https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

A Factorized Fourier Neural Operator Surrogate for Basin-Scale Tsunami Propagation

Jinyoung Kim[1], Myung Jin Koh[1], Seung-taek Oh[1], and Sangyoung Son[1]

[1]School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, South Korea

Correspondence: Sangyoung Son (sson@korea.ac.kr), Myung Jin Koh (myungj@korea.ac.kr)

Abstract. Tsunami models have been developed for several decades, and many have shown good agreement with observations from real world events. The model solves wave equations, but simulation is usually computationally expensive in a large-scale basin. To assess potential tsunami impacts, ensemble analysis is standard for sampling source uncertainties, but computational costs constrain the number of scenarios that can be evaluated. Machine-learning approaches have been developed to reduce the computational burden and accelerate typical tsunami-ensemble analyses. However, these surrogate models are usually task-specific; they emulate buoy signals, sensor inputs, and maximum water level maps. Recent advances in machine learning techniques, such as neural operators, allow learning full wave evolution from physics-based simulations. Here, we introduce a data-driven tsunami surrogate model based on a Factorized Fourier Neural Operator (F-FNO). Memory-efficient F-FNO supports higher Fourier mode capacity, enabling the tsunami surrogate model to learn scenario-based COMCOT simulations and generalize to unseen epicenter locations/extrapolated magnitudes. We designed logic tree-based COMCOT simulations for the East Sea (Sea of Japan) to construct a surrogate operator. The F-FNO learns tsunami propagation through a short sequence of wavefield states and creates a general operator function that generates future wave and velocity fields. From the logic tree, we hold out the largest magnitude (8.0) and one specific source location for model evaluation and to test the scalability of the neural operator. As a result, the surrogate predicted tsunami waves with root mean square errors in surface elevation of 2–8 cm and first-arrival timing errors of approximately 8–12 min. Running the F-FNO surrogate requires approximately 8.5-12 s per scenario on a single GPU, compared to 87.9-95.7 s of COMCOT simulation time. The computational efficiency of the operator and its potential to scale to larger scenario ensembles support more timely tsunami scenario analysis and can complement physics-based solvers in offshore applications.

1 Introduction

A submarine earthquake can generate a tsunami that travels across an ocean basin. In deep water, the wave is usually not noticeable. However, it can grow rapidly as it approaches distant coastlines. Severe damage can occur when the preparation and prediction for the tsunami-induced flooding failed (Röbke and Vött, 2017; Satake, 2014; Bernard and Titov, 2015). In many cases, we witnessed the limitations of deterministic risk assessment, which rely on historical records and expert judgement. The 2004 Indian Ocean tsunami struck coastlines with no prior scenario coverage (Synolakis and Bernard, 2006), and the 2011 Tōhoku tsunami exceeded expectations of engineered defenses (Mori et al., 2011). To mitigate this risk, the management system

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

shifted towards probabilistic hazard assessment. To capture the full range of uncertainties, hundreds to thousands of possible scenarios are usually evaluated (Annaka et al., 2007; Grezio et al., 2017; Selva et al., 2016; Davies et al., 2018).

This approach mainly relies on repeated basin-scale tsunami simulations. The outputs are used for probabilistic hazard assessment and design-basis evaluation for critical infrastructure (Lorito et al., 2015; Japan Society of Civil Engineers, 2016). The results are also used to build a database for real-time warning (Wei et al., 2003) and perform a sensitivity test. They also provide offshore boundary conditions for a resolved high-resolution model. Physics-based tsunami models such as COMCOT, MOST, and GeoClaw solve the depth-integrated shallow-water equations. These models have been tested for laboratory data and historical tsunami observations (Liu et al., 1995; Titov and Synolakis, 1998; Wang and Power, 2011; Synolakis et al., 2008). However, each simulation requires on the order of minutes on multi-core hardware at moderate resolution. A single run is affordable, but evaluating more than hundreds of scenarios to cover source uncertainties becomes inefficient.

For faster prediction and to alleviate the computational cost, several surrogate approaches have been introduced. Statistical emulators based on Gaussian process regression and PCA (principal component analysis) have been applied to surrogate maximum tsunami heights at specific coastal locations from parameterized source inputs (Sarri et al., 2012; Salmanidou et al., 2017; Guillas et al., 2018; Mulia et al., 2020b). Recurrent and convolutional neural networks have been shown to have good agreement in predicting tsunami time series at coastal water level gauges and offshore sensor observations (Makinoshima et al., 2021; Liu et al., 2021; Wang et al., 2023; Song and Cho, 2024). While most studies focus on predicting the time series of monitoring stations, other studies have adopted deep neural network architectures to focus on modeling the waveform of nearshore processes and the depth of inundation along a coastal region from seismic source parameters (Mulia et al., 2022; Song and Cho, 2024; Ragu Ramalingam et al., 2025; Fukutani and Motoki, 2025). However, almost all recent works focus on predicting a fixed set of outputs such as maximum water level and inundation maps. Thus, expanding the surrogate's general-purpose for basin-scale tsunami propagation is limited.

Neural Operator approaches provide an alternative way to learn tsunami wave propagation between function spaces (Lu et al., 2021; Kovachki et al., 2023). The Fourier Neural Operator (FNO) and its factorized variant (F-FNO) learn wavefields as function spaces rather than vectors. Thus, the operator-based surrogate can emulate and extrapolate some of the evolution of spatially distributed fields (Li et al., 2021; Tran et al., 2023). Neural operators' recent applications in regional ocean modeling have demonstrated their ability to reproduce dynamics of the ocean environment with high fidelity when trained on sufficient simulation databases (Chattopadhyay et al., 2024; Choi et al., 2024). Most of the neural operator-related studies have addressed ocean flow governed by persistent atmospheric and tidal forcing. However, basin-scale tsunami propagation differs from previous studies since the initial wavefield is generated impulsively during the rupture of the submarine fault, not by sustained forcing. The generated wave from the earthquake then propagates across the basins, undergoing repeated refraction due to varying bathymetry, and diffraction/reflection from surrounding coastlines. Whether FNO-variant architectures can capture these dynamics and their practicality has not yet been examined.

This study develops and evaluates an F-FNO-based tsunami surrogate model trained on a series of scenario-based wave simulation results from a COMCOT solver for the East Sea. We used 864 tsunami scenarios that combine the reference epicenters with variations such as strike angle, moment magnitude, and fault-dimension scaling relationships (Wells and Coppersmith, 1994;

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Leonard, 2010). The surrogate takes bathymetry, sea-surface elevation and velocity maps, and predicts the next propagation step. It achieves multi-step forecasting by repeating this process in a closed loop. Because prediction errors tend to accumulate over hundreds of prediction steps, we apply physics-guided loss terms to improve prediction stability. In particular, strengthening the continuity constraint and enforcing divergence control on the domain-averaged sea level help suppress error drift over long rollouts. We evaluate our approach in comparison with a standard FNO baseline. This study makes three contributions. First, we demonstrate that an F-FNO surrogate can reliably model a basin-scale tsunami propagation under impulsive source forcing. Second, the surrogate achieves stable, long-term predictions of tsunamis with physics-guided learning strategies. Third, by comparing the F-FNO against a standard FNO baseline and evaluating test scenarios with an unseen epicenter and an extrapolated magnitude, we show that the higher Fourier-mode capacity enabled by factorization is an important factor for the improvement, and that magnitude extrapolation remains a greater generalization challenge than spatial extrapolation.

2 Methods

2.1 Tsunami model and study domain

Fig. 1 summarizes our overall workflow to build and verify the neural operator used in this study. Seismic source parameters are first sampled via a logic tree including fault and magnitude uncertainties (Step 1). COMCOT then solves the shallow-water equations for each scenario and provides (η, u, v) temporal maps (Step 2). The simulation outputs are normalized with wet-cell masking and the scenarios are partitioned into training, validation, and test sets (Step 3). A 10-layer F-FNO is trained with loss functions (Step 4) and evaluated on test sets using different metrics (Step 5). The following subsections describe the simulation model and the surrogate.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

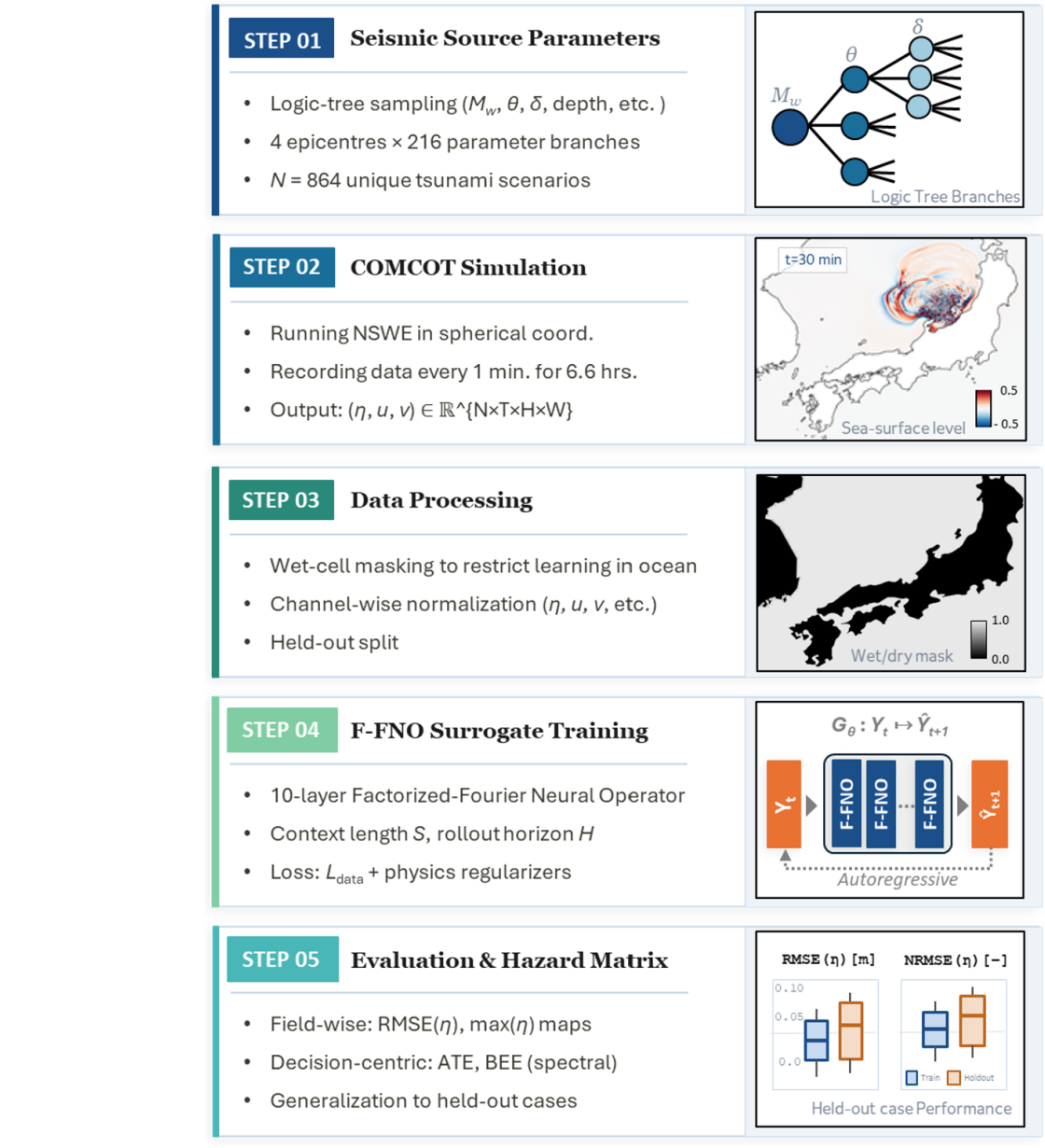

Figure 1. Overall workflow of the proposed F-FNO tsunami surrogate, from logic tree scenario design through COMCOT simulation, data preprocessing, surrogate training, and held-out scenario evaluation.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

COMCOT is a physics-based tsunami wave solver that simulates the shallow-water equations (Wang and Power, 2011). For the propagation in the basin considered here, wave amplitudes are small relative to water depth, and nonlinear effects are minor. We thus use the linear shallow-water equations with bottom friction:

$$\frac{\partial \eta}{\partial t} + \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} = 0, \tag{1}$$

$$\frac{\partial P}{\partial t} + gH\frac{\partial \eta}{\partial x} - fQ + F_x = 0, \tag{2}$$

$$\frac{\partial Q}{\partial t} + gH\frac{\partial \eta}{\partial y} + fP + F_y = 0, \tag{3}$$

where η is the free-surface elevation and $H = h(x, y)$ is the still-water depth (bathymetry). P and Q denote the depth-integrated volume fluxes in the x- and y-directions, g is gravitational acceleration, and f is the Coriolis parameter. The bottom-friction source terms are

$$F_x = \frac{gn^2}{H^{7/3}} P\sqrt{P^2 + Q^2}, \quad F_y = \frac{gn^2}{H^{7/3}} Q\sqrt{P^2 + Q^2}, \tag{4}$$

where $n = 0.025\,\mathrm{m}^{-1/3}\,\mathrm{s}$ is Manning's roughness coefficient. Relative to the full nonlinear system, advective terms in the momentum equations (e.g., $\partial(P^2/H)/\partial x$) are omitted, which is a common approximation for basin-scale tsunami propagation in deep water (Satake, 1995; Cho et al., 2007). The numerical configuration adopted in all simulations is summarized in Table 1.

As a study area, we selected the East Sea (Sea of Japan) for a test domain. As a semi-enclosed basin with maximum depths over 3,700 m and lateral dimensions of around 1,000 km, the basin shows spatially complex wave propagation due to bathymetric refraction/focusing and basin-scale reflection (Choi et al., 2003; Yoon et al., 2014). Historical records document damaging tsunamis from submarine earthquakes along the eastern margin: the 1983 central East Sea earthquake (M_w 7.7) generated waves of 3.5–4.0 m at Imwon Port on the Korean coast (Choi et al., 2003, 2008), and the 1993 Hokkaido Nansei-Oki earthquake (M_w 7.8), which produced runup exceeding 30 m on Okushiri Island and was recorded at 2.8 m in Sokcho, Korea (Shuto and Matsutomi, 1995; Choi et al., 2003). Most recently, the 2024 Noto Peninsula earthquake (M_w 7.5) produced damaging run-up along the Sea of Japan coast with minimal warning time (Heidarzadeh et al., 2024). Beyond these historical events, seismological studies have identified persistent seismic gaps along the eastern margin of the basin (Ohtake, 1995), and long-term seismicity evaluations suggest that the potential for future tsunamigenic earthquakes has not been exhausted (Earthquake Research Committee, 2003; Korea Peninsula Energy Development Organization, 1999). Existing tsunami studies for the East Sea have addressed individual historical events or site-specific hazard estimates (Son and Jung, 2022; Mulia et al., 2020a; Satake et al., 2022; Satake and Murotani, 2025), but none have constructed a systematic scenario database that spans fault-geometry and magnitude uncertainties at basin scale with full spatiotemporal wavefield output suitable for data-driven surrogate training. The computational domain covers 127–143° E, 30–50° N (Fig. 2). Four reference earthquake source locations (Ep 1–Ep 4) from the KEDO database are distributed along the eastern Japan margin between 38° N and 43° N, at distances of 800–1100 km

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Table 1. COMCOT numerical simulation configuration used to generate the training and test datasets.

