https://doi.org/10.5194/egusphere-2026-781
Preprint. Discussion started: 23 February 2026

# Physically Coherent Machine Learning for Tropical Cyclone Storm Surge Emulation

Hamish Wilkinson[1,2], Paul Bates[1,2], Chris Lucas[2], Niall Quinn[2], Ivan D. Haigh[3,4], Tom Collings[2], Peter Watson[1]

1. School of Geographical Sciences, University of Bristol, University Rd, Bristol, BS8 1SS, UK
2. Fathom, Floor 2 Clifton Heights, Clifton, Bristol, BS8 1EJ, UK
3. School of Ocean and Earth Science, National Oceanography Centre, University of Southampton, European Way, SO14 3ZH, UK
4. National Centre for Integrated Coastal Research and Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, USA

*Correspondence to:* Hamish Wilkinson (h.wilkinson.2020@bristol.ac.uk)

## Abstract

Climate change is projected to impact tropical cyclone magnitude and frequency, with high magnitude events becoming more common. The destructive nature of event derived storm surges and associated coastal flooding necessitates risk management. However, the historic record is too short and too sparse to assess risk effectively, resulting in incomplete probability distributions of surge heights, particularly for distribution tails. Hydrodynamic simulation can fill these gaps, but the number of simulations required, both spatially and under diverse climates, coupled with their high computational cost, is prohibitive. To address this, we present an Artificial Neural Network storm surge emulator, deployed in the northwest Gulf of Mexico. This is trained on a database of hydrodynamic simulations, and outputs spatially coherent time series of surge. Our model achieves an $R^2$ of 0.91, with a RMSE of 13cm when compared to an independent test set of hydrodynamic simulations, while exhibiting a computational gain factor of over 1500. Our approach is novel in its use of feature engineering to improve performance. Here variables which are physically relevant to surge are derived from commonly used features, such as wind and pressure, allowing us to maintain a simple model architecture, while steering the model towards physically coherent learning. Shapley Values are utilised for model interpretation and demonstrate that the model is making physically justified inference. Success is demonstrated by comparing our feature engineered model to a control, which engages in minimal feature engineering. The control achieves an $R^2$ of 0.71 and a RMSE of 23cm only.

## 1. Introduction

Coastal flooding is one of the deadliest and costliest hazards globally, with an estimated 34 million people affected in 2015, leading to economic costs of 0.3% of global GDP. This is projected to increase to between 119-246 million and 1.1-2.9% respectively by 2100, depending on adaptation scenario (Kirezci et al., 2023). While much of this increase in risk is driven by changes to mean sea level and socioeconomic change (Lincke et al., 2022) there are also indications that variations to storminess can have significant impacts regionally (Amarouche and Akpınar, 2021; Calafat et al., 2022; Mölter et al., 2016; Wood et al., 2023). Notable amongst these are changes to Tropical Cyclone (TC) dynamics, with the frequency of the most damaging and highest magnitude storms projected to increase, while overall frequency is projected to decrease, albeit with some spatial variation dependent on ocean basin (Emanuel, 2021; Marsooli et al., 2019; Webster et al., 2005). Regardless of magnitude, there are also indications that the intensification rate of TCs has undergone a recent increase, with links to climate change (Bhatia et al., 2022; Bhatia et al., 2019).

Even without climate change, the destructive nature and high potential cost of coastal flooding from landfalling TCs necessitates risk management, with hydrodynamic and statistical models developed to aid this goal (Pringle et al., 2021; Zhang et al., 2016). Hydrodynamic models can be used to simulate historic events, with boundary conditions taken from observations. However, the relative rarity of TCs means that the historical record is too short, and further too sparsely distributed spatially, to make robust estimates of return periods. In an attempt to offset these shortcomings, statistical methods (Jewson, 2024) and regional frequency analyses (Collings et al., 2024) have been utilised to provide more comprehensive views of risk, but these methods are constrained by large uncertainties emanating from difficulties that climate models have in representing TCs (Emanuel, 2013; Jewson, 2024), and also the limitations of the historic record, even where this is augmented by regional pooling(Collings et al., 2024). Strategies for overcoming these obstacles often focus on extending the historic record through the creation of many synthetic TC tracks (Bloemendaal et al., 2020; Loridan and Bruneau, 2025) and the simulation of storm surges created by these hypothetical systems using hydrodynamic models (Dawson et al., 2021; Johnson and Ahmadi, 2023). Further efforts have focused on downscaling CMIP6 projections (Emanuel, 2013), with a view to extend the probability distribution of TCs both spatially and in magnitude. However, simulation by hydrodynamic models is computationally expensive, particularly when done at high resolution, with computational cost increasing by approximately an order of magnitude for every halving of model grid size (Bates et al., 2021). The scale of the problem makes hydrodynamic modelling prohibitive, with many thousands of events required simply to achieve the spatial density needed for a comprehensive view of risk, and exponentially more required to determine the full probability distribution of climate, magnitude, sea level, and tidal phase.

This challenge has led to investigations into the use of machine learning (ML) techniques to emulate hydrodynamic models and simulate the required number of events at a much-reduced computational cost(Dong et al., 2022; Qin et al., 2023; Tiggeloven et al., 2021). Approaches vary from the use of Artificial Neural Networks (ANNs) (Bruneau et al., 2020), Long-Term-Short-Term Networks (LSTM) (Giaremis et al., 2024), Convolutional Neural Networks (CNNs) (Adeli et al., 2023), with complexity and data demands increasing sequentially. The advantage of these models is that once trained they can make inferences in a matter of minutes or less, compared to hydrodynamic model runtimes of hours to days (Dietrich et al., 2012). While the methodologies differ, the basic input for each of these ML models to predict surge tends to be similar, with pressure, wind, TC characteristics, coordinates, and bathymetry common (Dong et al., 2022; Qin et al., 2023). These requirements necessitate that data must be first available to train the ML model, either through the historic record, or from hydrodynamic model simulations. Due to these constraints the ability of ML models to generalise beyond their training dataset is paramount. Firstly, if trained on the historic record they must be able to generalise beyond observed storm magnitudes or risk failing to capture unobserved extremes. Additionally, in the case of both modelled synthetic and historic events ML models must be able to generalise spatially to new regions. If they cannot achieve these aims, they may both underestimate future risk in a non-stationary world and further will restrict improvements to areas which are comparatively data rich (Hernanz et al., 2024; Watson, 2022).

These issues are brought into sharp relief when considering that many current examples in the literature using ANNs (Bruneau et al., 2020; Qin et al., 2023; Ramos-Valle et al., 2021; Tadesse et al., 2020; Tiggeloven et al., 2021) employ a limited feature set, coupled with location specific information, such as gauges or coordinates. The issues here are twofold. Let us consider the issue of limited feature set. Commonly used features include wind vectors, pressure, and some TC characteristics such as the radius of maximum winds. However, even when all these processes are identical, the surge response can be vastly different depending on location, with these changes driven by other processes which are not typically included in the feature set – for example angle of approach (Ramos-Valle et al., 2021), fetch across the continental shelf, slope, and coastal geometry to name a few (Jelesnianski, 1972; Rego and Li, 2010). Therefore, we argue that the commonly applied feature set is insufficient to allow ML models to learn generalisable surge dynamics. This leads us to our second point: location specific learning. If we train a ML model on a single gauge, or a collection of gauges, with a gauge id as a feature, we can sidestep the above issues. Despite the lack of sufficient richness of features in the training dataset, the model is able to learn these implicitly for each location (Lockwood et al., 2022). However, when we try to apply this model to new locations, say in a data sparse area where there are no gauges to train on, its location specific implicit learning no longer applies, and its performance will be significantly degraded – a classic case of overfitting. Using spatial coordinates as a model input also suffers from the same issues. While coordinates are marginally more generalisable, with locations nearer to one another more likely to experience similar conditions, this still has only limited utility (Karimzadeh et al., 2025). Proximal locations can still experience vastly different surges if one is in an estuary and the other is open coast (Qian et al., 2024; Tiwari et al., 2025). Further, if we attempt to apply a model built in this way in an entirely new domain, its coordinate-based learning is redundant, and once again performance degrades (Meyer et al., 2019).

In some cases, this lack of generalisability is acceptable, for example location specific forecasting or sensitivity testing. While this may be true, we maintain that even in these cases performance can be improved if we provide a more comprehensive, and location invariant, feature set. These features allow the model to gain a richer understanding of surge drivers and apply these more generally. For example, if gauge *x* has not been impacted by a TC in the historical record, but gauge *y* has, which though spatially distant has very similar physical characteristics, then the model can apply the learning gleaned from gauge *y* to gauge *x*. However, this is only possible if the feature set captures these similarities. In essence, this ML approach applies many of the same principles developed for the regional frequency analysis that form part of extreme value statistics (Collings et al., 2024; Smith et al., 2015).

The process of expanding a feature set can be as simple as providing additional features which we know can impact storm surge. It can also relate to deriving new features from existing fields, for example slope from bathymetry, or spatial gradients from pressure fields. Additional techniques include manipulation of features into forms which ML models are more efficient at learning from, for example transforming data to a more normal distribution, or scaling data between a mean of 0 and 1 standard deviation (Standard Scaling). Collectively these changes are referred to as feature engineering (Koukaras and Tjortjis, 2025).

Note that there is a school of thought that maintains that with a sufficiently deep neural network feature engineering is unnecessary and these representations of existing features will be discovered. While there is some merit to this, by deliberately deriving new features we can reduce model complexity and training time, while guiding the model towards learning relationships that we know to be physically coherent, and so more generalisable (Mumuni and Mumuni, 2025). Further, by providing these features we believe that we can also improve model interpretability, using methods such as feature permutation importance, which allows us to rank feature contribution (Fisher et al., 2019). This process also enables feature set reduction to those which are most impactful in order to gain the best performing and simplest model, an important consideration, as adding more features also increases model complexity, and input data preprocessing.

In this paper we expand on the above themes and demonstrate that ANN models can be made more robust by engaging in significant feature engineering, in which we manipulate

https://doi.org/10.5194/egusphere-2026-781
Preprint. Discussion started: 23 February 2026

commonly available input data such as pressure, wind and bathymetry. This enhances the model's understanding of spatial and temporal characteristics which affect storm surge, thus negating the need to provide information such as coordinates, while enhancing model performance and improving model generalisability. Crucially, this creates greater feature diversity, without additional datasets. We use the northwest coast of the Gulf of Mexico as a case study region, as it is exposed to significant TC activity and modelled datasets of storm surge are readily available(Dawson et al., 2021). We hope this paper will serve as a feature engineering blueprint for future applications of ML in the field of storm surge, by highlighting which features are most impactful. While we use an ANN in this paper, we believe that the principles are fully applicable to other ML modelling frameworks such as LSTMs and CNNs.

## 2. Methods

Our approach is described in seven parts. Firstly, we describe the study area. Secondly, we give a brief overview of our choice of ML model. Thirdly we describe the methods and criteria used to select our storm set and train our model. Following we discuss the philosophy behind our approach to feature engineering. Next, we illustrate our neural network architecture and training. Following on we describe ML model interpretability and feature importance, and how this feeds into feature selection. Finally, we give an overview of the quantitative and qualitative methods used to assess model performance. The framework is illustrated in Fig. 2.

### 2.1 Study Area

The northwest of the Gulf of Mexico, and specifically the Texas coastline, has been chosen primarily due to data availability, with a large synthetic set of hydrodynamic simulations present (Dawson et al., 2021). Further, it provides a good test for our ANN model, as it includes a varied and complex coastal geometry including barrier bays and estuaries, as well as wide and narrow sections of continental shelf. Finally, this is an area of considerable applied interest, as it contains dense concentrations of population and assets in areas such as Houston. Figure 1 illustrates the study area and storm track distributions.

