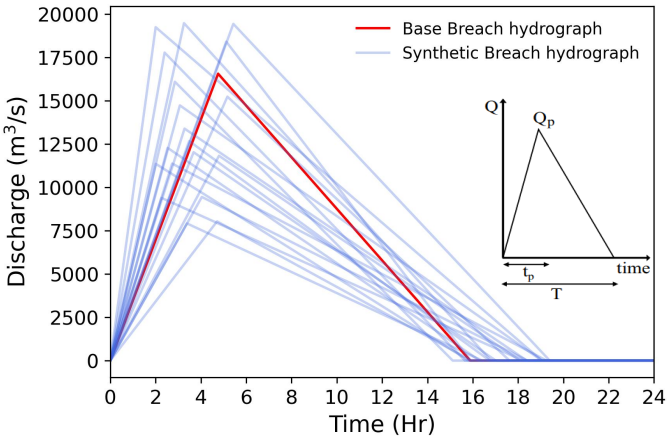


Figure 6. Overtopping breach hydrographs for the Fall River Lake Dam scenario. The red curve represents the reference breach configuration reported in Balachandran et al. (2025), while the blue curves show synthetic realizations obtained by sampling peak discharge (Q_p), time to peak (t_p), and total duration (T) within prescribed bounds (see Appendix C).

flood dynamics across the multi-anabranch network. The Magnifier module refined these predictions to the native $30\,\mathrm{m} \times 30\,\mathrm{m}$ resolution using a refinement ratio $f = 10$ and window size $N = 5$, enabling reconstruction of inundation boundaries influenced by subtle topographic gradients and channel connectivity.

3.4 Implementation and Training Configuration

The neural network was implemented in Python 3.12 using PyTorch 2.1 with tensor-based automatic differentiation. The coarse model follows the Fourier Neural Operator architecture of Li et al. (2020), comprising four Fourier layers with spectral convolutions truncated to 10 modes in each spatial (x, y) and temporal (t) coordinate. Inputs are lifted to 128 feature channels and projected back to physical depth through linear layers. The FiLM-RCNO Magnifier used FiLM layers with feature width 32 and three conditioning channels for mixing, followed by a convolutional mixer consisting of three convolutional layers with 32 feature maps each and a linear layer of width 32 projecting the latent representation to the physical residual space.

Training employed the AdamW optimizer with learning rates of 1×10^{-4} for the FNO and 5×10^{-4} for the FiLM-RCNO Magnifier, a weight-decay coefficient of 1×10^{-5}, and a batch size of 8. A random seed of 42 was used for model initialization, data shuffling, and patch sampling where applicable. The supervised training objective minimized the discrepancy between predicted water depth fields and ANUGA-generated reference depth fields. Each model was pretrained for 500 epochs, followed by 50 epochs of joint fine-tuning of the full multi-resolution framework. During training, a model checkpoint was saved after each epoch, and the checkpoint with the best performance on the training set was retained as the final model. The selected checkpoint was then evaluated on the held-out test set reserved exclusively for final assessment. Computations were performed on a node equipped with dual Intel Xeon Gold 6248 CPUs with 40 cores total and an NVIDIA Tesla V100S GPU with 32 GB memory. Data preprocessing was conducted on the CPU, while model training and inference used GPU acceleration.

3.5 Multi-resolution evaluation and performance metrics

The predictive performance of the proposed multi-resolution learning framework was evaluated against reference simulations generated using the two-dimensional ANUGA hydrodynamic solver (Roberts et al., 2015), which served as the benchmark solution throughout this study (Appendix D). Evaluation was conducted at two spatial resolutions corresponding to the architecture of the framework (Section 2.2 and Section 2.3). At the coarse scale, outputs from the global neural operator were compared with hydrodynamic solutions downsampled to the same resolution, enabling assessment of large-scale flood-wave propagation and basin-scale inundation dynamics. At the fine scale, magnifier predictions were evaluated against high-resolution hydrodynamic outputs to quantify the accuracy of reconstructed inundation boundaries and localized depth variations governed by topography. To further contextualize model performance, we additionally compared fine-scale magnifier predictions with simple geometric bilinear interpolation of the coarse outputs, thereby highlighting the inherent limitations of interpolation-based approaches in reconstructing high-resolution inundation structure, which constitutes one of the core motivations of the present work. To further examine hydrodynamic consistency, model accuracy was assessed separately across two depth regimes. Shallow inundation regions, defined by water depths less than $0.5\,\mathrm{m}$, were evaluated independently due to their sensitivity to wet–dry transitions and subgrid topographic effects (Kabir et al., 2020; Herath et al., 2025). Deeper flow regions with depths exceeding $0.5\,\mathrm{m}$ were analyzed to assess performance under fully developed hydraulic conditions. This stratified evaluation provides insight into model robustness under both marginal and fully inundated states.

For uncertainty-aware evaluation, we defined a structured topographic perturbation protocol for the dam-break case. With z denoting the baseline bathymetry, perturbed realizations are generated:

$$z^{(i)} = z + \epsilon^{(i)}, \quad i = 1, \ldots, 150, \tag{6}$$

where each $\epsilon^{(i)}$ is an independent, spatially-coherent, zero-mean random field. The perturbation magnitude is controlled by a fixed standard deviation σ, set to approximately 5% of the characteristic vertical scale to represent moderate yet meaningful terrain uncertainty. We used 100 realizations for training and 50 independent test realizations reserved only for final evaluation, while keeping all architecture and optimization settings identical to the deterministic configuration so that performance changes could be attributed to geometric uncertainty alone.

3.5.1 Depth-based error metrics:

Continuous depth predictions were evaluated using complementary dimensionless error and skill measures. The relative root mean squared error (relRMSE) is defined as

$$\mathrm{relRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2}}{\bar{O}}, \tag{7}$$

where O_i and P_i denote reference and predicted depth values, respectively, N is the number of evaluated samples, and $\bar{O}$ is the mean reference depth over the evaluated domain. The relRMSE provides a normalized measure of prediction error relative to characteristic flood magnitude.

Predictive skill relative to the reference hydrodynamic simulation is quantified using the Nash–Sutcliffe efficiency (NSE),

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} \left(O_i - \bar{O}\right)^2}. \tag{8}$$

An NSE value of unity indicates perfect agreement, values near zero indicate limited predictive skill, and negative values imply performance worse than using the mean reference depth as a predictor.

3.5.2 Peak magnitude and timing evaluation:

To assess the temporal fidelity of flood-wave propagation, peak magnitude and timing errors were evaluated using domain-integrated water volume.

$$H(t) = \int_{\Omega} h(x,y,t)\, dA \tag{9}$$

denotes the total water volume over the spatial domain Ω at time t.

The peak period is defined as the timesteps corresponding to the top 5% highest values of $H(t)$ obtained from the reference simulation. With $t^{\mathrm{pred}}_{\mathrm{peak},5\%}$ and $t^{\mathrm{ref}}_{\mathrm{peak},5\%}$ respectively denoting the vectors of timesteps associated with these highest values for the predicted and reference simulations, the average peak-period volumes are $\overline{H}^{\mathrm{pred}}_{\mathrm{peak},5\%}$ and$\overline{H}^{\mathrm{ref}}_{\mathrm{peak},5\%}$.

The relative peak value error is defined as

$$\mathrm{relPeakValErr} = \frac{\overline{H}^{\mathrm{pred}}_{\mathrm{peak},5\%} - \overline{H}^{\mathrm{ref}}_{\mathrm{peak},5\%}}{\overline{H}^{\mathrm{ref}}_{\mathrm{peak},5\%}}. \tag{10}$$

This metric quantifies the relative overestimation or underestimation of peak flood magnitude, providing a measure of how accurately the surrogate reproduces extreme event severity.

The average timesteps over the predicted and reference peak periods are respectively denoted by $\bar{t}^{\mathrm{pred}}_{\mathrm{peak},5\%}$ and $\bar{t}^{\mathrm{ref}}_{\mathrm{peak},5\%}$. Peak timing accuracy normalized by peak duration is computed as

$$\mathrm{relPeakTimeErr\text{-}1} = \frac{\bar{t}^{\mathrm{pred}}_{\mathrm{peak},5\%} - \bar{t}^{\mathrm{ref}}_{\mathrm{peak},5\%}}{\max\left(t^{\mathrm{ref}}_{\mathrm{peak},5\%}\right) - \min\left(t^{\mathrm{ref}}_{\mathrm{peak},5\%}\right)}. \tag{11}$$

This formulation evaluates temporal alignment within the peak plateau, indicating whether the predicted peak occurs too early or too late relative to the duration of peak conditions.

A second timing metric, normalized by the characteristic rising-limb duration, is defined as

$$\mathrm{relPeakTimeErr\text{-}2} = \frac{\bar{t}^{\mathrm{pred}}_{\mathrm{peak},5\%} - \bar{t}^{\mathrm{ref}}_{\mathrm{peak},5\%}}{\bar{t}^{\mathrm{ref}}_{\mathrm{peak},5\%} - t^{\mathrm{ref}}_{\mathrm{rise},10\%}}, \tag{12}$$

where $t^{\mathrm{ref}}_{\mathrm{rise},10\%}$ denotes the timestep corresponding to a 10% increase above the minimum reference volume, marking the onset of the rising limb. This metric evaluates the accuracy of flood-wave propagation timing relative to overall event development. Together, these measures characterize both the magnitude and temporal consistency of peak flood conditions while reducing sensitivity to short-term fluctuations by averaging over the highest 5% of event values (Fraehr et al., 2022).

3.5.3 Inundation extent metrics:

Binary inundation performance was evaluated using standard hydrologic classification metrics (Fraehr et al., 2022). The Probability of Detection (POD) quantifies detection capability:

$$\mathrm{POD}\ (\%) = 100 \times \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{13}$$

True positives denote correctly predicted flooded cells, false positives represent false alarms, and false negatives correspond to missed inundation.

False Alarm Ratio (FAR) measures false flooding predictions:

$$\mathrm{FAR}\ (\%) = 100 \times \frac{\text{False Positives}}{\text{True Positives} + \text{False Positives}}. \tag{14}$$

Critical Success Index (CSI) provides a balanced assessment penalizing both missed detections and false alarms:

$$\mathrm{CSI}\ (\%) = 100 \times \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives} + \text{False Positives}}. \tag{15}$$

3.5.4 Uncertainty calibration and probabilistic consistency:

In addition to deterministic accuracy metrics, we evaluated the probabilistic consistency of the learned operator under structured topographic uncertainty. Because the model produces an ensemble of solutions induced by bathymetric perturbations, inundation is interpreted in a probabilistic sense. For a prescribed depth threshold h_0, the model-predicted probability of inundation is defined as

$$P_{\mathrm{model}}(x, y, t) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(h_{\mathrm{model}}^{(i)}(x, y, t) > h_0 \right), \tag{16}$$

with an analogous definition for $P_{\mathrm{solver}}(x, y, t)$ obtained from the reference simulations.

Calibration was assessed using reliability diagrams that compare predicted probabilities with observed frequencies aggregated over all timesteps and spatial locations. Perfect calibration corresponds to alignment along the 1:1 diagonal. Deviations from this line indicate systematic overconfidence or underconfidence in predicted inundation probabilities.

To quantify probabilistic miscalibration, we compute the area-weighted Expected Calibration Error (ECE),

$$\mathrm{ECE} = \sum_{b=1}^{B} w_b \left| \mathrm{acc}_b - \mathrm{conf}_b \right|, \tag{17}$$

where B denotes the number of probability bins, conf_b is the mean predicted probability in bin b, acc_b is the corresponding observed inundation frequency from the solver, and w_b represents the fractional spatial area contained in that bin. This metric evaluates the extent to which predicted probabilities are statistically consistent with realized inundation outcomes across the domain. Low ECE values indicate that the learned operator not only reproduces deterministic depth fields but also preserves the probabilistic structure induced by geometric uncertainty.

4 Results

4.1 Spatiotemporal Reconstruction of Water Depth Dynamics

Continuous depth metrics across the 50 independent test scenarios for each study domain demonstrate that the proposed multi-resolution framework accurately reconstructed the spatiotemporal evolution of floodwater depth at both coarse and fine scales (Table 2). At coarse resolution, FNO preserved large-scale hydraulic structure with very low relRMSE and near-unity NSE, indicating that dominant wave propagation, volumetric redistribution, and basin-scale storage dynamics were captured with minimal distortion. These results indicate that the learned operator accurately reproduces the hydrodynamic-solver-generated depth response across distinct hydraulic regimes. Because the present surrogate predicts water depth rather than velocity or discharge, these metrics should be interpreted as evidence of solver-consistent depth and boundary emulation rather than full independent verification of shallow-water-equation dynamics.

Table 2. Aggregated depth prediction performance across the independent test scenarios for coarse (FNO) and fine (Magnifier) resolutions, with bicubic interpolation as the baseline. Relative root-mean-square error (relRMSE) and Nash–Sutcliffe Efficiency (NSE) are stratified by shallow ($h < 0.5\,\mathrm{m}$) and deep ($h \geq 0.5\,\mathrm{m}$) flow regimes.