Parameter	Value	Note
General settings		
Coordinate system	Spherical	
Governing equations	Linear shallow water	
Boundary condition	Open (radiating)	
Minimum water depth	5 m	Wet/dry threshold
Initial condition	Okada fault model	Cold start
Computational domain (Layer 1)		
Longitude range	127.0–142.98° E	
Latitude range	30.0–49.99° N	
Horizontal resolution	1.728 arcmin	≈ 3.2 km at domain center
Time step	1.0 s	CFL-stable
Bottom friction	$0.025\,\mathrm{m}^{-1/3}\,\mathrm{s}$	Manning's coefficient
Output		
Saved variables	η, u, v	Sea-surface elevation and depth-averaged velocities
Output interval	60 s	

from the Korean east coast. To monitor surrogate performance at representative propagation stages, we define three virtual offshore buoy stations (B080, B400, B720) at increasing distances from the source region, and three nearshore reference sites along the Korean coast (Imwon, Hanul, and Pohang) that correspond to locations of historical tsunami impact or critical coastal infrastructure. All station coordinates are snapped to the nearest wet grid cell. Detailed source parameters and the logic tree are described in Sect. 2.3 and Appendix A.

2.2 F-FNO Surrogate architecture and training

The training data includes space-time free-surface elevation η and depth-averaged velocities (u, v) field of COMCOT on a structured grid. As a preparation for training sets, we first define the wet index Ω_{wet} and apply a mask to train only wet cells. In the main experiments, wet cells were defined using a minimum still-water depth threshold of 5 m. This choice focuses training on the large scale ocean wave propagation where linear shallow water equations remain reliable, and to avoid numerical artifacts caused by velocity-field singularities in extremely shallow water. We tested minimum water depth thresholds of 1, 2, 3, and 5 m in a separate sensitivity test, and 5 m gave the lowest validation error (Table S4). Each training sample is constructed using a moving time window. The first $S = 10$ frames are used as input, and the model is trained to predict the tsunami evolution (η, u, v) for a forecast horizon of length $H_r = 40$. All dynamic variables are normalized channel-wise using a training set

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

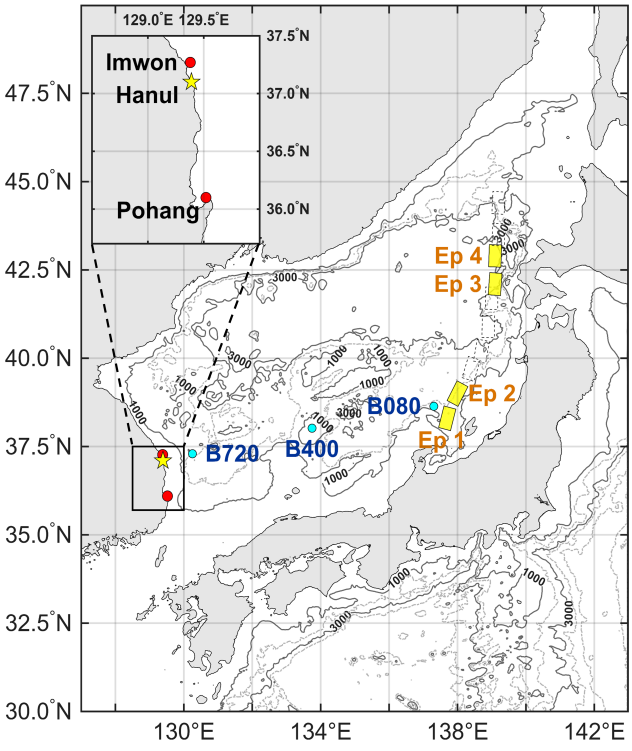

Figure 2. Bathymetry and COMCOT domain for the East Sea (127–142° E, 30–50° N). Gray areas indicate land. Four yellow boxes labeled Ep 1–Ep 4 represent four earthquake source locations used in this study from the KEDO database along the eastern Japan margin (38–43° N). Cyan circles show virtual buoy stations (B080, B400, B720) at representative water depths. The inset shows three nearshore reference sites along the Korean coast (inset: 129.0–129.5° E, 36.0–37.5° N): Imwon and Pohang are shown as red circles, whereas Hanul, located offshore of the Hanul Nuclear Power Plant, is shown as a yellow star. All station coordinates are snapped to the nearest wet grid cell.

from a 512-sample subset statistics computed over wet cells $\tilde{z} = (z - \mu_z)/\sigma_z$ to ensure the stable learning for autoregressive prediction. The same normalization is applied to validation and test sets. Bathymetry is provided as a static input channel and standardized using fixed depth statistics. Each COMCOT scenario is simulated for $N_{\mathrm{sim}} = 400$ output frames with $\Delta t = 60\,\mathrm{s}$ (total duration $\approx 6.6\,\mathrm{h}$). For training, we use a context of length $S = 10$ frames and apply multi-step rollout supervision over $H_r = 40$ steps, so that $S + H_r = 50$. The neural operator predicts the next step wave conditions and repeats this in a closed loop to build multi-step forecasts. Neural operator-based F-FNO has an advantage for wave evolution learning for roughly 1000 km length basin since the surrogate must represent long-range spatial interactions efficiently. F-FNO separates the spectral domain into one-dimensional Fourier transforms along each spatial axis (Tran et al., 2023). This factorization makes learning tsunami propagation more efficient. Unless otherwise stated, inference and benchmark results are reported using a shorter rollout horizon $H_r = 200$ steps ($\approx 3.3\,\mathrm{h}$) instead of the entire simulation time step of 400 minutes, as an efficient operational forecast window.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

The definition of a Fourier neural operator is as follows. At each time step, the dynamic state is $Y_t(\mathbf{x}) = [\eta_t(\mathbf{x}), u_t(\mathbf{x}), v_t(\mathbf{x})]$ defined on wet cells Ω_{wet}. Bathymetry $h(\mathbf{x})$ is provided as an additional input channel, and optional geographic coordinate channels (e.g., $\sin/\cos$ of latitude/longitude) can be concatenated to improve geometric awareness. The input state is first mapped to a higher-dimensional internal representation by a pointwise linear layer, producing $v^{(0)} \in \mathbb{R}^{H\times W\times d_v}$. Each F-FNO layer then combines a local pointwise pathway with a factorized spectral convolution:

$$v^{(\ell+1)} = \sigma\Big(W^{(\ell)} v^{(\ell)} + \mathcal{K}^{(\ell)}(v^{(\ell)})\Big), \tag{5}$$

where $\sigma(\cdot)$ denotes nonlinearity. The spectral operator $\mathcal{K}$ applies separated 1D Fourier transforms for each axis and adds up the contributions:

$$\mathcal{K}(v) = \mathcal{F}_y^{-1}(R_y \odot \mathcal{F}_y(v)) + \mathcal{F}_x^{-1}(R_x \odot \mathcal{F}_x(v)), \tag{6}$$

where R_x and R_y are learned spectral weights as complex values. They are truncated to prescribed Fourier modes. Factorization of each axis allows for substantially lower memory cost than the full 2D Fourier transforms in the standard FNO at formulation of (Fig. 3). This allows for increasing the mode number of the operator, and the model can resolve propagation patterns better while preserving the shorter-wavelength associated with coastal amplification. A lightweight decoder maps the final latent field to the next-step prediction $\hat{Y}_{t+1}$. The network outputs a residual increment $\Delta\hat{Y}_{t+1}$ so that:

$$\hat{Y}_{t+1} = Y_t + \Delta\hat{Y}_{t+1}, \tag{7}$$

which improves stability over long prediction sequences by biasing the model toward small updates during quiescent periods between wavefronts. In pilot experiments, the standard FNO often showed weaker stability in the predicted free-surface elevation during long recursive prediction. After adopting the factorized architecture with larger retained modes, the multi-step predictions became more stable.

All predicted states are used as input to the next prediction step. We leverage two essential properties in order to effectively counteract the effect of small modelling errors propagated over hundreds of steps. First, during training the model is gradually transitioned from using more ground-truth inputs to using more of its own predictions following Bengio et al. (2015). This way, the model learns to cope with its own prediction errors during training, rather than relying entirely on ground-truth inputs that would not be available at inference time. The specific decay schedule and parameters are given in Appendix C. Second, we compute multi-step supervision losses across the whole rollout, instead of merely comparing model output to the ground truth only at one step ahead as is typically done. This ensures that the model is directly optimized for long-horizon accuracy rather than one-step performance only. To limit memory use during multi-step training, we applied truncated backpropagation through time (TBPTT). Each rollout of length $H_r = 40$ was split into segments of length $K = 20$, and the computation results graph was detached between segments.

The following loss terms and their corresponding weights were selected to account for the unique physical characteristics of tsunami wavefield modeling.

$$L = L_{\text{data}} + \lambda_{\text{sw}} L_{\text{sw}} + \lambda_{\text{peak}} L_{\text{peak}} + \sum_{m\in\mathcal{R}} \lambda_m L_m. \tag{8}$$

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

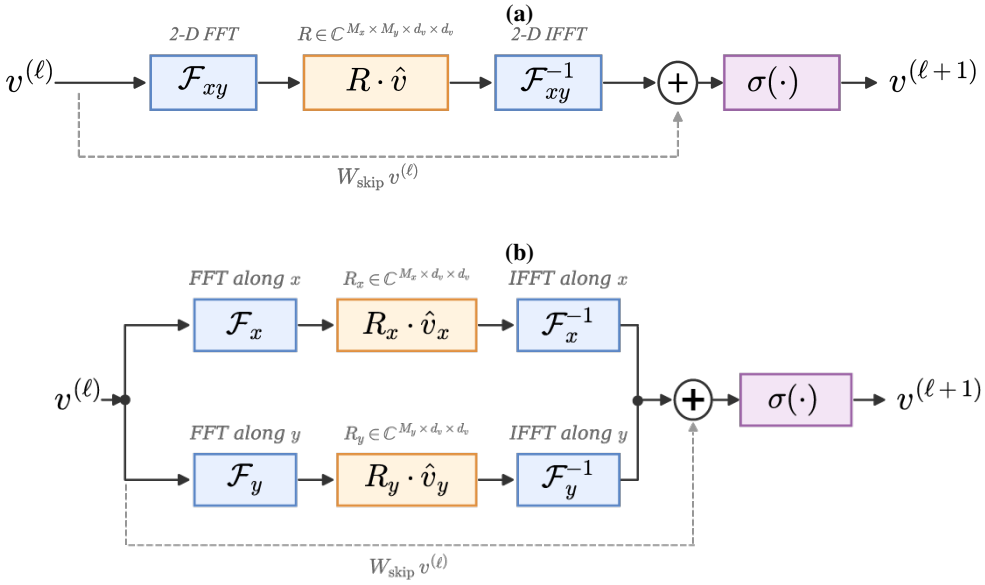

Figure 3. Architecture of a single spectral convolution layer. (a) The standard FNO applies a full 2D FFT followed by a learned complex spectral weight $R \in \mathbb{C}^{M_x \times M_y \times d_v \times d_v}$ and a 2D IFFT; the result is summed with a pointwise skip connection W_{skip} and passed through a nonlinearity σ. (b) The F-FNO replaces the single 2D spectral convolution with two independent 1D transforms along the x and y axes, each with its own spectral weight ($R_x \in \mathbb{C}^{M_x \times d_v \times d_v}$, $R_y \in \mathbb{C}^{M_y \times d_v \times d_v}$). This factorization reduces the learnable spectral parameters and memory cost from $\mathcal{O}(d_v^2\, M_x M_y)$ to $\mathcal{O}\big(d_v^2 (M_x + M_y)\big)$.

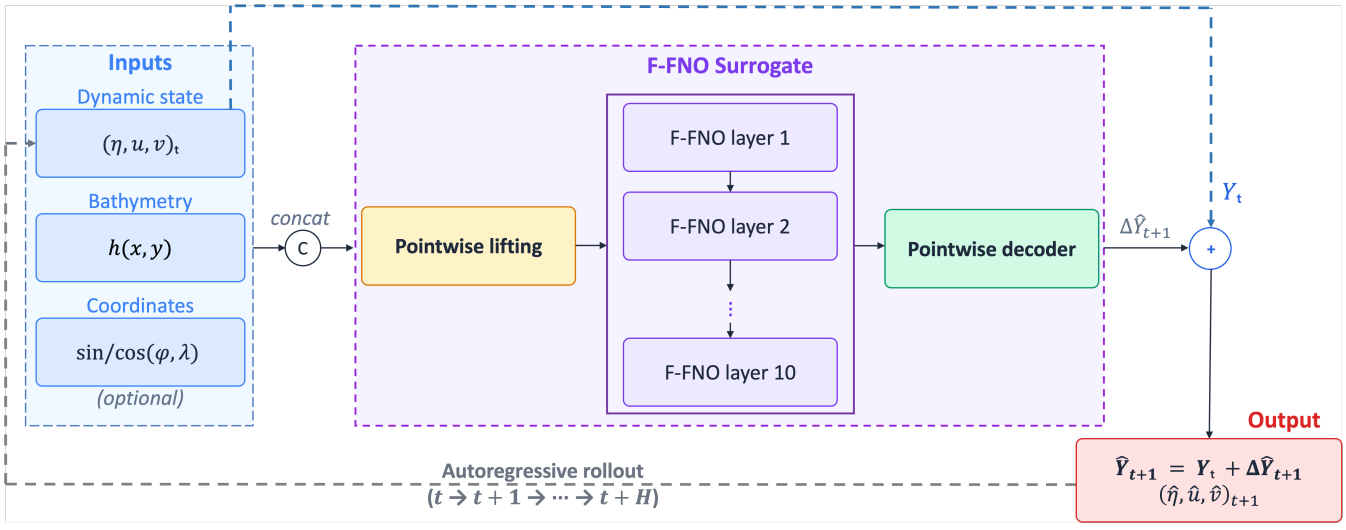

Figure 4. Overview of F-FNO autoregressive forecaster. With the dynamic state $Y_t = (\eta, u, v)$ and bathymetry h, the model predicts the next-step state at $t+1$ and is iterated in a closed loop for multi-step forecasting.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

The data term L_{data} is designed to measure the difference between the predicted and actual simulated (η, u, v) fields at all time locations. So it is a primary term to verify the accuracy of the prediction. The still-water suppression term L_{sw} is designed to remove the artificial noise ahead of the leading wave. The peak-aware time weighting (L_{peak}) gives extra importance to times when the water level is highest, rather than spreading training effort evenly across all steps. The mass-balance loss (L_{cont}) discourages violations of the continuity equation, and a drift control (L_{dc}) suppresses drift in the domain-averaged sea level. All loss functions are defined in Appendix B. For the Selected model, the weights of the loss functions are set to $\lambda_{\mathrm{sw}} = 0.1$, $\lambda_{\mathrm{peak}} = 0.05$, $\lambda_{\mathrm{cont}} = 10$, and $\lambda_{\mathrm{dc}} = 100$.

The 10-layer network with $\lambda_{\mathrm{cont}} = 10$ and $\lambda_{\mathrm{dc}} = 100$ was selected as the Selected model, as it had the lowest validation error for any of the configurations investigated. The training and validation errors of the Selected model are shown in Fig. 5. Both sets of errors are decreasing and appear to be converging to some stable value for all 20 training epochs. The graph shows gradual convergence, and the validation errors become stable around epoch 15. Near epoch 7-10, some fluctuations in error levels are recorded, and this is likely due to sample variability in the validation set. For clarity, we denote the default configuration as the Reference, the configuration without the divergence constraint as w/o DC, and the Fourier-mode increase variant as +M272.