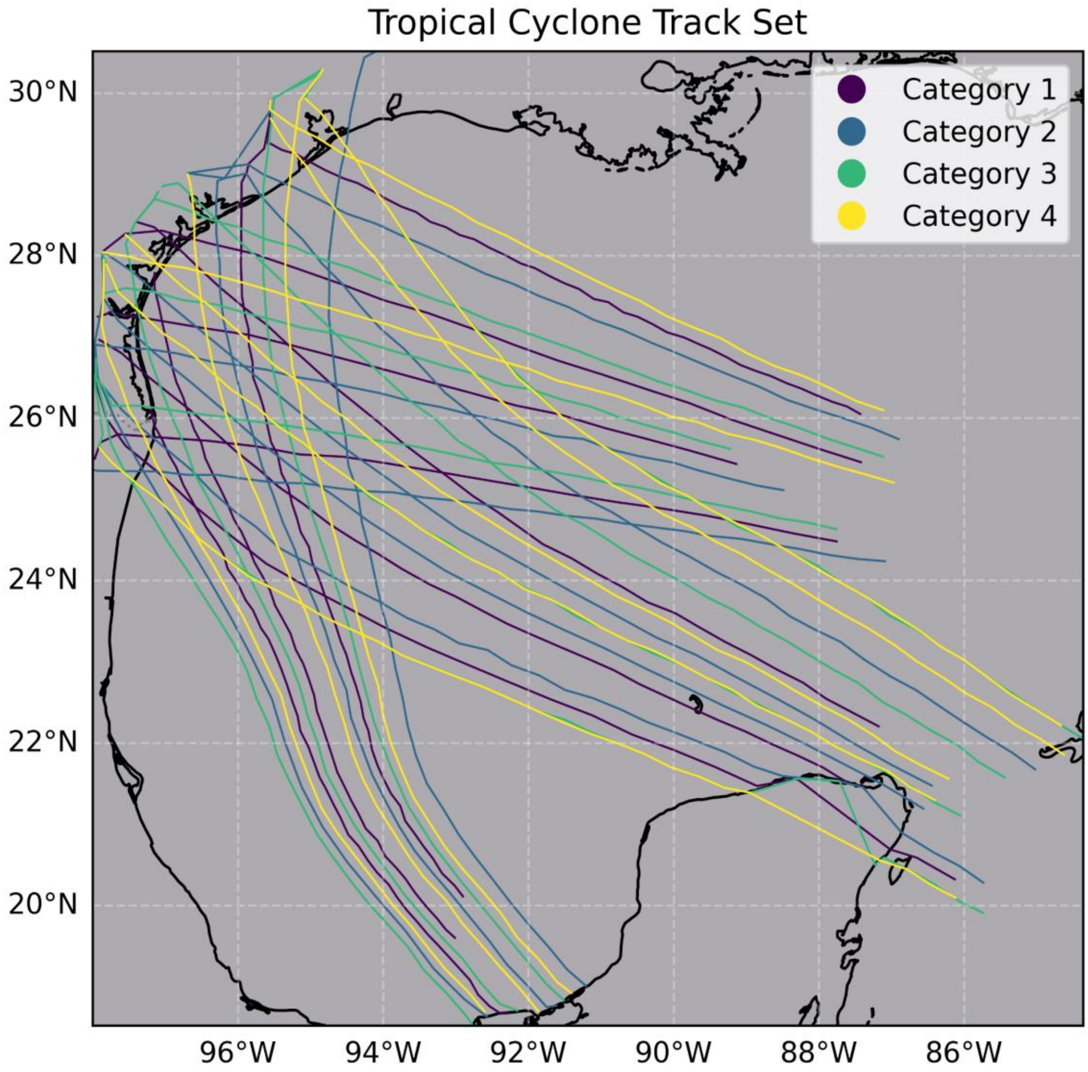


*Figure 1 - Plot illustrating the subset of tracks used in this study. Note there are 10 events from each category. Events have been selected to ensure good spatial coverage of the domain for each category. Note that these tracks are synthetic, and as such the paths are highly idealised. This limitation will be examined in the discussion.* 200

## 2.2 Choice of Machine Learning Model

As noted in Sect. 1, there are a wide variety of machine learning models. Here we specifically use deep learning neural networks (Rumelhart et al., 1986) (SI 1). For this paper we have 205 utilised an ANN model (Sazli, 2006). For our input layer we provide a 2D tensor array in which each row is a single point in space and time and each column is a feature. Our output for each row is a single continuous value representing the model's prediction of surge. Each row is independent, with multiple rows comprising a batch.

We have chosen an ANN for several reasons. Firstly, while there is a wealth of machine learning approaches for deep learning, an ANN is amongst the simplest, allowing us to focus efforts on feature engineering, rather than designing a complex model architecture. Secondly, all features are derived from unstructured meshes commonly used in ocean hydrodynamic models (Pain et al., 2005). These meshes exhibit higher resolution close to shore, and lower resolution offshore, providing our model with more data where we are most interested. ANNs here are well suited, as they are able to ingest data with this mixed resolution, unlike frameworks such as CNNs which require rectilinear grids (O'shea and Nash, 2015). Finally, by using a model which makes predictions for a single point in space and time, we are able to demonstrate the utility of feature engineering designed to provide awareness of surrounding points both spatially and temporally outside of using ML model frameworks such as CNNs and LSTMs (Krichen and Mihoub, 2025), which explicitly take consideration of spatial and temporal information respectively.

### 2.3 Data preparation

To train a ML model our data must include predictors, or features in ML parlance, and a target variable, surge in this case. The target variable is often referred to as the label. Therefore, we require a dataset including both surge data points, and variables which are known to cause surge – for example pressure and wind fields. In this section we will describe this data sourcing, and the processing required to allow input into our model framework.

Hydrodynamic simulations are sourced from the Texas Federal Emergency Management Agency (FEMA) Hurricane Winds and Surge collection of synthetic idealised storms (Dawson et al., 2021). This is a repository of 446 synthetic storms, developed by the United States Army Corp of Engineers (USACE) and FEMA. Storm tracks were synthesised by selecting probable combinations of radius of maximum winds, central pressure, forward speed, angle of approach and track location. Five primary tracks were synthesised, with other tracks derivative of these. Wind and pressure fields are then generated from these tracks using the Planetary Boundary Layer Model (PBL) provided by Oceanweather (Cardone and Cox, 2009; Dawson et al., 2021). PBL uses equations of horizontal motion vertically averaged, and is forced by storm location, minimum central pressure and maximum wind speed. Wind field radial structure was defined by the Holland B parameter (Holland, 2008), with values ranging from 1 to 1.3. Unsymmetrical wind flow is accounted for through use of a steering parameter. Wind fields are described at 10m and with pressure are input into ADCIRC to generate surge.

ADCIRC is a hydrodynamic model (Pringle et al., 2021), which solves the shallow water equations on a finite element unstructured triangular mesh. In this case it was run in 2D, with equations of motion averaged over water depth. However, it can also be run in 3D, with equations solved both vertically and horizontally (Luettich and Westerink, 2004). For the storm collection utilised here, ADCIRC uses the TX2007 mesh developed by FEMA, which contains more than 2 million computational nodes, with mesh element size ranging from ~30m at the coast to ~10,000m beyond the continental shelf. Bathymetry is an amalgamation of the ETOPO5 global relief model (National Geophysical Data, 1993) with Digital Nautical Chart values overlain where available (National Geospatial-Intelligence, 2026). Bathymetry is then updated using the National Oceanic and Atmospheric Administration's (NOAA) depth sounding database (Noaa, 2004). NOAA navigational charts were used for bathymetry validation with data deemed incorrect or missing replaced with values from these charts. Finally, grid scale filtering was used to assign bathymetry values to each mesh node. Bottom friction was defined by setting the Manning's roughness coefficient (Manning, 1891) to 0.02, with the key exception being on the Louisianan-Texas continental shelf, with Manning's coefficient reduced to 0.012 in water depths greater than 5m. This accounted for lower friction on the wide and muddy shelf (Kennedy et al., 2011). For further details see Dawson and Clinton (Dawson et al., 2021).

Wind vectors, atmospheric pressure at sea level, depth, and water surface elevation are extracted from the ADCIRC runs, with mesh structure retained. These are passed through a selection criteria designed to both save compute, and avoid training the ML model on weak storms, as this may compromise its ability to simulate the most extreme surges (Watson, 2022). Storms which fail to make landfall are discarded, as are those in which the minimum atmospheric pressure never falls below 1013mb. This data cleaning leaves 385 storms, varying in strength at landfall from Tropical Storm to Category 4. From these storms 40 were regularly sampled. To ensure a representative sample, tracks were selected to cover the full spatial distribution of storms, with storm magnitudes likewise regularly distributed (Fig. 1). Sensitivity testing (SI 2) indicated that model performance improvement continues with larger sample sizes, but concerns of overfitting on the relatively homogeneous synthetic set supported the sample size limitation of 40 storms.

Storm temporal length is subset based on atmospheric pressure. For example, if a storm enters the domain, but minimum atmospheric pressure is not below 1013mb on entry, all timesteps which are not below 1013mb will be discarded. However, from the time that the storm first falls below 1013mb, it will be included in an unbroken time series up to at least 6hrs after landfall, regardless of atmospheric pressure.

### 2.4 Feature Engineering

The purpose of this study is to produce a highly generalisable, and easily applicable model globally. As such, all extra features included in the model must be easily derived from commonly available datasets. For data training we have used the aforementioned ADCIRC simulations, but any atmospheric data coupled with surge simulations could be utilised. For example, as the features used to train this ML model are easily available, either through the generation of wind and pressure fields from track sets such as STORM (Bloemendaal et al., 2020), or from atmospheric data from climate models, it is possible to acquire the required features for ML model inference or training globally. Notwithstanding, the quality of the input will determine the quality of the output; climate models are generally accepted to be too coarse to resolve Tropical Cyclones, and so atmospheric data fed into the ML models described in this paper would produce surge values in line with the atmospheric input – i.e. at a lower magnitude (Camargo and Wing, 2016; Dulac et al., 2024).

In addition to ADCIRC simulations, we have used the Copernicus land mask (Fahrland et al., 2020), and the Geomorphology of the Oceans continental shelf dataset (Athanasiou et al., 2024). This provides us with 6 original features, namely; pressure at sea level, wind u and v velocity vectors, bathymetric depth, a land ocean mask, and fetch over the continental shelf. From these we then derive a further 45 features. These are designed to cover the full spectrum of forces which potentially drive surge (SI 3.1). For context these variables can be split into the following groups:

1. Atmospheric – these include wind and pressure fields, and features derived from them, such as wind stress.
2. Bathymetric – which includes features such as depth and slope, as well as fetch across the continental shelf and coastline geometry.
3. Temporal – which includes gradients of how features change over time.
4. Spatial – which includes transects and gradients of how features change through space.

Spatial and temporal features are key, as these provide context to the model, which makes deterministic predictions for a single point in space or time (SI 3.2). Additionally, each feature undergoes Yeo-Johnson normalisation (Yeo and Johnson, 2000) and standard scaling (Koukaras and Tjortjis, 2025) (SI 3.3).

### 2.5 Neural Network Architecture and Training

The neural network utilised here is a fully connected feed forward neural network, which is designed for the regression task of predicting surge. It has two hidden layers, decreasing from 128 to 64 neurons, and a final output later, consisting of a single neuron. The hidden layers each utilise batch normalisation and dropout (SI 4.1). Both hidden layers use ReLu as an activation function, allowing the model to learn non-linear relationships (Nair and Hinton, 2010) (SI 4.2).The final output layer produces a single continuous value located at the same location in space and time as the input – this is the surge prediction. The input is a 2D tensor, for example, each row is a single point in space and time, each column is a feature, with the label, surge in this case, also occupying a column. This label is what the ML model uses to derive losses during its training.

Model hyperparameters were defined initially using a simple grid search on the full feature set model. The selected learning rate was set to 0.0001 using the Adam optimiser (Kingma, 2014), batch size to 1024, and dropout probability to 0.3 (Srivastava et al., 2014). Mean squared error (MSE) (Equation 1) was used as the loss function, with MSE being derived between ML predicted surge and ADCIRC generated surge (SI 4.3). Following model size reduction using feature permutation importance (Sect. 2.6) these hyperparameters were again fine-tuned using grid search, with batch size increasing to 2048 and learning rate to 0.0003. Additionally, Smooth $L_1$ Loss (Girshick, 2015) (Equation 2), derived from Huber Loss (Huber, 1992), replaced MSE as the loss function, with a beta threshold of 1 utilised (SI 4.4).

As described in Sect. 2.3 forty representative storms were used in this paper. These were split into training/validation/test subsets using a ratio of 7/2/1 (SI 4.5). In the framework described here, where no improvement in validation loss had occurred over 5 epochs, the model ceases training, helping to prevent overfitting (Prechelt, 2002). To create these subsets storms are first randomly assigned to a subset. They are then merged into a single 2D tensor and shuffled. Effectively, this means that individual storms are not split across the subset, ensuring no data leakage.

Equation 1 describes Mean Squared Error. $Y$ refers the "True" value, in this case ADCIRC simulated surge, while $\hat{Y}$ refers to the ML predicted value. Equation 2 describes Smooth $L_1$ Loss, here $x$ refers to the predicted value, while $y$ refers to ADCRIC. $\beta$ refers to the beta threshold. Equation 3 describes Mean Absolute Error. $Y$ refers to the "True" value, in this case ADCIRC simulated surge, while $\hat{Y}$ refers to the ML predicted value.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{1}$$

$$L_\beta(x,y) = \begin{cases} \frac{0.5(x-y)^2}{\beta}, & \text{if } |x - y| < \beta \\ |x - y| - 0.5\beta, & \text{otherwise} \end{cases} \tag{2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| \tag{3}$$

## 2.6 Feature Interpretability and Selection

### 2.6.1 Interpretability

One of the criticisms of ML models is that they operate in an opaque manner, otherwise known as black box. The ML model may be taking shortcuts, and it may be implying causal processes which are not physically justifiable. Further, even assuming that the model is coming to its predictions in a justified manner, we are unable to see relative contributions of input features and so gain insights (Bostrom et al., 2024; Linardatos et al., 2021; Mcgovern et al., 2019). To remedy this, we utilise feature permutation importance through use of Shapley values (Lundberg and Lee, 2017; Shapley, 1953) (Eq. 4) (SI 5.1). The Shapley value is computed by summing marginal contributions and weighting them in accordance with the feature set combinations (Eq. 5). This Shapley value can be compared to a model baseline value – in this case the average surge across all predictions. The difference between the

https://doi.org/10.5194/egusphere-2026-781
Preprint. Discussion started: 23 February 2026

baseline and the Shapley value is the contribution of the target feature, and when taken across all features this explains the magnitude and sign of feature contribution to the prediction. This is a local explanation, at a single point in space and time (Eq. 6). However, when averaged across a representative sample this provides us with a global explanation, akin to feature permutation importance (SI 5.2).

