Domain-informed deep learning for downscaling tropical cyclone waves to hazard-relevant spatial scales

By

Eric Taylor

Department of Earth Science and Engineering

Imperial College London

This thesis is submitted in partial fulfilment of the requirements for the degree of MSc in Environmental Data Science and Machine Learning

August 2026

Abstract

This study investigates deep learning methods for downscaling tropical cyclone significant wave height (H_{m0}) from 8-to-1 km along the US East Coast. Deterministic U-Net and stochastic Wasserstein generative adversarial network (WGAN) architectures are evaluated under direct and residual formulations, conditioned on fine-resolution bathymetry and a bathymetry-relative directional encoding. Performance is assessed for unseen storms, unseen coastal geometries, and extreme wave conditions, emphasising nearshore prediction skill. The strongest performance was achieved by a residual-predicting U-Net using bathymetry-relative wave predictors, which successfully reconstructed fine-scale structure with substantially lower error than the WGANs. This U-Net reduced nearshore mean absolute error by 85.8% relative to bilinear interpolation. The final WGAN produced very little stochastic variability, with a mean pixel-wise ensemble standard deviation of 0.020 m. These results suggest that for 8-to-1 km H_{m0} downscaling, deterministic models can outperform stochastic generative approaches.

Acknowledgements

I am very grateful to those who have helped me with my independent research project.

Thank you to my primary supervisor, George Pouliasis, for his guidance and support throughout the project; Professor Matthew Piggott for his guidance as my secondary supervisor; and everyone in the Moody's office who made my time there enjoyable and gave informal feedback during presentations.

To those I have shared the MSc with, thank you for the conversations, encouragement, and friendship throughout the year – I hope my models are more useful than they are wrong.

Finally, I would like to thank my family and friends for their continued support throughout the year.

List of contents

1 Introduction

1.1 Background and motivation

Wave action is a major driver of coastal damage from tropical cyclones (Serafin et al., 2017). Critically, the wave hazard is highly spatially variable because processes such as bathymetric refraction, shoaling, and depth-limited breaking cause wave characteristics such as significant wave height ($\mathrm{H_{m0}}$), peak period ($\mathrm{T_p}$), and mean wave direction ($\mathrm{D_{mean}}$) to vary over spatial scales of hundreds of metres (Booij *et al.*, 1999), comparable to the scale of exposed assets. High-resolution spectral models, such as Simulating Waves Nearshore (SWAN; Booij *et al.*, 1999), can resolve wave variables at these scales, but at substantial computational cost. Such computational requirements limit their applicability to catastrophic risk modelling frameworks, which commonly rely on large stochastic sets of synthetic events to estimate risk (Emanuel, 2006; Hall and Jewson, 2007). Consequently, region-wide nearshore simulations at hazard analysis-relevant scales are unfeasible. Conversely, coarse-resolution wave models are more computationally affordable at regional scales but fail to resolve the nearshore detail.

Machine learning downscaling bridges more computationally efficient coarse-resolution models and high-resolution outputs (Jörges *et al.*, 2023; Kuehn *et al.*, 2023). This study focuses on the downscaling of $\mathrm{H_{m0}}$ from 8-to-1 km.

1.2 Related work

Deep learning has been increasingly applied to wave downscaling through two formulations: surrogate modelling of high-resolution numerical models (e.g., Jörges *et al.*, 2023; Kuehn *et al.*, 2023; Xu and Hao, 2026) and super-resolution of an existing coarse wave field (Kuehn *et al.*, 2023; Zhu *et al.*, 2023; Yu *et al.*, 2024). During tropical cyclones, waves can propagate far from their source, such that local conditions reflect both local forcing and remote wave generation (Tamizi and Young, 2020). This non-locality increases the complexity of full surrogate modelling, which must learn both wave generation and propagation history (Tamizi and Young, 2020, Kuehn *et al.*, 2023). Super-resolution instead retains this information from a coarse numerical model, thereby reducing the learning task to refining a physically based estimate, which can increase accuracy (Kuehn *et al.*, 2023). Consequently, this study adopts the super-resolution approach, with models either predicting the high-resolution field directly or learning the residual correction to an upsampled coarse estimate (Kim *et al.*, 2016).

Both formulations can incorporate physical information. Surrogate models have used bathymetry and wind as direct predictors (e.g., Michel *et al.*, 2022; Jörges *et al.*, 2023), whilst super-resolution models have conditioned coarse wave fields on additional high-resolution information, including bathymetry (Yu *et al.*, 2024) and wind (Wang *et al.*, 2026). Bathymetric effects become increasingly important nearshore, whilst wind is more influential offshore

(Booij *et al*., 1999). Some studies (e.g., Michel *et al.*, 2022; Yu *et al.*, 2025) transform wind predictors into variables more relevant for their given problem. For instance, Yu *et al*. (2025) rotate the wind vector into components parallel and perpendicular to the direction between given grid cells and target points. However, despite the importance of bathymetry for nearshore wave processes, less attention has been given to whether similar feature engineering to couple wave direction and bathymetry can improve predictions.

Model architecture and optimisation criteria also influence predictions. Architectures for downscaling waves have included CNNs (Jörges *et al.*, 2023; Zhu *et al*., 2023), ConvLSTMs and U-Nets for spatiotemporal downscaling (Yu *et al.*, 2025), and conditional GANs (Yu *et al.*, 2024). Super-resolution is ill-posed because a coarse input can represent many plausible fine-scale targets (Serifi *et al*., 2021). Deterministic models produce a single realisation, and training with pixelwise losses such as mean absolute error (MAE) or mean square error (MSE) can additionally produce smooth outputs (Vosper *et al.*, 2023). Stochastic generative models instead aim to represent the distribution of fine-resolution fields, allowing multiple plausible realisations to be generated from the same input (Leinonen *et al*., 2021).

Stochasticity has been particularly important in precipitation downscaling because coarse models leave substantial fine-scale variability unresolved. Generative architectures have therefore been used to produce stochastic high-resolution precipitation fields (Leinonen *et al*., 2021; Harris *et al.*, 2022; Vosper *et al.*, 2023). When directly comparing a deterministic U-Net, Vosper *et al*. (2023) found that a Wasserstein GAN (WGAN) better reproduced fine-scale spatial detail, extremes, and variability. Yu *et al*. (2024) used a conditional GAN to reconstruct high-resolution H_{m0} from coarse wave fields and physical conditioning variables. However, their analysis focused on reconstruction accuracy and spatial detail rather than whether stochastic inputs produced meaningful variability between high-resolution realisations. Therefore, whether these benefits transfer from precipitation to wave fields remains unclear.

1.3 Research aims

Section 1.2 established that deep learning can downscale H_{m0} through super-resolution and benefit from fine-resolution conditioning. However, comparatively little attention has been given to feature engineering to couple wave propagation with the bathymetry, which may improve predictions in shallow, nearshore regions. Physically meaningful predictor representations may improve generalisation to unseen regions, rather than encouraging memorisation of seen coastal geometries. Furthermore, whilst both deterministic and generative models have been used for wave downscaling, to the author's knowledge, no study has compared both under an equivalent super-resolution task.

Given these findings, this study aims to:

1. Develop physically informed representations of wave direction relative to local bathymetry and assess how predictor and target representation affect downscaling performance in nearshore regions.
2. Compare deterministic and stochastic generative architectures under an equivalent super-resolution task.
3. Assess model performance in unseen coastal geometries and extreme storms.

2 Methods and data

2.1 Problem setup

Let $\mathbf{W} \in \mathbb{R}^{C_w \times H_f \times W_f}$ and $\mathbf{X} \in \mathbb{R}^{C_w \times H_c \times W_c}$ denote the fine-resolution and coarse-resolution wave fields for a given spatial domain and timestep, respectively, where $H_f > H_c$ and $W_f > W_c$, and C_w is the number of wave variables. This study predicts fine-resolution $\mathrm{H_{m0}}$, such that the target $\mathbf{Y} \in \mathbb{R}^{1 \times H_f \times W_f}$.

Additional fine-resolution conditioning variables are collected in $\mathbf{Z} \in \mathbb{R}^{C_z \times H_f \times W_f}$, where C_z is the number of fine-resolution conditioning variables. These include the bathymetry and land-sea mask as well as feature-engineered predictors described in Section 2.3.

The coarse predictors are produced by applying the channel-wise operator f to $\mathbf{W}$:

$$\underbrace{\mathbf{X} = f(\mathbf{W})}_{\text{coarsening}}. \tag{1}$$

Downscaling seeks to infer a plausible $\mathbf{Y}$ from $\mathbf{X}$ and $\mathbf{Z}$. Since f is many-to-one, multiple fine-resolution fields can coarsen to the same coarse field, making the inverse problem ill-posed (Serifi *et al.*, 2021). The problem can therefore be expressed through the conditional distribution:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}). \tag{2}$$

A deterministic downscaling model (e.g., a U-Net) learns a parameterised mapping:

$$g_\theta : (\mathbf{X}, \mathbf{Z}) \mapsto \hat{\mathbf{Y}}. \tag{3}$$

Conversely, a stochastic conditional generative model (e.g., WGAN) introduces a noise input $\boldsymbol{\epsilon}$, such that:

$$G_\phi : (\mathbf{X}, \mathbf{Z}, \boldsymbol{\epsilon}) \mapsto \hat{\mathbf{Y}}. \tag{4}$$

Varying $\boldsymbol{\epsilon}$ allows the generator to produce multiple fine-resolution realisations for the same inputs, approximating $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$.

2.2 Data acquisition and preprocessing

Three study regions were defined across the US East Coast: Boston, New York (NY), and Virginia, each represented by a fixed bounding box (Figure 1). Wave data were obtained from the ultra-high-resolution SWAN dataset of Allahdadi et al. (2019), which uses an unstructured mesh with 200 m nearshore resolution. Tropical cyclones from 1979-2020 were identified using IBTrACS (Gahtan *et al.*, 2024), with scenes extracted whenever recorded events intersect a region. Hereafter, ‘event’ denotes a storm and ‘scene’ denotes a model output at one timestep. Modelled variables are $\mathrm{H_{m0}}$, $\mathrm{T_p}$, and $\mathrm{D_{mean}}$, hence $C_w = 3$.

Bathymetry was obtained from NOAA BlueTopo at 90 m resolution (NOAA, 2026). The unstructured SWAN outputs were interpolated to a regular 200 m grid and block-averaged to construct 1 km targets and 8 km inputs. NaNs were set to 0. Bathymetry and land-sea masks were reprojected and aligned to the target grid.

The dataset contains 36 Boston, 36 NY, and 75 Virginia events. To balance the regional datasets, a subset of 36 Virginia events was selected, retaining all storms also present in NY or Boston and randomly selecting the remainder. Hurricane Sandy (2012), present in NY and Virginia, was held out as an unseen extreme. Remaining storms were split 70/15/15% into training, validation, and testing, respectively, with storms occurring across multiple regions assigned consistently between splits.