All variants are implemented in PyTorch and share the same training schedule: 20 epochs, AdamW optimizer (learning rate 5×10^{-4}, weight decay 10^{-5}), cosine-annealing learning-rate schedule ($\mathrm{lr}_{\mathrm{min}} = 5 \times 10^{-5}$), gradient clipping (max norm = 1.0), automatic mixed-precision training (bfloat16 autocast, with spectral FFT and mode-mixing operations executed in float32 for numerical stability). All training runs were performed on a single GPU, mainly on an NVIDIA B200 GPU. Each full 20-epoch run takes approximately 37.5 hours, but this time may differ depending on the number of modes and layers in F-FNO.

Table 2. F-FNO model configurations evaluated in this study. All variants are implemented in PyTorch and share the same training schedule: 20 epochs, AdamW optimizer (learning rate 5×10^{-4}, weight decay 10^{-5}), cosine-annealing learning-rate schedule ($\mathrm{lr}_{\mathrm{min}} = 5 \times 10^{-5}$), gradient clipping (max norm = 1.0), automatic mixed-precision training (bfloat16 autocast, with spectral FFT and mode-mixing operations executed in float32 for numerical stability), batch size 1, 2 000 randomly sampled training windows per epoch, context length $S = 10$, rollout horizon $H_r = 40$, BPTT segment length $K = 20$, and random seed 42. Shared loss weights are $\lambda_{\mathrm{peak}} = 0.05$, $\lambda_{\mathrm{sw}} = 0.1$, $\lambda_{\mathrm{grad}} = 0.25$, $\lambda_{\mathrm{energy}} = 1.0$, $\lambda_{\eta\nabla} = 1.0$, $\lambda_{\mathrm{int}} = 0.01$. Differences from the Reference baseline are underlined.

	Reference	**Selected**	**w/o DC**	**+M272**
Width (d_v)	64	64	64	64
Depth (layers)	8	10	8	8
Modes (M_x, M_y)	(256,256)	(256,256)	(256,256)	(272,272)
λ_{cont}	0.5	10	0.5	0.5
λ_{dc}	100	100	0	100

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

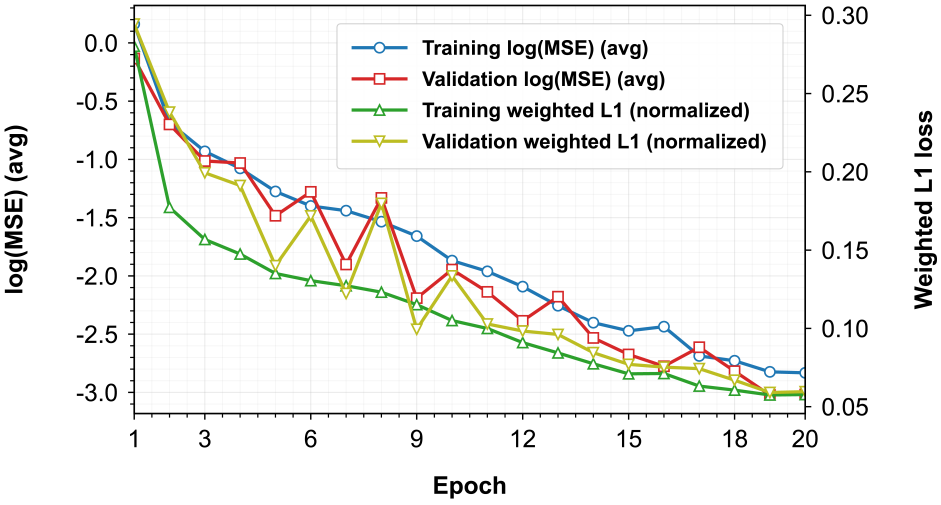

Figure 5. Training and validation convergence for the Selected model (10-layer, $\lambda_{\mathrm{cont}} = 10$, $\lambda_{\mathrm{dc}} = 100$). Left axis: Normalized $\log(\mathrm{MSE})$ (circles) for training (blue) and validation (red). Right axis: Normalized weighted L1 loss (triangles) for training (green) and validation (yellow).

2.3 Scenario and data splits

We generate a scenario database using COMCOT over the East Sea (Japan Sea) domain. Each scenario runs a different combination of source location and fault parameters in the logic tree split (Fig. 6). Fig. 6 shows the splits that were excluded for validation and to test the generalization to unseen cases (Test-M/E/EM). We denote the four source locations as Ep 1 – 4 in the main text. Appendix A reports their original KEDO case IDs. The full database consists of 864 scenarios. Out of these, 378 scenarios were reserved for the test and not used in the training, which included Test-M (162), Test-E (162), and Test-EM (54). The remaining 486 scenarios formed the development pool, from which 437 scenarios were used for training and 49 for validation. We selected the validation set at random and kept this split fixed throughout the study.

For training, COMCOT outputs are used for an input context of length S. The sequence is paired with a forecast horizon of length H_r. Specifically, for a given start index t_0, the model input is $\{\mathbf{x}_{t_0}, \ldots, \mathbf{x}_{t_0+S-1}\}$ and the target rollout is $\{\mathbf{y}_{t_0+S}, \ldots, \mathbf{y}_{t_0+S+H_r-1}\}$, where $\mathbf{x}_t, \mathbf{y}_t$ contain (η, u, v) on wet cells. All variables are normalized using training set statistics calculated from the grid. The same transformation is applied to the validation and test sets. Bathymetry is provided as an additional static input channel and is normalized using training set depth statistics. Data splits are performed at the scenario level to ensure that the evaluation reflects generalization to unseen source-parameter combinations. Random time-slice splitting is avoided because it would allow interpolation within the same scenario realization. In addition, we use a magnitude and epicenter hold-out split to test generalization to unseen magnitudes and/or epicenter locations. The reference COMCOT simulations were executed on a system equipped with an AMD Ryzen 9 7950X CPU. Training of the F-FNO surrogate was performed primarily

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

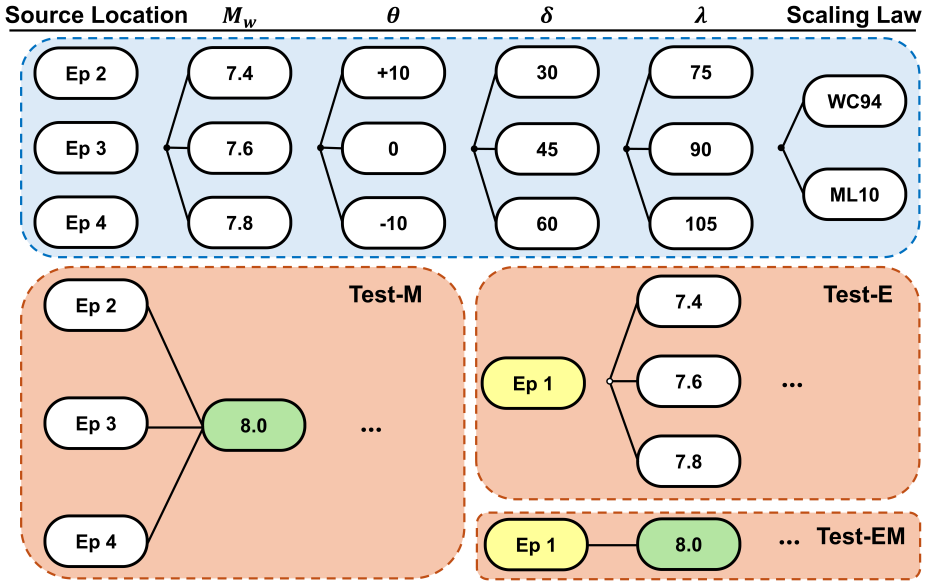

Figure 6. Schematic of the three-branch logic tree subset used in this study and the resulting held-out test splits. Training data use Ep 2 through Ep 4 with strike offsets $\{-10, 0, +10\}$, dip $\{30, 45, 60\}$, rake $\{75, 90, 105\}$, magnitudes M1 through M3, and two scaling laws (WC94/ML10). Held-out tests are defined as Test-M (Ep 2–Ep 4, M4 only), Test-E (Ep 1, M1–M3), and Test-EM (Ep 1, M4 only). Here M1–M4 correspond to $M_w = \{7.4, 7.6, 7.8, 8.0\}$, respectively.

on an NVIDIA B200 GPU, and inference was performed primarily on an NVIDIA RTX 5070 Ti GPU, unless otherwise stated. All main-text surrogate results were obtained in a single-GPU setting.

We generate a comprehensive scenario database using a logic tree that varies five fault parameters. Strike angle θ, dip angle δ, rake angle λ, moment magnitude M_w, and two different scaling relationships for fault geometry (Fig. A1). Four source locations (Ep 1–Ep 4) in the range of latitudes 38.3° N to 42.9° N are selected from the Korea Peninsula Energy Development Organization (KEDO) database. They are located near the eastern Japan margin at distances of 800–1100 km from the Korean east coast (Fig. 2). For each location, we consider three strike-angle variations ($\theta_0 - 10°$, θ_0, $\theta_0 + 10°$), three dip angles (30°, 45°, 60°), three rake angles (75°, 90°, 105°), and four moment magnitudes (M_w 7.4, 7.6, 7.8, 8.0), combined with two empirical scaling relationships: Wells and Coppersmith (1994) and Leonard (2010). The full logic tree yields 864 unique tsunami scenarios ($4 \times 3 \times 3 \times 3 \times 4 \times 2$). The scenarios cover the range of source uncertainties relevant to probabilistic tsunami hazard assessment. They also provide sufficient diversity as a training data set. Detailed parameter definitions, scaling-law formulations, and epicenter coordinates are provided in Appendix A.

2.4 Evaluation metrics

We evaluate the performance of the surrogate model using three groups of metrics including point-by-point errors of η, u, v for wet cells, domain mean RMSE over time, and quantities related to tsunami physics. For tsunami applications, free-surface

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

elevation is the primary hazard variable. Therefore, we report RMSE and normalized error measures for η, and additionally evaluate the precision of arrival-time patterns and amplitude evolution towards the Korean coast. To characterize basin-scale behavior, we also compare spatial maps of peak elevation:

$$\eta_{\max}(\mathbf{x}) = \max_{t\in[t_0,t_0+H]} \eta(\mathbf{x},t), \tag{9}$$

In addition, the error of $\eta_{\max}$ in wet cells is evaluated, which is directly related to the comparison of risk ranges.

2.4.1 Field-wise RMSE

Ω_w represents the union of wet masks for the evaluation region (excluding NaN), and $t = 1,\ldots,T$ is an index representing the prediction step. We calculate the water surface elevation RMSE as

$$\mathrm{RMSE}_\eta = \sqrt{\frac{1}{|\Omega_w|T}\sum_{t=1}^{T}\sum_{(i,j)\in\Omega_w}\left(\eta_t^{\mathrm{pred}}(i,j)-\eta_t^{\mathrm{true}}(i,j)\right)^2}. \tag{10}$$

2.4.2 Arrival timing diagnostics

We evaluate first-arrival timing at six virtual buoy locations. Each time series is first smoothed with a three-point moving average to suppress transient numerical spikes. The arrival time is defined as the earliest time index at which the smoothed absolute free-surface elevation continuously exceeds a fixed threshold $\theta = 0.05\,\mathrm{m}$ for at least 5 min (i.e. $\lceil 5\,\mathrm{min}/\Delta t\rceil$ consecutive steps). The 5 cm threshold follows the JMA NWPTAC operational definition of estimated tsunami arrival time used in its numerical forecast/database framework, in which arrival time is defined as the time when the estimated tsunami amplitude first exceeds 5 cm (Intergovernmental Oceanographic Commission, 2009). The additional 5 min persistence requirement is introduced in this study to suppress transient spikes and avoid false detections.

Using the same threshold and persistence criterion for both the reference and predicted signals, we obtain τ_g^{true} and τ_g^{pred}, respectively, at each gauge g. The per-gauge signed arrival lag is then defined as $e_g = (\tau_g^{\mathrm{pred}} - \tau_g^{\mathrm{true}})\,\Delta t$ (seconds), where $e_g < 0$ indicates early prediction and $e_g > 0$ late prediction. We report two complementary statistics. The mean absolute arrival-time error is

$$\mathrm{ATE} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}|e_g|, \tag{11}$$

which quantifies the typical magnitude of timing mismatch. The mean signed arrival lag is

$$\mathrm{SAL} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e_g, \tag{12}$$

which reveals systematic early or late bias. Here, $\mathcal{G}$ denotes the subset of gauges for which both τ_g^{true} and τ_g^{pred} are finite. A near-zero SAL does not imply a small ATE, because positive and negative lags can cancel. We also track the rate of buoys that are missing an arrival time and are excluded from the calculation.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

2.4.3 Band-wise spectral energy fraction error (BEE)

We verify the spectral accuracy of the field by calculating the 2D FFT power spectrum of η at each time step. After removing the mean over wet points and setting dry points to zero, we calculate the radial wavenumber $\kappa \in [0, 1]$ and the power $P(\kappa)$. Let E_{tot} be the total spectral energy within $\kappa \leq \kappa_{\mathrm{cut}}$ (here $\kappa_{\mathrm{cut}} = 0.50$), and let $f_b(t) = E_b(t)/E_{\mathrm{tot}}(t)$ be the energy fraction in band b. Using three bands (Low: $[0, 0.15]$, Mid: $(0.15, 0.35]$, High: $(0.35, 0.50]$), we report

$$\mathrm{BEE} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{3}\sum_{b\in\{L,M,H\}} |f_b^{\mathrm{pred}}(t) - f_b^{\mathrm{true}}(t)|. \tag{13}$$

The BEE is a measure of how well the surrogate model represents the wave energy. A lower BEE means the surrogate preserves this well. Low band covers the tsunami wavefront, Mid band covers the refraction pattern of the waves, and High band covers short-wavelength features from coastal reflection.

3 Results

3.1 Overall evaluation performance

We quantify the generalization in two aspects by considering (i) magnitude extrapolation and (ii) changes in the epicenter location. We train on data from earthquakes of lower magnitudes M1–M3 and only three epicenters out of four and use a portion of this data set for our validation. We then evaluate the trained emulator on three held-out test settings: Test-M (magnitude extrapolation) – Ep 2–Ep 4 with the hold-out magnitude M4 only. Test-E (unseen epicenter) – Ep 1 with in-distribution magnitudes (M1–M3). Test-EM (combined extrapolation) – Ep 1 with the hold-out magnitude M4. This design makes it possible to identify whether prediction changes result from magnitude extrapolation, location/orientation shift, or their combination.

Table 3 summarizes performance across four model configurations using field-wise errors (RMSE_η, $\mathrm{RMSE}_{\mathrm{avg}}$) and hazard-relevant diagnostics, namely mean absolute arrival-time error (ATE), signed arrival lag (SAL; used as a directional bias diagnostic), and spectral-band energy error (BEE) (see Sect. 2.4 for metric definitions). Because each test setting comprises multiple scenarios with varying source parameters, we report mean $\pm$ standard deviation over scenarios in each test list. For ATE, statistics are computed only over scenarios with a detected first arrival (non-NaN ATE), and we report the corresponding valid count N_{ATE} alongside the total scenario count N.