In Equation 4 $\phi$ is the marginal contribution. $A$ refers to a target feature. $s$ to a subset of features. $f$ to the ML model. $s_{+x_A}$ is the model including the target feature $s_{-\,x_A}$ is the model excluding it. $x_A$ is the value of the feature $A$ for the specific observation being explained. In Equation 5 $\phi$ is the Shapley value, $A$ is the target feature, with $x_A$ being the value of the feature $A$ for the specific observation being explained. $S$ represents all possible subsets of features. $w_s$ represents weights assigned to $s$. In Equation 6 $f(j)$ refers to output from the ML model at a single point in space and time. $\mu(f(J))$ refers to the average prediction.

$$\phi_s(x_A \;=\; f\left(s_{+x_A}\right) - f(s_{-\,x_A}) \tag{4}$$

$$\phi(x_A) = \; \sum_{s\epsilon S} w_s \phi_s(x_A) \tag{5}$$

$$f(j) - \mu\left(f(J)\right) = \sum_{k\epsilon K} \phi(x_A) + \ldots + \phi(x_K) \tag{6}$$

To calculate the marginal contributions for each Shapley value the prediction must be made a total of $2^{N-1}$ times, with N representing the number of features. This exponential growth is expensive, especially on models with large feature sets and many individual data points. To make this approach feasible we use SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), a computationally efficient approach described in equations 4-6. Further, we reduce feature set dimensionality using principal component analysis (Hotelling, 1933) (PCA), with the PCA set to capture 95% of variance, before calculating 500,000 centroids using k-means clustering (Arthur and Vassilvitskii, 2006). We then select the nearest neighbours to these centroids from the original feature set (SI 5.3).

### 2.6.2 Selection

Feature selection for reduced set models was undertaken with four separate methodologies. The first takes the results of the SHAP derived permutation feature importance (PFI) score and selects the 20 top ranked features. The second used a combined statistical approach using Spearman's Rank (Spearman, 1961), Mutual Information (Kraskov et al., 2004), and PFI. These are ranked, and the ranks summed. The 20 features with the lowest rank (i.e. the best performing across all three metrics) are then selected (SI 6).

The third approach relies on expert knowledge, coupled with support from SHAP scores and statistical measures. Further, it considers how closely related features are through the use of XGBoost redundancy (Lundberg et al., 2020). This works by training a neural network to use one feature in our set to predict another. The redundancy metric is the inverse of the resulting $R^2$ - so for example a feature which utterly fails to explain another would score 1, indicating no redundancy, while another which is able to perfectly predict a feature would score 0, indicating complete redundancy. Using this tool, highly redundant features have been excluded from the Expert set.

The fourth approach represents our benchmark. This is the control model, referred to as the 'Naive' model going forward. It uses only features commonly used in the literature, with no feature engineering beyond ML best practice – in this case using Yeo-Johnson normalisation and standard scaling, as described in Sect. 2.4. It consists of pressure at mean sea level, wind vectors and bathymetric depth.

Note that where possible the selection methods above were adhered to strictly, however where this would have resulted in incomplete information it has been altered. For example, two features are used specifically to define wind direction - cosine and sine fields. If cosine was ranked in the top 20, but sine was not, then sine would be included at the expense of the lowest ranked feature in the top 20. It should also be emphasised for the statistical reduced set that feature independence is assumed. However, as many of these features are defined from the same root variables this may not be the case (SI 3.1). As such it is possible that some features may be sharing importance. This includes paired features such as wind vectors, as these should be taken as the sum of their parts, but also highly correlated features such as coastline geometry and depth, whose relationship is more complex. While the pure ML and statistical approaches described above do not take this into account, the expert selection method does.

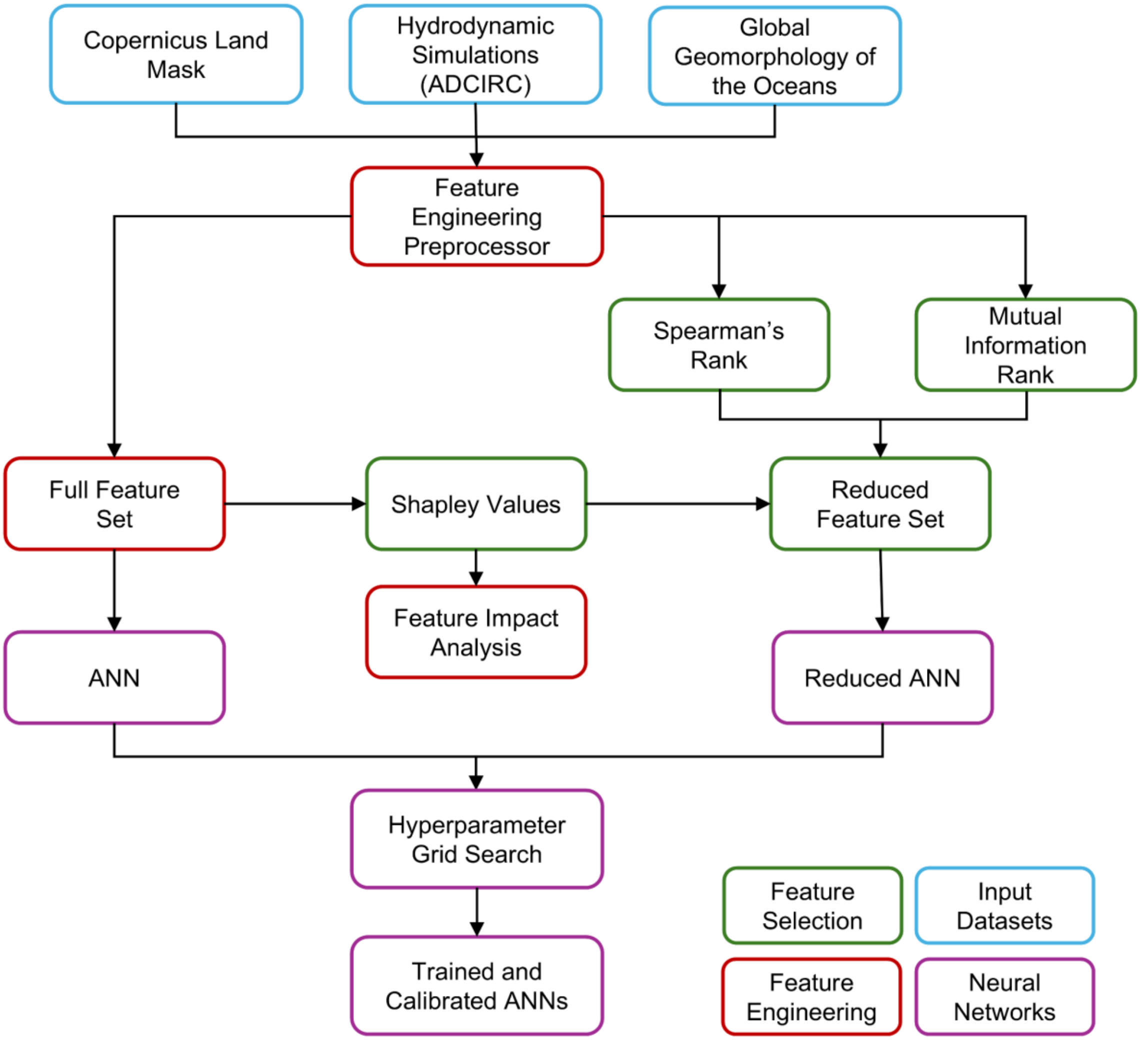


*Figure 2 - Schematic illustrating model framework. Input datasets are described in sect. 2.3, feature engineering in sect. 2.4, neural networks in 2.5, and feature selection in sect. 2.6.*

## 2.7 Model performance metrics

440 Model performance is evaluated using both quantitative and qualitative measures. Headline error statistics such as $R^2$, MAE, root mean square error (RMSE), and test loss (Huber) are utilised to provide quantitative information on error. Further, runtime and training time metrics are recorded. Qualitative analysis is performed with temporal plots of storm hydrographs, spatial maps of surge and surge anomaly, and density plots of ML predictions against

ADCIRC. Note that all results report on the test set – data that the ML model has not seen previously. Model performance has been evaluated across 3 domains: Firstly, from the 1m contour, representing performance across the full domain. Secondly, between the 1 and 5m contours, and finally from the 1 to 2m contours, illustrating the most difficult environment in which to make predictions. Finally, the internal function of the Fullset model has been analysed using SHAP values, to ensure that the model is building relationships that make physical sense.

## 3. Results

The results are presented in two sections. First, we examine the results of SHAP scores, this provides context on the internal workings of the model and indicates if it is performing in line with known storm surge dynamics. Secondly, we examine model's results against the test events using an independent hydrodynamic simulation produced by ADCIRC, with this illustrating model performance in emulating surge.

### 3.1 SHAP scores

Top performing features are plotted in Fig. 3, with the upper panel indicating mean absolute impact globally, while the bottom panel indicates local feature impact for all points in space and time. Broadly these results conform to what we would expect from storm surge physics, indicating that the model has learned which features are the most physically relevant. For example, wind is confirmed as being highly influential, with features relating to direction, speed, stress, and original vectors all being amongst the top twenty by SHAP value. The SHAP summary panel in Fig. 3 confirms that not only are wind features influential, but that their impact aligns well with theory. High wind speeds, vectors and stress all result in increased surge, while low values result in reduced surge. Wind direction features, indicated by cosine and sine fields, display an inverse relationship, indicating that these are working in an intertwined manner as intended. More difficult to explain are temporal and spatial gradients. High value spatial gradients for wind speed, indicating sharp decreases or increases in wind speed, result in decreased surge. Perhaps this is accurately capturing sheltering effects from land masses, but it may also be contributing to model difficulties in resolving the very high surges expected in the cyclone eye. This more complex interpretation of gradient extends to pressure (pressure at mean sea level (PRMSL)) features with the directionality of pressure spatial gradients producing a mixed signal. Pressure temporal gradient is clearer, with a high gradient producing reduced surge, correctly indicating that large changes in pressure produce corresponding changes in surge, effectively capturing storm eye movement over time.

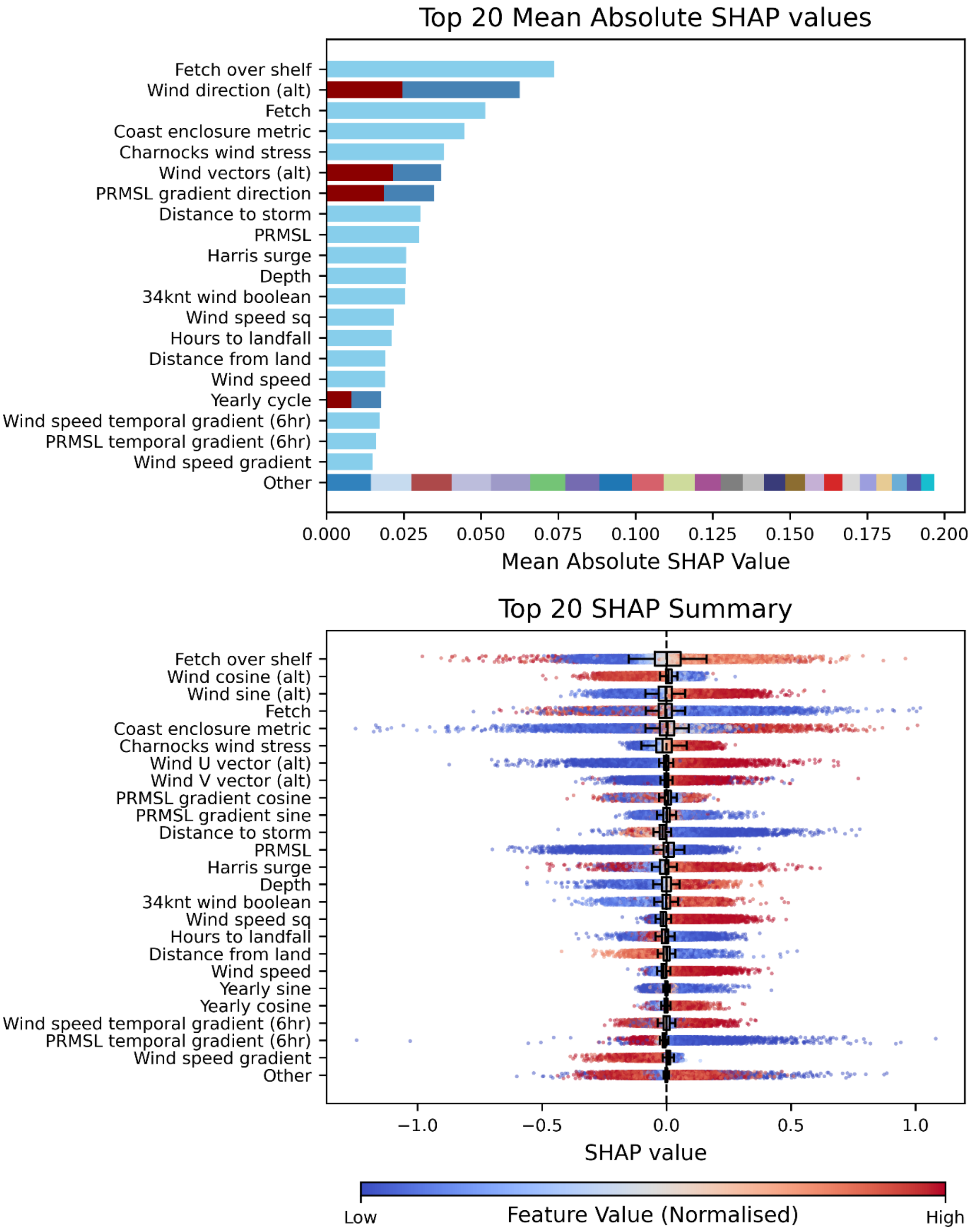


*Figure 3 - The top panel illustrates global SHAP explanations, plotting the mean absolute impact of each feature. Two tone bars indicate paired features, while the "Other category shows combined impact of features outside the top 20. The bottom panel illustrates the local impacts of a given feature. SHAP value is equivalent to surge, so a point at 1 on the x axis will have resulted in a surge increase of 1m. The colour scale indicates feature value, with feature values normalised. This is intuitive - a high pressure (PRMSL) value indicates high pressure.*