Two training regimes assessed geographical generalisation (Table 1). A two-region dataset contained only NY and Boston (51 bounding box events), allowing evaluation on the unseen Virginia geometry. A three-region dataset contained 17 randomly selected events per region (51 total), thereby approximately matching the training volume from the two-region regime so that differences in performance reflected exposure to additional geometry rather than more data.

Patch-based sampling was used to improve the number of training samples, reduce memory requirements, and expose the model to a wider range of coastal settings. Input and target patches were 16×16 and 128×128, respectively. 20 and 17 random patches (minimum 40% ocean pixels) were taken from each training scene for the restricted-geometry and full-geometry datasets, totalling 11,300 and 11,424 training patches, respectively. Patches per scene were chosen to approximately match the training volume between datasets. The validation and test datasets are constant across regimes.

		New York	Boston	Virginia
Regime 1	Train	25 (287)	26 (278)	-
	Validation	5 (70)	5 (40)	5 (98)
	Test	6 (61)	5 (46)	6 (47)
Regime 2	Train	17 (225)	17 (194)	17 (253)
	Validation	5 (70)	5 (40)	5 (98)
	Test	6 (61)	5 (46)	6 (47)

Table 1. Number of events per train/validation/test split per bounding box. The number of scenes per split is shown in brackets.

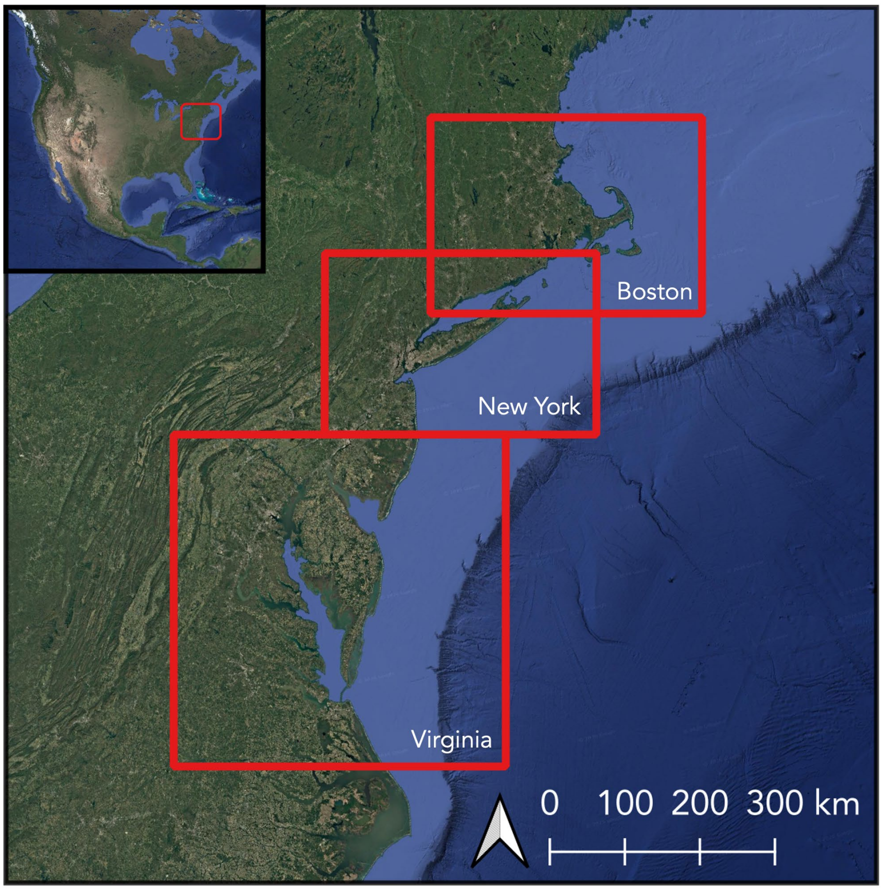

Figure 1. Bounding boxes for (top to bottom) Boston, New York, and Virginia. Their respective dimensions at 1 km resolution are 280 × 400, 240 × 360, and 440 × 440. The projection is EPSG:5070. Background imagery from Google Earth (Google, 2026).

2.3 Feature engineering and scaling

Several coarse variables were transformed before being passed to the models. Because $\mathrm{D}_{\mathrm{mean}}$ is cyclical, it was represented using sine and cosine components added to **X**:

$$\mathrm{D}_{\mathrm{mean,sin}} = \sin\left(\frac{\pi \mathrm{D}_{\mathrm{mean}}}{180}\right), \qquad \mathrm{D}_{\mathrm{mean,cos}} = \cos\left(\frac{\pi \mathrm{D}_{\mathrm{mean}}}{180}\right). \tag{5}$$

However, this retains wave direction relative to geographic North, whilst wave interaction depends on the propagation direction relative to the local seabed. Any model without additional feature engineering would have to learn this transformation implicitly. Therefore, it is useful to directly inform the model how wave direction and seabed slope interact.

X was first aligned with the fine grid using nearest-neighbour interpolation $\mathcal{U}_{\mathrm{NN}}$:

$$\widetilde{\mathbf{X}} = \mathcal{U}_{\mathrm{NN}}(\mathbf{X}) \in \mathbb{R}^{C_w \times H_f \times W_f}. \tag{6}$$

Preliminary experiments with the U-Net model architecture showed that nearest-neighbour upsampling yielded better performance than using bilinear interpolation and injecting **X** at the bottleneck. Hence, nearest-neighbour interpolation is used here.

The upsampled $\widetilde{H}_{\mathrm{m0}}$ and $\widetilde{\mathrm{D}}_{\mathrm{mean}}$ are represented as a directional wave vector:

$$\theta = \frac{\pi}{180}\widetilde{\mathrm{D}}_{\mathrm{mean}}, \qquad \mathbf{H} = \left(\widetilde{H}_{\mathrm{m0}} \cos\theta\,, \widetilde{H}_{\mathrm{m0}} \sin\theta\right), \tag{7}$$

and the corresponding local seabed orientation is obtained from the gradient of the bathymetry:

$$\nabla h = \left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right), \qquad |\nabla h| = \sqrt{h_x^2 + h_y^2}, \tag{8}$$

where h is the bathymetric depth. ∇h is estimated using second-order accurate central differences. The slope is then normalised:

$$\hat{\mathbf{s}} = \frac{\nabla h}{|\nabla h| + \varepsilon} = \left(\mathrm{s}_x, \mathrm{s}_y\right), \qquad \mathrm{s}_x = \frac{h_x}{|\nabla h| + \varepsilon}, \qquad \mathrm{s}_y = \frac{h_y}{|\nabla h| + \varepsilon}, \tag{9}$$

and its perpendicular component is:

$$\hat{\mathbf{s}}^{\perp} = \left(-\mathrm{s}_y, \mathrm{s}_x\right). \tag{10}$$

H is then projected onto these vectors:

$$H_{\mathrm{cross}} = \mathbf{H} \cdot \hat{\mathbf{s}}, \qquad H_{\mathrm{along}} = \mathbf{H} \cdot \hat{\mathbf{s}}^{\perp}, \tag{11}$$

where H_{cross} represents the wave component aligned with the local bathymetric gradient and H_{along} represents the perpendicular component. These components provide the models with physically meaningful information about wave propagation. These features are added to $\mathbf{Z}$.

Before model input, $\mathrm{H_{m0}}$ and $\mathrm{T_p}$ were divided by their respective training-set $q_{99.5}$ and q_{99}. Bathymetry was converted to positive depth, clipped at its training set q_{99}, and log-normalised. Model outputs were transformed back into metres before validation and testing.

2.4 Models

2.4.1 Bilinear Interpolation

Bilinear interpolation provides a non-learned deterministic baseline, estimating each fine-grid value by distance-weighted interpolation of the four neighbouring coarse-grid values. All model outputs henceforth are masked to the ocean mask and clipped at zero to enforce non-negativity.

2.4.2 U-Net

U-Nets are encoder-decoder CNNs with skip connections between corresponding resolution levels (Ronneberger *et al*., 2015). Two U-Net types were used: a 'Direct' U-Net and an otherwise equivalent 'Delta' U-Net. For experiments comparing direct and residual prediction, the two models use the same network architecture and inputs, differing only in the formulation of their final prediction. Figure 2 shows the model architectures.

$\widetilde{\mathbf{X}}$ and $\mathbf{Z}$ are concatenated channel-wise and passed to the U-Net, with predictor configurations varied between experiments (Section 2.5). No final activation function is used, allowing the Delta U-Net to predict both positive and negative residual corrections. No batch normalisation or dropout is used.

The Direct U-Net predicts the fine-resolution $\mathrm{H_{m0}}$ field directly:

$$\hat{\mathbf{Y}}_{\text{direct}} = g_{\theta_d}(\widetilde{\mathbf{X}}, \mathbf{Z}). \tag{12}$$

Residual learning instead predicts the correction to the baseline field, focusing the model's capacity on fine-scale correction (Kim *et al*., 2016). To test whether this improves wave downscaling, an otherwise identical U-Net is therefore trained. The Delta U-Net therefore predicts:

$$\hat{\mathbf{Y}}_{\text{delta}} = r_{\theta_r}(\widetilde{\mathbf{X}}, \mathbf{Z}) + \widetilde{\mathbf{X}}_{H_{m0}}, \tag{13}$$

where r_{θ_r} is the learned $\mathrm{H_{m0}}$ residual. Keeping the inputs and architecture constant isolates the effect of residual learning.

As the network is fully convolutional, full scenes were processed in a single forward pass after padding spatial dimensions for compatibility with the downsampling blocks. Padded pixels take on the value of the nearest edge pixel.

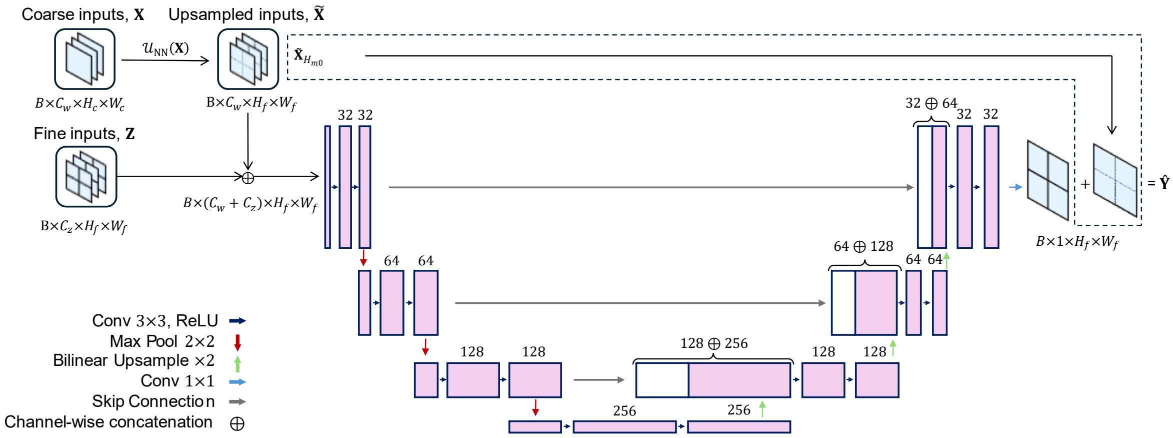

Figure 2. U-Net architecture. Each pink box corresponds to a multi-channel feature map. The number of channels is denoted above each box. White boxes represent feature maps propagated through skip connections, which are concatenated with the adjacent decoder feature maps. For the Delta U-Net, the network predicts a residual which is added to the upsampled $\tilde{\mathbf{X}}_{H_{m0}}$ to obtain the final prediction $\hat{\mathbf{Y}}$, as indicated by the dashed box.