Figs. 7–9 summarize generalization performance of the selected model across the three test splits. Across all three held-out splits, holdout scenarios exhibit degraded performance relative to the training-domain validation set in RMSE_η and NRMSE_η (panel (a), subpanels (a1)–(a2) in Figs. 7–9). The same pattern is also evident for MAE_η in Test-M and Test-EM, whereas the train–holdout difference is not statistically significant in Test-E ($p = 0.09$; subpanel (a3) of Fig. 7). For SAL, the train–holdout difference is not statistically significant in Test-E or Test-M ($p = 0.06$ and 0.79, respectively; subpanel (a4) of Figs. 7 and 8), but is significant in Test-EM ($p = 0.002$; subpanel (a4) of Fig. 9). Overall, the held-out splits separate the intended extrapolation regimes. Test-E results show the model's generalization for an unseen epicenter and orientation, but in the range of trained magnitudes. Selected has the lowest RMSE_η (0.0278 m, Table 3) and substantially reduces ATE to 7.5 min in comparison with

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

14.4 min for Reference. Relative to the 8-layer Reference model, the 10-layer Selected model better models the large-scale wave evolution and first-arrival structure for an unseen epicenter. However, Selected does not minimize BEE. Its mean BEE is the highest among the compared models (0.0387), whereas w/o DC gives the lowest value (0.0251). The SAL distribution is centered close to zero (Fig. 7a4), which indicates almost no early or late bias. The number of detected arrivals is lower than the total buoy count. Some of the lower-magnitude scenarios in Test-E produce water-level increases below 0.05 m at buoys near the Korean coast.

Test-M evaluates the model's ability to extrapolate an 8.0 magnitude from the same source locations. Selected shows the lowest mean RMSE_{η} (0.0578 m) and $\mathrm{RMSE}_{\mathrm{avg}}$ (0.0279 m). +M272 attains the lowest mean absolute ATE (7.7 min), but with a higher RMSE_{η} (0.0644 m) than Selected. In Fig. 8b–c, the validation cases ($M_w = 7.4$–7.8) are plotted together with the Test-M holdout cases ($M_w = 8.0$) to show the change across the full magnitude range. The buoy-level regression does not show a significant linear relationship ($R^2 = 0.001$, $p = 0.47$). However, in Test-M, the $M_w = 8.0$ holdout cases lie slightly on the positive side of zero SAL.

Test-EM is evaluating the most challenging predictions, which involve the combination of epicenter/orientation shift and magnitude extrapolation. Selected provides the lowest mean RMSE_{η} (0.0763 m) and $\mathrm{RMSE}_{\mathrm{avg}}$ (0.0382 m), while +M272 achieves the smallest mean absolute ATE (10.6 min) but with higher RMSE_{η} (0.0889 m). At the magnitude 8.0, some of the predicted wavefronts were detected earlier than the true results due to the overestimated wave amplitude.

We compare the degradation between splits. The median RMSE_{η} between validation and holdout sets does not change dramatically, going from about 0.02 m (Fig. 7a1) to about 0.03 m in Test-E. Larger changes are observed for Test-M and Test-EM with median RMSE_{η} values ranging from 0.05 to 0.07 m at $M_w = 8.0$ (Table 3). This indicates that magnitude extrapolation is a more difficult task than spatial generalization. All splits have RMSE_{η} that increases with magnitude (panel b), with the largest correlation coefficients being observed for Test-EM where both epicenter shift and magnitude extrapolation are combined. In contrast, the SAL in panel (c) shows no clear dependence on magnitude across all three splits. The median SAL remains close to zero in Test-E and Test-EM, whereas the $M_w = 8.0$ holdout cases in Test-M are shifted modestly toward positive SAL.

Some standard deviation values in Table 3 show more than 50% of the mean value for RMSE_{η}, driven by variation in source-parameter combinations. The large standard deviations in ATE also reflect variability across the test sets rather than numerical instability of the surrogate, because some cases have much larger timing mismatch than others. This produces a wide ATE distribution. N_{ATE} ranges from 85 to 101 in Test-E, reflecting the lower-magnitude detection issue noted above, while Test-M and Test-EM show near-complete detection (156–159 and 54 out of 54, respectively).

For reference, we also trained standard FNO baselines under the same training sets. Across the tests, the factorized model generally outperformed the standard FNO. The detailed comparisons are provided in the Supplement (Tables S1–S3). Unless otherwise noted, the remaining results below are reported for the Selected model.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

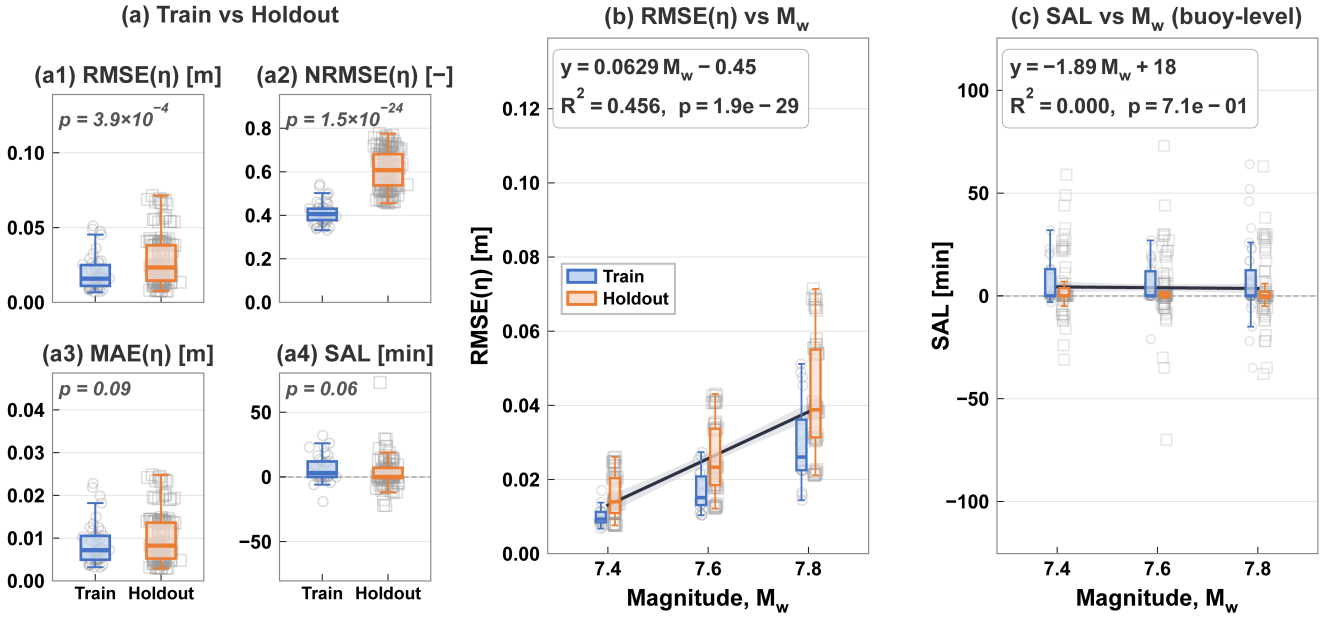

Figure 7. Generalization performance of the selected model on Test-E (Ep 1 holdout; $M_w = 7.4$–7.8). Panel (a) compares scenario-level error distributions between the training-domain validation set and the held-out test set. Subpanels (a1)–(a4) show RMSE_η, NRMSE_η, MAE_η, and SAL, respectively. Two-sided Mann–Whitney U p-values are reported. Semi-transparent open gray markers indicate individual samples overlaid on the boxplots (circles: training-domain validation; squares: held-out scenarios). Panel (b) shows RMSE_η as a function of M_w with fitted linear trends. Panel (c) shows buoy-level signed arrival lag e_g versus M_w. Negative values indicate earlier predicted arrivals than the reference, and positive values indicate later arrivals. The dashed horizontal line marks zero bias.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

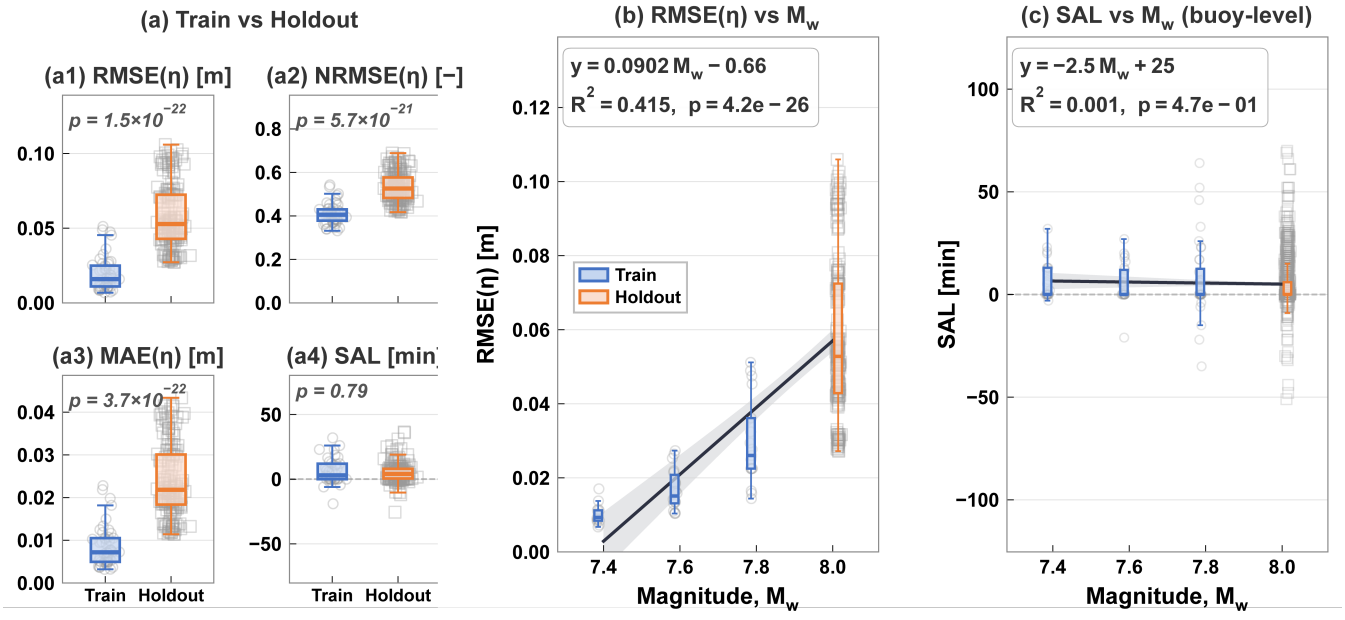

Figure 8. As in Fig. 7, but for Test-M. The holdout set consists of $M_w = 8.0$ scenarios at trained epicenters Ep 2–Ep 4, while panels (b) and (c) plot the training-domain validation cases ($M_w = 7.4$–7.8) together with the Test-M holdout cases ($M_w = 8.0$) to show the trend across the full magnitude range.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

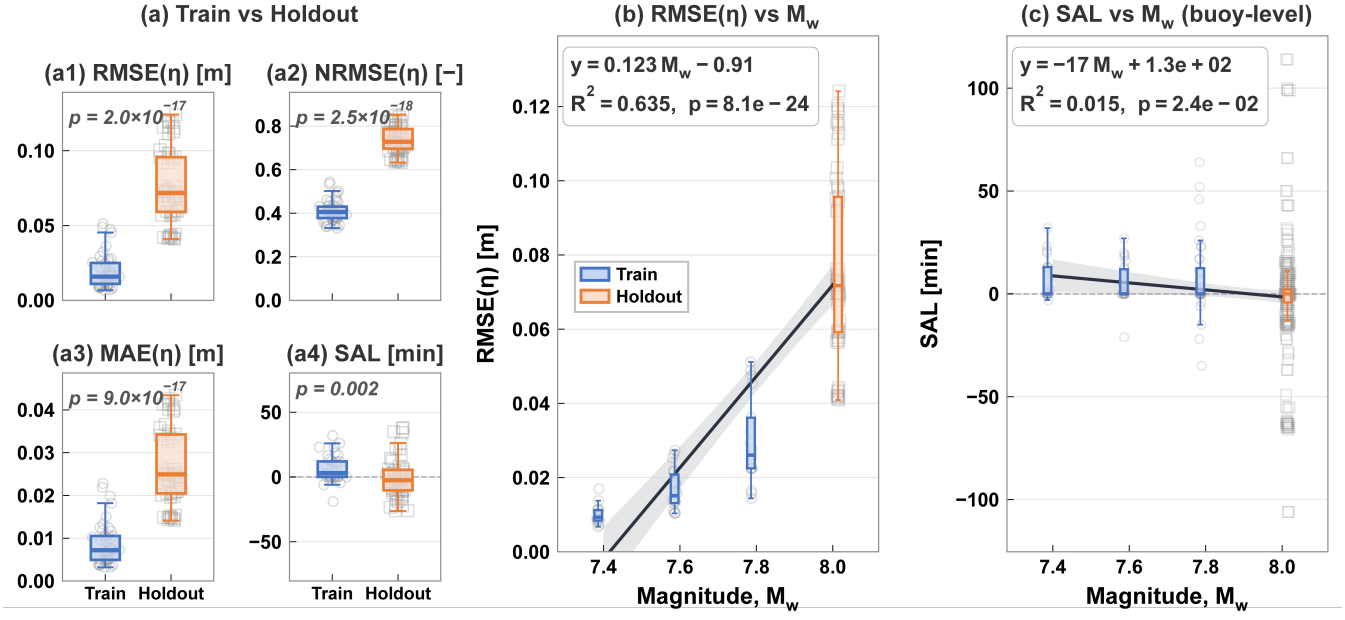

Figure 9. As in Fig. 7, but for Test-EM. The holdout set consists of unseen-epicenter Ep 1 scenarios at $M_w = 8.0$, while panels (b) and (c) plot the training-domain validation cases ($M_w = 7.4$–7.8) together with the Test-EM holdout cases ($M_w = 8.0$) to show the trend across the full magnitude range.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Table 3. Split-wise held-out test performance (mean ± std over test scenarios). For each metric, the best (lowest mean) value within each split is highlighted in bold. ATE (mean absolute arrival-time error) statistics are computed only over scenarios with a detected first arrival; N_{ATE} denotes the valid count and N denotes the number of scenarios for which at least one such valid gauge was available. Overall aggregates all three held-out test sets by concatenation ($N = 378 = 162 + 162 + 54$). Model configurations: Selected: 10-layer, $\lambda_{\mathrm{cont}} = 10$, $\lambda_{\mathrm{dc}} = 100$; Reference: 8-layer, $\lambda_{\mathrm{cont}} = 0.5$, $\lambda_{\mathrm{dc}} = 100$; w/o DC: 8-layer, $\lambda_{\mathrm{cont}} = 0.5$, $\lambda_{\mathrm{dc}} = 0$; +M272: 8-layer, $\lambda_{\mathrm{cont}} = 0.5$, $\lambda_{\mathrm{dc}} = 100$, M=272 modes.