Geomorphological features also provide insight into model internal mechanics, and, like the atmospheric features, the results are positive. The most powerful feature in the model is fetch over shelf, with this representing the distance a given node is from the edge of the continental shelf, with higher values indicating greater distances. The SHAP summary plot indicates a strong fit to theory (Rego and Li, 2010) with higher SHAP values generally resulting in higher surge, while lower values result in reduced surge. Likewise, the coast enclosure metric is also promising, with high values indicating confined coastal geometries such as bays, while low values indicate open coast. Here high values mainly result in increased surge, with the inverse true of low values - theoretically this is sound, with enclosed areas experiencing dynamic amplification as the surge wave propagates through a bay, while the open coast experiences reduced surge as water is able to dissipate (Tiwari et al., 2025; Qian et al., 2024). Notably this pattern is not absolute; on some occasions areas with high feature values for fetch across the shelf and coastal enclosure will result in reduced surge. Counterintuitively this effect is also encouraging, as it indicates that feature interaction is occurring, and as later results will indicate, this is happening in a physically coherent way. For example, if a given pixel has high feature values for wind speed, fetch across shelf and coastal enclosure it may still experience negative surge if the wind direction features indicate that water is being driven out of the bay. Unfortunately, while we can hypothesise on feature interaction based on results, these interactions are still largely hidden from us by the ANN. An example of this is depth, where, counterintuitively, shallower depths result in reduced surge, with the inverse true for deeper depths - an unphysical mechanic (Harris, 1963; Jelesnianski, 1965). Interestingly, in the Naive model, which lacks any other geomorphic features, this effect is flipped, indicating that some form of feature interaction is occurring in the Fullset model. However, what this interaction consists of is not immediately apparent.

Overall, these results provide strong indications that the model is performing in a physically consistent manner with regard to surge mechanics, while also serving to remind us that, even with feature interpretation tools such as SHAP, ML models remain to some extent black boxes. Despite this, these interpretability tools are valuable sense checks, and also enable us to produce reduced set models. However, as demonstrated in the upper panel of Fig. 3 while the contribution of less important features may be minor, en masse these are significant, and help capture greater nuance, as demonstrated later in the results.

## 3.2 Model Performance

### 3.2.1 Headline Error

Model performance across the full domain is illustrated in Table 1, which summarises model performance by headline error statistics. An analysis of spatial and temporal performance follows later in the text. It is notable that across all models, both including and excluding coordinates, that the Fullset model is the best performing, with a $R^2$ of 0.93. At the opposite end of the scale the Naive model is the worst performing, again across both coordinate modes, with a worst performing $R^2$ of 0.71. Notably the drop in performance when coordinates are removed is comparatively mild for the feature engineered models, with the Fullset model recording a drop in $R^2$ of just 0.02, with this pattern of mild decline mirrored across the other feature engineering models. In comparison the decline in the Naive model is comparatively precipitous, with $R^2$ dropping by 0.11. This strongly suggests that the Naive model lacks sufficient richness of features to accurately emulate surge when location specific information is removed. In contrast, the feature engineered models appear to have a sufficient richness of features to more accurately emulate surge across a diverse range of locations, even where these locations are not explicitly defined. Sensitivity testing on events outside the sample used in this paper displays complementary results (SI 2).

Beyond direct comparison between the Fullset and Naive models the Expert Selection model (referred to as Expert from here on), is the next best performing model across both coordinate modes, followed by the ML Selection and finally the Stat Selection. This result is perhaps unsurprising, with subject domain expertise, linked with theory regarded as best practice when building models. In the interest of brevity this paper will now limit analysis to the Fullset, Expert and Naive models, as this covers the full scope of our investigation, investigating as it does the performance of feature engineering and comparative performance of reduced set models. Further we will limit our analysis to models without coordinate information because, as we have already argued, coordinates lead to poor generalisation.

| **Model** | **XY** | **$R^2$** | **RMSE** | **MSE** | **MAE** |
|---|---|---|---|---|---|
| **Fullset** | TRUE | **0.93** | **0.11** | **0.01** | **0.06** |
| **Fullset** | FALSE | **0.91** | **0.13** | **0.02** | **0.06** |
| **Expert Selection** | TRUE | **0.91** | **0.13** | **0.02** | **0.06** |
| **ML Selection** | TRUE | **0.89** | **0.14** | **0.02** | **0.07** |
| **Expert Selection** | FALSE | **0.87** | **0.15** | **0.02** | **0.08** |
| **ML Selection** | FALSE | **0.87** | **0.15** | **0.02** | **0.08** |
| **Stat Selection** | TRUE | **0.85** | **0.16** | **0.03** | **0.08** |
| **Stat Selection** | FALSE | **0.85** | **0.17** | **0.03** | **0.09** |
| **Naive** | TRUE | **0.82** | **0.18** | **0.03** | **0.09** |
| **Naive** | FALSE | **0.71** | **0.23** | **0.05** | **0.12** |

*Table 1 - describes the headline error statistic of each model configuration. The XY column indicates if coordinates were provided as features to the model. Note the model has been sorted by $R^2$. This table indicates performance across the full domain.*

### 3.2.2 Computational Cost

Table 2 demonstrates the computational cost of running each model. The key takeaways are that training and inference time are not significantly affected by feature set size (for this specific feature set and task, note this is not always the case), but preprocess time is. This is problematic as while training only has to be performed once, preprocessing must be performed for every event. In this case, preprocessing refers to scaling, normalisation and transformation into tensors, but there is also a hidden cost associated with per storm feature generation which is not included. With this in mind, it is significant that the Fullset model is approximately a third more expensive than the Expert model, with the Naive model approximately a tenth more efficient than the Expert. Beyond reduced feature sets there are other methods to reduce computational cost, with reduced spatial scope being a leading candidate. Training and emulating only in areas where model use is required. For example, inundation models typically start nearshore, where the propagation of surge onto land is modelled (Leijnse et al., 2021). If the ML model is used to provide boundary conditions for inundation models, for example between the 1 and 5 meter contours, this reduces cost by around half for feature engineering model preprocessing, with more modest gains for the Naive model. Cost can be further reduced by only preprocessing and running inference (following model training) on a selection of points, for example for input into inundation models. This reduces inference time to ~1s across all models, with preprocess time also seeing a major decrease. As such, despite the lack of requirement for feature reduction from a model performance standpoint (SI 7), there is a clear prerogative for a reduced feature set model from a cost perspective, with a further reduction in spatial scope also desirable, dependent on model requirements.

In comparison to ADCRIC the computational gain factor is significant (the ML model is 688-1535x faster), both across the full domain, and with substantial gains when simulating reduced domains. Further, ADCIRC simulation time of ~20 minutes per simulation day requires a HPC of 1024 cores (Dietrich et al., 2012), compared to a single GPU for the ML models.

| **Model** | **Train time** | **Inference (per event)** | **Point inference (per event)** | **Preprocess time (per event)** | **Training Domain** | **Computational gain factor (ML / ADCIRC)** |
|---|---|---|---|---|---|---|
| **Expert Selection** | 24.34 | 0.72 | 0.003 | 0.25 | 1-5m | 1528 |
| **Expert Selection** | 92.56 | 1.51 | 0.003 | 0.50 | Full domain | 730 |
| **Fullset** | 30.54 | 0.72 | 0.003 | 0.37 | 1-5m | 1535 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Fullset** | 69.62 | 1.53 | 0.003 | 0.83 | Full domain | 721 |
| **Naive** | 19.26 | 0.78 | 0.003 | 0.22 | 1-5m | 1404 |
| **Naive** | 86.98 | 1.60 | 0.003 | 0.22 | Full domain | 688 |

*Table 2 - indicating runtimes, with all units referring to minutes. Note that the Train column refers to the time it took to train the model across a training and validation set of 37 events. All other columns indicate the time taken to run a single event. Inference refers to an event emulated across the full time series. Point inference refers to inference across the full time series for points spaced 1km along the 10m contour of the domain. Computational gain factor refers to the time it takes the ML model to make inferences across a given domain (the inference per event column) divided by the time it takes ADCIRC to simulate the event. ADCIRC simulation time is taken based on a HPC of 1024 cores working on a mesh of 3 million nodes, taking 20 minutes per simulation day (Dietrich et al., 2012).*

### 3.2.3 Model Fit

Scatter plots, illustrated in Fig. 4, describe model performance across both the full domain and between the 1 to 5 meter depth contours. As discussed previously, the Fullset model is the best performing, followed by the Expert, with the Naive model performing worst. In terms of surge magnitude distribution, all models perform best in the middle of the surge range and worst in the extremes, a common critique of ML surge models. However, the feature engineered models perform better in this regard, albeit with a continued propensity to under predict the high extremes while overpredicting the low extremes. These issues are more moderate when considering the full domain, but this is to be expected as emulating surge offshore is an easier task than emulating surge nearshore, where there are a wealth of confounding factors such as coastal geometry. The Naive model exhibits the same trends with regard to the extremes but with these errors more pronounced, with a particular difficulty in emulating negative surge. This is an indicator that the model is struggling to capture underlying physics, a point which will be elaborated on when we examine spatial performance. These patterns are highly deleterious in the Naive 1 to 5m model, with $R^2$ dropping from 0.71 to 0.48. Here the poor performance in the extremes increases, with the model now also

struggling to resolve more moderate percentiles where the feature engineered models perform
605 relatively well.

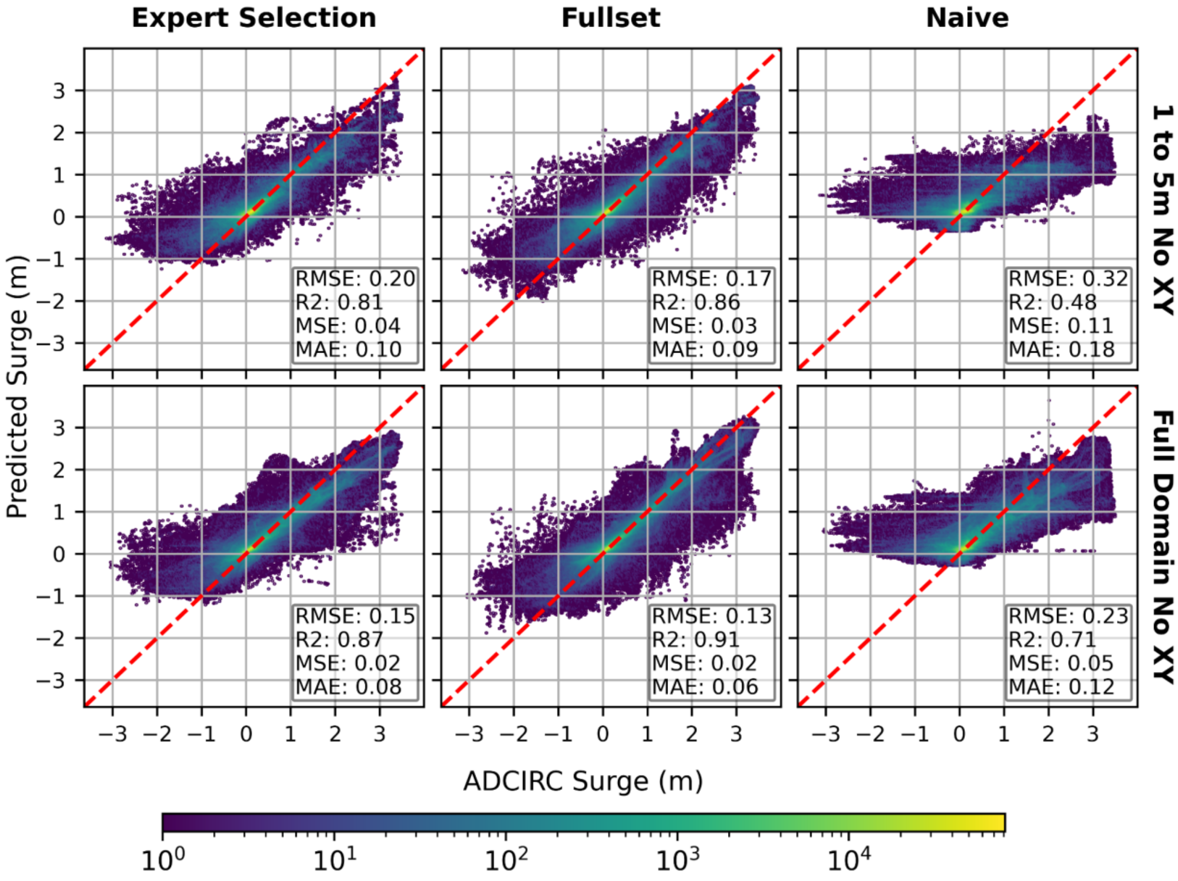


*Figure 4 - Plotting error distribution of each model across both the 1 to 5m depth contour and the full domain. Note that this is a density plot, with the number of points described by the log scale. The red line indicates perfect*
610 *fit between ADCIRC, here used as "truth" and a given model.*

Figure 5 plots the maximum and minimum spatial anomaly of each ML model against ADCIRC, with maximum and minimum here calculated across the full time series. Histograms of error are plotted on a logarithmic scale on the bottom row, and these are taken across all
615 time steps. Focussing firstly on maximum anomaly, both feature engineered models struggle to capture peak offshore surge, as indicated by the blue found along the storm track.