2.4.3 WGAN

The WGAN architecture is adapted from Vosper *et al.* (2023), based on the precipitation-downscaling framework of Leinonen *et al.* (2021) and Harris *et al*. (2022). The WGAN is a stochastic generative model, based on a conditional GAN (Mirza and Osindero, 2014). Figure 3 shows the generator and critic architectures. To adapt their original $10\times$ downscaling architecture to $8\times$ downscaling, the two bilinear upsampling were changed from $(5, 2)$ to $(4, 2)$, with corresponding changes to the generator and critic downsampling blocks.

We also keep the fine-channel input from Harris *et al.* (2022), which was unused by Vosper *et al*. (2023). Whilst Harris *et al.*'s intention for the fine-channel input was that these were spatially static inputs, we also supply the dynamic feature-engineered $H_{\text{along/cross}}$ here. Four Gaussian noise channels are concatenated with $\mathbf{X}$. Thus, unlike the U-Nets, $\mathbf{X}$ and $\mathbf{Z}$ are input at different points in the network. $\mathbf{X}$ and $\boldsymbol{\epsilon}$ enter at the coarse input stage, and $\mathbf{Z}$ is injected at two points – downsampled onto the coarse grid at the input and at the full resolution after the upsampling blocks. The WGAN output for realisation k is:

$$\hat{\mathbf{Y}}^{(k)}_{\text{direct}} = G_{\phi_d}\left(\mathbf{X}, \mathbf{Z}, \boldsymbol{\epsilon}^{(k)}\right), \quad \boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \,, \tag{14}$$

A 'Delta' WGAN was also implemented, with the final softplus activation removed to permit negative residual corrections. The Delta WGAN prediction is:

$$\hat{\mathbf{Y}}^{(k)}_{\text{delta}} = R_{\phi_r}\left(\mathbf{X}, \mathbf{Z}, \boldsymbol{\epsilon}^{(k)}\right) + \tilde{\mathbf{X}}_{H_{m0}}, \tag{15}$$

Ensembles were generated by independently sampling $\boldsymbol{\epsilon}$ while keeping $\mathbf{X}$ and $\mathbf{Z}$ fixed.

The generator is trained adversarially against a critic which outputs a scalar score for observed or generated fields:

$$D_\psi\colon (\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \mapsto \mathbb{R}, \qquad D_\psi\colon \left(\hat{\mathbf{Y}}^{(k)}, \mathbf{X}, \mathbf{Z}\right) \mapsto \mathbb{R}. \tag{16}$$

Unlike a conventional GAN discriminator (Goodfellow *et al.*, 2014), the critic does not output a probability and has no final activation. Training is described in Section 2.6. The generator and critic are also fully convolutional.

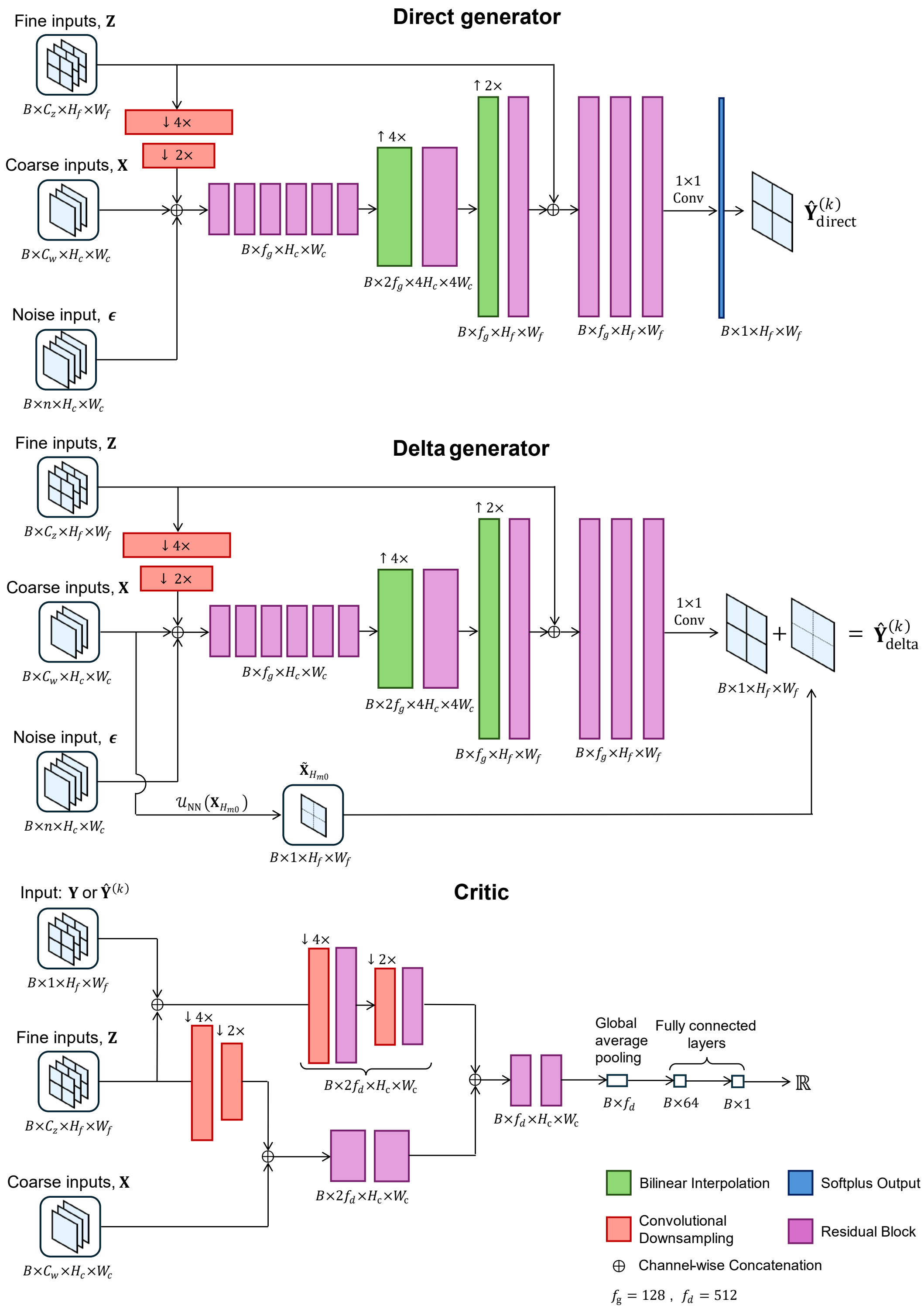

Figure 3. GAN generator and critic architectures for 8× downscaling. The output tensor dimensions of each architectural stage are shown below their respective stages. Pink boxes represent Residual Blocks (He *et al.*, 2015). Figures adapted from Vosper et al. (2023, p.5) and Harris et al. (2022, p. 6), with modifications as specified in Section 2.4.3.

2.5 Experimental design

Figure 4 summarises the experimental design. Two predictor configurations were compared while holding model architecture otherwise fixed. The first uses $\mathrm{H_{m0}}$, $\mathrm{T_p}$, and the sine and cosine encodings of $\mathrm{D_{mean}}$. The second retains $\mathrm{H_{m0}}$ and $\mathrm{T_p}$, but replaces the absolute directional encoding with H_{cross} and H_{along}. These are hereafter referred to as 'AllWaves' and 'RelativeWaves', respectively. Both configurations are used for both the Direct and Delta model variants to assess the effects of predictor representation, target formulation, and architecture.

Experiments were initially conducted under the two-region training regime. The best validation model (Section 2.7) was then retrained under the three-region regime (Section 2.2), isolating the effect of exposure to additional coastal geometries on geographical generalisation.

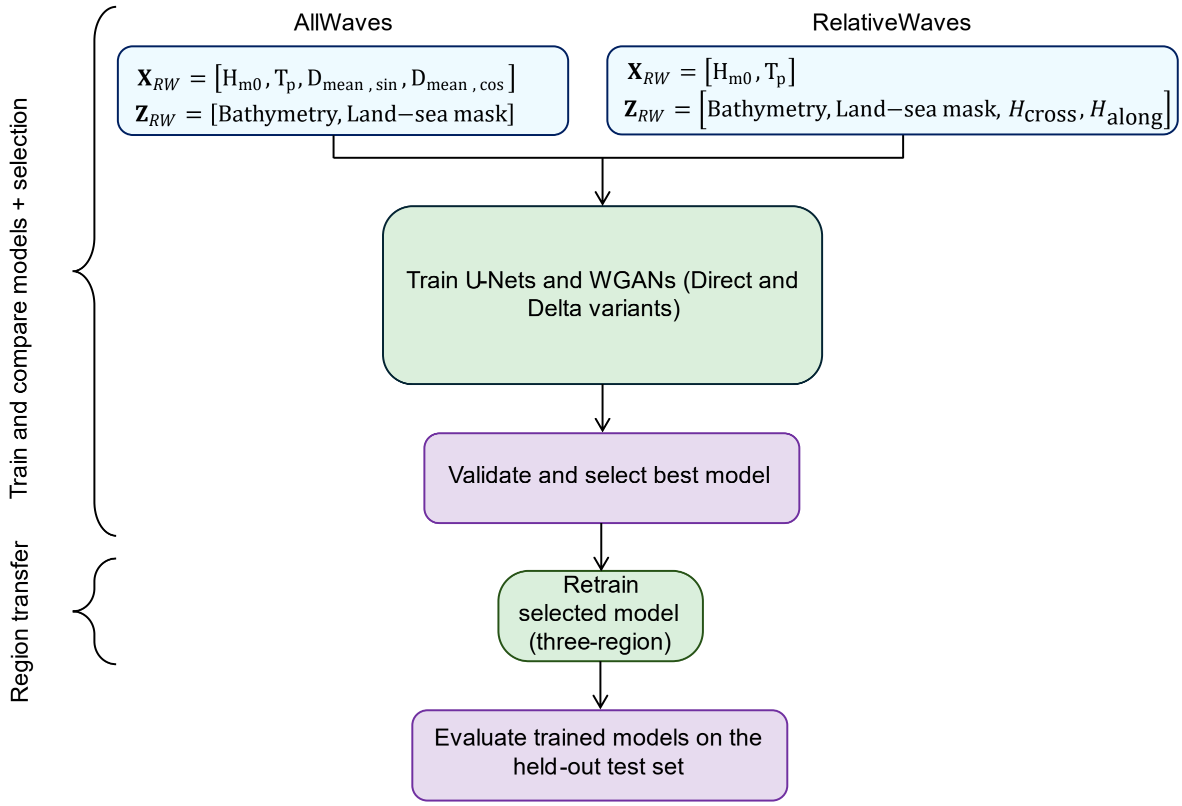

Figure 4. Flowchart showing experimental workflow. Square brackets represent channel-wise concatenation.