Case	Scale	relRMSE	NSE	Shallow ($h < 0.5$ m)		Deep ($h \geq 0.5$ m)	
				relRMSE	NSE	relRMSE	NSE
(i) Neuse River	Coarse (FNO)	0.0219	0.9999	0.1831	0.9977	0.0049	0.9999
	Fine (Magnifier)	0.3315	0.9900	1.1839	0.9295	0.1076	0.9666
	Fine (Interpolation)	1.3875	0.8257	9.9874	-3.9941	0.3988	0.5421
(ii) Fall River Dam Break	Coarse (FNO)	0.0130	0.9999	0.3874	0.9955	0.0032	0.9999
	Fine (Magnifier)	0.0375	0.9999	2.1503	0.9413	0.0093	0.9997
	Fine (Interpolation)	1.0557	0.9080	44.9142	-24.5191	0.2989	0.7361
(iii) Chowilla floodplain	Coarse (FNO)	0.0228	0.9999	0.1969	0.9959	0.0088	0.9997
	Fine (Magnifier)	0.0368	0.9998	0.4647	0.9838	0.0118	0.9996
	Fine (Interpolation)	1.1168	0.7747	9.6478	-5.8420	0.4508	0.4815

To complement the aggregated metrics (Table 2), we plot the event-wise distributions of relRMSE and NSE across the 50 independent test scenarios (Figure 7). This distributional view demonstrates that the reported aggregate performance is consistent across unseen scenarios rather than being dominated by a small number of favorable cases.

At fine resolution, the Neural Residual Magnifier maintained strong agreement with the reference solution while enhancing spatial fidelity relative to the coarse prediction. Although relRMSE increased slightly due to resolution refinement and sensitivity to local gradients, NSE values remained consistently above 0.92, confirming that FNO-magnifier joint training did not degrade global predictive skill. In contrast, simple interpolation exhibited substantially larger relative errors and reduced efficiency across all domains.

Although it preserved coarse-scale mean values, it assumes linear variation between neighboring cells and neglects the nonlinear coupling between depth, momentum, and terrain gradients inherent to the shallow water equations. Consequently, in-

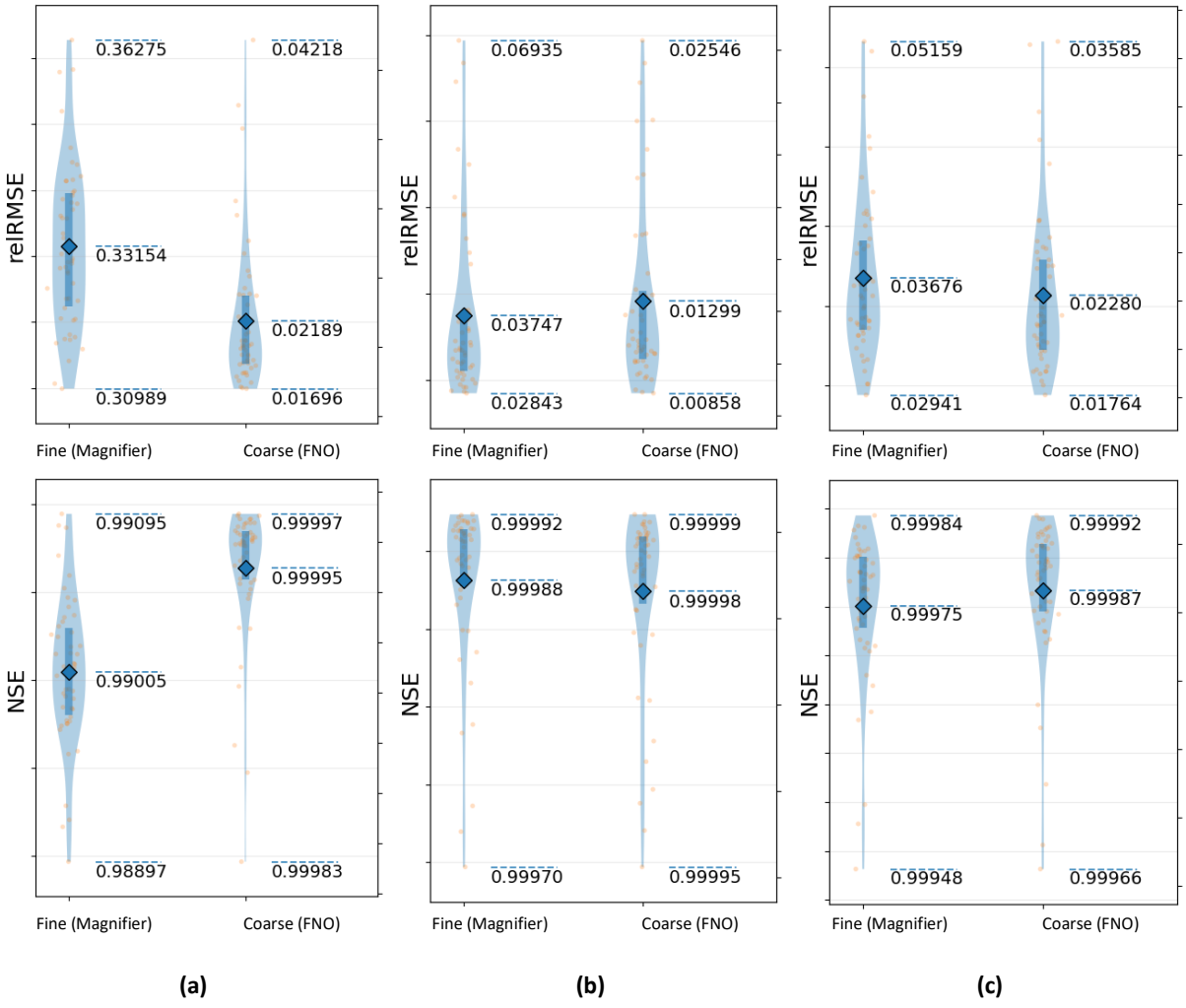


Figure 7. Distribution of event-wise depth-prediction performance across the independent test scenarios. Violin plots show the variability of relative root-mean-square error (relRMSE; top row) and Nash–Sutcliffe efficiency (NSE; bottom row) for the three study domains: (a) Neuse River, (b) Fall River Lake, and (c) Chowilla floodplain. For each violin, the top, middle, and bottom annotated values respectively indicate the maximum, mean, and minimum. The upper and lower limits of the inner rectangle respectively indicate the 75th and 25th percentiles. The diamond denotes the mean, and the yellow dots represent the individual test-scenario values.

terpolation cannot recover subgrid hydraulic structure or terrain-controlled flow dynamics. This limitation became particularly evident near inundation boundaries and shallow-flow regions, where wet–dry transitions are governed by fine-scale topographic thresholds and small elevation differences strongly influence inundation extent. In such regions, purely geometric upscaling amplifies relative errors and produces boundary smearing and increased scatter dispersion, reflecting structural rather than random

error. These observations demonstrate that geometric upscaling alone is insufficient for hydraulically-faithful, high-resolution depth reconstruction, underscoring the necessity of terrain-aware residual refinement.

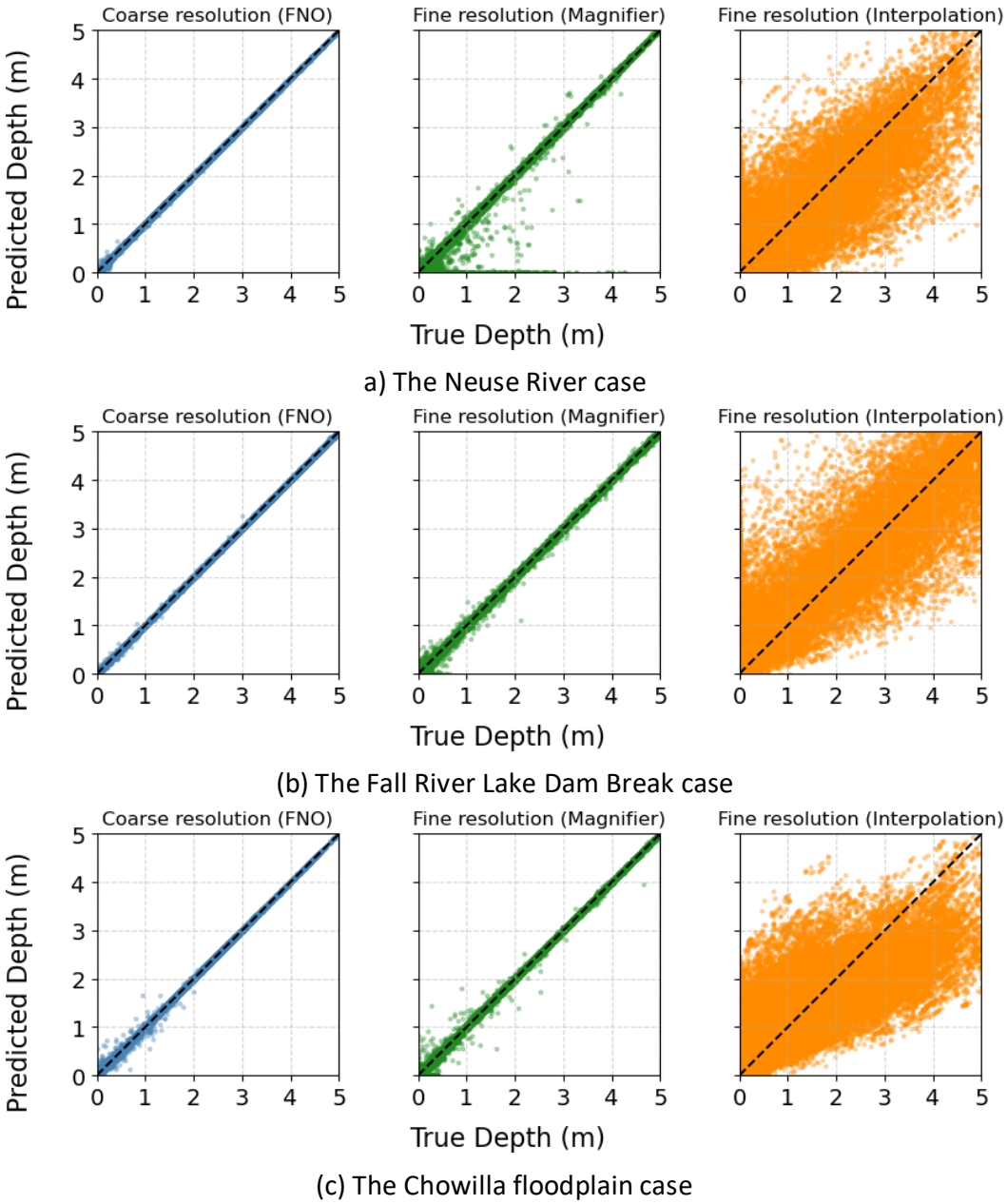


Figure 8. True versus predicted water depth for the three study areas at coarse resolution (FNO) and fine resolution via both the Magnifier operator and simple geometric bilinear interpolation. The dashed line indicates the 1:1 reference.

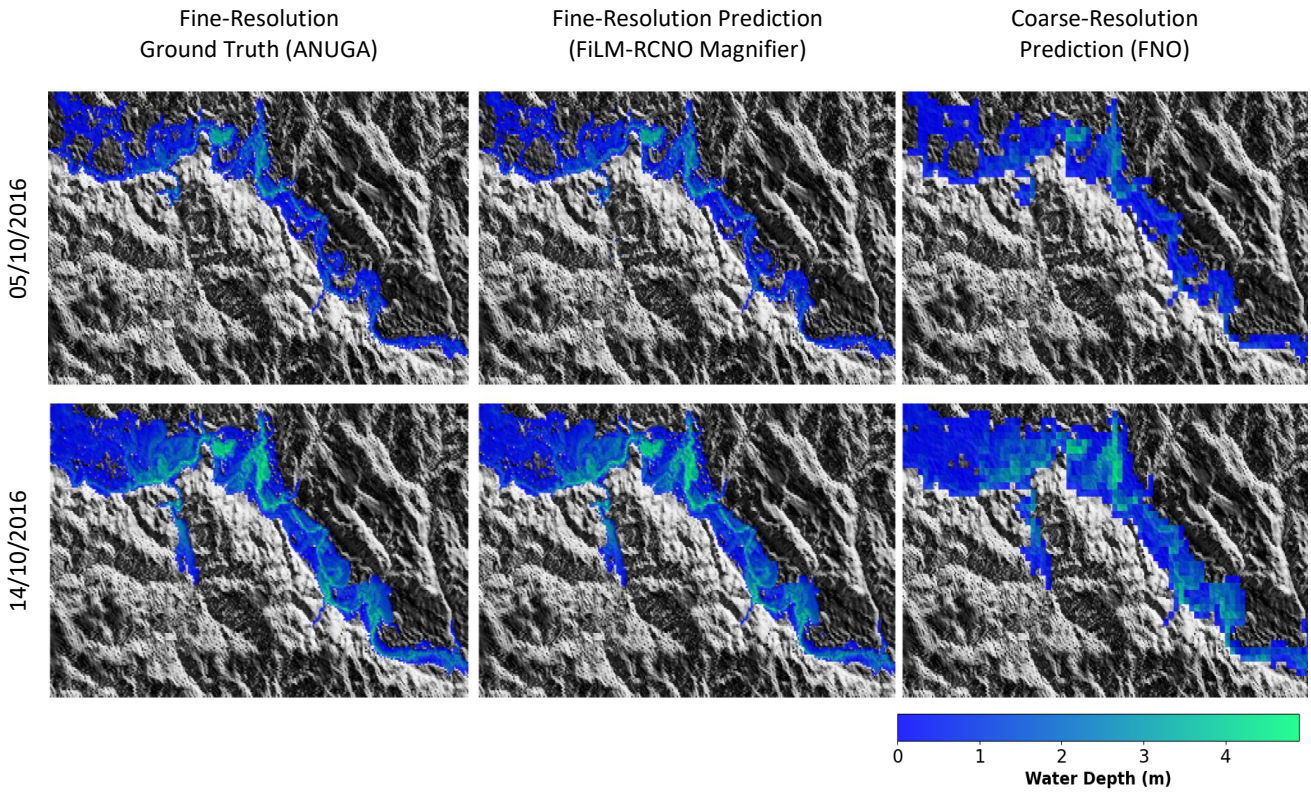


Figure 9. Multi-resolution flood depth comparison for the Neuse River during the reference Hurricane Matthew event (5–16 October 2016).