Additionally, they each underestimate surge at landfall location, with this effect more localised in the Fullset model, and more spatially expansive in the Expert. Feature engineered models also do a comparatively better job at capturing peak surge within enclosed coastal geometry, 620 indicating that these models are able to emulate dynamic amplification, with this theme later expanded on with reference to Fig. 6.

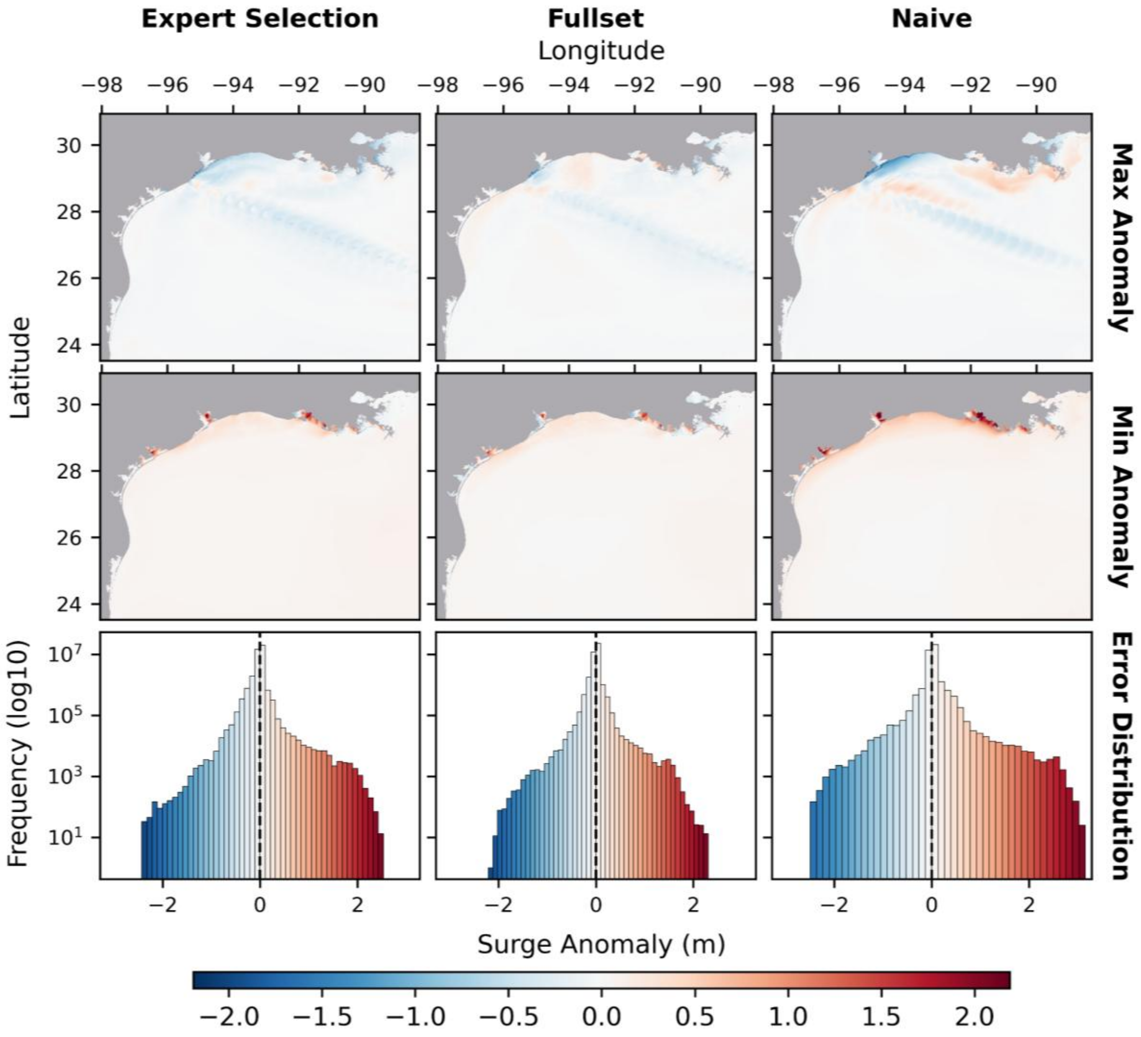


*Figure 5 - Demonstrating ML model surge anomaly against ADCIRC in the first two rows. Max/min surge is generated by taking the max/min value for each node across the time dimension. For anomalies ADCIRC surge is subtracted from the ML models. The 3rd row plots error distribution at histograms from each model timestep with frequencies on a logarithmic scale.* 625

https://doi.org/10.5194/egusphere-2026-781
Preprint. Discussion started: 23 February 2026

In contrast to the feature engineered models, the Naive model has a less spatially correlated anomaly field with a leading ‘false track’ of overprediction to the north of the true track underprediction. Further, it has a swath of overprediction extending northeast of the landfall location, with this pattern indicating that the Naive model has difficulties in replicating observed tropical cyclone behaviour, in which greater surge is generated to the front left of storm forward direction. Surge at landfall is also underpredicted, with this being both more acute in terms of magnitude and more spatially dispersed than in the feature engineered models. Finally, there is an indication that the Naive model fails to emulate surge propagation into Trinity and Galveston Bays (hereafter referred to as Houston Bays), suggesting that dynamic amplification is poorly captured.

Turning our attention to minimum anomalies the patterns first identified in the scatter plots (Fig. 4) are spatially illustrated, with all models failing to greater or lesser extents to capture negative surge within areas of enclosed coastal geometry. This indicates difficulties in emulating water being driven out of enclosed areas, such as bays, by strong wind forcing. In line with the scatter plots (Fig. 4), the Fullset model has the lowest magnitude and smallest spatial spread of this error, with the Naive model displaying the inverse. This disparity will be discussed in more detail in conjunction with Fig. 6, as this spatiotemporal plot panel gives us a more detailed view of surge propagation.

The anomalies which we have discussed spatially are further quantified by the histograms shown in Fig. 5. Note that the frequency count here is displayed on a logarithmic scale as otherwise the chart is dominated by the very large number of errors within a few centimetres of zero. These log scaled histograms serve to show that the feature engineered models have a narrow error range. The Fullset model shows limited signs of skew, and therefore bias, along with the most constrained error range, and with the majority of error near zero. The Expert model exhibits signs of increased bias, with a positive skew evident indicating that the model is prone to overprediction, largely of negative surges. The trend of model performance degradation with feature reduction continues, with the Naive model displaying both the widest error range, and the most significant bias. Please note that all plots and analysis in reference to Fig. 5 apply to models full domain performance.

Figure 6 (also available in more detail as a GIF in SI 8.1) shows the spatiotemporal evolution of surge as the generating single example storm moves across the domain. Focusing firstly on the ADCIRC column we can see the storm approaching, making landfall and finally dissipating. Notably as the storm approaches landfall, illustrated from timesteps 4 to 16, we see surge building on the east side of the Mississippi Delta, while simultaneously we see negative surge on the west side of the same peninsula. This effect makes theoretical sense - the winds of a tropical cyclone, in the Northern Hemisphere, circulate around the eye in an anticlockwise direction. Accordingly, we see water being driven onshore to the east of the delta, where it is unable to dissipate and so surge builds. In contrast the wind is blowing offshore to the east of the delta, driving water offshore where it is able to dissipate, and so producing negative surge. Examining our ML models we can see that this phenomenon is captured in both the feature engineered models, with the Fullest capturing the spatial pattern and magnitudes most faithfully. Contrastingly, the Naive model fails to capture these dynamics, with a positive surge on each side of the peninsula. This indicates that the feature engineered models are able to capture the effect of wind direction, and also crucially the relationship this has with landmasses, in a manner in which the Naive model cannot.

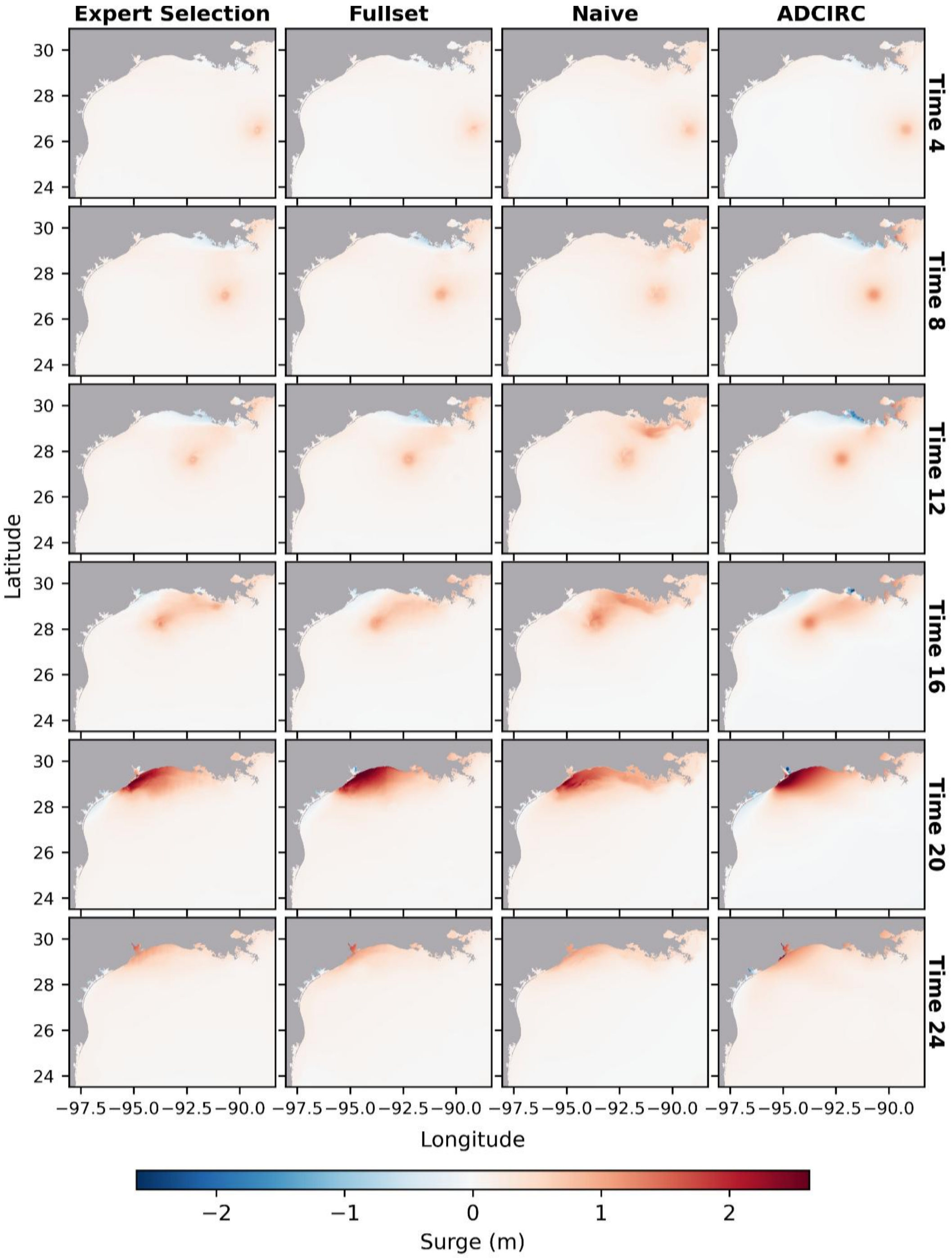


*Figure 6 - illustrating the spatiotemporal propagation of surge throughout the domain. Each column represents a model and each row a timestep, with time progressing from top to bottom. The models are trained to emulate ADCIRC. A GIF of propagation is shown in SI 8.1, and an anomaly version of this plot is shown in SI 8.2.*

Turning our attention to landfall (timesteps 20 and 24), and again focusing firstly on ADCIRC we see a large buildup of surge centering on the Galveston region, with a slight reduction of surge to the west. Within Houston Bays we see negative surge in the northeast, while positive surge reflects the surge wave propagating through the bay in the southwest. As the storm dissipates at timestep 24 we see elevated surge levels continue in the bay when compared to the coast, reflecting the lag as water egress from the bay is slowed by the constricting barrier island coastal geometry. In the feature engineered models we see this pattern largely replicated, albeit with caveats. In timestep 20 we see that the Fullset model captures negative surge in the northeast of the bay, while the expert model begins to capture surge wave propagation in the southeast. However, there is evidence of a temporal lag between each ML model and ADCIRC, with this discussed more fully when examining Fig. 7. Both feature engineered models correctly capture elevated surge in contrast to the open ocean at timestep 24 as the storm dissipates, indicating that they once again resolve the effects of coastal geometry. In the Naive model at timestep 20, it is clear that the spatial footprint of surge is flawed, and further, there does not appear to be evidence of dynamic amplification occurring in Houston Bay. This error is compounded at timestep 24, where we see no evidence of elevated surge in the self-same bay. These patterns provide further evidence that the Naive model lacks the necessary information to capture surge dynamics, particularly in the complex nearshore areas. Please note that these trends are more visible in the accompanying GIF in SI 8.1.