2.6 Training procedure and hyperparameters

Training was conducted on a G4 GPU in Google Colab using Adam (Kingma and Ba, 2017) with a batch size of 16. U-Nets were trained for 90 epochs with a learning rate of 1×10^{-4}, minimising ocean-masked MSE.

WGAN training followed Vosper *et al*. (2023), using the Wasserstein-GAN gradient-penalty framework (Arjovsky *et al*., 2017; Gulrajani *et al*., 2017). The generator and critic had learning rates of 1×10^{-5} over 100 epochs, with five critic updates per generator update (Kurach *et al*., 2019). Following Vosper *et al*. (2023), the content-loss term (Ravuri *et al.*, 2021) used by Harris *et al*. (2022) was omitted because it produced blurred outputs. Full definitions of the adversarial and gradient-penalty losses are available in Harris *et al*. (2022, pp. 8-10).

2.7 Model selection and evaluation

Model selection used full-scene validation MAE, binned by distance from the coast, with particular emphasis on the ≤1 km distance-to-coast bin. MAE was appropriate because it provides a measure of prediction error in the same units as $\mathrm{H_{m0}}$, while being less strongly influenced by a small number of large errors than squared-error metrics. To prevent storms with more scenes from contributing disproportionally, MAE was calculated separately for each storm within each bin and then averaged. The storm-to-storm variability is reported as the standard deviation of these MAEs.

Tail performance was assessed using q_{95}, q_{99}, and $q_{99.9}$ $\mathrm{H_{m0}}$ thresholds calculated from the training dataset within each distance-to-coast bin. Tail MAE was calculated using only pixels whose target $\mathrm{H_{m0}}$ exceeded the corresponding training-derived threshold. Again, particular attention was paid to the ≤1 km bin. Because exceedances become increasingly sparse at higher percentiles, tail MAE was pooled across all qualifying pixels rather than averaging by storm.

WGAN predictions used the mean of 20 independently generated ensemble members for comparison with the deterministic U-Nets. Validation results informed model and predictor selection. The test set was used for final evaluation, with Hurricane Sandy assessed separately.

Following Vosper *et al.* (2023), the power spectra of model outputs were calculated to quantify the spatial quality of predictions.

3 Results

3.1 Model selection

Validation performance depended on predictor representation, prediction formulation, and architecture. For the Direct U-Nets, AllWaves slightly outperformed RelativeWaves in the nearshore. Under the Delta formulation, RelativeWaves outperformed AllWaves. The Delta-RelativeWaves U-Net achieved the strongest overall U-Net performance, reducing $\leq$1 km MAE by 83.5% and 84.0% in NY and Boston relative to bilinear interpolation. Hereafter, percentage reductions are provided relative to the bilinear interpolation baseline, unless specified. It also achieved the lowest $>q_{99,1km}$ MAE in both regions, at 0.124 and 0.126 m, respectively – an 84.4% and 88.7% reduction. Only six pixels in the validation dataset were $>q_{99.9,1km}$, hence, $>q_{99.9,1km}$ was not used for model comparison.

For the WGANs, AllWaves generally performed best, with the Delta-AllWaves configuration achieving the lowest MAE across most distance bins. However, its nearshore MAE reductions were only 65.1% in NY and 69.3% in Boston, substantially below the U-Net. Therefore, the Delta-RelativeWaves U-Net was carried forward to the three-region experiment.

Broadly, similar model rankings were observed for unseen Virginia. The U-Nets reduced MAE relative to the baseline, whilst the WGANs were consistently outperformed by the baseline. Full validation statistics are shown in Appendix 1.

3.2 Held-out test performance

Hurricane Sandy is evaluated separately in Section 3.3.2. Here, we compare the best-performing U-Net and WGAN, referring to them as such. Validation rankings persisted on the test set. The selected Delta-RelativeWaves U-Net outperformed bilinear interpolation across all distance-to-coast bins, with a consistent 85% reduction in MAE up to 5 km offshore over both NY and Boston (Table 2). Improvements decreased further offshore but remained positive throughout.

Meanwhile, the strongest WGAN, Delta-AllWaves, also improved upon the baseline, but remained substantially weaker than the U-Net, reducing ≤1 km MAE by 65.4% and 65.7% in NY and Boston, respectively. The performance decreased offshore and fell below the baseline beyond 20 km in NY and 50 km in Boston.

Despite the strong U-Net performance in NY and Boston, model performance fell considerably in unseen Virginia. The U-Net ≤1 km MAE was only reduced by 14.2%, whilst the WGAN increased MAE by 62.6%. The U-Net remained better than the baseline across all distance bins.

Tail performance followed a similar pattern in the 1 km bin (Table 3). Above q_{99}, the U-Net reduced MAE by 81.4% and 90.9% for NY and Boston, respectively, compared to 52.3% and 80.2% for the WGAN. The U-Net's performance above $q_{99.9}$ remained strong, reducing MAE by 78.6% and 90.6%, respectively. Whilst the validation dataset only had 6 pixels for this bin in Boston, the test dataset had 182 and 135 pixels, respectively. Interestingly, in unseen-Virginia, despite poorer general MAE performance, the WGAN outperformed the U-Net. It reduced MAE by 41.8% and 73.8% above q_{95} and q_{99}, respectively, compared to 23.9% and 56.3% for the U-Net. No pixels in Virginia surpassed $q_{99.9}$.

For the selected Delta-RelativeWaves U-Net, full-scene inference on an NVIDIA A100 GPU required, on average, 8.6 ms, 11.1 ms, and 32.4 ms per test scene for NY, Boston, and Virginia, respectively.

		Model	<=1km	1-2km	2-5km	5-10km	10-20km	20-50km	>50km
Seen geography	New York	**Bilinear Interpolation**	0.374 ± 0.24	0.350 ± 0.19	0.299 ± 0.19	0.140 ± 0.11	0.058 ± 0.05	0.031 ± 0.03	0.015 ± 0.01
		ΔU-Net (RelativeWaves)	85.8 ± 1.40%	86.1 ± 1.0%	87.5 ± 1.3%	77.6 ± 3.8%	51.5 ± 10.3%	37.2 ± 14.3%	23.8 ± 13.4%
		ΔWGAN (AllWaves)	65.4 ± 6.60%	63.6 ± 6.5%	68.2 ± 5.1%	57.4 ± 6.0%	12.6 ± 16.4%	-9.1 ± 20.0%	-45.5 ± 24.6%
	Boston	**Bilinear Interpolation**	0.297 ± 0.12	0.269 ± 0.11	0.208 ± 0.09	0.081 ± 0.04	0.034 ± 0.02	0.027 ± 0.02	0.015 ± 0.01
		ΔU-Net (RelativeWaves)	84.7 ± 1.3%	85.1 ± 1.5%	86.3 ± 1.8%	73.7 ± 4.5%	49.7 ± 10.1%	45.6 ± 16.3%	26.8 ± 19.4%
		ΔWGAN (AllWaves)	65.7 ± 9.0%	63.8 ± 9.1%	68.8 ± 7.4%	51.8 ± 7.2%	12.4 ± 15.1%	8.4 ± 19.6%	-30.3 ± 17.7%
Unseen	Virginia	**Bilinear Interpolation**	0.128 ± 0.09	0.120 ± 0.06	0.135 ± 0.08	0.085 ± 0.06	0.048 ± 0.04	0.036 ± 0.03	0.015 ± 0.01
		ΔU-Net (RelativeWaves)	14.2 ± 14.0%	27.3 ± 4.3%	51.8 ± 3.3%	39.9 ± 5.0%	11.5 ± 3.9%	18.6 ± 2.8%	17.3 ± 15.6%
		ΔWGAN (AllWaves)	-62.7 ± 43.4%	-75.3 ± 43.9%	-5.0 ± 23.9%	-10.4 ± 18.0%	-34.4 ± 18.1%	-8.5 ± 12.9%	-40.8 ± 35.8%

Table 2. Mean absolute error per bounding box, binned by distance from the coast, over the held-out test set. Statistics are shown for the bilinear interpolation baseline and the best-performing U-Net and WGAN. MAE is reported as the average MAE per storm. The error is defined as the standard deviation across test storms. The baseline is provided in absolute units (metres), with the U-Net and WGAN performance reported as the percentage reduction in MAE, relative to the baseline for ease of interpretation. Entries in red do not beat the baseline. Full results are shown in Appendix 2.

Region		Model	$>q_{95}$	$>q_{99}$	$>q_{99.9}$
Seen geography	New York	**Bilinear**	0.621	0.909	1.041
		ΔU-Net (RelativeWaves)	81.0%	81.4%	78.6%
		ΔWGAN (AllWaves)	61.2%	52.3%	43.2%
	Boston	**Bilinear**	0.725	1.081	1.399
		ΔU-Net (RelativeWaves)	88.2%	90.9%	90.6%
		ΔWGAN (AllWaves)	80.0%	80.2%	77.9%
Unseen	Virginia	**Bilinear**	0.533	0.890	-
		ΔU-Net (RelativeWaves)	23.9%	56.3%	-
		ΔWGAN (AllWaves)	41.8%	73.8%	-

Table 3. Mean absolute error per bounding box in the ≤1 km from the coast bin over the held-out test set, calculated over pixels where the corresponding target pixel is above the percentile calculated from the respective distance bin from the two-region dataset. The baseline is provided in absolute units (metres), with the U-Net and WGAN performance reported as the percentage reduction in MAE, relative to the baseline for ease of interpretation. For the ≤1 km from the coast bin, these values are 1.71 m, 3.05 m, 4.73 m, for each respective percentile. Full results are shown in Appendix 2.

3.2.1 Visual and spatial analysis

Power spectra broadly followed the observations at low wave numbers, where power is already present in the coarse input (Figure 5). At higher wave numbers, bilinear interpolation increasingly lost power, falling to 80-90% of the observed spectrum across the three regions. Both learned models retained much more fine-scale power in NY and Boston, with the U-Net remaining closest to observations. The WGAN showed excess fine-scale power in Boston. Performance in Virginia was weaker, however the U-Net remained closer to the observed spectrum. Thus, the U-Net was able to recreate much of the observed fine-scale spectrum.

These differences are also reflected visually in the downscaled spatial fields (Figure 6). The better representation of fine detail is clear in Barry (Boston, 2007), where fine detail in the shallow coastal region in the bottom right is recovered by both learned models. The U-Net more closely reconstructs a thin band of higher H_{m0} in the top left not captured by any other method. The learned models also more closely reproduce the target coastal gradients along Long Island in Danny (1997) and Cape Cod in Floyd (1999). The WGAN also produces small-scale artefacts, which are mostly absent or less evident in the U-Net.