In fully developed flow regions ($h \geq 0.5\,\mathrm{m}$), both coarse and fine predictions achieved very low error and efficiency values approaching unity, whereas relative errors increased for all methods as expected in shallow-flow regions ($h < 0.5\,\mathrm{m}$). This confirms that wet–dry transitions and subgrid topographic variability dominate error behavior. Nevertheless, the Magnifier consistently outperformed interpolation in this regime, maintaining NSE values that were stable and > 0.96 while interpolation frequently exhibited degraded or even negative efficiency. The amplification of interpolation error in shallow zones arises from low mean depths and the absence of topographic conditioning, leading to unstable depth reconstruction near inundation boundaries.

These trends were confirmed by the true-versus-predicted scatter plots (Figure 8), which show that the Magnifier preserved near one-to-one depth scaling while reducing dispersion relative to interpolation, particularly in shallow-flow regimes. Coarse predictions aligned closely with the 1:1 reference line with minimal dispersion, confirming accurate large-scale depth scaling. Fine-resolution predictions retained this alignment with modest spread concentrated near low-depth values, consistent with shallow-regime sensitivity. In contrast, interpolation displayed pronounced heteroscedasticity and systematic deviation from the reference line, particularly at intermediate depths, indicating structural reconstruction error rather than simple smoothing bias. Spatial comparisons in Figures 9–10 corroborate the statistical findings. The coarse operator captures global inundation

extent and wave arrival patterns but exhibits boundary staircasing and smoothed channel structure due to grid coarsening. The Magnifier refines these boundaries, restores channelized flow geometry, and aligns inundation edges with fine-scale topography while preserving volumetric coherence. Importantly, no artificial depth amplification or spurious isolated flooding is observed, indicating that residual-based refinement redistributes water in a terrain-consistent manner rather than introducing nonphysical corrections.

Collectively, these results confirm that the multi-resolution architecture achieves robust depth reconstruction. The coarse operator accurately emulates basin-scale wave dynamics, while the Magnifier provides necessary fine-scale refinement beyond geometric interpolation. The pronounced degradation observed in interpolation-based reconstruction—particularly in shallow-flow regimes—demonstrates that residual-informed operator refinement is essential for hydraulically-reliable, high-resolution flood mapping.

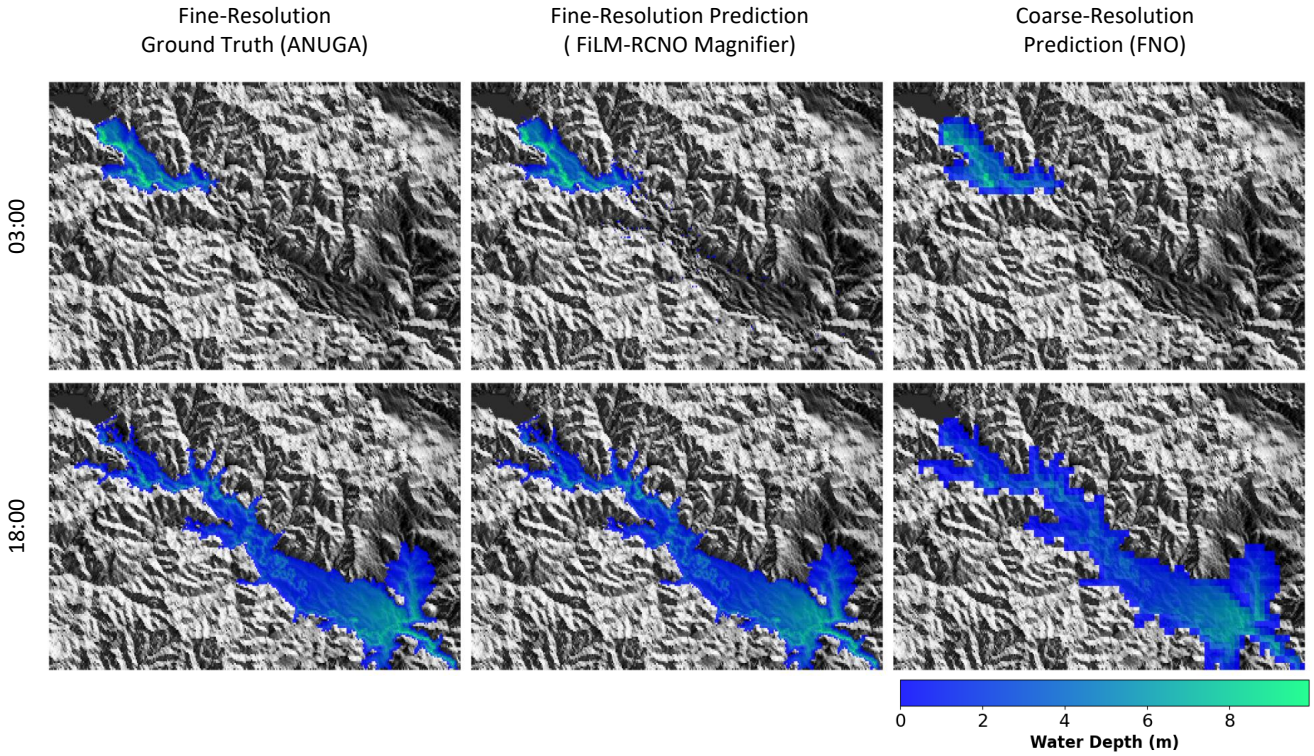


Figure 10. Multi-resolution flood depth comparison for the reference dam-break overtopping scenario at selected times during the 24-hour simulation.

4.2 Flood Extent Delineation and Boundary Consistency

Building on the continuous depth reconstruction results presented in Section 4.1, we now evaluate how accurately the predicted water depth fields translate into reliable wet–dry delineation, which could be a crucial prediction influencing flood evacuation

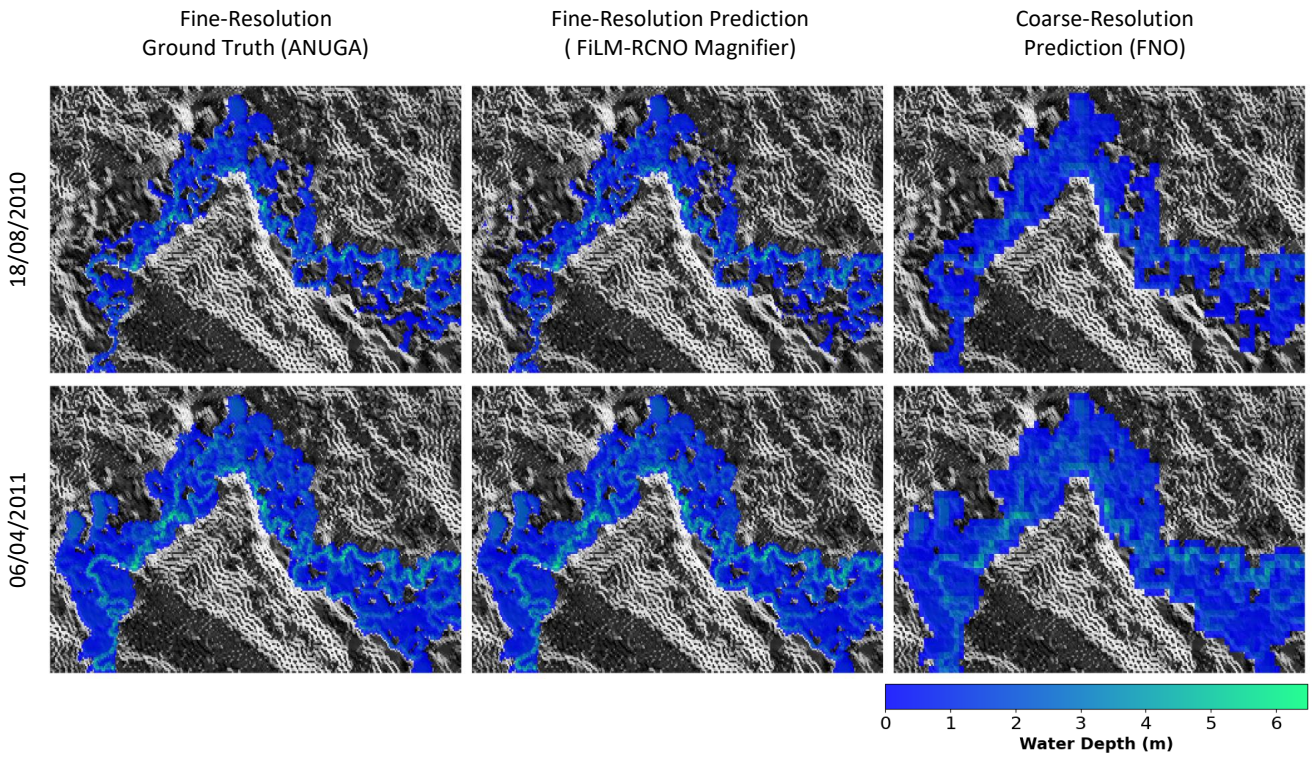


Figure 11. Multi-resolution flood depth comparison for the Chowilla floodplain under reference regulated Murray–Darling Basin inflow conditions (5 August 2010–1 June 2011).

recommendations. While continuous metrics quantify volumetric and amplitude agreement, binary inundation performance directly assesses the framework's ability to reproduce flood boundaries and spatial extent.

Using a binary inundation threshold of $h > 0.01\,\mathrm{m}$, the model retained strong delineation skill across Probability of Detection (POD, the fraction of flooded cells captured, or "recall"), False Alarm Rates (FAR), and Critical Success Index (CSI, overall skill; Section 3.5.3) (Table 3). At coarse resolution, the FNO consistently achieved high POD and CSI with low FAR, indicating that the dominant flood envelope and large-scale inundation connectivity were captured reliably but the predicted extent was too large. Basin-scale activation of floodplain storage and primary flow pathways were well represented. However, due to grid coarsening, inundation boundaries were approximated with staircase artifacts and smoothed transitions, limiting spatial precision near terrain-controlled margins where small depth variations determine wet–dry status.

To complement the aggregated binary inundation metrics (Table 3), we plot the event-wise distribution of the Critical Success Index (CSI) across the 50 independent test scenarios (Figure 12). This distributional view helps assess whether the reported inundation skill is consistently maintained across unseen scenarios rather than being controlled by a small number of favorable cases.

Table 3. Aggregated binary inundation performance across the independent test scenarios. Probability of Detection (POD, the fraction of flooded cells captured), False Alarm Rate (FAR), and Critical Success Index (CSI, overall skill) are reported for coarse (FNO) and fine (Magnifier) resolutions, along with a bicubic interpolation baseline at fine scale. Binary classification was performed using a depth threshold of $h > 0.01\,\mathrm{m}$.

Case	Scale	Binary Inundation Metrics		
		POD (%)	FAR (%)	CSI (%)
(i) Neuse River	Coarse (FNO)	99.71	0.80	98.92
	Fine (Magnifier)	94.88	2.06	93.03
	Fine (Interpolation)	96.45	23.05	74.82
(ii) Fall River Dam Break	Coarse (FNO)	98.82	0.40	98.43
	Fine (Magnifier)	98.37	2.01	96.42
	Fine (Interpolation)	99.28	24.38	75.21
(iii) Chowilla floodplain	Coarse (FNO)	99.74	0.47	99.28
	Fine (Magnifier)	98.84	0.91	97.95
	Fine (Interpolation)	99.21	20.08	79.41

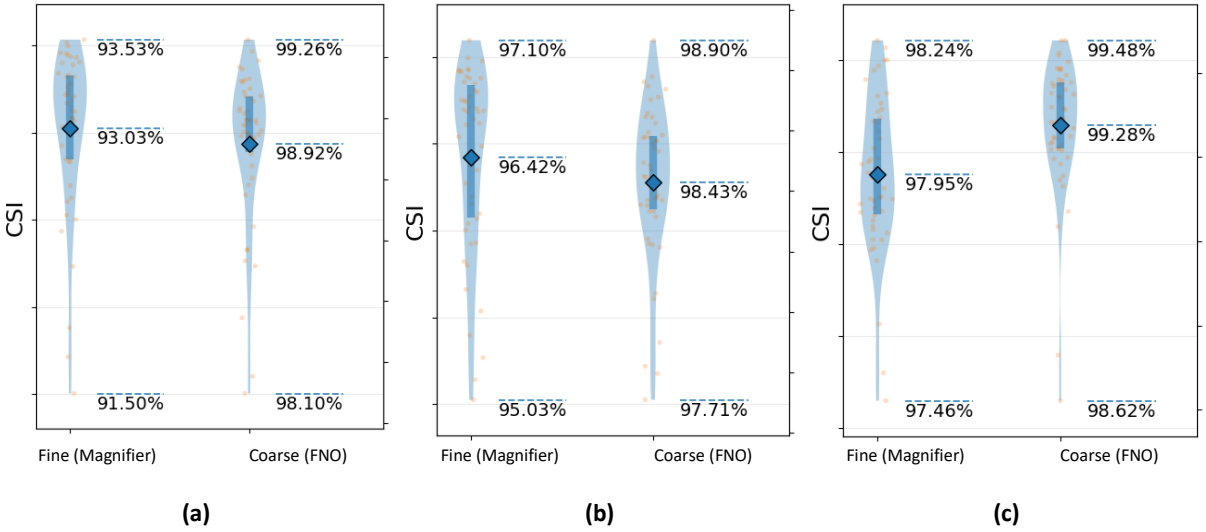


Figure 12. Distribution of event-wise binary inundation performance across the independent test scenarios. Violin plots show the variability of the Critical Success Index (CSI) for the three study domains: (a) Neuse River, (b) Fall River Lake, and (c) Chowilla floodplain. For each violin, the top, middle, and bottom annotated values respectively indicate the maximum, mean, and minimum. The upper and lower limits of the inner rectangle respectively indicate the 75th and 25th percentiles. The diamond denotes the mean, and the yellow dots represent the individual test-scenario values.