Test split	Model	RMSE_η (m)	ATE (min)	BEE	$\mathrm{RMSE}_{\mathrm{avg}}$	N	N_{ATE}
Test-E (Ep 1, M1–M3)	Selected	**0.0278 ± 0.0157**	**7.5 ± 10.2**	0.0387 ± 0.0143	**0.0125 ± 0.0072**	162	86
	Reference	0.0289 ± 0.0164	14.4 ± 17.7	0.0287 ± 0.0109	0.0129 ± 0.0076	162	101
	w/o DC	0.0292 ± 0.0170	16.1 ± 18.7	**0.0251 ± 0.0081**	0.0130 ± 0.0078	162	85
	+M272	0.0287 ± 0.0178	14.9 ± 19.5	0.0297 ± 0.0123	0.0130 ± 0.0082	162	99
Test-M (Ep 2–Ep 4, M4 only)	Selected	**0.0578 ± 0.0214**	8.5 ± 11.3	**0.0164 ± 0.0066**	**0.0279 ± 0.0122**	162	156
	Reference	0.0751 ± 0.0281	9.7 ± 9.1	0.0306 ± 0.0114	0.0349 ± 0.0146	162	159
	w/o DC	0.0738 ± 0.0274	8.2 ± 8.6	0.0271 ± 0.0063	0.0344 ± 0.0143	162	158
	+M272	0.0644 ± 0.0236	**7.7 ± 6.9**	0.0281 ± 0.0124	0.0307 ± 0.0130	162	158
Test-EM (Ep 1, M4 only)	Selected	**0.0763 ± 0.0248**	12.1 ± 14.4	**0.0312 ± 0.0107**	**0.0382 ± 0.0123**	54	54
	Reference	0.0836 ± 0.0257	11.7 ± 10.0	0.0360 ± 0.0057	0.0414 ± 0.0127	54	54
	w/o DC	0.0850 ± 0.0297	12.1 ± 9.8	0.0392 ± 0.0073	0.0418 ± 0.0146	54	54
	+M272	0.0889 ± 0.0322	**10.6 ± 11.1**	0.0351 ± 0.0125	0.0438 ± 0.0158	54	54
Overall (concat. all held-out tests)	Selected	**0.0476 ± 0.0268**	**8.9 ± 11.7**	0.0281 ± 0.0152	**0.0227 ± 0.0141**	378	296
	Reference	0.0565 ± 0.0336	11.5 ± 12.8	0.0306 ± 0.0108	0.0264 ± 0.0167	378	314
	w/o DC	0.0563 ± 0.0337	11.2 ± 13.0	**0.0280 ± 0.0086**	0.0263 ± 0.0168	378	297
	+M272	0.0526 ± 0.0318	10.5 ± 13.3	0.0298 ± 0.0125	0.0250 ± 0.0162	378	311

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

3.2 Wavefield and Buoy Metrics

For a representative test event, Fig. 10 compares spatial snapshots of the free-surface elevation $\eta(\mathbf{x}, t)$ (Ep 1, $M_w = 8.0$, an unseen epicenter location and extrapolated magnitude) at three propagation stages (10, 30, and 60 minutes post-rupture). Columns correspond to the three propagation stages ($t = 10, 30$, and 60 min). Rows show the COMCOT reference, the F-FNO prediction, and the spatial error (Prediction − Truth), respectively. The surrogate captures the main wave patterns well even under Test-EM conditions. Errors at $t = 10$ minutes are limited to near the source location and expand along the propagation zone as the wave propagates (t=30–60 min). The error field alternates between positive and negative values, but does not show a one-sided amplitude bias.

To see how well the surrogate performs at specific locations, we compare the full time series of free-surface elevation at virtual buoys. Fig. 11 shows $\eta(t)$ at three offshore (B080, B400, and B720) and three nearshore buoys (Hanul, Imwon, and Pohang). The surrogate reproduces the main wave patterns and peak timing well from offshore to nearshore locations. Some discrepancies appear as small differences in peak free-surface elevation. The Pohang shows slightly larger errors, probably due to the geometry of the harbor near the site. Velocity comparisons at the same stations are provided in the Supplement (Figs. S1–S2).

Arrival-time and spectral diagnostics for the Test-EM cases are presented in Fig. 12. The scatter plot in Fig. 12a shows $R^2 = 0.736$ with a mean absolute error of 13.3 minutes. The mean signed bias of -2.4 minutes indicates a weak tendency toward early arrival prediction. This is expected from the near-zero signed lag observed in the analysis (Figs. 7–9). Scatter increases for arrival times beyond approximately 1.5 hours, because those gauges are farther from the source. Fig. 12b shows the Test-EM cases' high-frequency energy fraction averaged over all test scenarios. Both the COMCOT truth and the surrogate show more high-frequency content as time goes on, produced by shorter-wavelength features from coastal reflection and diffraction. The surrogate tracks the overall trend of the reference throughout the 3.3-hour rollout, though with moderate overestimation between approximately 0.5 and 1.5 hours. After 2 hours, the two signals reach similar levels (∼0.25–0.30). The larger ± 1 standard deviation envelope for the surrogate indicates greater scenario-to-scenario variability in high-frequency content.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

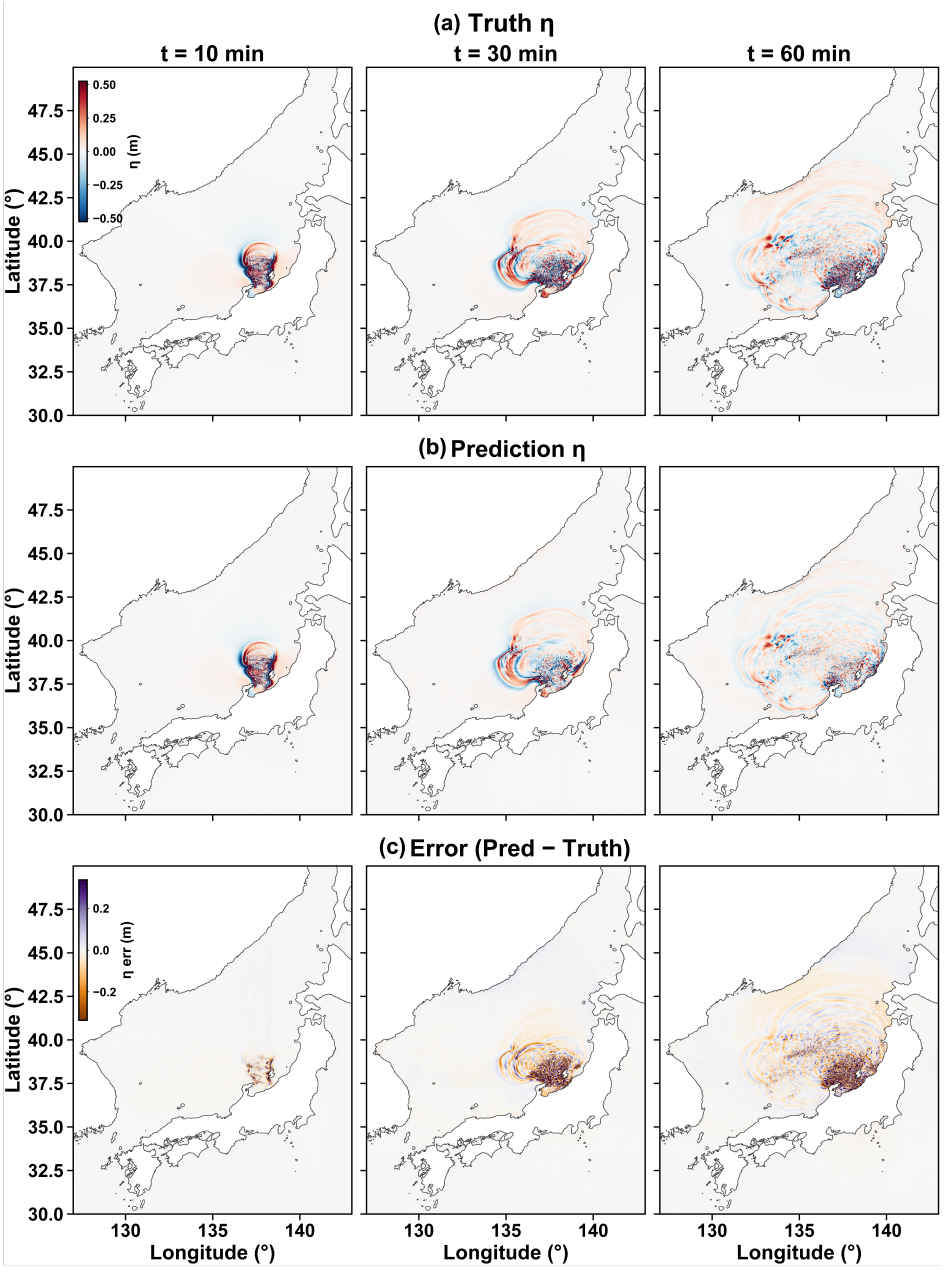

Figure 10. Snapshots of free-surface elevation η at $t = 10$, 30, and 60 min for a test event (Ep 1, $M_w = 8.0$). Columns show the three propagation stages, and rows show the COMCOT reference, the F-FNO prediction, and the pointwise error (Prediction − Truth), respectively. Larger differences appear near the source region at early times and spread across the propagation zone at later times.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

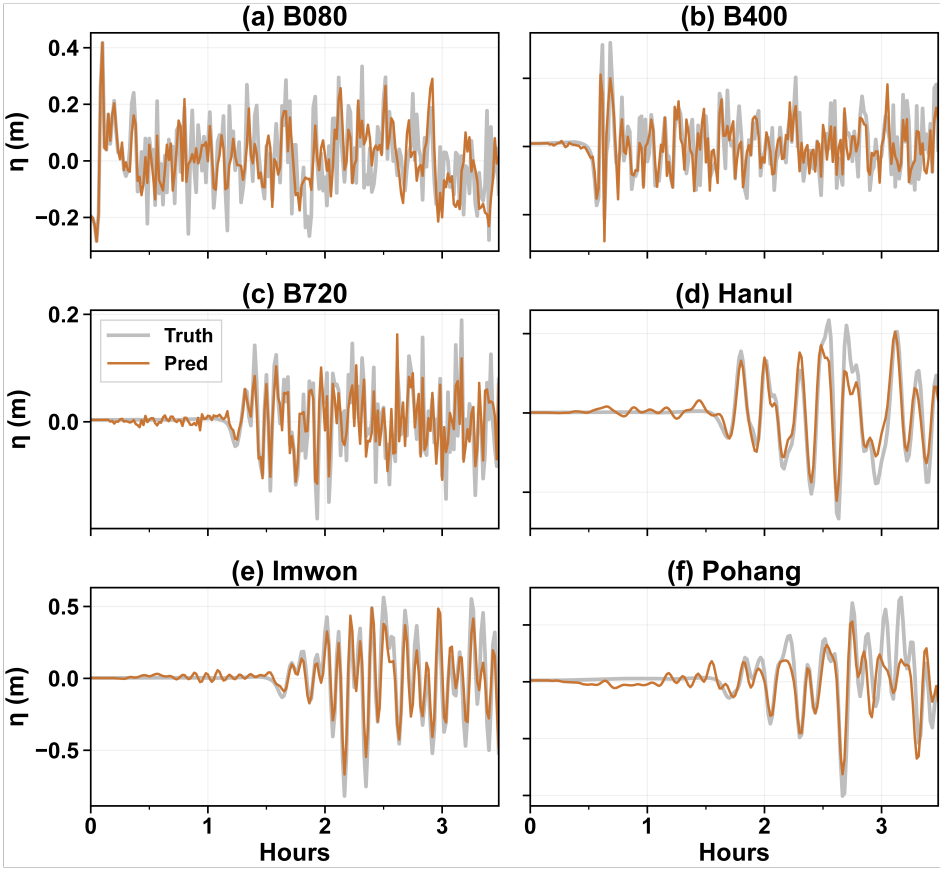

Figure 11. The surrogate prediction (Pred) versus the COMCOT truth (True) at three offshore virtual buoys (B080, B400, B720) and three nearshore virtual buoys (Hanul, Imwon, Pohang).

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

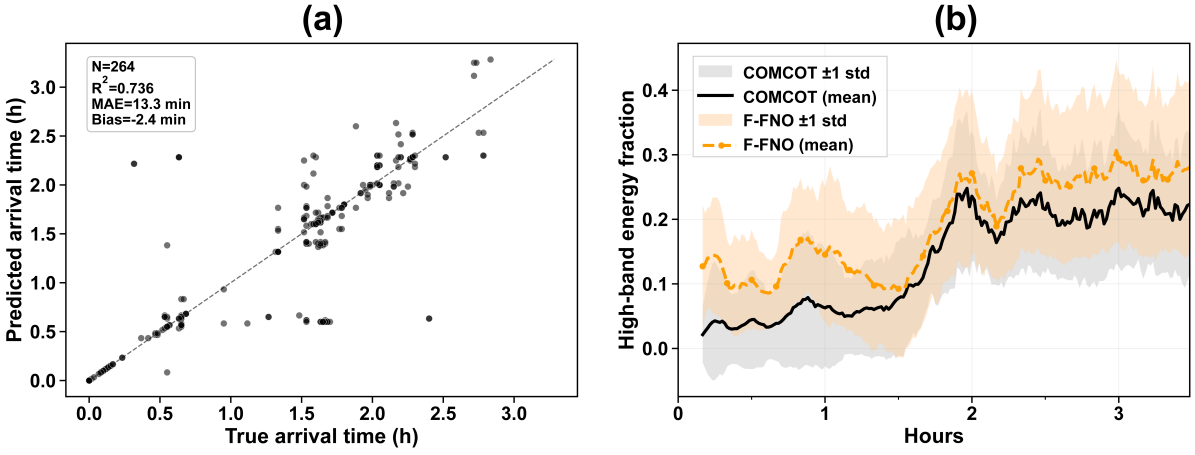

Figure 12. Arrival-time and spectral diagnostics for the Test-EM cases. (a) Comparison of predicted and true first-arrival times across valid buoy–scenario pairs. Here, $N = 264$ denotes the number of valid buoy–scenario pairs included in the scatter plot, for which both the reference and surrogate arrival times were detected under the 0.05 m threshold and 5 min persistence criterion. (b) Mean high-band energy fraction as a function of time, averaged over the buoy set for the Test-EM cases; shaded regions is ± 1 standard deviation for COMCOT and F-FNO.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

3.3 Spatial pattern of peak elevation and error

Fig. 13 shows the peak free-surface elevation $|\eta|_{\max}$ over the 3.3-hour simulation. This event (Ep 1, $M_w = 8.0$) belongs to Test-EM. The surrogate matches the spatial pattern of peak elevation well (Fig. 13a–b). In the source region around ∼38° N, 138° E, both the reference model and the surrogate show peak amplitudes more than ∼1.5 m. The surrogate also captures the concentration of wave energy toward the southeastern Korean coast (∼35–36° N), where peak elevations reach around 0.8–1.0 m. The error map (Pred–Truth) in Fig. 13c shows where the errors are largest. (i) In the open ocean, errors are generally smaller than those near the coast. (ii) Along the Japanese coast near the Ep 1 (139–141° E), the surrogate underestimates $|\eta|_{\max}$ by about 0.2–0.4 m where the COMCOT results show large peaks. (iii) Along the eastern Korean coast, errors are small and show no one-sided bias. For this event, the $\mathrm{RMSE}_{|\eta|_{\max}}$ is 0.04 m and the $\mathrm{MAE}_{|\eta|_{\max}}$ is 0.03 m.