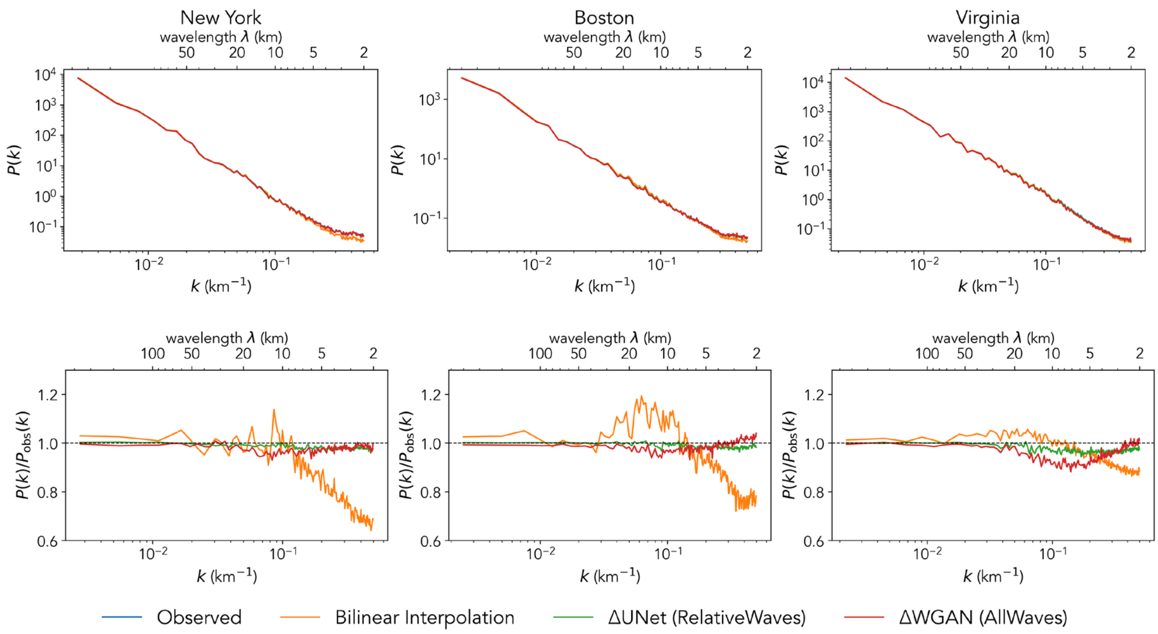

Figure 5. Spectral power curves of the predictions from bilinear interpolation, the Delta-RelativeWaves U-Net and the Delta-AllWaves WGAN. The top row shows the absolute values. The bottom row shows the values normalised by the observed spectra.

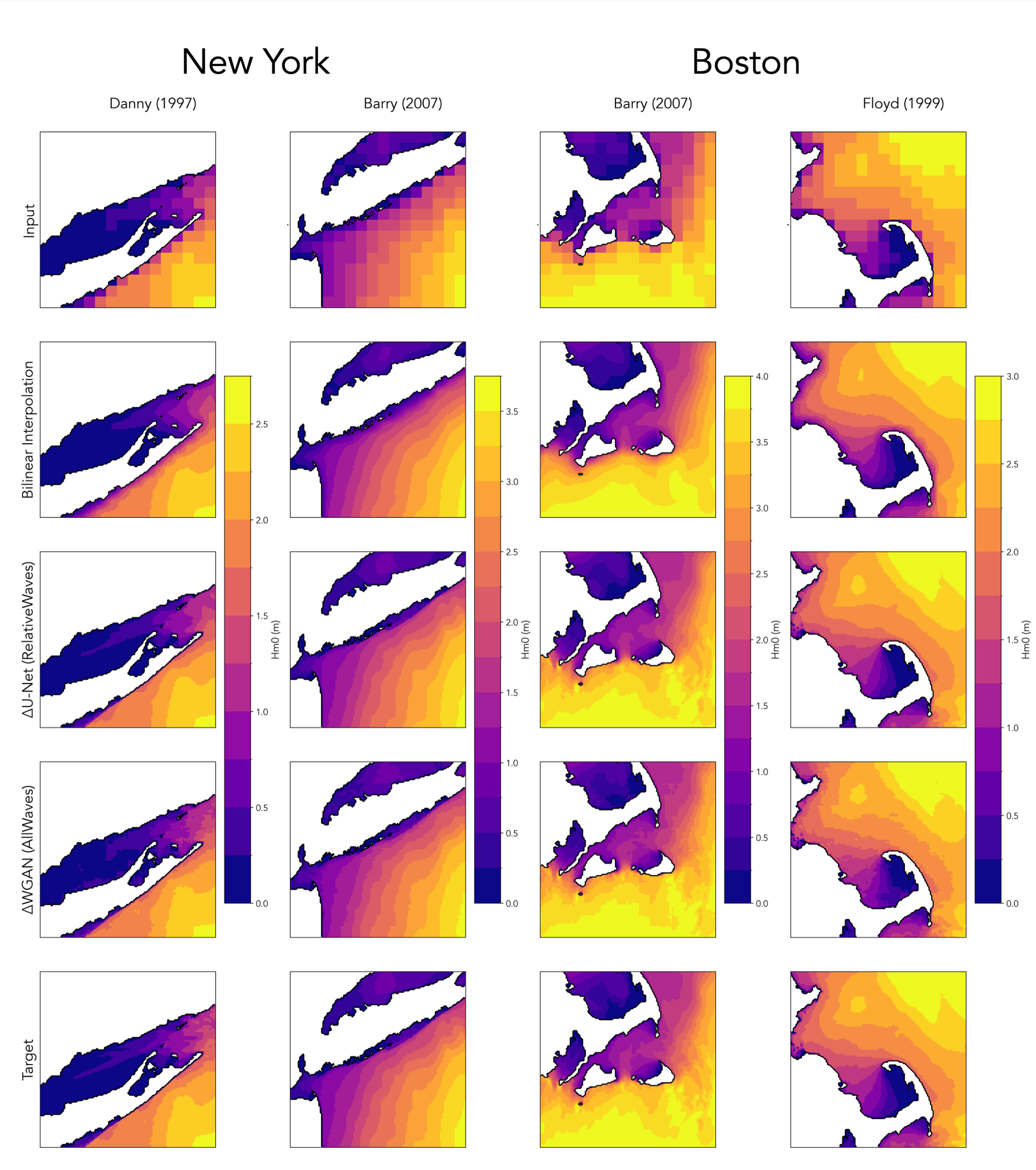

Figure 6. Example test patches showing the coarse input, and model outputs for the bilinear interpolation baseline, Delta-RelativeWaves U-Net, and Delta-AllWaves WGAN. The target is shown in the bottom row. Patches were sampled randomly from the test scenes by first randomly choosing a storm, then a random scene, then a random patch. Patches are the same size used in training – 128 km × 128 km. Patches are projected onto EPSG:5070. Geographic North is the y-direction. Note the different colour bar scale for each plot.

3.3 Generalisation to unseen conditions

3.3.1 Geographical transfer

Retraining the Delta-RelativeWaves U-Net under the three-region regime yielded substantial improvements in Virginia, with ≤1 km MAE reduced by 64.5% relative to its two-region counterpart to 0.029 m. This gain was accompanied by modest increases in MAE in NY (0.040 to 0.051 m) and Boston (0.045 to 0.058 m), where improvements relative to bilinear interpolation decreased from 85.8% to 82.0% and from 84.7% and 80.4%, respectively. Relative to bilinear interpolation, the Virginia MAE was reduced by 69.5%. Thus, exposure to Virginia largely closed the geographical performance gap while retaining strong performance for the original training regions, despite reducing their training data by a third.

3.3.2 Extreme-event transfer (Hurricane Sandy)

Hurricane Sandy was used as a manually held-out extreme-event test to evaluate generalisation to extreme events. Sandy's global $\mathrm{H_{m0}}$ $q_{99.9}$ threshold was 13.8 m, compared with 10.5 m in training. Within ≤1 km, the corresponding values were 5.29 m and 4.73 m.

For the two-region U-Net, in NY, the ≤1 km MAE above the training-derived q_{99} and $q_{99.9}$ thresholds were 0.270 and 0.236 m, respectively. In Virginia, MAE $>q_{99}$ was 0.430 m, while no pixels were $>q_{99.9}$. For the three-region model, the Virginia $>q_{99}$ MAE was reduced by 67.4% to 0.140 m. In NY, $>q_{99}$ MAE was unchanged (+0.179%), although $>q_{99.9}$ MAE increased by 28.7%. Thus, adding Virginia to training substantially improved Virginia's extreme-event transfer, but did not improve NY.

Figure 7 shows model outputs for Sandy in NY and Virginia and the difference between the output and the target. Both models reduced coastal error in NY, while the three-region model substantially reduced errors along Virginia's Eastern coastline.

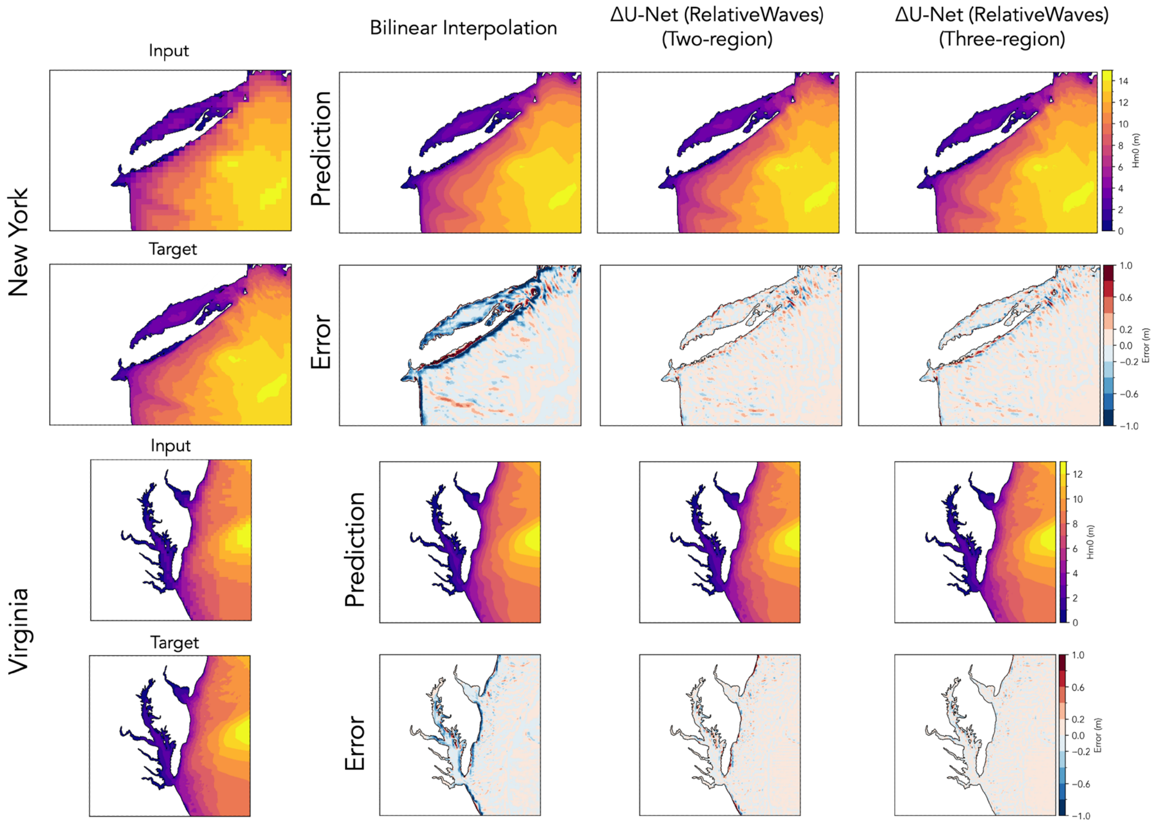

Figure 7. Bilinear interpolation and the two- and three-region Delta-RelativeWaves U-Net outputs for scene 5 in Hurricane Sandy (2012) for New York and Virginia. New York's scenes are 240 km × 360 km. Virginia's scenes are 440 km × 440 km. Error is defined as the prediction minus target. Patches are projected onto EPSG:5070. Geographic North is the y-direction. Note the different colour bar scales between New York and Virginia.