To evaluate sensitivity to the wet–dry threshold, CSI was recomputed using $h > 0.01$, 0.05, 0.10, and $0.20\,\mathrm{m}$ (Figure 13).
490 This tests whether coarse FNO and fine Magnifier inundation skill remains stable under stricter flood definitions. Figure 13

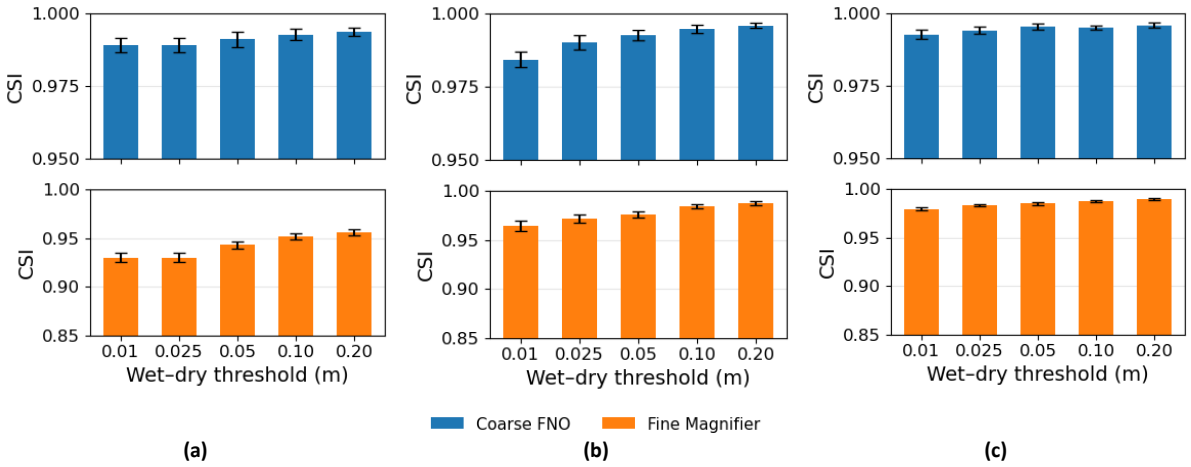


Figure 13. Critical Success Index (CSI) sensitivity to wet–dry threshold across the independent test scenarios. CSI was recomputed using thresholds of $h > 0.01$, 0.05, 0.10, and $0.20\,\mathrm{m}$ for (a) Neuse River, (b) Fall River Lake, and (c) Chowilla floodplain. Bars indicate the mean CSI across the independent test scenarios, and black error bars indicate $\pm$ one standard deviation.

shows that CSI remains generally stable across the tested thresholds, indicating that the binary inundation performance is not solely dependent on the default $0.01\,\mathrm{m}$ threshold.

Fine-resolution reconstruction addresses this limitation and improves boundary realism. The Magnifier maintains comparable detection skill while refining boundary placement and reducing discretization artifacts. Because shallow-flow regions are particularly sensitive to small topographic gradients, terrain-aware refinement of depth near the inundation edge directly improves wet–dry delineation. Although the coarse solution identifies which regions are broadly inundated, fine-scale prediction determines where the flood boundary intersects terrain features such as levees, road embankments, and localized depressions. These distinctions are critical for exposure assessment and operational flood mapping. The consistently strong CSI values confirm that boundary refinement was achieved without sacrificing overall detection accuracy.

In contrast, bilinear interpolation exhibited persistently elevated FAR despite often achieving high POD, reflecting systematic overprediction of inundated cells along flood margins. Because interpolation lacks dynamic conditioning, it spreads water linearly between coarse cells without respecting terrain-controlled activation thresholds; as a result, inundation artificially expands into shallow depressions and along gentle slopes. These structural errors are also visible in Figure 15, where interpolation generated widespread false alarms while the Magnifier produced predominantly correct wet and dry classifications and concentrated residual errors near complex terrain transitions.

Flooded-area time series show that both coarse- and fine-resolution predictions tracked overall growth and recession dynamics (Figure 14), confirming consistency with the reconstructed basin-scale depth evolution. However, fine-resolution predictions aligned more closely with the reference during peak and transitional phases, when boundary placement is most sensitive to local topography and shallow-depth gradients. Detection metrics over time confirm that the Magnifier maintained stable

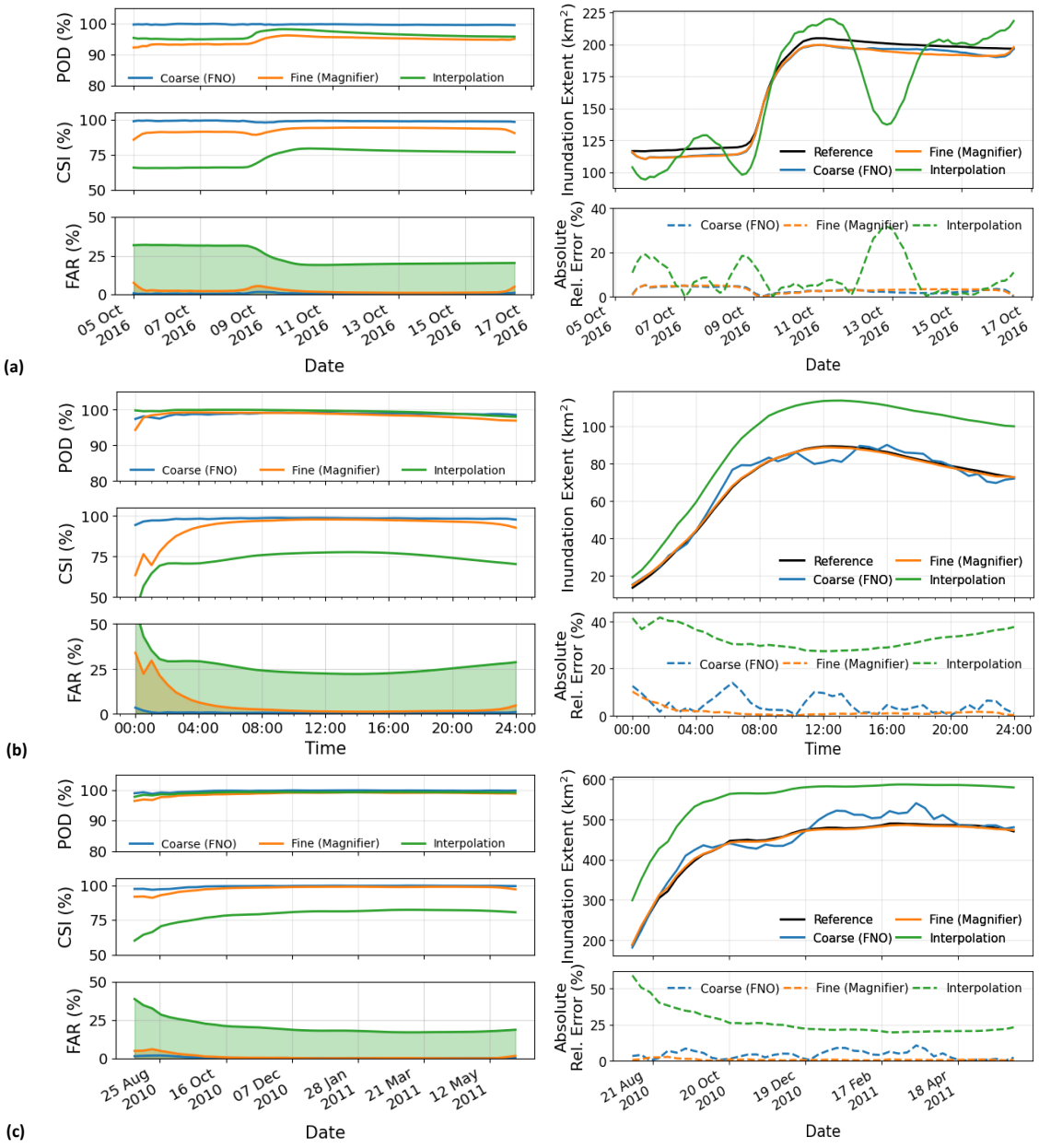
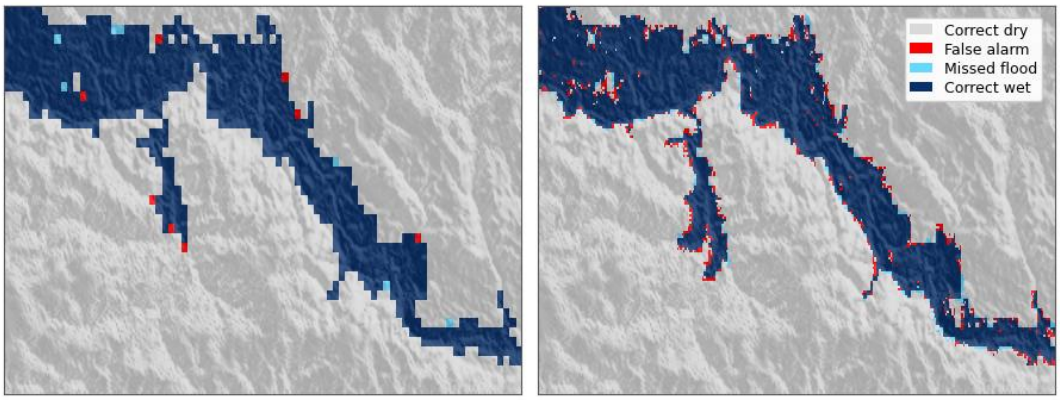
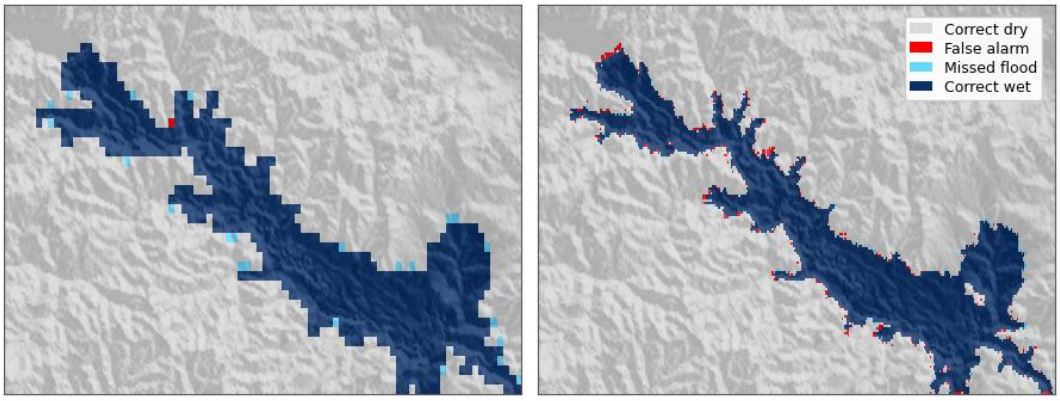


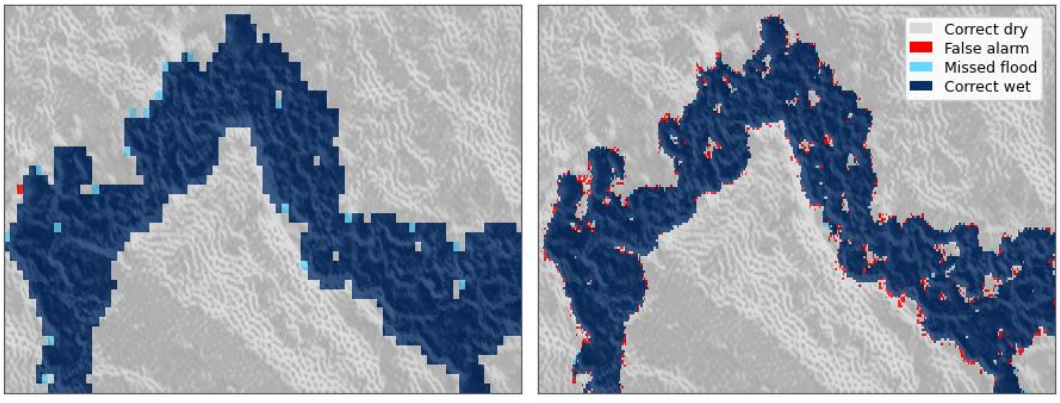
Figure 14. Time-resolved evaluation across the three case studies. Left panels show binary inundation metrics on standard upward axes: Probability of Detection (POD), Critical Success Index (CSI), and False Alarm Rate (FAR). Right panels present inundation extent dynamics, including reference and predicted inundation extent in the upper panels and absolute relative error in the lower panels. Panels correspond to (a) the Neuse River case, (b) the Fall River Dam Break case, and (c) the Chowilla floodplain case, respectively.



a) The Neuse River case



(b) The Fall River Lake Dam Break case



(c) The Chowilla floodplain case

Figure 15. Spatial contingency maps aggregated over the full event duration, comparing coarse-resolution FNO predictions (left column) and fine-resolution Magnifier predictions (right column). Binary classification was performed using a depth threshold of $h > 0.01\,\mathrm{m}$.