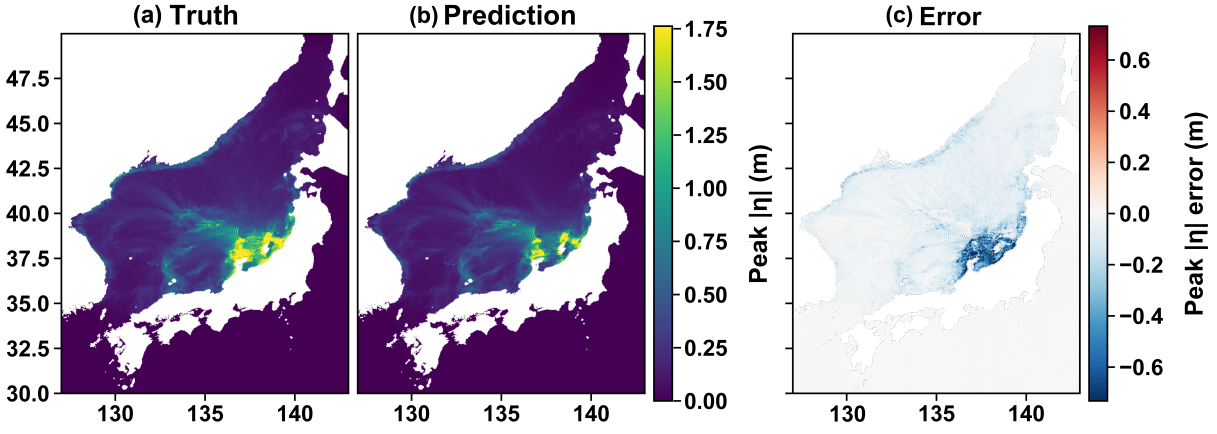

Figure 13. Peak free-surface elevation over the 3.3-hour for a test event (Ep 1, $M_w = 8.0$). (a) COMCOT $|\eta|_{\max}$ computed at each grid cell as $\max_{t\in[0,T]} |\eta(\mathbf{x}, t)|$ with $T = 3.3$ h. (b) Surrogate prediction of $|\eta|_{\max}$. (c) Error (Pred − Truth). Errors are mainly located near the source region, where the surrogate slightly underestimates peak elevation.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

3.4 Prediction stability

Because the surrogate predicts one step at a time and feeds each output back as the next input, errors can accumulate over long predictions. Keeping this error low is important for building a reliable surrogate model. Fig. 14 shows the RMSE_η evolution across all 103 validation and Test-EM scenarios for the Selected model. Error levels vary across scenarios with different source locations and magnitudes, but all 103 cases do not diverge over the full 200-step rollout. We find that error growth plateaus after about step 100 (corresponding to 100 minutes into the forecast). Error growth therefore slows down as wave energy dissipates and the basin becomes calmer. The RMSE_η at step 200 is 0.076 m (25th–75th percentile: 0.048–0.105 m). Most scenarios are in the range of centimeter-to-decimeter error at the end of the 3.3-hour forecast (rollout from $t_0 + S$ to $t_0 + S + H_r$ with inference context length $S = 10$).

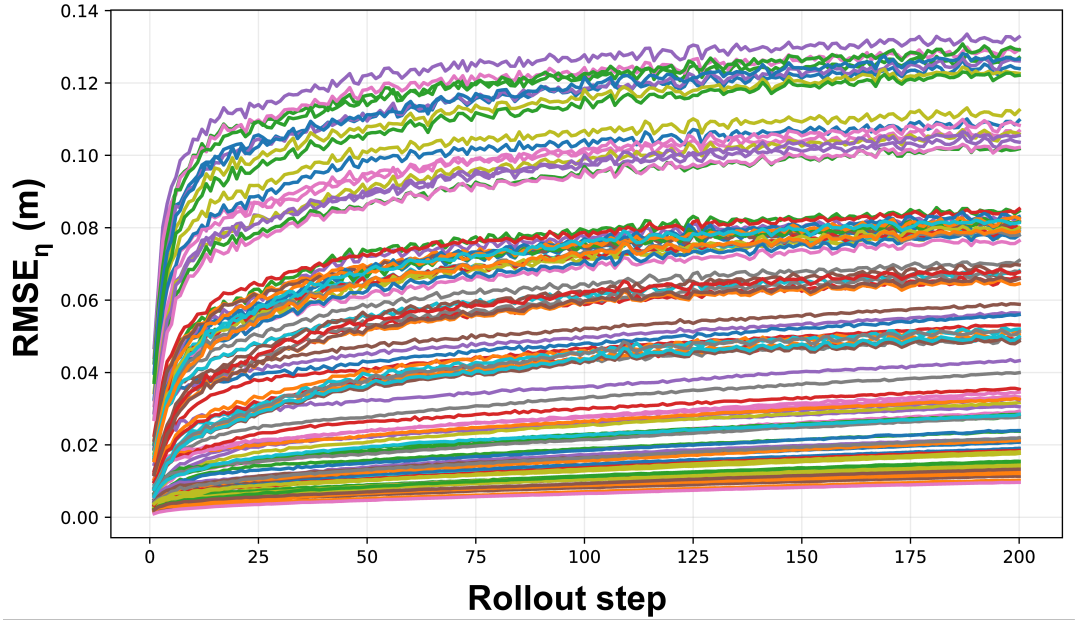

Figure 14. RMSE_η evolution over the 200-step prediction horizon for the Selected model. Each line represents a single scenario from the validation set and Test-EM split (103 scenarios total).

3.5 Computational efficiency

Results shown here are like-for-like runtime comparisons of compute-only simulation times measured with file I/O disabled for COMCOT and the surrogate. We selected ten different sample scenarios from both Ep 3 and Ep 4, covering a wide range of parameter settings and examined (i) the COMCOT wall clock simulation time on an AMD Ryzen 9 7950X CPU when no output is written, and (ii) the surrogate rollout times on two different GPUs NVIDIA RTX 5070 Ti and NVIDIA B200 when all output writing is disabled. Note that for the single-layer (Layer 1) model used here, COMCOT runs with single-core CPU. The

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

given surrogate rollout times correspond to the times of the pure autoregressive forecast when we include the context burn-in time (H_{bi}) plus $H_r = 200$ forecast steps. We repeated the surrogate rollout for each scenario 3 times and we ignored the first one which serves as a GPU warm-up. The time averages are taken over the remaining two simulations for each scenario. The runtime for the ten different scenarios for all the models considered in this study is shown in Table 4.

Table 4. Compute-only runtime comparison (H_r = 200 rollout, file I/O disabled for all methods). COMCOT was executed on an AMD Ryzen 9 7950X CPU in single-core mode (the single-layer configuration used here does not employ parallelization); the F-FNO surrogate was evaluated on NVIDIA RTX 5070 Ti and NVIDIA B200 GPUs. Surrogate times are averages over two post-warm-up runs per scenario. PCOMCOT timing is a single-scenario reference on different hardware (see table footnote).

Method	Platform	Time (s)	Speedup
COMCOT	CPU (Ryzen 9 7950X; single core)	91.8 ± 3.9	—
PCOMCOT[†]	CPU (Xeon E5-2695 v3; 16 cores)	632.0	—
F-FNO	GPU (NVIDIA RTX 5070 Ti)	12.0 ± 0.2	7.6×
F-FNO	GPU (NVIDIA B200)	8.5 ± 0.4	10.7×

[†]Single-scenario reference timing. PCOMCOT solves Boussinesq-type dispersive equations (Zhu et al., 2024) and was run on older hardware than COMCOT; its runtime is not directly comparable but is included to illustrate the computational cost of higher-fidelity numerical solvers.

The GPU-accelerated surrogate has mean inference times of 12.0 s (RTX 5070 Ti) and 8.5 s (B200) per scenario. Relative to COMCOT's mean runtime of 91.8 s, the surrogate provides approximately 7.6× and 10.7× speedup, respectively. Both methods have low runtime across all test cases (coefficient of variation <5 %), as expected for deterministic computations. For reference, a single equivalent scenario computed with the Boussinesq-type dispersive solver PCOMCOT (Zhu et al., 2024) on using the 16 CPU cores, the actual time required for the calculation was about 632 s, which is about 50–75 times slower than with our surrogate (see Table 4). This simulation should be considered approximate due to differences between different computer systems. This run time reference should be treated with caution as there are differences between the equations of motion and numerical schemes in PCOMCOT compared to COMCOT. A brief description of the two codes is given in the Supplement (Table S5). Such a speedup makes it possible to test ensemble forecasts of hundreds of scenarios within operational time scales.

4 Discussion

4.1 Surrogate fidelity and speed

The model can propagate a tsunami wave for a sustained time with low error of key hazard parameters such as first-arrival timing (8–12 min) and peak elevation ($\mathrm{RMSE}_{|\eta|_{\max}}$ of 0.04 m). The F-FNO reduces RMSE_{η} by 24–36 % and spectral energy distribution error by 75–84 % compared to a standard FNO model trained under identical training samples. This is primarily through higher Fourier-mode retention (M=256 vs. M=64) of F-FNO within same GPU memory constraints. Full results are reported in the Supplement (Tables S1–S3). The F-FNO model fidelity suggests that the surrogate can be used as a screening

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

tool for hazard assessment and as a rapid scenario test within the training domain. The surrogate emulated 3.3 hours of tsunami propagation at 8.5–12 s compared to 91 s of COMCOT simulation time. The surrogate can evaluate the full 864-scenario database in less than 3 hours, compared with approximately 22 hours for COMCOT. This acceleration makes ensemble analysis more feasible and comprehensive by testing an extended range of scenarios.

The surrogate operator can be adopted as a fast outer grid propagation model that provides offshore boundary conditions for nearshore simulations along the Korean coast. Basin-scale propagation is first simulated to provide wavefields at the boundaries of refined nearshore grids, which resolve local processes such as shoaling and inundation. For site-specific hazard assessments along the Korean east coast, the surrogate can be used to run a large number of scenarios effectively. The surrogate is also differentiable and thus could be used for inversions or parameter estimation although this was not explored here. The semi-enclosed geometry of the East Sea produces complex reflection and interference patterns, which the F-FNO's global spectral mixing is well suited to capture. However, the same geometry limits the range of propagation distances and directivity patterns represented in the training data.

We cannot compare quantitatively our results with other works on tsunami surrogate models because each work is characterized by different computational settings, such as the spatial domain, discretization and validation criteria. We therefore only provide here a qualitative assessment of the relative merits of our approach. As mentioned before, Sarri et al. (2012) and Guillas et al. (2018) in a more recent version, have been able to achieve centimeter accuracy for the peak wave height in a pre-defined position, offshore or at the shore, by using statistical emulators for tsunami but they do not compute the wavefield. Mulia et al. (2022) used a CNN to compute the inundation depth with a mean absolute error of 0.2–0.5 m for megathrust slip sources. In another study, Fukutani and Motoki (2025) directly estimated inundation depth distributions from fault slip parameters and obtained an RMSE of about 1.2 m for validation. Here, they deal with onshore inundation which is a quite different quantity than the quantities of interest to us in offshore wave propagation. The main difference here is that others compute a specific quantity, such as the peak wave height at a given point, the inundation depth or time series at a given fixed point in space, while here we compute the full wavefield in space and time. Achieving the full wavefield requires to generate a large number of training data and we used 864 full-field simulations in our work.

4.2 Generalization and coastal limitations

Application to open-ocean areas requires training results based on scenario databases specialized to the study area. The current scenario database spans magnitudes of M_w 7.4 to 8.0 across four epicenter locations along the eastern Japan margin. However, four epicenters cover a limited area only, since all source locations are distributed around the eastern Japan margin. To achieve a more general tsunami propagation emulator in this area, a variety of source locations and extrapolated magnitude ranges should be included in the training samples. In particular, magnitude extrapolation produces larger errors than spatial generalization (Table 3), which suggests that extending the training database to higher magnitudes would be necessary to reliably predict scenarios beyond the current training range.

The surrogate is designed for basin-scale tsunami propagation, so it is not intended to emulate nearshore tsunami propagation due to several limitations from the initial training design setting. The 5 m depth threshold for the training set is adopted to focus

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

on the basin-scale tsunami propagation immediately after fault generation, but this threshold can impact how coastal reflection is considered. Since the surrogate is not trained with the exact shoreline, the model may not learn the full wave reflection process. A flux based formulation (momentum Hu, Hv) can solve this issue. Using a nested higher-resolution grid is the most appropriate direction to emulate such reflection correctly.

Currently, a horizontal grid spacing of approximately 3.2 km represents the major wavelengths of deep-sea tsunamis, ranging from tens to hundreds of kilometers, by a factor of ten or more. To retrain surrogate models on higher-resolution grids, more Fourier modes must be included to capture finer bathymetry details, and the memory capacity required will also increase. On the other hand, if we were to use a coarser horizontal resolution the bathymetric gradients that steer the wavefront through refraction would be somewhat smoothed and this may lead to a loss of accuracy in coastal locations for the modeling of tsunami arrivals. This is a topic that we will need to investigate further in the future, in particular when modeling tsunamis entering the shelf where larger gradients are present, or modeling tsunamis passing through narrow channels with steeper channel sides. To evaluate the scalability of the Fourier neural operator for general tsunami surrogates, future work should extend the training database to other basin geometries, such as open-ocean domains where propagation distances are longer and boundary reflections play a smaller role.

4.3 Physical consistency

The loss terms help the surrogate produce more physically reasonable results, but they do not force the model to follow the governing equations exactly. In particular, the continuity loss function helps reduce mass-balance errors, but how much it helps depends on how strongly it is weighted. Because the model predicts one step at a time, small errors can build up as the rollout becomes longer. The RMSE of η increases from about 0.02 m at step 10 to 0.076 m at step 200 (Fig. 14). After about step 100, the error growth begins to level off as the wave energy decreases. The rollout therefore remains stable, although part of this leveling-off is also related to the decay of the tsunami itself. The error level also varies with source location and earthquake magnitude.

The Selected model gave the lowest RMSE for Test-E but the largest BEE. This suggests that it is possible to achieve a good representation of the spatial and temporal water level distribution without necessarily having an equal accuracy in the spectral representation. By contrast, the model without the divergence constraint (w/o DC) gave the lowest BEE, but its overall field errors were larger (Table 3). There is a balance between accuracy of the overall field prediction versus the spectral accuracy in the test cases. In cases where a basin-scale surrogate model is desired for nearshore accuracy, the results would then be combined with nested higher-resolution local models that use surrogate boundary conditions.

The evaluation implies that the various accuracy measures reflect different aspects of the errors. The choice of minimizing ATE and RMSE for the tsunami wave path errors needs to be done with caution, since a reduction in one measure often leads to an increase in the other. The training and evaluation was made against a single reference solution made available by the open-source finite difference code COMCOT. In order to get some idea of the uncertainty of the model, it would be desirable to compare with other codes using the Boussinesq equation. This is particularly important for short wavelengths, since it is at these wavelengths that the Boussinesq approximation is expected to be most relevant.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

The present version of COMCOT solves the linear shallow-water equations and hence frequency dispersion is not considered. While dispersion effects are relatively small in the deep basin of the East Sea (depths over 3000 m), they can become non-negligible over the Korean continental shelf where depths decrease to 100–200 m. Although the surrogate reproduces the solver output accurately, it retains the same physical assumptions embedded in COMCOT. If more accurate modeling of dispersive effects is required, the surrogate would need to be retrained on output from a more accurate solver. The errors presented in this paper are therefore surrogate-vs-solver errors rather than absolute prediction errors.

The surrogate does not currently give any quantitative measure of the uncertainty of the field predictions. In operational use of this tool for modeling tsunamis impacting shorelines to determine the extent of the area at risk, this would be important. The confidence in the model could be established by performing an ensemble of simulations with small amounts of noise injected into the initial conditions (e.g. at the source with small variations in the magnitude and location of the epicenter, the focal depth, etc.). Alternatively, a set of surrogates could be developed, and the final prediction of the effect of the tsunami for the study location could be determined by ensemble averaging.