4 Discussion

The following sections consider the implications of the results for generative downscaling, physically informed feature engineering, and geographical generalisation.

4.1 Are generative architectures necessary for $\mathbf{H_{m0}}$ downscaling?

Generative architectures can be useful for general downscaling because the inverse problem is ill-posed (Section 2.1). They can represent multiple plausible fine-resolution fields for a given coarse input, reproduce finer-scale spatial structure without the smoothing associated with deterministic point prediction, and they can be better at representing tail events (Vosper *et al.*, 2023). However, despite these theoretical advantages, the WGAN did not outperform the U-Nets and produced very little variability between ensemble members. The mean pixel-wise standard deviation across members from the test dataset was 0.020 m, which did not vary substantially between regions. There are two plausible diagnoses for this result. First, the generator learned to ignore its stochastic input. Second, the conditioning variables may constrain the fine-resolution solution sufficiently to make the conditional distribution $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$ narrow.

Ignoring noise is a known failure mode of generators, causing them to learn a deterministic mapping (Isola *et al.*, 2018; Ramasinghe *et al.*, 2021). However, Vosper *et al.* (2023) found that their ensemble predictions had good but slightly systematically underspread variability. Therefore, the same architecture (except for 8× rather than 10× downscaling) and training have generated meaningful stochastic variability in tropical-cyclone precipitation downscaling, suggesting that the architecture is at least capable of exploiting stochastic inputs. Thus, the noise-ignoring cannot be solely attributed to the architecture.

The limited ensemble spread may also reflect the diversity of the training data. Although the WGAN was trained on 11,300 patches (565 scenes), spatial overlap between patches and repeated sampling from the same events mean that these patches do not constitute an equivalent number of independent samples. The generator may, therefore, have had limited evidence to learn a broad conditional distribution. Increasing the number of patches would not necessarily resolve this limitation owing to their overlap. Rather, greater diversity in independent storms could be more informative. Therefore, it cannot be ruled out that a theoretical, fully optimised WGAN could learn an ensemble spread.

Nevertheless, several results support a narrow conditional distribution. The U-Nets outperformed the WGANs on MAE and in recreating fine spatial structure. This contrasts with precipitation-downscaling, where deterministic U-Nets lost substantial spectral power at fine scales, whereas generative models endured (Vosper *et al.*, 2023). The strong U-Net performance, coupled with the low WGAN ensemble spread, is consistent with a narrow conditional distribution.

This difference is physically plausible given the different spatial characteristics of $\mathrm{H_{m0}}$ and precipitation. Precipitation is highly variable in space, with key processes occurring below the coarse-resolution scale, leaving substantial fine-scale variability to be reconstructed during downscaling (Wang *et al.*, 2021; Harris *et al.*, 2022), hence motivating the GAN. Conversely, $\mathrm{H_{m0}}$ is a bulk measure of the sea state, whose spatial evolution is affected by wave propagation, thereby producing smoother wave fields (Booij *et al.*, 1999; Obakrim *et al.*, 2023). Consequently, deterministic U-Net smoothing may incur relatively little loss of physically meaningful structure for this downscaling problem. This result, however, should not be generalised to all wave downscaling problems. Other wave variables, or different scales, may require a broader conditional distribution and the generation of sharper structures.

4.2 Importance of input and target representation

Predictor representation was not universally beneficial. AllWaves slightly outperformed RelativeWaves for direct U-Net prediction, whereas this order reversed for the Delta U-Net, where RelativeWaves produced the best overall performance across all models. In contrast, AllWaves achieved the strongest performance in the WGANs.

One possible explanation is that RelativeWaves is well aligned with residual-prediction. Where residuals contain systematic structure associated with physical predictors, providing these predictors can help the learning of residuals (Tedesco *et al.*, 2023). If the fine-scale correction to the coarse $\mathrm{H_{m0}}$ is related to the local bathymetry through wave processes (Booij *et al.*, 1999), then representing the wave direction relative to the bathymetry may provide information specifically relevant for this correction. Between both Delta U-Net variants, RelativeWaves produced its largest improvements closest to the coastline, where bathymetric effects on $\mathrm{H_{m0}}$ are expected to be strongest, whilst improvements weakened offshore. However, this interpretation does not explain why the same predictor representation did not improve direct prediction. This result therefore suggests that the RelativeWaves encoding is better for residual-style learning in the U-Net, but not that it is necessarily a superior method of directional encoding for model predictors. RelativeWaves's small improvement in unseen Virginia also suggests potential value for geographical transfer.

Interpreting the WGAN comparison is less straightforward. The U-Net directional predictors enter through the same input pathway regardless of their representation. Conversely, in the WGAN, changing the predictor representation also changes where information is injected into the networks. AllWaves directional predictors entered through the coarse-input pathway, whereas, under RelativeWaves, $H_{\mathrm{along/cross}}$ enter through the fine conditioning pathway. Therefore, the poorer RelativeWaves WGAN performance may reflect an interaction between the predictor representation and how conditioning is done within the network, rather than the representation alone.

These results indicate that the effectiveness of residual learning and predictor representation depends on the prediction objective and architecture. For this study, bathymetry-relative directional predictors coupled with a residual target had the best performance. Predictor and target representation should therefore be considered as non-independent modelling choices.

4.3 Limitations and areas for future research

This study has acknowledged several limitations. This section summarises them and provides directions for future research.

The negligible WGAN ensemble spread could be due to the training dataset containing a limited number of independent samples. Further research should therefore evaluate generative models using larger event datasets before determining whether the limited ensemble variability is symptomatic of the underlying downscaling problem or limitations in the trained generator.

This study only considered one type of generative model. Consequently, the poor performance of the WGAN should not be interpreted as evidence that generative architectures are unnecessary for H_{m0} downscaling. Future work should test alternative probabilistic models to assess whether they produce greater ensemble spread and test model performance at finer target resolutions.

This study used an artificially coarsened dataset, derived from fine-resolution simulations. However, true coarse-resolution simulations are different to spatially averaged fine-resolution simulations (Kuehn *et al.*, 2023). Consequently, performance may worsen when the trained models are applied to true coarse-resolution inputs. Further work should evaluate the models using paired coarse- and fine-resolution simulations generated independently at their respective resolutions.

Finally, the U-Net architecture and optimisation were not exhaustively tuned. Alternative channel widths, network depths, introducing drop-out, or optimisation criteria may improve its results. The reported U-Net performance should therefore be interpreted as evidence that a deterministic architecture performs sufficiently for this task, rather than as a fully optimised downscaling model.

5 Conclusions

This study has applied deep learning methods for downscaling significant wave height from an 8 km grid to a 1 km grid. This is important because such approaches provide a computationally efficient means of generating high-resolution wave information for catastrophic risk modelling frameworks where direct fine-resolution numerical modelling is computationally expensive. Once a coarse field was available, the selected U-Net required an average of 16.6 ms per full-scene inference on an NVIDIA A100 GPU. This study contributes to the literature by comparing deterministic and generative super-resolution models conditioned on fine-resolution bathymetry and physically informed representations of the wave state relative to the local seabed.

Bathymetry-relative predictors improved residual prediction but not direct prediction, demonstrating that predictor and target representation are not independent modelling choices. The same representation also produced a small improvement when transferring to unseen coastal geometry. When including the previously unseen Virginia region in training, performance over Virginia substantially improved while retaining strong performance in New York and Boston.

Deterministic U-Nets outperformed the WGANs across all metrics and produced better fine-scale spatial structure. The WGAN additionally produced little ensemble variability, consistent with a relatively narrow conditional distribution. Therefore, unlike precipitation-downscaling, the expected advantages of stochastic generative modelling did not transfer to this downscaling problem.

6 References

Allahdadi, M.N. *et al.* (2019) 'Development and validation of a regional-scale high-resolution unstructured model for wave energy resource characterization along the US East Coast', *Renewable Energy*, 136, pp. 500–511. Available at: https://doi.org/10.1016/j.renene.2019.01.020.

Arjovsky, M., Chintala, S. and Bottou, L. (2017) 'Wasserstein GAN'. arXiv. Available at: https://doi.org/10.48550/arXiv.1701.07875.

Booij, N., Ris, R.C. and Holthuijsen, L.H. (1999) 'A third-generation wave model for coastal regions: 1. Model description and validation'. Available at: https://doi.org/10.1029/98JC02622.

Emanuel, K. (2006) 'Climate and Tropical Cyclone Activity: A New Model Downscaling Approach'. Available at: https://doi.org/10.1175/JCLI3908.1.

Gahtan, J. *et al.* (2024) 'International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4.01'. NOAA National Centers for Environmental Information. Available at: https://doi.org/10.25921/82ty-9e16.

Goodfellow, I.J. *et al.* (2014) 'Generative Adversarial Networks'. arXiv. Available at: https://doi.org/10.48550/arXiv.1406.2661.

Google (2026) 'Google Earth'. Available at: https://earth.google.com/.

Gulrajani, I. *et al.* (2017) 'Improved Training of Wasserstein GANs'. arXiv. Available at: https://doi.org/10.48550/arXiv.1704.00028.

Hall, T.M. and Jewson, S. (2007) 'Statistical modelling of North Atlantic tropical cyclone tracks', *Tellus*, 59(4). Available at: https://tellusjournal.org/articles/10.1111/j.1600-0870.2007.00240.x (Accessed: 11 June 2026).

Harris, L. *et al.* (2022) 'A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts', *Journal of Advances in Modeling Earth Systems*, 14(10), p. e2022MS003120. Available at: https://doi.org/10.1029/2022MS003120.

He, K. *et al.* (2015) 'Deep Residual Learning for Image Recognition'. arXiv. Available at: https://doi.org/10.48550/arXiv.1512.03385.

Isola, P. *et al.* (2018) 'Image-to-Image Translation with Conditional Adversarial Networks'. arXiv. Available at: https://doi.org/10.48550/arXiv.1611.07004.