POD and low FAR throughout the event, whereas interpolation exhibited consistently higher false alarm rates. Spatially, the Magnifier produceed predominantly correct wet and correct dry classifications, with errors concentrated near complex terrain transitions where small depth differences control inundation (Figure 15). The coarse prediction retained connectivity but displayed boundary staircasing due to resolution constraints. Interpolation, by contrast, generated widespread false alarms along inundation margins and shallow terrain features, consistent with its inability to resolve terrain-governed wet–dry thresholds.

Taken together, these results demonstrate that accurate basin-scale depth reconstruction is necessary but not sufficient for operational flood mapping. Fine-resolution, terrain-aware refinement is required to translate dynamically-plausible depth fields into boundary-consistent inundation extent. The Magnifier achieves this by preserving hydrodynamic coherence while resolving subgrid topographic controls, whereas purely geometric interpolation lacks the dynamic structure required for topography-consistent flood delineation.

4.3 Computational efficiency and marginal inference cost

A central motivation of the proposed framework is to reduce the computational cost of repeated flood simulations after the surrogate has been trained. Table 4 summarizes the approximate runtime required for one complete event simulation using ANUGA and the trained surrogate.

Table 4. Runtime comparison between ANUGA and the trained surrogate for one complete event simulation.

Case study	ANUGA runtime	Surrogate runtime	Approx. speedup
Neuse River	$\sim$2.5 h	$\sim$1 s	$\sim$9,000$\times$
Fall River dam-break	$\sim$3 h	$\sim$1 s	$\sim$10,800$\times$
Chowilla floodplain	$\sim$10 h	$\sim$1 s	$\sim$36,000$\times$

While the high-fidelity solver remains necessary for generating training data and reference simulations, the trained surrogate reduces the marginal cost of evaluating additional scenarios to approximately one forward pass. This distinction is important: the reported speedup does not imply that hydrodynamic solvers are unnecessary, but rather that repeated ANUGA-like scenario evaluations can be performed much more efficiently once the surrogate has learned the solver-generated input–output relationship. This reduction in marginal inference cost is especially relevant for large-ensemble simulation, uncertainty analysis, sensitivity testing, and rapid scenario screening, where many repeated high-resolution simulations would otherwise be required.

4.4 Peak Magnitude and Timing Accuracy

To evaluate large-scale volume dynamics, we examined the domain-integrated water-volume time series $H(t)$ defined in Equation 9. This diagnostic assesses whether the model captures the temporal evolution of total stored water volume across the floodplain.

Table 5. Peak flood magnitude and timing performance across all events. Relative peak value error, and relative peak timing error normalized by (1) peak duration and (2) characteristic rising-limb duration, are reported as means for coarse (FNO) and fine (Magnifier) resolutions. Peak statistics were computed over the highest 5% of domain-integrated volume values.

Case	Scale	Peak Metrics (mean)		
		relPeakValErr	relPeakTimeErr-1	relPeakTimeErr-2
(i) Neuse River	Coarse (FNO)	-2.30×10^{-4}	$+1.33 \times 10^{-2}$	$+2.58 \times 10^{-3}$
	Fine (Magnifier)	-5.46×10^{-3}	$+1.33 \times 10^{-2}$	$+2.58 \times 10^{-3}$
(ii) Fall River Dam Break	Coarse (FNO)	$+2.20 \times 10^{-4}$	-4.00×10^{-2}	-1.86×10^{-3}
	Fine (Magnifier)	-3.11×10^{-4}	-2.00×10^{-2}	-8.89×10^{-4}
(iii) Chowilla floodplain	Coarse (FNO)	-5.35×10^{-4}	$+2.00 \times 10^{-1}$	$+7.38 \times 10^{-3}$
	Fine (Magnifier)	-2.63×10^{-3}	$+2.20 \times 10^{-1}$	$+8.01 \times 10^{-3}$

Table 5 summarizes the mean signed peak-magnitude and peak-timing errors for the coarse FNO and fine Magnifier predictions. These signed values are interpreted directionally: negative peak-magnitude errors indicate underestimation of the peak response, while positive values indicate overestimation. Similarly, negative timing errors indicate that the predicted peak occurred earlier than the reference, whereas positive timing errors indicate a delayed predicted peak. Therefore, the values in Table 5 should be interpreted as mean directional bias rather than absolute event-level error magnitude. At coarse resolution, the FNO produced consistently small signed peak-magnitude errors across all case studies, demonstrating its ability to accurately capture the dominant flood-volume response even without explicitly enforcing strict mass conservation. Peak-timing performance was also generally strong, indicating that the learned operator effectively captured the dominant wave-propagation behavior. Fine-resolution reconstruction with the Magnifier preserved this strong performance, maintaining small peak-magnitude errors and comparable peak-timing accuracy across all three cases. In particular, the mean timing errors normalized by the characteristic rising-limb duration remained below 1% in magnitude at both resolutions. The close agreement between coarse and fine peak responses demonstrates that the Magnifier effectively preserves the dominant temporal and volumetric characteristics of the flood dynamics while refining local depth distributions and spatial inundation patterns over complex terrain.

The domain-integrated water-volume time series further confirms this behavior (Figure 16). Across all study areas, both coarse and fine predictions closely followed the reference response during the rising, peak, and recession phases. Relative errors remained small and oscillatory for the coarse FNO rather than systematically biased. The fine-resolution prediction produced a smoother volume response while preserving the same temporal dynamics. No artificial amplification or attenuation of the peak stored volume was observed after fine-scale refinement, indicating that the residual correction preserves volumetric consistency.

Together, these findings show that the multi-resolution framework achieves consistent reconstruction not only in spatial structure but also in domain-integrated volume dynamics. The coarse operator captures basin-scale flood wave evolution, while the Magnifier refines local structure without compromising peak magnitude or timing fidelity. This consistency across spatial,

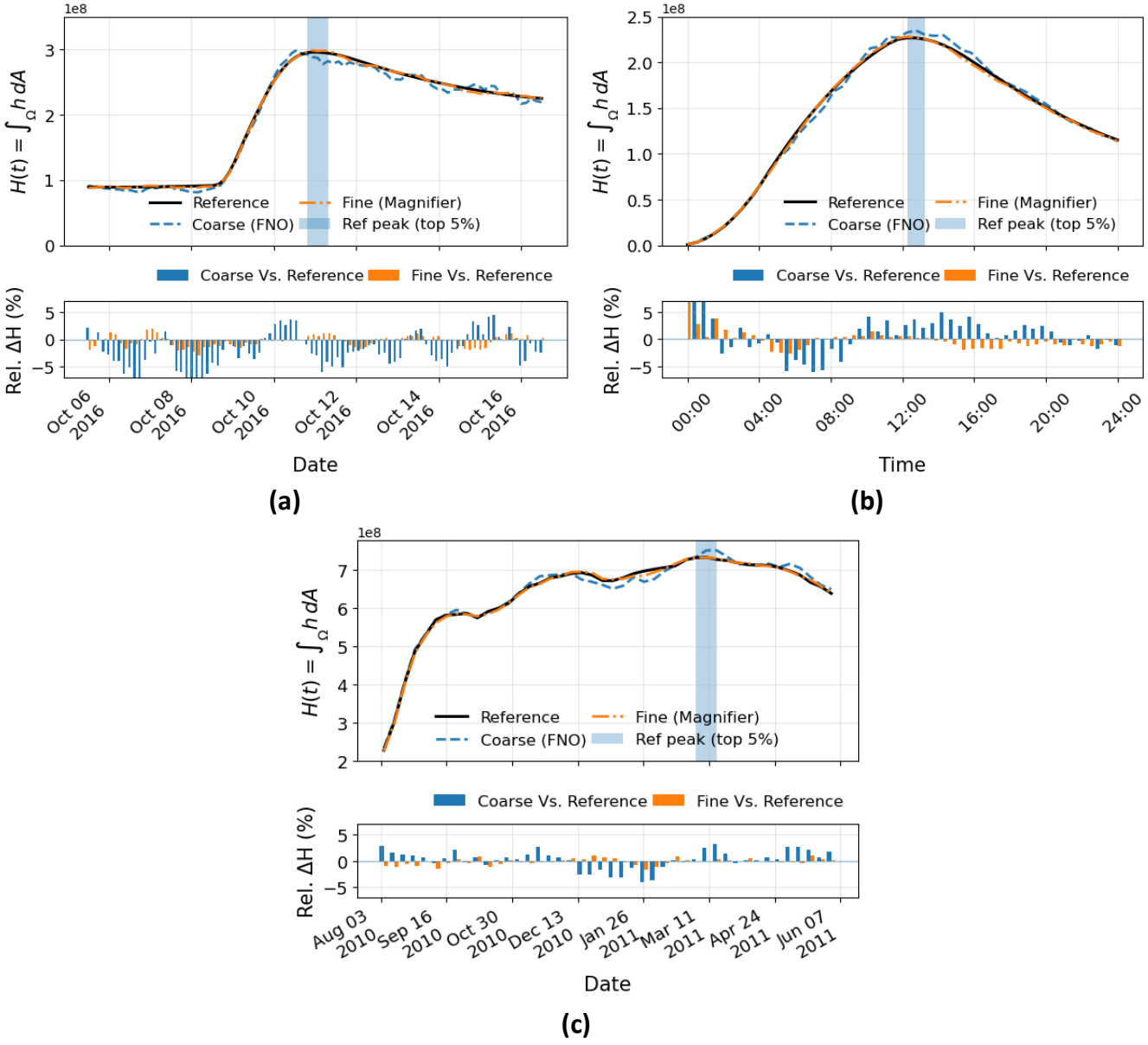


Figure 16. Domain-integrated water-volume time series showing the temporal evolution of total floodwater volume $H(t)$ for (a) Neuse River, (b) Fall River Dam Break, and (c) Chowilla floodplain.

boundary, and integrated metrics supports the interpretation that the framework preserves geometric realism and volume-consistent behavior.

4.5 Robustness Under Structured Topographic Uncertainty

We next evaluated whether the surrogate remained decision-relevant under bed topography perturbations and could represent modest changes in bathymetry, rather than using a fixed geometry. The structured topographic perturbations in magnitude, realization generation, and train/test partition (Section 3.5) led to significant differences in responses. Our FNO takes topography as an input and learns to resolve its impact after training on many topographies. In this section, we focus on skill and calibration under that uncertainty setting, with three guiding questions: (i) Does the learned operator retain accuracy under held-out, realistically perturbed bathymetries? (ii) Can it approximate solver responses across a distribution of plausible terrain realizations rather than a single deterministic bed configuration? And (iii) Does it preserve the spatial structure of hydraulic sensitivity induced by terrain perturbations, particularly near wet–dry transitions governed by fine-scale topographic thresholds?

Table 6. Ensemble-averaged continuous depth performance under structured bathymetric uncertainty for the dam-break case. relRMSE is relative root-mean-square error and NSE is Nash-Sutcliffe Efficiency.

Model	Structured Uncertainty Ensemble		Clean Bed (Baseline)	
	relRMSE ($\mu \pm \sigma$)	NSE ($\mu \pm \sigma$)	relRMSE	NSE
Coarse (FNO)	0.0197	0.9980	0.0130	0.9999
Fine (Magnifier)	0.0348	0.9984	0.0375	0.9999

The framework remained highly accurate under moderate bathymetric perturbations (Table 6). Relative error increased versus the deterministic baseline (e.g., coarse relRMSE from 0.0131 to 0.0271), but NSE remained near unity (≈ 0.997–0.999). Fine-resolution results remained comparable, showing that Magnifier refinement does not amplify sensitivity to geometric noise. This indicates that the surrogate remains accurate for the tested perturbation range and that the FNO can represent the influence of bed topography as an input within this structured uncertainty setting. The bathtub assumption and dynamic correction in the Magnifier serve to stabilize the sensitivity to the topography.

Probabilistic calibration remained strong under topographic uncertainty (Figure 17). The reliability diagrams compare predicted inundation probabilities with observed solver frequencies aggregated over all spatial locations and timesteps. Deviations from the 1:1 line remained relatively small across probability bins, and the area-weighted Expected Calibration Error (ECE) increased from approximately 0.09% for the coarse operator to 0.67% for the Magnifier. Despite this modest increase, the Magnifier maintained good overall probabilistic calibration at fine resolution. Both models demonstrated good overall probabilistic calibration, with relatively small deviations from perfect calibration across the probability range, particularly at higher predicted probabilities.