Moving from the large-scale spatiotemporal trends illustrated in Fig. 6, we now analyse how models resolve surge locally by examining storm hydrographs at a number of locations illustrated in Fig. 7. Here nodes most closely matching a given ADCIRC quantile are plotted. Notably all ML models have relatively smooth temporal evolution, with only quantile 0.4 showing signs of inconsistency, and this being located both within a sheltered bay distal to the storm. Further, hydrograph phase is well captured by all models, albeit with noticeable lag in the Naive model. When forcing inundation models this lag would likely have very limited impact on the flood footprint, but is nevertheless penalised in error metrics. Beyond lag, the Naive model continues to present the failings we have identified previously, with the worst resolved negative surge in quantiles 0.01 and 0.05, along with the lowest peak surges, relative to ADCIRC, in quantiles 0.05, 0.1 and 0.3. While the feature engineered models emulate peak surge more faithfully, they still fail to capture the high extremes as seen in quantile 0.99, with the ML models ability to emulate peak surge more closely examined in Fig. 8. Focusing on negative surge, the Fullset model exaggerates negative surge, as seen in quantiles 0.4, 0.8 and 0.99. These are all located in areas of confined coastal geometry, where negative surges might be expected. However, it is clear the Fullset model lacks a sufficiently detailed understanding of negative surge dynamics to reliably emulate its occurrence, with this issue being problematic for the use of emulated surge in onshore inundation models, as significant overprediction of negative surge would change flood extents.

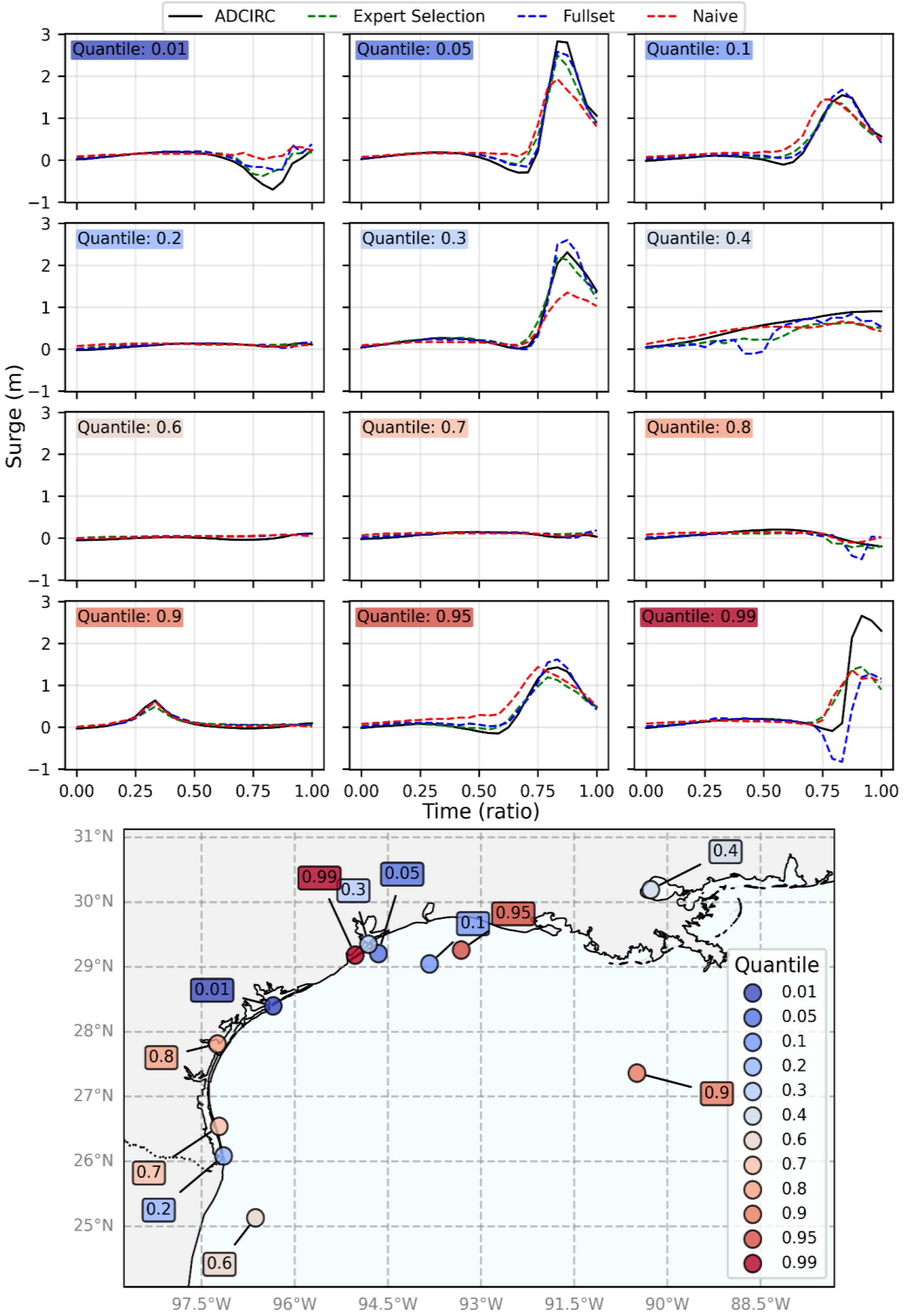
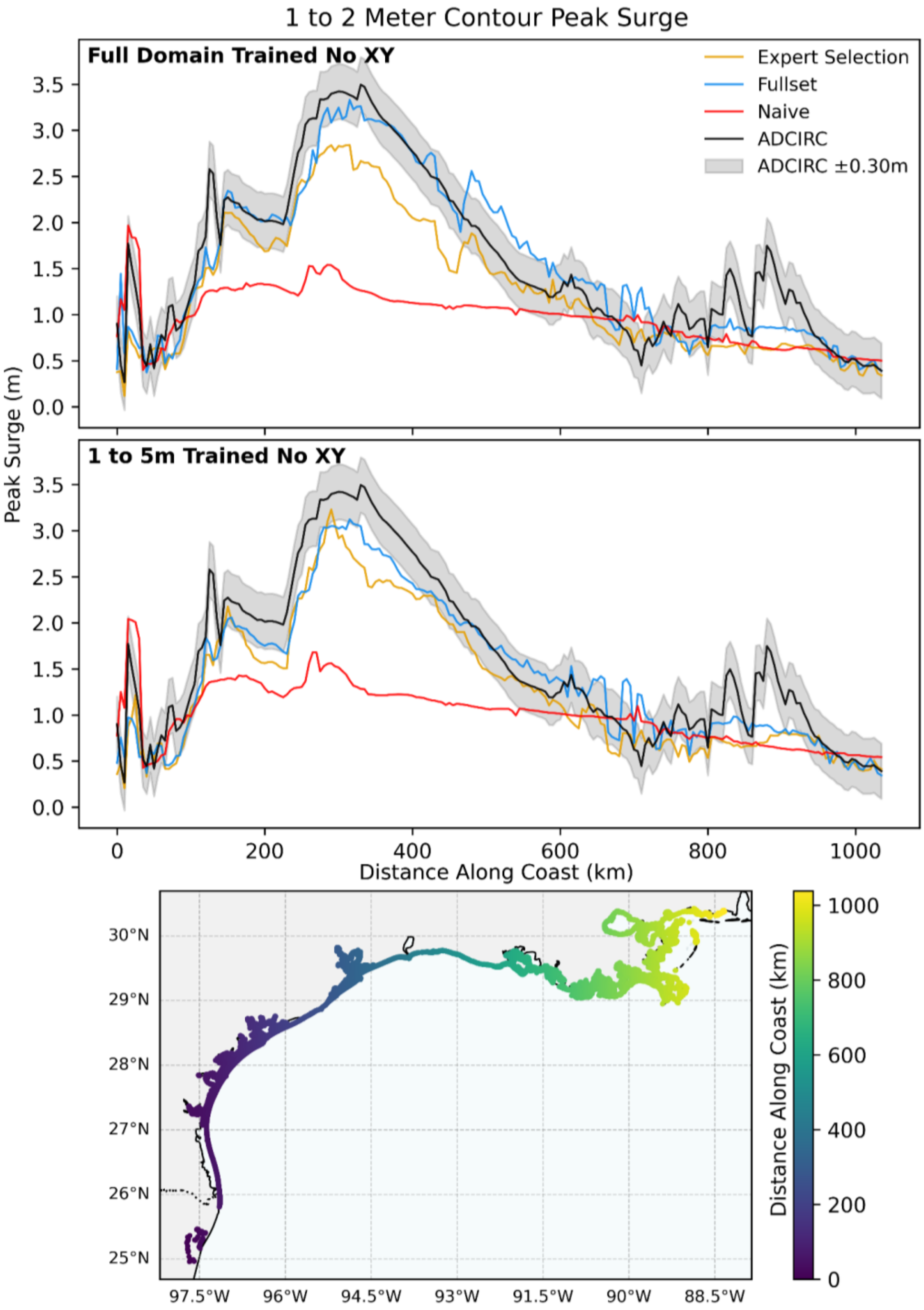


*Figure 7 - describes storm hydrographs at a representative sample of nodes, with these selected by taking the nearest ML model surge, at any point in time, to ADCIRCs surge quantiles, again selected from any point in time.*

The map in Fig. 7, illustrates that the quantiles with the largest errors are located in the areas where surge dynamics are most difficult to emulate - generally the nearshore and principally in bays and behind barrier islands. Unfortunately, this is where model performance, for the purpose of forcing inundation models, is most critical. To further examine model performance at this critical juncture, we have plotted peak surge along the coast between the 1 and 2m depth contours in Fig. 8, using both models trained on the full domain, and models trained only between the 1 and 5m contour. ADCIRC is plotted as "truth" with this the target of ML model emulation. Around ADCIRC's time series is the vertical error range of typical bathymetric LiDAR - 30cm (Eren et al., 2019; Westfeld et al., 2016). This range is relevant, as it illustrates the best possible vertical height error within hydrodynamic models. ML predictions following within this range can be determined to be within ADCIRC error derived from terrain, and so arguably functionally equivalent, particularly where total model error is significantly larger, as it includes forcing and numerical error (Bates et al., 2010; Hunter et al., 2007).The Fullset model, across both domains, shows the best fit  to ADCIRC, while the Naive model performs the worst by a substantial margin. The Expert model exhibits underprediction of the highest peak surges across both domains, but improves when trained within the 1 to 5m contours. In contrast the Fullset model is more prone to overprediction when trained on the full domain, while more prone to underprediction when trained on the nearshore domain. The overprediction when trained on the full domain is something of a rarity in ML models, which often struggle to capture extreme event peaks, and is an indication that the improvement of fit seen in the density plots of Fig. 4 are not an artefact of better prediction of high surge offshore events only, but an improvement of nearshore representation. However, this improvement of fit to high extremes does appear to come at the cost of more overprediction, and indeed more noise, with this particularly apparent between 400 and 600km along the coast. Putting these results into spatial context is illuminating. For example, we can see that between 0 and 200km distance along the coastline results are noisy, and that the ML models struggle to emulate ADCIRC, with the Naive model performing best. This noise may be due to the high number of barrier islands in this section of coast, and the corresponding large difference in surge profile therein. The improved performance of the Naive model is perhaps surprising here as it often fails to capture dynamic amplification, instead presenting a depressed surge profile, as demonstrated in both the spatiotemporal and anomaly plots. The relatively worse performance of the feature engineered models is perhaps then best explained by the storm hydrographs (Fig. 7). Here overprediction of negative surges in the Fullset model perhaps results in reduced peak surges. In contrast, the Naive model is unburdened by this limitation. Between 300-400km, we see the largest peak surges, with this corresponding to Houston Bay. As noted when discussing the anomaly and spatiotemporal plots, we see that surge deep inside the bay is best resolved by the feature engineered models, indicating the successful emulation of dynamic amplification. As also noted previously, we see the depressed surge levels of the Naive model, with dynamic amplification poorly resolved.



765 *Figure 8 - illustrates peak surge across the full test set of events. Peak surge is determined by selecting the maximum surge within the nearest 50 nodes from each test event. All nodes considered are within the 1 to 2m depth contour. This approach was chosen to improve visual clarity while maintaining peaks.*

Between 700 and 900km we see an increase in noise, and a corresponding decrease in performance across all models. Spatially, this relates to complex bathymetry and coastal topography of the Mississippi Delta, with this complexity perhaps relating to poorer model performance. However, it is worth considering that the worse performance of the feature engineered models is seen at the edges of the training domain. These areas, as illustrated in the storm track map, are impacted by comparatively few events, and in the case of the Mississippi Delta never struck directly. Perhaps poorer performance in peak surge emulation is linked to limited learning opportunities of high surge in these areas, and is further skewed by the large number of low surge situations which these areas' particular combination of features are exposed to.