Jörges, C. *et al.* (2023) 'Spatial ocean wave height prediction with CNN mixed-data deep neural networks using random field simulated bathymetry', *Ocean Engineering*, 271, p. 113699. Available at: https://doi.org/10.1016/j.oceaneng.2023.113699.

Kim, J., Lee, J.K. and Lee, K.M. (2016) ‘Accurate Image Super-Resolution Using Very Deep Convolutional Networks’. arXiv. Available at: https://doi.org/10.48550/arXiv.1511.04587.

Kingma, D.P. and Ba, J. (2017) ‘Adam: A Method for Stochastic Optimization’. arXiv. Available at: https://doi.org/10.48550/arXiv.1412.6980.

Kuehn, J. *et al.* (2023) ‘A deep learning super-resolution model to speed up computations of coastal sea states’, *Applied Ocean Research*, 141, p. 103776. Available at: https://doi.org/10.1016/j.apor.2023.103776.

Kurach, K. *et al.* (2019) ‘A Large-Scale Study on Regularization and Normalization in GANs’. arXiv. Available at: https://doi.org/10.48550/arXiv.1807.04720.

Leinonen, J., Nerini, D. and Berne, A. (2021) ‘Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields with a Generative Adversarial Network’, *IEEE Transactions on Geoscience and Remote Sensing*, 59(9), pp. 7211–7223. Available at: https://doi.org/10.1109/TGRS.2020.3032790.

Michel, M. *et al.* (2022) ‘Deep learning for statistical downscaling of sea states’, *Advances in Statistical Climatology, Meteorology and Oceanography*, 8(1), pp. 83–95. Available at: https://doi.org/10.5194/ascmo-8-83-2022.

Mirza, M. and Osindero, S. (2014) ‘Conditional Generative Adversarial Nets’. arXiv. Available at: https://doi.org/10.48550/arXiv.1411.1784.

NOAA Office of Coast Survey (2026) ‘BlueTopo: A Curated Collection of High-Resolution Seafloor Models’. National Oceanic and Atmospheric Administration (NOAA). Available at: https://nauticalcharts.noaa.gov/data/bluetopo.html.

Obakrim, S. *et al.* (2023) ‘Statistical modeling of the space–time relation between wind and significant wave height’, *Advances in Statistical Climatology, Meteorology and Oceanography*, 9(1), pp. 67–81. Available at: https://doi.org/10.5194/ascmo-9-67-2023.

Ramasinghe, S. *et al.* (2021) ‘Rethinking conditional GAN training: An approach using geometrically structured latent manifolds’. arXiv. Available at: https://doi.org/10.48550/arXiv.2011.13055.

Ravuri, S. *et al.* (2021) ‘Skillful Precipitation Nowcasting using Deep Generative Models of Radar’, *Nature*, 597(7878), pp. 672–677. Available at: https://doi.org/10.1038/s41586-021-03854-z.

Ronneberger, O., Fischer, P. and Brox, T. (2015) ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1505.04597.

Serifi, A., Günther, T. and Ban, N. (2021) 'Spatio-Temporal Downscaling of Climate Data Using Convolutional and Error-Predicting Neural Networks', *Frontiers in Climate*, 3. Available at: https://doi.org/10.3389/fclim.2021.656479.

Tamizi, A. and Young, I.R. (2020) 'The Spatial Distribution of Ocean Waves in Tropical Cyclones', *Journal of Physical Oceanography*, 50(8), pp. 2123–2139. Available at: https://doi.org/10.1175/JPO-D-20-0020.1.

Tedesco, P. *et al.* (2023) 'Bias Correction of Operational Storm Surge Forecasts Using Neural Networks'. arXiv. Available at: https://doi.org/10.48550/arXiv.2301.00892.

Vosper, E. *et al.* (2023) 'Deep Learning for Downscaling Tropical Cyclone Rainfall to Hazard-Relevant Spatial Scales', *Journal of Geophysical Research: Atmospheres*, 128(10). Available at: https://doi.org/10.1029/2022JD038163.

Wang, J. *et al.* (2021) 'Fast and accurate learned multiresolution dynamical downscaling for precipitation', *Geoscientific Model Development*, 14(10), pp. 6355–6372. Available at: https://doi.org/10.5194/gmd-14-6355-2021.

Wang, Y. *et al.* (2026) 'A multi-resolution Feature fusion network for significant wave height downscaling', *Ocean Engineering*, 343, p. 123401. Available at: https://doi.org/10.1016/j.oceaneng.2025.123401.

Xu, S. and Hao, Q. (2026) 'Downscaling prediction of significant wave height in nearshore waters', *Ocean Dynamics*, 76(4), p. 47. Available at: https://doi.org/10.1007/s10236-026-01804-9.

Yu, M. *et al.* (2024) 'Deep learning approach for downscaling the significant wave height based on CBAM_CGAN', *Ocean Engineering*, 312, p. 119169. Available at: https://doi.org/10.1016/j.oceaneng.2024.119169.

Yu, M. *et al.* (2025) 'Wind-driven nearshore wave downscaling based on a coupled deep learning model', *Ocean Engineering*, 339, p. 122151. Available at: https://doi.org/10.1016/j.oceaneng.2025.122151.

Zhu, X., Wu, K. and Huang, W. (2023) 'Deep learning approach for downscaling of significant wave height data from wave models', *Ocean Modelling*, 185, p. 102257. Available at: https://doi.org/10.1016/j.ocemod.2023.102257.

7 Appendix

Appendix 1

Full validation results are presented here to accompany Section 3.1.

	Bilinear Interpolation			U-Net (AllWaves)			U-Net (RelativeWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	0.238 ± 0.086	0.306 ± 0.162	0.088 ± 0.051	0.040 ± 0.012	**0.048 ± 0.026**	0.081 ± 0.037	0.042 ± 0.010	0.053 ± 0.029	**0.071 ± 0.026**
1-2km	0.234 ± 0.077	0.276 ± 0.136	0.087 ± 0.045	0.045 ± 0.013	0.049 ± 0.024	0.076 ± 0.029	0.048 ± 0.016	0.057 ± 0.031	0.067 ± 0.029
2-5km	0.183 ± 0.066	0.213 ± 0.116	0.093 ± 0.055	0.030 ± 0.010	0.035 ± 0.018	0.057 ± 0.020	0.031 ± 0.010	0.040 ± 0.023	0.047 ± 0.023
5-10km	0.077 ± 0.027	0.083 ± 0.045	0.055 ± 0.036	0.022 ± 0.007	0.026 ± 0.013	0.039 ± 0.015	0.023 ± 0.007	0.026 ± 0.013	0.033 ± 0.017
10-20km	0.031 ± 0.011	0.038 ± 0.022	0.027 ± 0.015	0.018 ± 0.007	0.022 ± 0.011	0.026 ± 0.013	0.019 ± 0.007	0.022 ± 0.012	0.027 ± 0.014
20-50km	0.019 ± 0.008	0.030 ± 0.018	0.020 ± 0.011	0.013 ± 0.006	0.019 ± 0.012	0.017 ± 0.009	0.014 ± 0.006	0.019 ± 0.012	0.018 ± 0.010
>50km	0.010 ± 0.005	0.016 ± 0.010	0.009 ± 0.005	0.010 ± 0.005	0.014 ± 0.009	0.009 ± 0.003	0.010 ± 0.004	0.013 ± 0.008	0.009 ± 0.004

	ΔU-Net (AllWaves)			ΔU-Net (RelativeWaves)			WGAN (AllWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	0.043 ± 0.015	0.058 ± 0.028	0.077 ± 0.028	**0.035 ± 0.011**	0.049 ± 0.028	0.074 ± 0.027	**0.080 ± 0.026**	0.130 ± 0.045	**0.119 ± 0.017**
1-2km	0.043 ± 0.015	0.055 ± 0.027	0.074 ± 0.029	**0.036 ± 0.011**	**0.045 ± 0.026**	**0.063 ± 0.026**	0.098 ± 0.027	0.136 ± 0.036	0.157 ± 0.018
2-5km	0.032 ± 0.010	0.039 ± 0.020	0.050 ± 0.022	**0.026 ± 0.008**	**0.032 ± 0.017**	**0.044 ± 0.021**	0.087 ± 0.016	0.117 ± 0.020	0.172 ± 0.013
5-10km	0.022 ± 0.008	0.027 ± 0.014	**0.031 ± 0.015**	**0.021 ± 0.006**	**0.024 ± 0.011**	0.031 ± 0.016	0.069 ± 0.010	0.094 ± 0.013	0.138 ± 0.008
10-20km	0.018 ± 0.006	0.020 ± 0.011	0.026 ± 0.013	**0.017 ± 0.006**	**0.020 ± 0.010**	**0.025 ± 0.012**	0.056 ± 0.008	0.078 ± 0.010	0.078 ± 0.011
20-50km	0.013 ± 0.006	0.017 ± 0.011	**0.016 ± 0.009**	**0.012 ± 0.005**	**0.017 ± 0.010**	0.017 ± 0.009	0.043 ± 0.009	0.064 ± 0.021	0.043 ± 0.008
>50km	**0.008 ± 0.003**	**0.010 ± 0.006**	0.008 ± 0.003	0.008 ± 0.003	0.011 ± 0.006	**0.008 ± 0.002**	0.037 ± 0.017	0.062 ± 0.050	0.039 ± 0.011

	WGAN (RelativeWaves)			ΔWGAN (AllWaves)			ΔWGAN (RelativeWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	0.106 ± 0.010	0.127 ± 0.032	0.157 ± 0.013	0.083 ± 0.015	**0.094 ± 0.027**	0.143 ± 0.030	0.092 ± 0.015	0.109 ± 0.030	0.175 ± 0.022
1-2km	0.112 ± 0.007	0.118 ± 0.017	0.200 ± 0.011	**0.086 ± 0.011**	**0.090 ± 0.021**	**0.156 ± 0.028**	0.086 ± 0.012	0.098 ± 0.026	0.161 ± 0.025
2-5km	0.077 ± 0.006	0.081 ± 0.011	0.155 ± 0.009	0.059 ± 0.010	**0.063 ± 0.021**	0.097 ± 0.029	**0.057 ± 0.011**	0.064 ± 0.025	**0.094 ± 0.020**
5-10km	0.042 ± 0.008	0.051 ± 0.014	0.089 ± 0.008	**0.036 ± 0.010**	**0.041 ± 0.020**	0.059 ± 0.024	0.036 ± 0.011	0.041 ± 0.021	**0.051 ± 0.019**
10-20km	0.030 ± 0.010	0.039 ± 0.017	0.047 ± 0.016	**0.028 ± 0.009**	**0.033 ± 0.018**	**0.038 ± 0.016**	0.029 ± 0.010	0.033 ± 0.018	0.041 ± 0.019
20-50km	0.025 ± 0.013	0.037 ± 0.028	0.033 ± 0.018	**0.019 ± 0.007**	**0.027 ± 0.016**	**0.022 ± 0.010**	0.023 ± 0.010	0.031 ± 0.020	0.030 ± 0.015
>50km	0.049 ± 0.033	0.043 ± 0.048	0.116 ± 0.086	**0.014 ± 0.005**	**0.019 ± 0.011**	**0.013 ± 0.005**	0.023 ± 0.010	0.029 ± 0.020	0.028 ± 0.011

Appendix Table 1. Mean absolute error (metres) per bounding box, binned by distance from the coast, over the validation dataset per model. MAE is reported as the average MAE per storm. The error is defined as the standard deviation across storms. Entries in bold represent the lowest error amongst their architecture group (U-Net or WGAN).