Collectively, Table 6 and Figure 17 show that the multi-resolution framework maintains stable agreement with hydrodynamic-solver-generated depth and extent fields under the tested structured topographic perturbations, rather than overfitting to a single deterministic geometry. We must note here that the FNO is only tested on small perturbations (not in training) around the

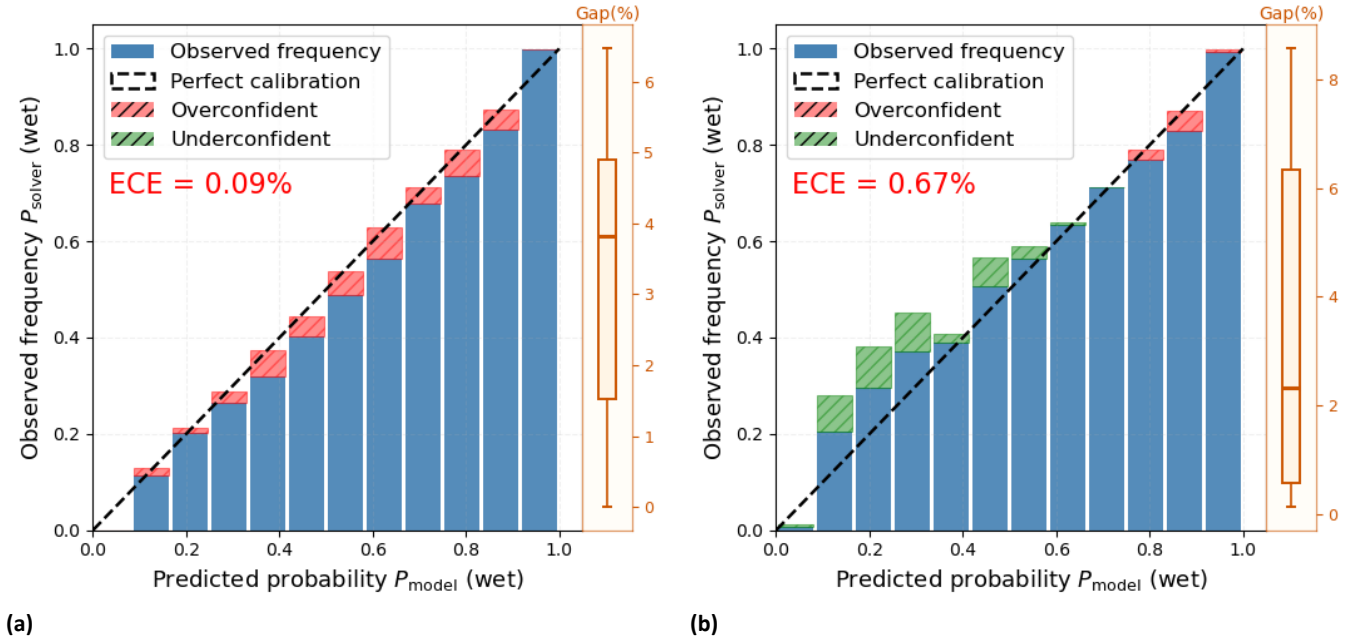


Figure 17. Reliability diagrams for probabilistic inundation prediction under structured bathymetric uncertainty in the dam-break case. (a) Coarse-resolution neural operator (FNO); (b) Fine-resolution Magnifier prediction. Blue bars denote the observed inundation frequency from the reference solver conditioned on predicted probability bins, and the dashed line indicates perfect calibration. Red and green shaded regions represent overconfidence and underconfidence, respectively. The orange box plot summarizes the distribution of the absolute calibration gap across probability bins; the center line denotes the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. ECE denotes the Expected Calibration Error.

baseline bathymetry. Larger scale changes, i.e., fundamentally different bed topography, may still lead to large errors with the surrogate model. This is currently a fundamental challenge due to the high-dimensional nature of topography as an input and its highly nonlinear impact. Hence the use case here is limited to uncertainty estimation and limited scenario analysis, not arbitrary simulations. As discussed earlier, very few surrogate models for the shallow water equation can in fact correctly handle different topography, and thus our work here still represents a rare advance. We envision this problem is ultimately addressable with training on much more extensive datasets (possibly using a foundation model) and with more uncertainty suppression techniques, so that cheap surrogate models can eventually remain accurate under wide changes in topography.

5 Conclusions

This study introduced a multi-resolution deep learning framework for rapid shallow water equation surrogate emulation of hydrodynamic-solver-generated flood depth and extent fields. The framework is designed as a fast, physically anchored surrogate for repeated scenario simulation and uncertainty-aware analysis, rather than as a replacement for high-fidelity hydrody-

namic solvers. By coupling a coarse-scale neural operator with a terrain-aware refinement mechanism, the approach captures basin-scale flood-wave evolution while resolving fine-scale inundation boundaries after training on solver-generated reference simulations.

Comprehensive evaluation across riverine, dam-break, and complex floodplain systems demonstrates accurate reconstruction of continuous water-depth fields, wet–dry delineation, and peak flood magnitude and timing. Across all evaluation dimensions—continuous depth accuracy, boundary consistency, domain-integrated water-volume behavior, and probabilistic consistency under structured bathymetric uncertainty—the framework closely reproduces hydrodynamic-solver-generated depth and extent responses across spatial scales. The agreement in domain-integrated water volume provides an additional diagnostic of volume-consistent behavior, although it should not be interpreted as full independent verification of shallow-water-equation dynamics because the present model does not directly predict velocity, discharge, or momentum.

The results show that, once trained, the surrogate substantially reduces the marginal cost of repeated simulations, making it useful for rapid scenario evaluation, large-ensemble analysis, sensitivity testing, and uncertainty-aware flood hazard assessment. High-fidelity solvers such as ANUGA, HEC-RAS, Delft3D, FLOW-3D, and related hydrodynamic models remain essential for detailed hydraulic modeling, calibration, validation, and training-data generation. The independent test set used in this study evaluates unseen forcing realizations within the sampled perturbation envelope; validation on independent historical flood events and deliberately withheld forcing ranges remains important future work. Future extensions may also include additional physical variables, such as velocity and discharge, broader uncertainty quantification, and data assimilation.

Code and data availability. The source code implementing the multi-resolution flood modeling framework presented in this study is openly available through GitHub at https://github.com/AMBehroozi/Flood_Neural_operator. The datasets and pretrained neural network models used in this study, including the Neuse River flooding, Fall River Lake dam-break, and Chowilla floodplain case studies, are publicly available through Zenodo at https://doi.org/10.5281/zenodo.23990987 (Behroozi and Shen, 2026).

Author contributions. A.M. Behroozi: conceived the study; developed the multi-resolution flood modeling framework; performed simulations; conducted data analysis; prepared figures; managed code and datasets; led manuscript writing.

C. Shen: supervised the research; contributed to study design and conceptual development; provided guidance on methodology, analysis, and interpretation; reviewed and revised the manuscript.

K. Lawson: reviewed and revised the manuscript.

All authors: reviewed and approved the final manuscript.

Competing interests. C. Shen and K. Lawson have financial interests in HydroSapient, Inc., a company that could potentially benefit from the results of this research. These interests have been reviewed and managed by The Pennsylvania State University in accordance with its conflict of interest policies to ensure the objectivity and integrity of the research. A.M. Behroozi declares no conflicts of interest for this manuscript.

Disclaimer. AI tools were used for grammatical corrections and refinement of the manuscript.

Acknowledgements. This work was primarily supported by subaward A23-0249-S001 from the Cooperative Institute for Research to Operations in Hydrology (CIROH) under the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement (Grant No. NA22NWS4320003). The statements, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NOAA. This work was also partially supported by the U.S. Department of Energy, Office of Science, under Award No. DE-SC0021979.

References

Alipour, A., Jafarzadegan, K., and Moradkhani, H.: Global sensitivity analysis in hydrodynamic modeling and flood inundation mapping, Environmental Modelling & Software, 152, 105 398, 2022.

Arcement, G. J. and Schneider, V. R.: Guide for selecting Manning's roughness coefficients for natural channels and flood plains, Tech. rep., USGPO; For sale by the Books and Open-File Reports Section, US Geological . . . , 1989.

Balachandran, S., Nikrou, P., Nemnem, A. M., Alipour, R. S., Cohen, S., Li, X., Goharian, E., Imran, J., and Burian, S. J.: Rapid flood inundation mapping for dam failure and operations, Journal of Hydrology, p. 134759, 2025.

Baruah, A., Dhital, S., Cohen, S., Tran, T. N. D., Elhaddad, H., Watts, C. L., Devi, D., Chen, Y., and Pruitt, C.: FIMserv v. 1.0: A tool for streamlining Flood Inundation Mapping (FIM) using the United States operational hydrological forecasting framework, Environmental Modelling & Software, p. 106581, 2025.

Bass, B. and Bedient, P.: Surrogate modeling of joint flood risk across coastal watersheds, Journal of Hydrology, 558, 159–173, 2018.

Bates, P. D.: Flood inundation prediction, Annual Review of Fluid Mechanics, 54, 287–315, 2022.

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

Behroozi, A. and Shen, C.: Datasets and Trained Models for Physically Anchored Multi-Resolution Neural Operator Framework for Flood Inundation Prediction, https://doi.org/10.5281/zenodo.23990987, 2026.

Behroozi, A. M., Meier, C. I., and Vaghefi, M.: Radial basis function-based differential quadrature for dam break problems, Engineering Analysis with Boundary Elements, 155, 307–322, 2023.

Behroozi, A., Shen, C., and Kifer, D.: Sensitivity-constrained Fourier neural operators for forward and inverse problems in parametric differential equations, in: International Conference on Learning Representations, vol. 2025, pp. 12 916–12 937, 2025.

Bentivoglio, R., Isufi, E., Jonkman, S. N., and Taormina, R.: Deep learning methods for flood mapping: a review of existing applications and future research directions, Hydrology and Earth System Sciences Discussions, 2022, 1–50, 2022.

Bubeck, P., Otto, A., and Weichselgartner, J.: Societal impacts of flood hazards, in: Oxford Research Encyclopedia of Natural Hazard Science, Oxford University Press, https://doi.org/10.1093/acrefore/9780199389407.013.281, 2017.

Buttinger-Kreuzhuber, A., Konev, A., Horváth, Z., Cornel, D., Schwerdorf, I., Blöschl, G., and Waser, J.: An integrated GPU-accelerated modeling framework for high-resolution simulations of rural and urban flash floods, Environmental Modelling & Software, 156, 105 480, 2022.

Cappato, A., Baker, E. A., Reali, A., Todeschini, S., and Manenti, S.: The role of modeling scheme and model input factors uncertainty in the analysis and mitigation of backwater induced urban flood-risk, Journal of Hydrology, 614, 128 545, 2022.

Cohen, S., Baruah, A., Nikrou, P., Tian, D., and Liu, H.: Toward robust evaluations of flood inundation predictions using remote sensing derived benchmark maps, Authorea Preprints, 2024.

Dewitz, J.: National land cover database (NLCD) 2019 products (ver. 3.0, february 2024), US Geological Survey (USGS) Data Release, p. 624, 2021.

Fraehr, N., Wang, Q. J., Wu, W., and Nathan, R.: Upskilling low-fidelity hydrodynamic models of flood inundation through spatial analysis and Gaussian process learning, Water Resources Research, 58, e2022WR032 248, 2022.

Fraehr, N., Wang, Q. J., Wu, W., and Nathan, R.: Development of a fast and accurate hybrid model for floodplain inundation simulations, Water Resources Research, 59, e2022WR033 836, 2023.

Fraehr, N., Wang, Q. J., Wu, W., and Nathan, R.: Assessment of surrogate models for flood inundation: The physics-guided LSG model vs. state-of-the-art machine learning models, Water Research, 252, 121 202, 2024.

George, R. J., Zhao, J., Kossaifi, J., Li, Z., and Anandkumar, A.: Incremental spatial and spectral learning of neural operators for solving large-scale PDEs, arXiv preprint arXiv:2211.15188, 2022.

Guo, K., Guan, M., and Yu, D.: Urban surface water flood modelling–a comprehensive review of current models and future challenges, Hydrology and Earth System Sciences Discussions, 2020, 1–27, 2020.

Guo, Z.-H. and Li, H.-B.: MgFNO: Multi-grid architecture Fourier neural operator for parametric partial differential equations, arXiv preprint arXiv:2407.08615, 2024.

He, J., Zhang, L., Xiao, T., Wang, H., and Luo, H.: Deep learning enables super-resolution hydrodynamic flooding process modeling under spatiotemporally varying rainstorms, Water Research, 239, 120 057, 2023.

Herath, H. M. V. V., Marshall, L., Saha, A., Rasnayaka, S., and Seneviratne, S.: Subgrid informed neural networks for high-resolution flood mapping, Journal of Hydrology, p. 133329, 2025.

Idowu, J. A. and Alfahid, A.: Open problems in uncertainty quantification for flood modelling: A systematic review, Environmental Modelling & Software, p. 106799, 2025.

Jakubauskas, M., deNoyelles, J., and Martinko, E. A.: Bathymetric and Sediment Survey of Fall River Reservoir, Greenwood County, Kansas, Tech. Rep. Report No. 2010-02, Applied Science and Technology for Reservoir Assessment (ASTRA) Program, Lawrence, Kansas, https://biosurvey.ku.edu/pub/bathymetric-and-sediment-survey-fall-river-reservoir-greenwood-county-kansas, 2011.

John, T. J. and Nagaraj, R.: Prediction of floods using improved PCA with one-dimensional convolutional neural network, International Journal of Intelligent Networks, 4, 122–129, 2023.

Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., and Pender, G.: A deep convolutional neural network model for rapid prediction of fluvial flood inundation, Journal of Hydrology, 590, 125 481, 2020.