## 4. Discussion

Throughout this study we have demonstrated that through use of feature engineering ML surge emulation models can be improved beyond those using classic inputs previously demonstrated within the literature (Dong et al., 2022; Qin et al., 2023). Further, we have demonstrated that feature engineered models maintain performance when coordinates are removed, demonstrating an explicit understanding of generalised surge dynamics and not simply location-specific characteristics (Meyer et al., 2019). This improved understanding is critical, as coordinates, or other location specific features, are poorly generalisable and so are unsuited to extrapolation beyond the physical confines of their training domain (Karimzadeh et al., 2025). Additionally, as demonstrated in Table 1, feature engineered models without coordinate information outperform Naive models, even where Naive models include coordinates, indicating that even ML models intended to be limited to a single geographic domain, for example for forecasting, would benefit from feature engineering.

The ultimate goal this work is aimed at is the development of an ML model which can be deployed globally, across any spatial domain, and under any climate scenario. This would enable quantification of the full distribution of tropical cyclone risk (Bloemendaal et al., 2020; Loridan and Bruneau, 2025). Such quantification would require the ML model to be run hundreds of thousands of times, and in such an endeavour compute costs are critical, necessitating the development of ML emulation models, as conventional hydrodynamic models would be exorbitantly expensive (Bates et al., 2021). This is demonstrated in Table 2, in which we demonstrate that ML models can be up to ~1500 times faster when running only at the nearshore for input into inundation models, and ~750 times faster when running for the full domain. Further, the ML models do not require a HPC environment, with it being possible to train and run them on a single desktop PC GPU. However, even accounting for improvements in compute demonstrated by Table 2, across hundreds of thousands of runs even relatively minor cost increases per event can cumulatively be highly significant. With this limitation in mind, this paper utilised SHAP values (Lundberg and Lee, 2017), amongst other statistical measures, to inform selection of the most impactful features for surge emulation. This allows for the production of reduced feature set models, which in turn reduces preprocessing costs. However, beyond the computational improvements, this method has its own inherent costs, with reduced set ML models suffering a decline in performance, particularly in the high extremes (Fig. 4). Therefore, it is perhaps more beneficial to retain the full feature set, and instead reduce domain scope. Fig. 8 indicates that the Fullset model, when trained on the full domain, more accurately emulates nearshore surge. However, once trained there are no constraints on model deployment, allowing targeted inference. For example, restricting inference to within the 1 to 5m domain, the area most critical for onshore risk analysis (Leijnse et al., 2021), reduces preprocess time by approximately half. Further, reducing inference to only boundary condition points, for input into inundation models, takes inference per event to below a second (Table 2). While these improvements to inference are noteworthy, they must be caveated by the fact that many spatial features utilised require spatial connectivity between nodes - for example for the creation of spatial gradient features. This requirement limits gains to inference, as preprocessing still must be performed on a wider domain.

Beyond computational cost, process driven challenges remain, even when considering the best performing Fullset model. The first, and perhaps most worrying, is that the model shows signs of performance degradation at the edges of the training domain, as illustrated by Fig. 8. This suggests the model may be overfit to central areas of the training domain, as this is where the majority of tracks make landfall. Perhaps more pertinently it may indicate that the model has reached training saturation. This supposition is supported by the training bank of storms being limited to only 5 primary tracks (Dawson et al., 2021), from which the remainder are perturbations of magnitude and location. This lack of variety being a constraint on model performance is underscored by sensitivity testing (SI 2) indicating minor performance improvements beyond a sample size of 40. Additionally, when comparing the track set utilised in this study (Fig. 1) to historic tracks, it is evident that the synthetic set of events is highly idealised, and lacks the chaotic nature of the historic (Kaiser et al., 2023) (SI 9), with this conceivably limiting the training potential of the dataset. To address this challenge historic events could be utilised as training data, but preliminary work (SI 10) indicates that this record is too sparse, both spatially and in magnitude (Kaiser et al., 2023). These issues are compounded by the lack of both tides (Horsburgh and Wilson, 2007; Rego and Li, 2010) and waves (Almar et al., 2021) in the training dataset used within this study, with both these processes key in flood dynamics and risk management. Therefore, a target for future studies would ideally be to produce a more diverse training dataset.

Separate from dataset considerations, the best performing feature engineered model, while much improved over the Naive model, still shows indications of struggling with spatial autocorrelation. Firstly, surge predictions deriving from the ML models are not as smooth as those produced by ADCIRC, resulting in unphysical spatial discontinuities (Fig. 6). Further, peak surge on the open coast at landfall is typically underestimated (Fig. 8), while in barrier bays there is a trend for the overprediction of negative surges (Fig. 5), although more broadly the ability of the model to capture negative surge is encouraging (Fig. 6). These issues, particularly in areas of complex coastal geometry, suggest limits to provision of spatial and temporal information through feature engineering alone, although with the caveat that there are likely more sophisticated modes of feature engineering which have been neglected in this study. This notwithstanding, a key consideration is what performance is acceptable for an ML emulator. An emulator which is unbiased and falls within the target hydrodynamic model's total error, is arguably indistinguishable from the hydrodynamic model itself (Bates et al., 2010; Hunter et al., 2007), although note that by this metric hydrodynamic and ML model error may be cumulative. The ADCIRC simulations used within this study reported errors of ~0.5-1m when compared to historic events (Dawson et al., 2021). This indicates that ML model performance within the training domain and dataset is satisfactory, and extrapolating beyond the training domain is the next step.

Nevertheless, more complex ML architectures can be utilised which inherently provide spatial and temporal information. Leading candidates include CNNs and LSTMs, or these can be coupled to produce ConvLSTMs (Adeli et al., 2023), that would be spatially and/or temporally aware by design. However, CNNs require rectilinear grids as input (O'shea and Nash, 2015), and transforming unstructured meshes to rectilinear grids results in a huge increase in computational cost, storage, RAM and processing, as they lose the size efficiency of unstructured meshes (Pain et al., 2005), while also losing information if coarsened to counteract this loss of efficiency. Therefore, Graph Neural Networks (GNNs), which work on the basis of node connectivity, may be a preferred avenue, as through explicit connections to their neighbours they enforce spatial awareness, while preserving the advantages of an unstructured mesh. This would preserve temporal awareness, as GNNs can be coupled with LSTMs(Wu et al., 2020).

Despite described improvements to performance derived from more complex model architectures, more complex architectures come at a cost, with increases to training and inference times expected, with potential overfitting risks (Thompson et al., 2020). Further, in the case of spatially aware models such as GNNs and CNNs, these also place limits on inference, with emulation of a single point in space and time, using only data at that single point in space and time, no longer possible (O'shea and Nash, 2015; Wu et al., 2020). However, improved inherent handling of spatiotemporal information within these model structures, when coupled with an LSTM, may allow for further feature reduction while maintaining or even improving performance, reducing the computational cost of preprocessing. Additionally, the inherent handling of spatiotemporal relationships in these model architectures may further improve generalisability (Krichen and Mihoub, 2025; O'shea and Nash, 2015; Wu et al., 2020). When combined with feature engineering, and the provision of a more diverse training dataset, the improvement to generalisability could be significant.

## 5. Conclusion

This study has demonstrated that feature engineering can significantly improve the accuracy of ML surge emulation, with $R^2$ increasing from 0.71 to 0.91, alongside much improved representation of surge dynamics. Additionally, ML models are between ~700 to 1500 times faster than hydrodynamic simulation, with reduced domain inference more effective at compute reduction than reducing the feature set i.e. only inferring boundary points for an inundation model. Further, ML model RMSE of 0.13m sits within the ~0.5-1m error range reported for hydrodynamic simulations used for training within this study. Where ML predictions are unbiased, and fall within this range, it can be argued that performance is sufficient, although ML and hydrodynamic model error can be cumulative if error sign is the same. Nevertheless, this indicates that future work should build upon these findings by testing the generalisability of feature engineered emulation models beyond their training domain, rather than myopically focusing on small performance gains. Future work on generalisability will likely require a more realistic and diverse hydrodynamic model training set, as opposed to the relatively homogenous hydrodynamic set used in this study. The utilisation of more sophisticated machine learning architectures may improve emulation fidelity and generalisability further, with a specific focus on coupled Long Short Term Memory and Graph Neural Networks, as these inherently capture spatiotemporal information. Finally, waves and tide-surge interactions are key factors in risk analysis, and so any future work must strive to include these processes.

## *Code/Data Availability*

*ADCIRC synthetic runs are available from:*

*https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-2968*

*ADCIRC storm surge hindcasts are available from:*

*https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-3932*

*Geomorphology of the Oceans is available from:*

*https://bluehabitats.org/?page_id=58*

*TPXO Atlas 10 is available from:*

*https://www.tpxo.net/global/tpxo10-atlas*

*The Copernicus land mask is available from: https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM*

*The models discussed in this paper are available to academic institutions for non-commercial research upon request by contacting the corresponding author.*

## *Author contribution*

***Hamish Wilkinson**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft and Reviewing and Editing, Visualisation. **Paul Bates**: Conceptualization, Methodology, Writing - Reviewing and Editing, Supervision. **Chris Lucas**: Conceptualization, Methodology, Writing - Reviewing and Editing, Supervision. **Niall Quinn**: Conceptualization, Methodology, Writing - Reviewing and Editing, Supervision. **Ivan Haigh**: Conceptualization, Methodology, Writing - Reviewing and Editing, Funding acquisition, Supervision. **Tom Collings**: Conceptualization, Methodology. **Peter Watson:** Conceptualization, Methodology, Supervision.*

## *Competing Interests*

*The authors declare no competing interests.*

## *Disclaimer*

*Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications* *makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher*

***Financial support***

*This work was supported by the Natural Environment Research Council funded Doctoral Training for Resilient Flood Futures (FLOOD-CDT). NERC Reference: NE/Y006364/1*

***Use of AI Acknowledgement***

*During the preparation of this work the authors used Gemini during coding. All ideas and methods were human derived, with Gemini only used for improved implementation. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.*

## Reference List

Adeli, E., Sun, L., Wang, J., and Taflanidis, A. A.: An advanced spatio-temporal convolutional recurrent neural network for storm surge predictions, Neural Computing and Applications, 35, 18971-18987, 10.1007/s00521-023-08719-2, 2023.

Almar, R., Ranasinghe, R., Bergsma, E. W. J., Diaz, H., Melet, A., Papa, F., Vousdoukas, M., Athanasiou, P., Dada, O., Almeida, L. P., and Kestenare, E.: A global analysis of extreme coastal water levels with implications for potential coastal overtopping, Nature Communications, 12, 3775, 10.1038/s41467-021-24008-9, 2021.

Amarouche, K. and Akpınar, A.: Increasing Trend on Storm Wave Intensity in the Western Mediterranean, Climate, 9, 11, 2021.

Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, Stanford, 2006.

Athanasiou, P., van Dongeren, A., Pronk, M., Giardino, A., Vousdoukas, M., and Ranasinghe, R.: Global Coastal Characteristics (GCC): a global dataset of geophysical, hydrodynamic, and socioeconomic coastal indicators, Earth Syst. Sci. Data, 16, 3433-3452, 10.5194/essd-16-3433-2024, 2024.

Bates, P. D., Horritt, M. S., and Fewtrell, T. J.: A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling, Journal of Hydrology, 387, 33-45, https://doi.org/10.1016/j.jhydrol.2010.03.027, 2010.

Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., Savage, J., Olcese, G., Neal, J., Schumann, G., Giustarini, L., Coxon, G., Porter, J. R., Amodeo, M. F., Chu, Z., Lewis-Gruss, S., Freeman, N. B., Houser, T., Delgado, M., Hamidi, A., Bolliger, I., E. McCusker, K., Emanuel, K., Ferreira, C. M., Khalid, A., Haigh, I. D., Couasnon, A., E. Kopp, R., Hsiang, S., and Krajewski, W. F.: Combined Modeling of US Fluvial, Pluvial, and

Coastal Flood Hazard Under Current and Future Climates, Water Resources Research, 57, e2020WR028673, https://doi.org/10.1029/2020WR028673, 2021.

Bhatia, K., Baker, A., Yang, W., Vecchi, G., Knutson, T., Murakami, H., Kossin, J., Hodges, K., Dixon, K., Bronselaer, B., and Whitlock, C.: A potential explanation for the global increase in tropical cyclone rapid intensification, Nature Communications, 13, 6626, 10.1038/s41467-022-34321-6, 2022.

Bhatia, K. T., Vecchi, G. A., Knutson, T. R., Murakami, H., Kossin, J., Dixon, K. W., and Whitlock, C. E.: Recent increases in tropical cyclone intensification rates, Nature Communications, 10, 635, 10.1038/s41467-019-08471-z, 2019.

Bloemendaal, N., Haigh, I. D., de Moel, H., Muis, S., Haarsma, R. J., and Aerts, J. C. J. H.: Generation of a global synthetic tropical cyclone hazard dataset using STORM, Scientific Data, 7, 40, 10.1038/s41597-020-0381-2, 2020.