5 Conclusions

This study develops a basin-scale tsunami surrogate based on a Factorized Fourier Neural Operator (F-FNO), trained on a systematically constructed COMCOT scenario database comprising 864 logic tree scenarios for the East Sea. Across held-out test splits, the surrogate achieves RMSE_η of 2–8 cm and first-arrival timing errors of 8–12 min, including scenarios with unseen epicenter locations and extrapolated Mw 8.0. At approximately 8.5–12 s per scenario on a single GPU, the surrogate achieves up to 10.7× speedup relative to COMCOT. This speedup allows the full 864-scenario ensemble to be evaluated in under 3 hours. With the surrogate, we can sample epistemic uncertainties far more densely than would be practical with the physics-based solver alone.

The surrogate functions as a rapid offshore boundary condition generator, facilitating the creation of nested grids that enhance the resolution of site-specific hazard studies. By accelerating the outer grid propagation, it enables site-specific hazard studies to apply highly resolved nearshore models to the identified critical scenarios without limiting scenario coverage due to computational constraints. The peak offshore elevation maps and first-arrival times produced here offer a basis for coastal emergency planning at the selected reference points.

Accurately reproducing the space–time wave field does not necessarily ensure correct partitioning of spectral energy. Both aspects therefore need to be assessed independently in future applications. An important next step will be testing the method's robustness on larger domains and more extreme earthquake scenarios, which will provide understanding into its scalability and reliability. Coupling with coastal inundation models is another future step.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Code availability. The F-FNO surrogate code, including training, inference, and evaluation scripts, pretrained model weights, scenario parameter table (864 logic tree configurations), COMCOT control file template, and train/val/test split list files, is archived on Zenodo (Kim et al., 2026, https://doi.org/10.5281/zenodo.19198928).

Data availability. The Test-EM evaluation dataset (54 NetCDF scenarios; unseen epicenter combined with extrapolated M_w 8.0, 44.1 GB) is included in the same Zenodo archive (Kim et al., 2026, https://doi.org/10.5281/zenodo.19198928) and is sufficient to reproduce all Test-EM inferences reported in this paper. The full training dataset (approximately 642 GB, 864 scenarios) can be regenerated from the provided scenario parameters using COMCOT and is available from the corresponding author upon reasonable request. The bathymetric grid was derived from GEBCO (https://www.gebco.net/).

Author contributions. JK: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing (original draft, review and editing). MJK: Data curation, Investigation, Methodology, Software, Visualization, Writing (original draft, review and editing). STO: Data curation, Investigation, Software. SS: Conceptualization, Funding acquisition, Project administration, Resources, Software, Supervision, Visualization, Writing (review and editing).

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors used AI-assisted tools for coding and debugging during model development and for wording adjustments. All analysis, interpretation, and final editing were carried out by the authors.

Financial support.

This research was supported by the Ministry of Science and ICT, Republic of Korea, through the National Research Foundation of Korea (grant Nos. RS-2024-00356663 and RS-2024-00444224) and the Advanced GPU Utilization Support Program (no. 02-26-01-0368).

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

References

Annaka, T., Satake, K., Sakakiyama, T., Yanagisawa, K., and Shuto, N.: Logic-tree approach for probabilistic tsunami hazard analysis and its applications to the Japanese coasts, Pure and Applied Geophysics, 164, 577–592, https://doi.org/10.1007/s00024-006-0174-3, 2007.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks, Advances in neural information processing systems, 28, 2015.

Bernard, E. and Titov, V.: Evolution of tsunami warning systems and products, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 2015.

Chattopadhyay, A., Gray, M., Wu, T., Lowe, A. B., and He, R.: OceanNet: a principled neural operator-based digital twin for regional oceans, Scientific Reports, 14, 21 181, 2024.

Cho, Y.-S., Sohn, D.-H., and Lee, S. O.: Practical modified scheme of linear shallow-water equations for distant propagation of tsunamis, Ocean Engineering, 34, 1769–1777, 2007.

Choi, B. H., Pelinovsky, E., Hong, S. J., and Woo, S. B.: Computation of tsunamis in the East (Japan) Sea using dynamically interfaced nested model, Pure and Applied Geophysics, 160, 1383–1414, 2003.

Choi, B. H., Pelinovsky, E., Kim, D. C., Kim, K. O., and Kim, K. H.: Three-dimensional simulation of the 1983 central East (Japan) Sea earthquake tsunami at the Imwon Port (Korea), Ocean Engineering, 35, 1545–1559, 2008.

Choi, B.-J., Jin, H. S., and Lkhagvasuren, B.: Applications of the Fourier neural operator in a regional ocean modeling and prediction, Frontiers in Marine Science, 11, 1383 997, 2024.

Davies, G., Griffin, J., Løvholt, F., Glimsdal, S., Harbitz, C., Thio, H. K., Lorito, S., Basili, R., Selva, J., Geist, E., and Baptista, M. A.: A global probabilistic tsunami hazard assessment from earthquake sources, Geological Society, London, Special Publications, 456, 219–244, https://doi.org/10.1144/SP456.5, 2018.

Earthquake Research Committee: Long-term Evaluation of Seismic Activity along the Eastern Margin of the Sea of Japan, Tech. rep., Headquarters for Earthquake Research Promotion, Government of Japan, (in Japanese), 2003.

Fukutani, Y. and Motoki, M.: A neural network-based surrogate model for efficient probabilistic tsunami inundation assessment, Coastal Engineering, 200, 104 767, 2025.

Grezio, A., Babeyko, A., Baptista, M. A., Behrens, J., Costa, A., Davies, G., Geist, E. L., Glimsdal, S., González, F. I., Griffin, J., Harbitz, C. B., LeVeque, R. J., Lorito, S., Løvholt, F., Omira, R., Mueller, C., Paris, R., Parsons, T., Polet, J., Power, W., Selva, J., Sørensen, M. B., and Thio, H. K.: Probabilistic tsunami hazard analysis: Multiple sources and global applications, Reviews of Geophysics, 55, 1158–1198, https://doi.org/10.1002/2017RG000579, 2017.

Guillas, S., Sarri, A., Day, S. J., Liu, X., and Dias, F.: Functional emulation of high resolution tsunami modelling over Cascadia, The Annals of Applied Statistics, 12, 2023 – 2053, https://doi.org/10.1214/18-AOAS1142, 2018.

Heidarzadeh, M., Ishibe, T., Gusman, A. R., and Miyazaki, H.: Field surveys of tsunami runup and damage following the January 2024 Mw 7.5 Noto (Japan Sea) tsunamigenic earthquake, Ocean Engineering, 307, 118 140, https://doi.org/10.1016/j.oceaneng.2024.118140, 2024.

Intergovernmental Oceanographic Commission: Operational Users Guide for the Pacific Tsunami Warning and Mitigation System (PTWS), IOC Technical Series 87, UNESCO, Intergovernmental Oceanographic Commission, Paris, first edition, January 2009, 2009.

Japan Society of Civil Engineers: Tsunami Assessment Method for Nuclear Power Plants in Japan, Tech. rep., Japan Society of Civil Engineers, Tokyo, Japan, revised edition (original 2002), 2016.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Kim, J., Koh, M. J., Oh, S.-t., and Son, S.: F-FNO basin-scale tsunami surrogate model (code, weights and training data sample), https://doi.org/10.5281/zenodo.19198928, zenodo, version 1.0.0, 2026.

Korea Peninsula Energy Development Organization: Estimation of tsunami height for KEDO LWR Project, Tech. rep., Korea Power Engineering Company, Inc., South Korea, 1999.

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.

Leonard, M.: Earthquake fault scaling: Self-consistent relating of rupture length, width, average displacement, and moment release, Bulletin of the Seismological Society of America, 100, 1971–1988, https://doi.org/10.1785/0120090189, 2010.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A. M., and Anandkumar, A.: Fourier neural operator for parametric partial differential equations, arXiv preprint, 2021.

Liu, C. M., Rim, D., Baraldi, R., and LeVeque, R. J.: Comparison of machine learning approaches for tsunami forecasting from sparse observations, Pure and Applied Geophysics, 178, 5129–5153, 2021.

Liu, P. L.-F., Cho, Y.-S., Briggs, M. J., Kanoglu, U., and Synolakis, C. E.: Runup of solitary waves on a circular island, Journal of Fluid Mechanics, 302, 259–285, https://doi.org/10.1017/S0022112095004095, 1995.

Lorito, S., Selva, J., Basili, R., Romano, F., Tiberti, M. M., and Piatanesi, A.: Probabilistic hazard for seismically induced tsunamis: accuracy and feasibility of inundation maps, Geophysical Journal International, 200, 574–588, https://doi.org/10.1093/gji/ggu408, 2015.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E.: Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators, Nature machine intelligence, 3, 218–229, 2021.

Makinoshima, F., Oishi, Y., Yamazaki, T., Furumura, T., and Imamura, F.: Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks, Nature communications, 12, 2253, 2021.

Mori, N., Takahashi, T., Yasuda, T., and Yanagisawa, H.: Survey of 2011 Tohoku earthquake tsunami inundation and run-up, Geophysical Research Letters, 38, L00G14, https://doi.org/10.1029/2011GL049210, 2011.

Mulia, I. E., Asano, T., and Nagashima, F.: Regional probabilistic tsunami hazard assessment associated with active faults along the eastern margin of the Sea of Japan, Earth, Planets and Space, 72, https://doi.org/10.1186/s40623-020-01256-5, 2020a.

Mulia, I. E., Gusman, A. R., and Satake, K.: Applying a deep learning algorithm to tsunami inundation database of megathrust earthquakes, Journal of Geophysical Research: Solid Earth, 125, e2020JB019 690, 2020b.

Mulia, I. E., Ueda, N., Miyoshi, T., Gusman, A. R., and Satake, K.: Machine learning-based tsunami inundation prediction derived from offshore observations, Nature Communications, 13, 5489, 2022.

Ohtake, M.: A seismic gap in the eastern margin of the Sea of Japan as inferred from the time-space distribution of past seismicity, Island Arc, 4, 128–138, https://doi.org/10.1111/j.1440-1738.1995.tb00140.x, 1995.

Ragu Ramalingam, N., Johnson, K., Pagani, M., and Martina, M. L. V.: Advancing nearshore and onshore tsunami hazard approximation with machine learning surrogates, Natural Hazards and Earth System Sciences, 25, 1655–1679, https://doi.org/10.5194/nhess-25-1655-2025, 2025.

Röbke, B. and Vött, A.: The tsunami phenomenon, Progress in Oceanography, 159, 296–322, 2017.

Salmanidou, D., Guillas, S., Georgiopoulou, A., and Dias, F.: Statistical emulation of landslide-induced tsunamis at the Rockall Bank, NE Atlantic, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 473, 20170 026, 2017.

Sarri, A., Guillas, S., and Dias, F.: Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification, Natural Hazards and Earth System Sciences, 12, 2003–2018, 2012.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Satake, K.: Linear and nonlinear computations of the 1992 Nicaragua earthquake tsunami, Pure and Applied Geophysics, 144, 455–470, https://doi.org/10.1007/BF00874378, 1995.

Satake, K.: Advances in earthquake and tsunami sciences and disaster risk reduction since the 2004 Indian ocean tsunami, Geoscience Letters, 1, 15, 2014.

Satake, K. and Murotani, S.: Tsunami heights on the Korean east coast from submarine active faults along the eastern margin of the Sea of Japan, Coastal Engineering Journal, 67, 812–821, 2025.

Satake, K., Ishibe, T., Murotani, S., Mulia, I. E., and Gusman, A. R.: Effects of uncertainty in fault parameters on deterministic tsunami hazard assessment: examples for active faults along the eastern margin of the Sea of Japan, Earth, Planets and Space, 74, 36, https://doi.org/10.1186/s40623-022-01594-6, 2022.

Selva, J., Tonini, R., Molinari, I., Tiberti, M. M., Romano, F., Grezio, A., Melini, D., Piatanesi, A., Basili, R., and Lorito, S.: Quantification of source uncertainties in seismic probabilistic tsunami hazard analysis (SPTHA), Geophysical Journal International, 205, 1780–1803, https://doi.org/10.1093/gji/ggw107, 2016.

Shuto, N. and Matsutomi, H.: Field survey of the 1993 Hokkaido Nansei-Oki earthquake tsunami, Pure and Applied Geophysics, 144, 649–663, https://doi.org/10.1007/BF00874388, 1995.

Son, S. and Jung, T.: Statistical analysis of tsunamis from multiple faults' sequential failure with different time intervals and geographical layouts, Ocean Engineering, 250, 110 720, 2022.

Song, M.-J. and Cho, Y.-S.: Early warning for maximum tsunami heights and arrival time based on an artificial neural network, Coastal Engineering, 192, 104 563, 2024.

Synolakis, C. E. and Bernard, E. N.: Tsunami science before and beyond Boxing Day 2004, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 364, 2231–2265, 2006.

Synolakis, C. E., Bernard, E. N., Titov, V. V., Kânoğlu, U., and Gonzalez, F. I.: Validation and verification of tsunami numerical models, Pure and Applied Geophysics, 165, 2197–2228, https://doi.org/10.1007/s00024-004-0427-y, 2008.

Titov, V. V. and Synolakis, C. E.: Numerical modeling of tidal wave runup, Journal of Waterway, Port, Coastal, and Ocean Engineering, 124, 157–171, https://doi.org/10.1061/(ASCE)0733-950X(1998)124:4(157), 1998.

Tran, A., Mathews, A., Xie, L., and Ong, C. S.: Factorized Fourier Neural Operators, in: International Conference on Learning Representations (ICLR), OpenReview.net, 2023.

Wang, X. and Power, W.: COMCOT: A tsunami generation, propagation and run-up model, GNS Science Report 2011/43, GNS Science, 2011.

Wang, Y., Imai, K., Miyashita, T., Ariyoshi, K., Takahashi, N., and Satake, K.: Coastal tsunami prediction in Tohoku region, Japan, based on S-net observations using artificial neural network, Earth, Planets and Space, 75, 154, 2023.

Wei, Y., Cheung, K. F., Curtis, G. D., and McCreery, C. S.: Inverse algorithm for tsunami forecasts, Journal of Waterway, Port, Coastal, and Ocean Engineering, 129, 60–69, https://doi.org/10.1061/(ASCE)0733-950X(2003)129:2(60), 2003.

Wells, D. L. and Coppersmith, K. J.: New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, Bulletin of the Seismological Society of America, 84, 974–1002, https://doi.org/10.1785/BSSA0840040974, 1994.

Yoon, S. B., Kim, S. C., Baek, U., and Bae, J. S.: Effects of bathymetry on the propagation of tsunamis towards the east coast of Korea, Journal of Coastal Research, 70, 332–337, https://doi.org/10.2112/SI70-056.1, 2014.

Zhu, Y., An, C., Yu, H., Zhang, W., and Chen, X.: High-resolution tsunami hazard assessment for the Guangdong-Hong Kong-Macao Greater Bay Area based on a non-hydrostatic tsunami model, Science China Earth Sciences, 67, 2326–2351, 2024.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Appendix A: Earthquake source parameters and logic tree design

This appendix documents the earthquake source parameters and scaling relationships used to construct the 864-scenario database. The East Sea is a semi-enclosed marginal sea with a maximum depth exceeding 3700 m, bounded by the Korean Peninsula, Japan, and Russia. We selected four virtual earthquake sources from the Korea Peninsula Energy Development. The KEDO database was searched for seismic events occurring within the latitude range 38.3° N to 42.9° N and at distances between 800 km and 1100 km from the east coast of the Korean Peninsula. The potential tsunamigenic seismic events were thus restricted to those occurring in the eastern margin of the Korean Peninsula. Table A1 summarizes the reference fault parameters for each source location.