Karim, F., Armin, M. A., Ahmedt-Aristizabal, D., Tychsen-Smith, L., and Petersson, L.: A review of hydrodynamic and machine learning approaches for flood inundation modeling, Water, 15, 566, 2023.

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Neural operator: Learning maps between function spaces with applications to pdes, Journal of Machine Learning Research, 24, 1–97, 2023.

Leijnse, T., van Ormondt, M., Nederhoff, K., and van Dongeren, A.: Modeling compound flooding in coastal systems using a computationally efficient reduced-physics solver: Including fluvial, pluvial, tidal, wind-and wave-driven processes, Coastal Engineering, 163, 103 796, 2021.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier neural operator for parametric partial differential equations, arXiv preprint arXiv:2010.08895, 2020.

Luo, P., Luo, M., Li, F., Qi, X., Huo, A., Wang, Z., He, B., Takara, K., Nover, D., and Wang, Y.: Urban flood numerical simulation: Research, methods and future perspectives, Environmental modelling & software, 156, 105 478, 2022.

Merwade, V., Olivera, F., Arabi, M., and Edleman, S.: Uncertainty in flood inundation mapping: Current issues and future directions, Journal of hydrologic engineering, 13, 608–620, 2008.

Ming, X., Liang, Q., Xia, X., Li, D., and Fowler, H. J.: Real-time flood forecasting based on a high-performance 2-D hydrodynamic model and numerical weather predictions, Water Resources Research, 56, e2019WR025 583, 2020.

Morales-Hernández, M., Sharif, M. B., Kalyanapu, A., Ghafoor, S. K., Dullo, T. T., Gangrade, S., Kao, S.-C., Norman, M. R., and Evans, K. J.: TRITON: A Multi-GPU open source 2D hydrodynamic flood model, Environmental Modelling & Software, 141, 105 034, 2021.

Mosavi, A., Ozturk, P., and Chau, K.-w.: Flood prediction using machine learning models: Literature review, Water, 10, 1536, 2018.

Musser, J., Watson, K., and Gotvald, A.: Characterization of peak streamflows and flood inundation at selected areas in North Carolina following Hurricane Matthew, October 2016 (ver. 2.0, August 2017), https://doi.org/10.3133/ofr20171047, 2017.

Neal, J., Fewtrell, T., and Trigg, M.: Parallelisation of storage cell flood models using OpenMP, Environmental Modelling & Software, 24, 872–877, 2009.

Noh, S. J., Lee, J.-H., Lee, S., Kawaike, K., and Seo, D.-J.: Hyper-resolution 1D-2D urban flood modelling using LiDAR data and hybrid parallelization, Environmental Modelling & Software, 103, 131–145, 2018.

Pappenberger, F., Beven, K. J., Ratto, M., and Matgen, P.: Multi-method global sensitivity analysis of flood inundation models, Advances in water resources, 31, 1–14, 2008.

Qiu, J., Cheng, L., Zhou, L., Yang, Y., Wu, K., Zhang, J., Tian, W., and Liu, P.: An explainable and transferable deep learning framework for spatiotemporal urban flood prediction by integrating Vision Transformer and U-Net, Water Research, p. 125504, 2026.

Rahman, A., Weinmann, P. E., Hoang, T. M. T., and Laurenson, E. M.: Monte Carlo simulation of flood frequency curves from rainfall, Journal of Hydrology, 256, 196–210, 2002.

Rak, A., Mewis, P., and Guthe, S.: Accelerating flash flood simulations: An efficient GPU implementation for a slim shallow water solver, Environmental Modelling & Software, 177, 106 030, 2024.

Rivera-Casillas, P., Dutta, S., Cai, S., Loveland, M., Nath, K., Shukla, K., Trahan, C., Lee, J., Farthing, M., and Dawson, C.: A Neural Operator-Based Emulator for Regional Shallow Water Dynamics, arXiv preprint arXiv:2502.14782, 2025.

Roberts, S., Nielsen, O., Gray, D., Sexton, J., and Davies, G.: ANUGA user manual, Commonwealth of Australia (Geoscience Australia) and the Australian National University, 127, 2015.

Sanders, B. F. and Schubert, J. E.: PRIMo: Parallel raster inundation model, Advances in Water Resources, 126, 79–95, 2019.

Sridharan, B., Bates, P. D., Sen, D., and Kuiry, S. N.: Local-inertial shallow water model on unstructured triangular grids, Advances in Water Resources, 152, 103 930, 2021.

Sun, A. Y., Jiang, P., Shuai, P., and Chen, X.: Bridging hydrological ensemble simulation and learning using deep neural operators, Water Resources Research, 60, e2024WR037 555, 2024.

Taghizadeh, M., Zandsalimi, Z., Nabian, M. A., Goodall, J. L., and Alemazkoor, N.: FloodForecaster: A domain-adaptive geometry-informed neural operator framework for rapid flood forecasting, Journal of Hydrology, p. 134512, 2025.

Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F., Dutta, D., and Kim, S.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, Environmental modelling & software, 90, 201–216, 2017.

Thomas Steven Savage, J., Pianosi, F., Bates, P., Freer, J., and Wagener, T.: Quantifying the importance of spatial resolution and other factors through global sensitivity analysis of a flood inundation model, Water Resources Research, 52, 9146–9163, 2016.

Tran, A., Mathews, A., Xie, L., and Ong, C. S.: Factorized fourier neural operators, arXiv preprint arXiv:2111.13802, 2021.

Turkoglu, M. O., Becker, A., Gündüz, H. A., Rezaei, M., Bischl, B., Daudt, R. C., D'Aronco, S., Wegner, J., and Schindler, K.: Film-ensemble: Probabilistic deep learning via feature-wise linear modulation, Advances in neural information processing systems, 35, 22 229–22 242, 2022.

Watson, K. M. and Musser, J. W.: Flood Inundation, Flood Depth, and High-Water Marks for Selected Areas in North Carolina from the October 2016 Flood, https://doi.org/10.5066/F75X276T, u.S. Geological Survey data release, published May 5, 2017, revised October 12, 2017, 2017.

Wing, O. E., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., and Erickson, T. A.: Validation of a 30 m resolution flood hazard model of the conterminous U nited S tates, Water Resources Research, 53, 7968–7986, 2017.

Xu, Q., De Vos, L. F., Shi, Y., Rüther, N., Bronstert, A., and Zhu, X. X.: Urban flood modeling and forecasting with deep neural operator and transfer learning, Journal of Hydrology, p. 133705, 2025.

Zhao, G., Yamazaki, D., Tanaka, Y., Zhou, X., Li, S., Hu, Y., Hirabayashi, Y., Neal, J., and Bates, P.: Developing a levee module for global flood modeling with a reach-level parameterization approach, Water Resources Research, 61, e2024WR039 790, 2025.

Appendix A: Neural Operators and Fourier Neural Operators

Neural operators provide a learning framework for approximating mappings between infinite-dimensional function spaces, extending classical neural networks beyond finite-dimensional vector-to-vector regression. This formulation is particularly well suited for physical systems governed by partial differential equations (PDEs), where both inputs and outputs are naturally represented as continuous fields defined over space and time.

Let $\mathcal{A}$ denote a space of admissible input functions (e.g., initial conditions, forcing terms, material properties), and let $\mathcal{U}$ denote the corresponding space of solution functions. The goal of operator learning is to approximate a nonlinear solution operator

$$\mathcal{G}: \mathcal{A} \to \mathcal{U}, \tag{A1}$$

such that for any input function $a \in \mathcal{A}$, the operator produces the corresponding solution $u = \mathcal{G}(a)$ (Kovachki et al., 2023). Unlike traditional surrogate models that are tied to a fixed discretization, neural operators aim to learn $\mathcal{G}$ in a discretization-invariant manner, enabling evaluation on grids of different resolutions without retraining.

A1 Integral Operator Formulation

Many neural operator architectures are motivated by the representation of nonlinear operators as integral transforms. A common formulation updates a latent representation v through layers of the form

$$(K(a)v)(x) = \int_D \kappa(x, y, a(x), a(y))\, v(y)\, dy, \tag{A2}$$

where D denotes the spatial domain and κ is a kernel encoding nonlocal interactions. In practice, the input function a is first lifted into a higher-dimensional latent space via a learnable operator P, producing $v_0 = P(a)$. The latent field is then evolved through a sequence of operator layers before being projected back to the solution space by a decoding operator.

This formulation captures long-range spatial dependencies and global coupling effects that are difficult to model using purely local convolutional architectures.

A2 Resolution Invariance and Practical Limitations

A key advantage of the FNO is that its learned operator acts on continuous function representations rather than discrete grid indices. As a result, once trained, the operator can be evaluated on spatial grids different from those used during training, provided the input functions are appropriately discretized.

Despite these advantages, the direct application of FNOs to very large, high-resolution domains presents practical challenges. The computational cost of FFT-based convolutions scales with the total number of grid points, and memory requirements grow rapidly with increasing spatial resolution (Tran et al., 2021; Guo and Li, 2024). In addition, the assumption of global periodicity implicit in standard FFT implementations can be restrictive for domains with complex boundaries or strongly heterogeneous conditions (George et al., 2022).

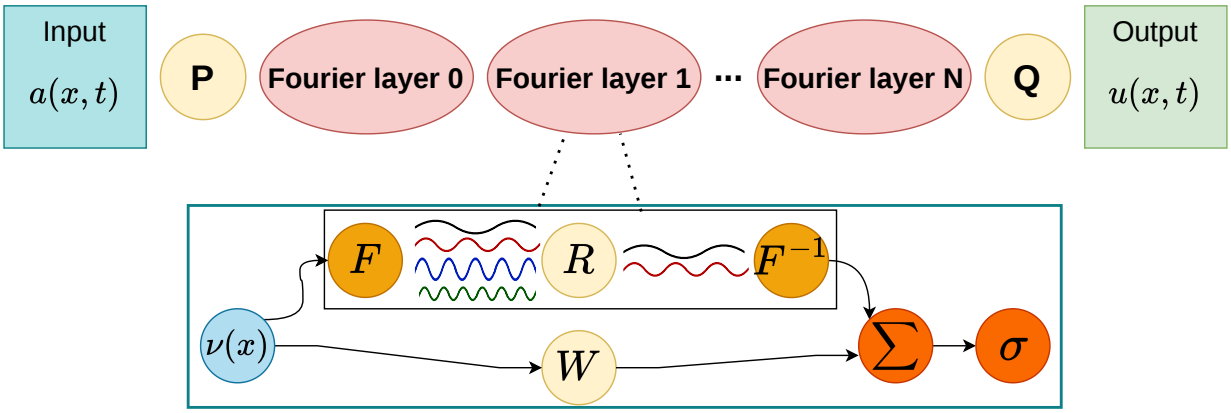


Figure A.1. Schematic architecture of the Fourier Neural Operator (FNO). The input function $a(x,t)$ is first mapped to a higher-dimensional latent representation through the lifting layer P. The resulting latent features are propagated through a sequence of Fourier layers and are finally mapped to the output function $u(x,t)$ through the projection layer Q. The enlarged Fourier layer illustrates the two parallel operations applied to the latent representation $v(x)$: a spectral convolution branch, in which the Fourier transform $\mathcal{F}$ maps the features to the frequency domain, the learnable operator R acts on the retained Fourier modes, and the inverse transform $\mathcal{F}^{-1}$ returns the result to the spatial domain; and a local linear transformation W. The outputs of the two branches are summed and passed through a nonlinear activation function σ.

These limitations motivate hybrid modeling strategies that combine coarse-resolution global operator learning with localized, memory-efficient refinement mechanisms. The framework proposed in this work follows this philosophy by employing the FNO as a global hydrodynamic surrogate and delegating fine-scale, terrain-controlled inundation structure to a separate localized refinement model.

Appendix B: Spatiotemporal Forcing Construction from Point-Scale Observations

This appendix describes the construction of spatiotemporal forcing fields used as inputs to the global neural operator. The objective is to transform sparse point-scale discharge observations into a smooth, spatially continuous forcing field defined on the coarse computational grid, while preserving mass consistency and stabilizing temporal dynamics for learning.

B1 Point-Source Observations

Let $\{s_k = (x_k, y_k)\}_{k=1}^{K}$ denote the spatial locations of K discharge gauges, and let $q_k(t)$ denote the corresponding discharge time series sampled at discrete times $t = 1, \ldots, T$. These observations are assumed to represent localized inflow or forcing contributions that must be distributed over the spatial domain.

B2 Spatial Projection via Gaussian Kernel Weighting

To obtain a spatially continuous forcing field, each point-source discharge signal is projected onto the computational grid using a radially symmetric Gaussian kernel. For a spatial location (x, y), the unnormalized contribution of gauge k is given by

$$\tilde{\phi}_k(x,y) = \exp\left(-\frac{\|(x,y) - s_k\|^2}{2\sigma^2}\right) \mathbb{I}(\|(x,y) - s_k\| \le r), \tag{B1}$$

where σ controls the spatial spread of influence, r is a finite cutoff radius beyond which contributions are truncated, and $\mathbb{I}(\cdot)$ denotes the indicator function.

To ensure that relative discharge magnitudes are preserved while maintaining spatial locality, the kernel weights are normalized pointwise:

$$\phi_k(x,y) = \frac{\tilde{\phi}_k(x,y)}{\sum_{j=1}^{K} \tilde{\phi}_j(x,y)}. \tag{B2}$$

The continuous spatiotemporal forcing field is then defined as

$$F(x,y,t) = \sum_{k=1}^{K} q_k(t)\, \phi_k(x,y). \tag{B3}$$

This formulation ensures that (i) forcing remains localized near gauge locations, (ii) overlapping influence regions are smoothly blended, and (iii) the spatial distribution of discharge is independent of grid resolution.