Distance: <=1km	Bilinear Interpolation			U-Net (AllWaves)			U-Net (RelativeWaves)		
	NY	Boston	Virginia	NY	Boston	Virginia	NY	Boston	Virginia
>q95	0.508	0.684	0.538	0.171	0.094	**0.278**	0.145	0.118	0.371
>q99	0.796	1.119	0.565	0.219	0.106	0.291	0.225	0.164	0.240
>q999	-	2.065	-	-	0.152	-	-	**0.151**	-

Distance: <=1km	ΔU-Net (AllWaves)			ΔU-Net (RelativeWaves)			WGAN (AllWaves)		
	NY	Boston	Virginia	NY	Boston	Virginia	NY	Boston	Virginia
>q95	0.112	0.107	0.359	**0.098**	**0.092**	**0.357**	0.262	0.205	0.330
>q99	0.128	0.142	0.135	**0.124**	**0.126**	0.089	0.406	0.195	0.410
>q999	-	0.217	-	-	0.075	-	-	0.106	-

Distance: <=1km	WGAN (RelativeWaves)			ΔWGAN (AllWaves)			ΔWGAN (RelativeWaves)		
	NY	Boston	Virginia	NY	Boston	Virginia	NY	Boston	Virginia
>q95	0.293	0.195	0.373	**0.186**	0.134	0.285	0.197	**0.127**	**0.276**
>q99	0.414	0.243	0.097	**0.285**	0.185	0.426	0.317	**0.168**	0.564
>q999	-	**0.565**	-	-	0.247	-	-	0.261	-

Appendix Table 2. Mean absolute error (metres) over the validation dataset per bounding box in the ≤1 km from the coast bin, calculated over pixels where the corresponding target pixel is above the percentile calculated from the respective distance bin from the two-region dataset. For the ≤1 km from the coast bin, these values are 1.71 m, 3.05 m, 4.73 m, for each respective percentile. Entries in grey are based on only six pixels. Only Boston had values for >$q_{99.9}$. Entries in bold represent the lowest error amongst their architecture group (U-Net or WGAN).

Appendix 2

Full test results are presented here to accompany Table 2 and Table 3 in Section 3.2.

	Bilinear Interpolation			**U-Net (AllWaves)**			**U-Net (RelativeWaves)**		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	0.374 ± 0.236	0.297 ± 0.124	0.128 ± 0.079	84.4 ± 1.6%	84.6 ± 1.7%	9.7 ± 12.0%	82.9 ± 2.5%	83.3 ± 2.8%	**16.2 ± 12.8%**
1-2km	0.350 ± 0.187	0.269 ± 0.106	0.120 ± 0.063	81.7 ± 1.2%	82.6 ± 1.4%	15.4 ± 10.6%	81.7 ± 1.2%	81.3 ± 2.2%	23.0 ± 3.2%
2-5km	0.299 ± 0.194	0.208 ± 0.089	0.135 ± 0.084	84.9 ± 1.8%	84.6 ± 1.9%	37.9 ± 12.7%	84.6 ± 1.6%	83.5 ± 2.1%	50.3 ± 3.7%
5-10km	0.140 ± 0.108	0.081 ± 0.039	0.085 ± 0.063	76.0 ± 4.0%	71.4 ± 4.8%	26.6 ± 13.2%	74.8 ± 3.9%	71.4 ± 4.7%	38.8 ± 5.3%
10-20km	0.058 ± 0.049	0.034 ± 0.018	0.048 ± 0.035	49.0 ± 9.3%	43.4 ± 10.3%	5.8 ± 4.4%	47.6 ± 10.2%	45.4 ± 8.7%	5.0 ± 4.4%
20-50km	0.031 ± 0.027	0.027 ± 0.016	0.036 ± 0.027	33.9 ± 9.9%	38.9 ± 12.9%	15.1 ± 3.0%	31.9 ± 13.2%	40.0 ± 14.3%	13.7 ± 4.6%
>50km	0.015 ± 0.011	0.015 ± 0.010	0.015 ± 0.012	2.7 ± 9.1%	12.3 ± 6.2%	5.1 ± 16.7%	5.8 ± 12.9%	11.7 ± 13.0%	9.8 ± 12.7%
	ΔU-Net (AllWaves)			**ΔU-Net (RelativeWaves)**			**WGAN (AllWaves)**		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	83.4 ± 1.6%	80.7 ± 2.1%	12.7 ± 14.4%	**85.8 ± 1.4%**	**84.7 ± 1.3%**	14.2 ± 14.0%	**67.7 ± 2.4%**	54.9 ± 11.3%	**-36.4 ± 43.8%**
1-2km	83.6 ± 1.3%	79.7 ± 2.3%	16.4 ± 7.4%	**86.1 ± 1.0%**	**85.1 ± 1.5%**	**27.3 ± 4.3%**	62.5 ± 3.6%	48.1 ± 18.1%	**-70.5 ± 54.9%**
2-5km	84.1 ± 1.8%	82.0 ± 1.8%	44.9 ± 6.3%	**87.5 ± 1.3%**	**86.3 ± 1.8%**	**51.8 ± 3.3%**	59.3 ± 7.2%	40.8 ± 27.3%	-82.6 ± 75.7%
5-10km	75.1 ± 3.6%	70.8 ± 3.3%	39.0 ± 5.2%	**77.6 ± 3.8%**	**73.7 ± 4.5%**	**39.9 ± 5.0%**	25.7 ± 15.4%	-19.6 ± 53.7%	-161.5 ± 120.4%
10-20km	50.0 ± 11.2%	47.5 ± 9.4%	6.4 ± 4.9%	**51.5 ± 10.3%**	**49.7 ± 10.1%**	**11.5 ± 3.9%**	-54.3 ± 37.1%	-131.4 ± 90.2%	-170.9 ± 89.9%
20-50km	**37.6 ± 11.9%**	45.5 ± 14.4%	19.1 ± 2.4%	37.2 ± 14.3%	**45.6 ± 16.3%**	**18.6 ± 2.8%**	-200.8 ± 234.0%	-161.1 ± 158.0%	-112.1 ± 74.4%
>50km	**25.8 ± 11.2%**	**30.9 ± 15.7%**	18.6 ± 12.8%	23.8 ± 13.4%	26.8 ± 19.4%	**17.3 ± 15.6%**	-348.9 ± 216.6%	-309.6 ± 285.6%	-279.9 ± 109.4%
	WGAN (RelativeWaves)			**ΔWGAN (AllWaves)**			**ΔWGAN (RelativeWaves)**		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
<=1km	55.9 ± 8.9%	54.5 ± 13.8%	-85.4 ± 65.3%	65.4 ± 6.6%	**65.7 ± 9.0%**	-62.7 ± 43.4%	62.4 ± 6.6%	59.9 ± 10.3%	-104.3 ± 66.6%
1-2km	53.9 ± 11.3%	52.7 ± 17.5%	-129.6 ± 79.5%	63.6 ± 6.5%	**63.8 ± 9.1%**	-75.3 ± 43.9%	**64.6 ± 5.7%**	60.7 ± 9.0%	-80.7 ± 48.9%
2-5km	60.1 ± 11.8%	58.2 ± 16.7%	-72.8 ± 73.5%	68.2 ± 5.1%	**68.8 ± 7.4%**	-5.0 ± 23.9%	**70.3 ± 4.9%**	68.6 ± 5.5%	**-2.4 ± 29.1%**
5-10km	51.6 ± 9.0%	38.3 ± 17.0%	-73.2 ± 76.1%	57.4 ± 6.0%	**51.8 ± 7.2%**	-10.4 ± 18.0%	**57.9 ± 5.6%**	51.2 ± 6.8%	**1.2 ± 16.3%**
10-20km	9.4 ± 11.5%	-0.1 ± 17.0%	-59.4 ± 33.9%	**12.6 ± 16.4%**	**12.4 ± 15.1%**	**-34.4 ± 18.1%**	12.5 ± 17.7%	10.8 ± 15.2%	-42.1 ± 16.3%
20-50km	-35.7 ± 16.1%	-19.6 ± 27.5%	-64.2 ± 24.9%	**-9.1 ± 20.0%**	**8.4 ± 19.6%**	**-8.5 ± 12.9%**	-26.3 ± 28.6%	-6.1 ± 22.4%	-43.2 ± 19.3%
>50km	-425.9 ± 163.3%	-129.4 ± 53.6%	-1303.1 ± 421.0%	**-45.5 ± 24.6%**	**-30.3 ± 17.7%**	**-40.8 ± 35.8%**	-126.4 ± 42.6%	-86.8 ± 22.1%	-212.1 ± 71.4%

Appendix Table 3. Mean absolute error per bounding box, binned by distance from the coast, over the held-out test set. The baseline is provided in absolute units (metres), with trained model performance reported as the percentage reduction in MAE, relative to the baseline for ease of interpretation. Entries in bold represent the lowest error amongst their architecture group (U-Net or WGAN).

Distance: <=1km	Bilinear Interpolation			U-Net (AllWaves)			U-Net (RelativeWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
>q95	0.621	0.725	0.533	0.189	0.093	**0.340**	0.165	0.108	0.425
>q99	0.909	1.081	0.890	0.253	0.124	0.297	0.228	0.157	0.269
>q999	1.041	1.399	-	0.341	0.122	-	**0.288**	**0.176**	-

Distance: <=1km	ΔU-Net (AllWaves)			ΔU-Net (RelativeWaves)			WGAN (AllWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
>q95	0.149	0.111	0.431	**0.118**	**0.086**	**0.406**	0.362	0.244	0.358
>q99	0.231	0.151	0.482	**0.169**	**0.098**	0.389	0.560	0.294	0.202
>q999	0.365	0.212	-	0.223	0.131	-	0.621	0.495	-

Distance: <=1km	WGAN (RelativeWaves)			ΔWGAN (AllWaves)			ΔWGAN (RelativeWaves)		
	NY	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**	**NY**	**Boston**	**Virginia**
>q95	0.365	0.221	0.498	**0.241**	0.145	0.310	0.241	**0.142**	**0.286**
>q99	0.613	0.343	0.468	**0.434**	0.214	0.233	0.448	**0.182**	0.138
>q999	**1.015**	**0.632**	-	0.591	0.309	-	0.793	0.348	-

Appendix Table 4. Mean absolute error for all models per bounding box in the ≤1 km from the coast bin over the held-out test set, calculated over pixels where the corresponding target pixel is above the percentile calculated from the respective distance bin from the two-regime dataset. For the ≤1 km from the coast bin, these values are 1.71 m, 3.05 m, 4.73 m, for the respective percentiles. Entries in bold represent the lowest error amongst their architecture group (U-Net or WGAN).