To represent uncertainty in the earthquake source, we constructed a logic tree using strike angle (θ), dip angle (δ), rake angle (λ), moment magnitude (M_w), and fault-geometry scaling relationship. Focal depth was fixed for each reference source, whereas fault length, fault width, fault area, and average slip were derived from M_w using the selected scaling relationship. The source-slip pattern used to generate the initial condition was prescribed and was not treated as an independent branching variable in the logic tree.

For each source location, the strike angle was varied at three levels. A range of three possible azimuths for the reference angle θ_0 that will cover any uncertainty in the fault plane orientation relative to this reference angle: $\theta_0 - 10°$, θ_0, and $\theta_0 + 10°$. A range of possible dip angles (30°, 45°, 60°) is used that are thought to be more appropriate for the East Sea. We suspect that thrust faulting occurred under compressional tectonics in the East Sea, and for this analysis, we have used the three possible rake angles (75°, 90°, and 105°) to calculate the stress differences from thrust faulting mechanisms.

The four moment magnitudes (M_w) 7.4, 7.6, 7.8 and 8.0 are selected for this location because they represent a reasonable interval for credible tsunamigenic events in this area, taking into account the historical seismicity of the area and the results of the paleoseismic investigation. For these moment magnitudes, two empirical relationships, the Wells and Coppersmith (1994) (Wells and Coppersmith, 1994) and Leonard (2010) (Leonard, 2010) are applied involving fault length (L), width (W) and average slip (D). Wells and Coppersmith (Wells and Coppersmith, 1994) often cited in the literature concerning tsunami hazard assessment studies:

$$\log L = a_L + b_L M_w, \qquad \log W = a_W + b_W M_w, \qquad \log D = a_D + b_D M_w, \tag{A1}$$

Here we have coefficients a_L, b_L, a_W, b_W, a_D, b_D which depend on the type of fault. For reverse faults this formula gives length L in km, width W in km and average slip D in meters.

The Leonard (2010) scaling relations provide an alternative formula based on the following constraints:

$$W = C_1 L^{\beta}, \qquad D = C_2 A^{\gamma}, \tag{A2}$$

where $\beta \approx 2/3$, $A = L \times W$ is fault area. The other constants C_1, C_2, and γ are empirical. This form is for the seismic moment $M_0 = \mu A D$ and uses the shear modulus $\mu = 3 \times 10^{10}$ Pa for crustal faults.

The strike modeling has been carried out using a combination of 4 fault locations (see fault locations) and 3 strike angles, with 10 degree deviations on either side of the nominal strike angle. We have 3 dip angles (30°, 45°, 60°), 3 rake angles (75°,

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

630 90°, 105°), 4 moment magnitudes between 7.4 and 8.0, and 2 fault-geometry scaling relationships. This results in a total of 864 tsunami scenarios ($4 \times 3 \times 3 \times 3 \times 4 \times 2 = 864$).

Table A1. Earthquake source parameters for four potential tsunami scenarios. Longitude and latitude indicate epicenter location; H is focal depth; the strike angle θ_0 at the reference location, δ the dip angle and λ rake angle. Parameter variations described in the logic tree (Fig. A1).

Epicenter	**Long.** (° E)	**Lat.** (° N)	H **(km)**	θ_0 (°)	δ (°)	λ (°)
01	137.70	38.30	1	14.5	45	90
02	138.00	39.00		27.5		
03	139.10	42.10		4.0		
04	139.10	42.90		2.0		

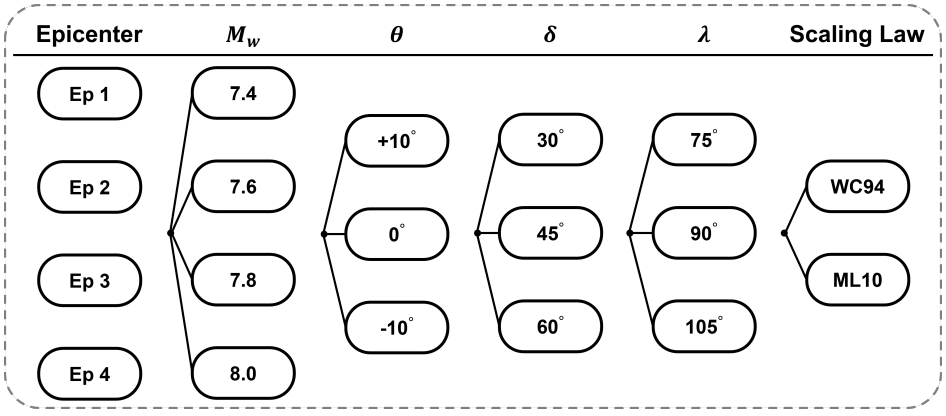

Figure A1. Logic tree for systematic generation of 864 tsunami scenarios. Our source scenarios are derived from 4 locations (Ep 1–4) with 5-dimensional parameter spaces defined by strike angle perturbations ($\theta_0 \pm 10°$ and θ_0), dip angle (δ = 30°, 45°, and 60°), rake angle (λ = 75°, 90°, and 105°), moment magnitude (M_w = 7.4, 7.6, 7.8, and 8.0), and scaling relationship (WC94: Wells and Coppersmith 1994; ML10: Leonard 2010).

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Appendix B: Training loss definitions

This appendix describes the loss terms explained in Eq. 8. The losses and their physical roles are listed in Table B1, together with the hyperparameters used in the Selected configuration. The loss weights were tuned with short pilot runs on the validation set. The coefficients cannot be compared across loss terms because the raw terms have different numerical scales. They were chosen from pilot runs using the validation weighted L1 loss as the main selection criterion.

Spatial averages are computed over the wet mask Ω_{wet}. For terms that are sensitive near the wet–dry boundary, we use an interior mask $\Omega_{\mathrm{int}} \subset \Omega_{\mathrm{wet}}$ obtained by removing edge cells from the wet cells, instead of Ω_{wet}. Here, Y is the COMCOT reference vector $Y = (\eta, u, v)$, with $\hat{Y}$ denoting the model prediction. The prediction horizon is H_r with discrete steps $k = 1, \ldots, H_r$.

B1 Training loss terms

Masked weighted data loss. The main misfit uses a weighted L1 error over wet cells:

$$L_{\mathrm{data}} = \frac{\sum_{k=1}^{H_r} \sum_{\mathbf{x}\in\Omega_{\mathrm{wet}}} \sum_{c\in\{\eta,u,v\}} w_k^{(c)}(\mathbf{x}) \left| \hat{Y}_{t_0+k}^{(c)}(\mathbf{x}) - Y_{t_0+k}^{(c)}(\mathbf{x}) \right|}{\sum_{k=1}^{H_r} \sum_{\mathbf{x}\in\Omega_{\mathrm{wet}}} \sum_{c\in\{\eta,u,v\}} w_k^{(c)}(\mathbf{x})}. \tag{B1}$$

Channel weights $w^{(u)} = w^{(v)} = 1.5$ are fixed, while $w^{(\eta)}$ uses an amplitude-dependent weighting based on the robust 95th-percentile scale.

Still-water suppression. Spurious oscillations during near-rest intervals are penalized by

$$L_{\mathrm{sw}} = \langle \mathbb{I}(\text{still}) \, |\hat{\eta} - \eta| + \gamma_{\mathrm{uv}} \, \mathbb{I}(\text{still}) \, (|\hat{u}| + |\hat{v}|) \rangle_{\Omega_{\mathrm{wet}},\, k=1..H_r}, \tag{B2}$$

where $\mathbb{I}(\text{still})$ is activated when $|\eta| < \varepsilon_{\mathrm{sw}} = 10^{-3}$ m, $|u| < 10^{-2}$ m s^{-1}, and $|v| < 10^{-2}$ m s^{-1}, all in physical units. The threshold $\varepsilon_{\mathrm{sw}}$ is small compared to typical tsunami amplitudes of 0.1–1 m.

Peak-aware time-weighted loss.

$$L_{\mathrm{peak}} = \frac{1}{3} \sum_{c\in\{\eta,u,v\}} \sum_{k=1}^{H_r} w_{k,c} \left\langle \left| \hat{Y}_{t_0+k}^{(c)} - Y_{t_0+k}^{(c)} \right| \right\rangle_{\Omega_{\mathrm{wet}}}, \tag{B3}$$

where the time weight is $w_{k,c} = \exp(a_{k,c}/T) \big/ \sum_j \exp(a_{j,c}/T)$ with $a_{k,c} = \langle |Y_{t_0+k}^{(c)}| \rangle_{\Omega_{\mathrm{wet}}}$ and temperature $T = 0.5$. This concentrates the training signal on high-amplitude time steps corresponding to the hazard-critical peak period.

Continuity residual loss. We denote the elevation of the input state at step k by $\eta_{t_0+k-1}^{\mathrm{in}}$. The local continuity residual, with local depth $h(\mathbf{x})$ and time step Δt, is

$$r_{t_0+k} = \frac{\hat{\eta}_{t_0+k} - \eta_{t_0+k-1}^{\mathrm{in}}}{\Delta t} + \nabla_s \cdot (h\,\hat{u}_{t_0+k},\, h\,\hat{v}_{t_0+k}), \tag{B4}$$

and the continuity loss is

$$L_{\mathrm{cont}} = \langle |r_{t_0+k}| \rangle_{\Omega_{\mathrm{int}},\, k=1..H_r}. \tag{B5}$$

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

This penalty reduces mismatch with the linearized shallow-water continuity equation.

Divergence control (mean-offset suppression).

$$L_{\mathrm{dc}} = \frac{1}{H_r}\sum_{k=1}^{H_r}\left(\bar{\hat{\eta}}_{t_0+k} - \bar{\eta}_{t_0+k}\right)^2, \qquad \bar{\eta}_{t_0+k} = \langle \eta_{t_0+k}\rangle_{\Omega_{\mathrm{wet}}}. \tag{B6}$$

This term prevents gradual drift in the spatially averaged sea level, which can accumulate over hundreds of prediction steps and bias peak-elevation and arrival-time diagnostics.

Intermediate one-step supervision. To stabilize autoregressive rollout training, we additionally penalize the error at each intermediate prediction step. The intermediate supervision loss is

$$L_{\mathrm{int}} = \frac{1}{H_r}\sum_{k=1}^{H_r}\left\langle \sum_{c\in\{\eta,u,v\}}\left(\hat{Y}^{(c)}_{t_0+k} - Y^{(c)}_{t_0+k}\right)^2\right\rangle_{\Omega_{\mathrm{wet}}}. \tag{B7}$$

The term penalizes errors at every prediction step, not only at the final horizon.

Surface-elevation energy proxy loss. We define a simple energy proxy from the free-surface elevation as

$$E(\eta) = \frac{1}{2}\eta^2. \tag{B8}$$

Using this quantity, the auxiliary energy loss is

$$L_{\mathrm{energy}} = \langle |E(\hat{\eta}_{t_0+k}) - E(\eta_{t_0+k})| \rangle_{\Omega_{\mathrm{wet}},\, k=1..H_r}. \tag{B9}$$

This term ensures consistency in the amplitude distribution of the predicted free surface. It should be interpreted as a surface-elevation energy proxy, not as a full mechanical-energy diagnostic.

Amplitude and gradient regularizers. Two auxiliary terms target different aspects of field structure. The first penalizes mismatch in a surface-elevation energy proxy $E(\eta) = \frac{1}{2}\eta^2$:

$$L_{\mathrm{energy}} = \langle |E(\hat{\eta}_{t_0+k}) - E(\eta_{t_0+k})| \rangle_{\Omega_{\mathrm{wet}},\, k=1..H_r}, \tag{B10}$$

which helps preserve the amplitude distribution of the predicted free surface. Note that $E(\eta)$ is not a full mechanical-energy diagnostic. The second enforces spatial sharpness by matching gradient magnitudes across all three channels:

$$L_{\mathrm{grad}} = \frac{1}{3}\sum_{c\in\{\eta,u,v\}}\left\langle \left|\left|\nabla_s \hat{Y}^{(c)}_{t_0+k}\right| - \left|\nabla_s Y^{(c)}_{t_0+k}\right|\right|\right\rangle_{\Omega_{\mathrm{int}},\, k=1..H_r}. \tag{B11}$$

This all-channel gradient term is distinct from $L_{\eta\nabla}$ (defined below), which applies only to η and penalizes the gradient of the prediction error rather than the difference in gradient magnitudes.

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

B2 Hyperparameter selection

For the Selected model, the key weights are $\lambda_{\mathrm{sw}} = 0.1$, $\lambda_{\mathrm{peak}} = 0.05$, $\lambda_{\mathrm{cont}} = 10$, and $\lambda_{\mathrm{dc}} = 100$. Table B1 summarizes all values.

Table B1. Loss terms and hyperparameters (Selected model configuration). All thresholds are evaluated in physical units by decoding normalized tensors. Code-level parameter names are documented in the project training script.

Symbol	Role	Value	Units
Main loss terms (Eq. 8)			
λ_{peak}	Peak-period time weighting	0.05	–
T	Peak-weighting concentration	0.5	–
λ_{sw}	Still-water suppression	0.1	–
$\varepsilon_{\mathrm{sw}}$	Still-water η threshold	10^{-3}	m
$\varepsilon_u, \varepsilon_v$	Velocity still-mask thresholds	10^{-2}	$\mathrm{m\,s^{-1}}$
γ_{uv}	Velocity penalty scale	1.0	–
Data-loss channel weights			
w_u, w_v	Constant multipliers for u, v	1.5	–
α_η (high/low)	η amplitude up-weighting	1.5/0.3	–
Auxiliary regularizers			
λ_{cont}	Continuity residual	10	–
λ_{dc}	Mean-offset control ($\bar{\eta}$)	100	–
$\lambda_{\mathrm{energy}}$	Surface-elevation energy proxy	1.0	–
λ_{grad}	All-channel gradient-magnitude matching	0.25	–
$\lambda_{\eta\nabla}$	Eta-only error-gradient penalty	1.0	–
λ_{int}	Intermediate one-step supervision	0.01	–
Δt	Time step (continuity)	60	s

https://doi.org/10.5194/egusphere-2026-1909
Preprint. Discussion started: 22 May 2026

Appendix C: Rollout stabilization schedule

We use scheduled sampling during training to stabilize multi-step prediction. The teacher-forcing probability at epoch e is written as $p_{\mathrm{TF}}^{(e)}$:

$$p_{\mathrm{TF}}^{(e)} = \begin{cases} p_0, & e \le E_{\mathrm{warm}}, \\ \max\left(p_{\min}, p_0 \alpha^{\,e - E_{\mathrm{warm}}}\right), & e > E_{\mathrm{warm}}, \end{cases} \tag{C1}$$

where $p_0 = 0.5$, $E_{\mathrm{warm}} = 2$, $\alpha = 0.9$, and $p_{\min} = 0$.

For rollout steps that are not forced into free-run mode, the next input is chosen as

$$X_{t_0+k} = \begin{cases} Y_{t_0+k}, & \text{with probability } p_{\mathrm{SS}}^{(e)}, \\ \hat{Y}_{t_0+k}, & \text{with probability } 1 - p_{\mathrm{SS}}^{(e)}, \end{cases} \tag{C2}$$

with

$$p_{\mathrm{SS}}^{(e)} = \max\left(p_{\mathrm{SS,min}}, \beta\, p_{\mathrm{TF}}^{(e)}\right), \tag{C3}$$

where $p_{\mathrm{SS,min}} = 0.05$ and $\beta = 0.20$.

In each TBPTT segment, the last $K_{\mathrm{FR}} = 2$ steps always use the model's own previous predictions as input. At inference, the model runs in full closed loop.