B3 Temporal Aggregation and Block Summation

Raw discharge time series often contain high-frequency variability that can hinder stable operator learning and increase computational cost. To mitigate this, temporal aggregation is applied via non-overlapping block summation.

Let m denote the temporal aggregation factor. The aggregated forcing field is defined as

$$\tilde{F}(x,y,\tau) = \sum_{i=1}^{m} F\big(x,y,(\tau-1)m + i\big), \qquad \tau = 1, \ldots, \left\lceil \frac{T}{m} \right\rceil. \tag{B4}$$

If T is not divisible by m, zero-padding is applied to the final block. This aggregation preserves cumulative discharge over each temporal block while reducing the effective temporal resolution and improving numerical stability during training.

B4 Neural-Operator Input Tensor Assembly

The aggregated forcing field $\tilde{F}(x, y, \tau)$ is discretized on a uniform coarse grid of size $M \times M$ and stacked with static spatial features to form the neural-operator input tensor

$$\mathbf{A} \in \mathbb{R}^{M \times M \times n_t \times C}, \tag{B5}$$

where $n_t = \lceil T/m \rceil$ is the number of aggregated time steps and C denotes the total number of input channels.

For the present experiments, the channel set is fixed to $C = 2$: the aggregated forcing field $\tilde{F}(x, y, \tau)$ and the static bed-topography field $B(x, y)$.

This construction yields a smooth, localized, and mass-consistent representation of point-source forcing that is well suited for operator learning on coarse grids, while remaining independent of the downstream localized refinement stage.

Appendix C: Synthetic Boundary Hydrograph Generation

A single historical hydrograph represents only one realization of boundary forcing and is therefore insufficient for learning a robust operator mapping between inflow conditions and flood response. To promote generalization beyond a single deterministic trajectory, we construct a controlled ensemble of synthetic boundary hydrographs that systematically expands the admissible forcing range while preserving hydrologic realism and numerical stability of the shallow-water-equation solver.

The formulation of synthetic inflow depends on the physical nature of the boundary condition. For riverine flood events, inflow is obtained by perturbing an observed hydrograph while preserving its large-scale temporal morphology. For overtopping dam-break scenarios, inflow is generated from a physically parameterized breach-release formulation. In both cases, the objective is to broaden the dynamically consistent inflow space while maintaining hydraulically admissible discharge evolution. The independent test hydrographs were generated separately from the training hydrographs but within the same controlled perturbation envelope; therefore, they evaluate generalization to unseen forcing realizations within the designed scenario space, rather than strict extrapolation beyond that space.

C1 Riverine Flood Cases (Neuse River and Chowilla Floodplain)

For riverine flood cases, synthetic inflow hydrographs are constructed by systematically perturbing an observed reference event. Rather than modifying discharge independently at each time step—which may introduce nonphysical oscillations incompatible with shallow water dynamics—we apply a low-dimensional parametric transformation to the entire hydrograph.

This strategy preserves the event-scale temporal structure of the reference signal, including multi-stage rising behavior, peak plateaus, secondary peaks where present, and recession dynamics, while allowing controlled variation in peak magnitude, timing, and limb steepness. The resulting ensemble spans a substantially broader range of inflow conditions than the original record, yet retains physically coherent discharge evolution suitable for hydraulic simulation.

Let $Q_{\mathrm{ref}}(t)$ denote the observed hydrograph at a given upstream boundary, where t is time and discharge is measured in m^3/s. The initial baseflow is defined as

$$Q_b = Q_{\mathrm{ref}}(0).$$

To preserve the characteristic event shape while allowing amplitude scaling, the hydrograph is normalized by removing baseflow and scaling by the peak excursion:

$$\hat{Q}(t) = \frac{Q_{\mathrm{ref}}(t) - Q_b}{Q_{\mathrm{ref,peak}} - Q_b}, \tag{C1}$$

where $Q_{\mathrm{ref,peak}}$ denotes the maximum observed discharge. This normalization yields a dimensionless curve satisfying $\hat{Q}(0) = 0$ and $\max_t \hat{Q}(t) = 1$.

For each synthetic realization k, the peak discharge and time to peak are modified as

$$Q_{\mathrm{peak}}^{(k)} = Q_{\mathrm{ref,peak}}(1+\epsilon_k), \qquad t_p^{(k)} = t_{p,\mathrm{ref}}(1+\eta_k), \tag{C2}$$

where ϵ_k controls relative variation in peak magnitude, η_k controls variation in peak timing, and $t_{p,\mathrm{ref}}$ denotes the reference time to peak.

To vary the steepness of the rising and recession limbs while preserving the large-scale event structure, a temporal warping function $\tau_k(t)$ is introduced:

$$\tau_k(t) = \begin{cases} t_p^{(k)} \left(\dfrac{t}{t_p^{(k)}} \right)^{\gamma_{\mathrm{rise}}^{(k)}}, & t \le t_p^{(k)}, \\ t_p^{(k)} + (T - t_p^{(k)}) \left(\dfrac{t - t_p^{(k)}}{T - t_p^{(k)}} \right)^{\gamma_{\mathrm{rec}}^{(k)}}, & t > t_p^{(k)}, \end{cases} \tag{C3}$$

where $\gamma_{\mathrm{rise}}^{(k)}$ and $\gamma_{\mathrm{rec}}^{(k)}$ regulate the steepness of the rising and recession limbs, and T denotes the total event duration. Values of $\gamma > 1$ produce more gradual limbs, whereas $\gamma < 1$ produce sharper transitions.

The synthetic hydrograph is reconstructed as

$$Q^{(k)}(t) = Q_b + \left(Q_{\mathrm{peak}}^{(k)} - Q_b\right) \hat{Q}(\tau_k(t)), \qquad Q^{(k)}(t) \ge 0. \tag{C4}$$

In this study, peak discharge was varied within $\pm 50\%$ of the observed maximum, time to peak was varied within $\pm 80\%$ of the reference value, and limb-steepness parameters were selected within prescribed bounded intervals to ensure hydraulically admissible discharge evolution.

C2 Overtopping Dam-Break Case (Fall River Lake)

For the overtopping failure scenario, inflow was defined using a physically parameterized breach hydrograph based on the overtopping configuration described in Balachandran et al. (2025). The discharge time series is represented by a triangular release function:

$$Q(t) = \begin{cases} \dfrac{Q_p}{t_p}\, t, & 0 \le t \le t_p, \\ Q_p \left(1 - \dfrac{t - t_p}{T - t_p} \right), & t_p < t \le T, \\ 0, & t > T, \end{cases} \tag{C5}$$

where Q_p denotes the maximum breach discharge, t_p the time required to reach peak flow, and T the total release duration, consistent with overtopping breach development dynamics.

The reference breach scenario was defined within the parameter ranges $Q_p \in [6500, 20000]$ $\mathrm{m}^3 \cdot \mathrm{s}^{-1}$, $t_p \in [2.0, 6.0]$ h, and $T \in [15.0, 20.0]$ h, reflecting physically plausible overtopping dynamics at the reservoir scale.

Appendix D: Hydrodynamic Solver (ANUGA)

Flood dynamics in this study are simulated using the ANUGA (Roberts et al., 2015) hydrodynamic model, developed by Geoscience Australia and the Australian National University, which solves the depth-averaged two-dimensional Shallow Water Equations (SWE) on complex topography. The model is formulated in conservative form with state vector

$$\mathbf{U} = \begin{bmatrix} h \\ uh \\ vh \end{bmatrix}, \tag{D1}$$

where h is water depth and (u, v) are depth-averaged velocities in the x and y directions.

The governing equations are written as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} = \mathbf{S}(\mathbf{U}), \tag{D2}$$

where the flux vectors are

$$\mathbf{E} = \begin{bmatrix} uh \\ u^2 h + \frac{1}{2} g h^2 \\ uvh \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} vh \\ vuh \\ v^2 h + \frac{1}{2} g h^2 \end{bmatrix}, \tag{D3}$$

and the source term accounts for bed slope and friction:

$$\mathbf{S} = \begin{bmatrix} 0 \\ -gh(z_x + S_{fx}) \\ -gh(z_y + S_{fy}) \end{bmatrix}. \tag{D4}$$

Here, $z(x, y)$ is the bed elevation and g is gravitational acceleration. Bottom friction is represented using Manning's formulation:

$$S_{fx} = \frac{u \eta^2 \sqrt{u^2 + v^2}}{h^{4/3}}, \quad S_{fy} = \frac{v \eta^2 \sqrt{u^2 + v^2}}{h^{4/3}}, \tag{D5}$$

where η is the Manning roughness coefficient.

Spatial discretization is performed using a finite-volume method on an unstructured triangular mesh, enabling accurate representation of complex terrain and localized refinement. The computational domain is partitioned into triangular control volumes T_i with area A_i, and conserved quantities are stored at cell centroids. The semi-discrete update for each cell is given by

$$\frac{d\mathbf{U}_i}{dt} + \frac{1}{A_i} \sum_{j \in \mathcal{N}(i)} \mathbf{H}_{ij} \, l_{ij} = \mathbf{S}_i, \tag{D6}$$

where $\mathcal{N}(i)$ denotes neighboring cells, l_{ij} is the edge length, and $\mathbf{H}_{ij}$ is the numerical flux computed using a central-upwind (Riemann-based) scheme.

To achieve second-order spatial accuracy, a piecewise linear reconstruction is applied within each cell, combined with slope limiting to prevent non-physical oscillations near discontinuities such as hydraulic jumps. The finite-volume formulation naturally accommodates wetting and drying processes, making it well-suited for floodplain inundation.

Time integration is performed using an explicit scheme with adaptive time stepping constrained by a Courant–Friedrichs–Lewy (CFL) condition:

$$\Delta t \leq \min\left(\frac{r}{|\lambda_{\max}|}\right), \tag{D7}$$

where r is a characteristic cell length and $\lambda_{\max}$ is the maximum wave speed across cell interfaces.

The model assumes hydrostatic pressure and neglects vertical accelerations and viscous effects, which is appropriate for large-scale flood inundation problems. Despite these simplifications, the solver robustly captures key hydrodynamic processes including wave propagation, momentum exchange, and flow over dry terrain, providing high-fidelity spatiotemporal flood fields used as reference data in this study.

Appendix E: Training-Size Sensitivity Analysis

To evaluate the sensitivity of model performance to the number of training realizations, we conducted an additional learning-curve analysis using progressively larger training sets. The purpose of this analysis was to determine whether increasing the number of synthetic hydrodynamic simulations provides substantial additional predictive skill, or whether model performance reaches a plateau once the main forcing variability has been sampled.

For each case study, models were trained using different numbers of training realizations, while evaluation was always performed on the same independent test set containing 50 unseen scenarios. Therefore, changes in performance reflect the effect of training-set size rather than changes in the evaluation data. The same test scenarios were kept fixed across all training-size experiments to ensure a consistent comparison.

Table E.1(a) summarizes the resulting relRMSE values for the coarse FNO and fine Magnifier predictions. Across all three study domains, increasing the number of training realizations produces a consistent reduction in relRMSE, indicating that additional simulations improve the learned input–output mapping. However, the improvement becomes progressively smaller as the training set grows. Most of the gain occurs between the smallest training set and the intermediate training sizes, while the difference between the larger training sets is relatively minor.

Table E.1(b) shows the corresponding CSI values for binary inundation prediction. Similar to the relRMSE results, CSI improves with additional training realizations but approaches a plateau as the training set increases. This indicates that both continuous depth accuracy and wet–dry classification skill benefit from additional training data, but the marginal improvement becomes limited once the dominant forcing-response patterns are represented.

This saturation behavior suggests that, within the sampled forcing envelope, the synthetic hydrograph generation procedure already captures the dominant variability controlling the hydrodynamic response. Additional training realizations provide useful refinement but do not fundamentally change the learned operator. These results support the use of the selected training-set size in the main experiments, while also showing that the framework can benefit modestly from larger training libraries when additional hydrodynamic simulations are available.

Table E.1. Training-size sensitivity across independent test scenarios for increasing numbers of training realizations.

(a) relRMSE

Case	Scale	50	100	200	300
(i) Neuse River	Coarse (FNO)	0.0317	0.0219	0.0213	0.0211
	Fine (Magnifier)	0.4448	0.3315	0.3261	0.3227
(ii) Fall River Dam Break	Coarse (FNO)	0.0206	0.0130	0.0126	0.0125
	Fine (Magnifier)	0.0589	0.0375	0.0369	0.0364
(iii) Chowilla floodplain	Coarse (FNO)	0.0346	0.0228	0.0225	0.0221
	Fine (Magnifier)	0.0562	0.0368	0.0359	0.0357

(b) CSI (%)

Case	Scale	50	100	200	300
(i) Neuse River	Coarse (FNO)	97.29	98.92	99.07	99.10
	Fine (Magnifier)	90.67	93.03	93.29	93.51
(ii) Fall River Dam Break	Coarse (FNO)	96.18	98.43	98.61	98.74
	Fine (Magnifier)	94.87	96.42	96.53	96.57
(iii) Chowilla floodplain	Coarse (FNO)	97.68	99.28	99.41	99.47
	Fine (Magnifier)	96.75	97.95	97.99	98.07