Bostrom, A., Demuth, J. L., Wirz, C. D., Cains, M. G., Schumacher, A., Madlambayan, D., Bansal, A. S., Bearth, A., Chase, R., Crosman, K. M., Ebert-Uphoff, I., Gagne II, D. J., Guikema, S., Hoffman, R., Johnson, B. B., Kumler-Bonfanti, C., Lee, J. D., Lowe, A., McGovern, A., Przybylo, V., Radford, J. T., Roth, E., Sutter, C., Tissot, P., Roebber, P., Stewart, J. Q., White, M., and Williams, J. K.: Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences, Risk Analysis, 44, 1498-1513, https://doi.org/10.1111/risa.14245, 2024.

Bruneau, N., Polton, J., Williams, J., and Holt, J.: Estimation of global coastal sea level extremes using neural networks, Environmental Research Letters, 15, 074030, 10.1088/1748-9326/ab89d6, 2020.

Calafat, F. M., Wahl, T., Tadesse, M. G., and Sparrow, S. N.: Trends in Europe storm surge extremes match the rate of sea-level rise, Nature, 603, 841-845, 10.1038/s41586-022-04426-5, 2022.

Camargo, S. J. and Wing, A. A.: Tropical cyclones in climate models, WIREs Climate Change, 7, 211-237, https://doi.org/10.1002/wcc.373, 2016.

Cardone, V. J. and Cox, A. T.: Tropical cyclone wind field forcing for surge models: critical issues and sensitivities, Natural Hazards, 51, 29-47, 10.1007/s11069-009-9369-0, 2009.

Collings, T. P., Quinn, N. D., Haigh, I. D., Green, J., Probyn, I., Wilkinson, H., Muis, S., Sweet, W. V., and Bates, P. D.: Global application of a regional frequency analysis to extreme sea levels, Nat. Hazards Earth Syst. Sci., 24, 2403-2423, 10.5194/nhess-24-2403-2024, 2024.

Dawson, C. N., Del-Castillo-Negrete, C., Shukla, A., Pachev, B., Kaiser, C., and Kutanoglu, E.: ADCIRC Simulation of Synthetic Storms in the Gulf of Mexico, Designsafe-CI [dataset], 10.17603/DS2-68A9-0S64, 2021.

Dietrich, J. C., Tanaka, S., Westerink, J. J., Dawson, C. N., Luettich, R. A., Zijlema, M., Holthuijsen, L. H., Smith, J. M., Westerink, L. G., and Westerink, H. J.: Performance of the Unstructured-Mesh, SWAN+ADCIRC Model in Computing Hurricane Waves and Surge, Journal of Scientific Computing, 52, 468-497, 10.1007/s10915-011-9555-6, 2012.

Dong, C., Xu, G., Han, G., Bethel, B. J., Xie, W., and Zhou, S.: Recent Developments in Artificial Intelligence in Oceanography, Ocean-Land-Atmosphere Research, 2022, doi:10.34133/2022/9870950, 2022.

Dulac, W., Cattiaux, J., Chauvin, F., Bourdin, S., and Fromang, S.: Assessing the representation of tropical cyclones in ERA5 with the CNRM tracker, Climate Dynamics, 62, 223-238, 10.1007/s00382-023-06902-8, 2024.

Emanuel, K.: Response of Global Tropical Cyclone Activity to Increasing CO2: Results from Downscaling CMIP6 Models, Journal of Climate, 34, 57-70, https://doi.org/10.1175/JCLI-D-20-0367.1, 2021.

Emanuel, K. A.: Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century, Proceedings of the National Academy of Sciences, 110, 12219-12224, doi:10.1073/pnas.1301293110, 2013.

Eren, F., Jung, J., Parrish, C. E., Sarkozi-Forfinski, N., and Calder, B. R.: Total vertical uncertainty (TVU) modeling for topo-bathymetric LIDAR systems, Photogrammetric Engineering & Remote Sensing, 85, 585-596, 2019.

Fahrland, E., Jacob, P., Schrader, H., and Kahabka, H.: Copernicus digital elevation model product handbook, Airbus Defence and Space—Intelligence: Potsdam, Germany, 2024-2006, 2020.

Fisher, A., Rudin, C., and Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, Journal of Machine Learning Research, 20, 1-81, 2019.

Giaremis, S., Nader, N., Dawson, C., Kaiser, C., Nikidis, E., and Kaiser, H.: Storm surge modeling in the AI era: Using LSTM-based machine learning for enhancing forecasting accuracy, Coastal Engineering, 191, 104532, https://doi.org/10.1016/j.coastaleng.2024.104532, 2024.

Girshick, R.: Fast R-CNN, 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 Dec. 2015, 1440-1448, 10.1109/ICCV.2015.169,

Harris, D. L.: Characteristics of the hurricane storm surge, 48, Weather Bureau.1963.

Hernanz, A., Correa, C., Sánchez-Perrino, J.-C., Prieto-Rico, I., Rodríguez-Guisado, E., Domínguez, M., and Rodríguez-Camino, E.: On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues, Atmospheric Science Letters, 25, e1195, https://doi.org/10.1002/asl.1195, 2024.

Holland, G.: A Revised Hurricane Pressure–Wind Model, Monthly Weather Review, 136, 3432-3445, https://doi.org/10.1175/2008MWR2395.1, 2008.

Horsburgh, K. J. and Wilson, C.: Tide-surge interaction and its role in the distribution of surge residuals in the North Sea, Journal of Geophysical Research: Oceans, 112, https://doi.org/10.1029/2006JC004033, 2007.

Hotelling, H.: Analysis of a complex of statistical variables into principal components, Journal of educational psychology, 24, 417, 1933.

Huber, P. J.: Robust estimation of a location parameter, in: Breakthroughs in statistics: Methodology and distribution, Springer, 492-518, 1992.

Hunter, N. M., Bates, P. D., Horritt, M. S., and Wilson, M. D.: Simple spatially-distributed models for predicting flood inundation: A review, Geomorphology, 90, 208-225, https://doi.org/10.1016/j.geomorph.2006.10.021, 2007.

Jelesnianski, C. P.: A numerical calculation of storm tides induced by a tropical storm impinging on a continental shelf, Monthly Weather Review, 93, 343-358, 1965.

Jelesnianski, C. P.: SPLASH:(Special Program to List Amplitudes of Surges from Hurricanes).. I, Landfall Storms, National Weather Service1972.

Jewson, S.: Projecting future tropical cyclone frequencies by combining uncertain empirical estimates of baseline frequencies with climate model estimates of change, The Journal of Catastrophe Risk and Resilience, 23, 10.63024/m9wk-3420, 2024.

Johnson, D. and Ahmadi, M.: ADCIRC Simulations of Synthetic Tropical Cyclones Impacting Coastal Louisiana, Designsafe-CI [dataset], 10.17603/DS2-0KSB-YY40, 2023.

Kaiser, C., Dawson, C. N., Nikidis, E., and Fleming, J. G.: ADCIRC/SWAN Hindcasts for Historical Storms 2003-2023, in CERA / ADCIRC Storm Surge Hindcasts: Historical Storms 2003-2024, Designsafe-CI [dataset], 10.17603/DS2-B5GH-CE94, 2023.

Karimzadeh, M., Wang, Z., and Crooks, J. L.: Performance and generalizability impacts of incorporating location encoders into deep learning for dynamic PM2.5 estimation, GIScience & Remote Sensing, 62, 2594797, 10.1080/15481603.2025.2594797, 2025.

Kennedy, A. B., Gravois, U., Zachry, B. C., Westerink, J. J., Hope, M. E., Dietrich, J. C., Powell, M. D., Cox, A. T., Luettich Jr, R. A., and Dean, R. G.: Origin of the Hurricane Ike forerunner surge, Geophysical research letters, 38, 2011.

Kingma, D. P.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kirezci, E., Young, I. R., Ranasinghe, R., Lincke, D., and Hinkel, J.: Global-scale analysis of socioeconomic impacts of coastal flooding over the 21st century, Frontiers in Marine Science, Volume 9 - 2022, 10.3389/fmars.2022.1024111, 2023.

Koukaras, P. and Tjortjis, C.: Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices, AI, 6, 257, 2025.

Kraskov, A., Stögbauer, H., and Grassberger, P.: Estimating mutual information, Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 69, 066138, 2004.

Krichen, M. and Mihoub, A.: Long Short-Term Memory Networks: A Comprehensive Survey, AI, 6, 215, 2025.

Leijnse, T., van Ormondt, M., Nederhoff, K., and van Dongeren, A.: Modeling compound flooding in coastal systems using a computationally efficient reduced-physics solver: Including fluvial, pluvial, tidal, wind- and wave-driven processes, Coastal Engineering, 163, 103796, https://doi.org/10.1016/j.coastaleng.2020.103796, 2021.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy, 23, 18, 2021.

Lincke, D., Hinkel, J., Mengel, M., and Nicholls, R. J.: Understanding the Drivers of Coastal Flood Exposure and Risk From 1860 to 2100, Earth's Future, 10, e2021EF002584, https://doi.org/10.1029/2021EF002584, 2022.

Lockwood, J. W., Lin, N., Oppenheimer, M., and Lai, C.-Y.: Using Neural Networks to Predict Hurricane Storm Surge and to Assess the Sensitivity of Surge to Storm Characteristics, Journal of Geophysical Research: Atmospheres, 127, e2022JD037617, https://doi.org/10.1029/2022JD037617, 2022.

Loridan, T. and Bruneau, N.: Reask UTC: a machine learning modeling framework to generate climate-connected tropical cyclone event sets globally, Nat. Hazards Earth Syst. Sci., 25, 2863-2884, 10.5194/nhess-25-2863-2025, 2025.

Luettich, R. A. and Westerink, J. J.: Formulation and numerical implementation of the 2D/3D ADCIRC finite element model version 44. XX, R. Luettich Chapel Hill, NC, USA2004.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature machine intelligence, 2, 56-67, 2020.

Manning, R.: On the flow of water in open channels and pipes, Transactions, Institution of Civil Engineers of Ireland, 20, 161, 1891.

Marsooli, R., Lin, N., Emanuel, K., and Feng, K.: Climate change exacerbates hurricane flood hazards along US Atlantic and Gulf Coasts in spatially varying patterns, Nature Communications, 10, 3785, 10.1038/s41467-019-11755-z, 2019.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, Bulletin of the American Meteorological Society, 100, 2175-2199, https://doi.org/10.1175/BAMS-D-18-0195.1, 2019.

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T.: Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction, Ecological Modelling, 411, 108815, https://doi.org/10.1016/j.ecolmodel.2019.108815, 2019.

Mölter, T., Schindler, D., Albrecht, A. T., and Kohnle, U.: Review on the Projections of Future Storminess over the North Atlantic European Region, Atmosphere, 7, 60, 2016.

Mumuni, A. and Mumuni, F.: Automated data processing and feature engineering for deep learning and big data applications: A survey, Journal of Information and Intelligence, 3, 113-153, https://doi.org/10.1016/j.jiixd.2024.01.002, 2025.

Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, Proceedings of the 27th international conference on machine learning (ICML-10), 807-814,

National Geophysical Data, C.: 5-minute Gridded Global Relief Data (ETOPO5), National Geophysical Data Center, NOAA [dataset], 10.7289/V5D798BF, 1993.

National Geospatial-Intelligence, A.: Digital Nautical Chart (DNC) Portal, 2026.

NOAA: Multibeam Bathymetry Database (MBBDB), NOAA [dataset], 10.7289/V56T0JNC, 2004.

O'Shea, K. and Nash, R.: An Introduction to Convolutional Neural Networks, ArXiv e-prints, 2015.

Pain, C. C., Piggott, M. D., Goddard, A. J. H., Fang, F., Gorman, G. J., Marshall, D. P., Eaton, M. D., Power, P. W., and de Oliveira, C. R. E.: Three-dimensional unstructured mesh ocean modelling, Ocean Modelling, 10, 5-33, https://doi.org/10.1016/j.ocemod.2004.07.005, 2005.

Prechelt, L.: Early stopping-but when?, in: Neural Networks: Tricks of the trade, Springer, 55-69, 2002.

Pringle, W. J., Wirasaet, D., Roberts, K. J., and Westerink, J. J.: Global storm tide modeling with ADCIRC v55: unstructured mesh design and performance, Geosci. Model Dev., 14, 1125-1145, 10.5194/gmd-14-1125-2021, 2021.

Qian, X., Hwang, S., and Son, S.: A Study on Key Determinants in Enhancing Storm Surges Along the Coast: Interplay Between Tropical Cyclone Motion and Coastal Geometry, Journal of Geophysical Research: Oceans, 129, e2023JC020400, https://doi.org/10.1029/2023JC020400, 2024.

Qin, Y., Su, C., Chu, D., Zhang, J., and Song, J.: A Review of Application of Machine Learning in Storm Surge Problems, Journal of Marine Science and Engineering, 11, 1729, 2023.

Ramos-Valle, A. N., Curchitser, E. N., Bruyère, C. L., and McOwen, S.: Implementation of an Artificial Neural Network for Storm Surge Forecasting, Journal of Geophysical Research: Atmospheres, 126, e2020JD033266, https://doi.org/10.1029/2020JD033266, 2021.

Rego, J. L. and Li, C.: Nonlinear terms in storm surge predictions: Effect of tide and shelf geometry with case study from Hurricane Rita, Journal of Geophysical Research: Oceans, 115, https://doi.org/10.1029/2009JC005285, 2010.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323, 533-536, 10.1038/323533a0, 1986.

Sazli, M. H.: A brief review of feed-forward neural networks, Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, 50, 0-0, 10.1501/commua1-2_0000000026, 2006.

Shapley, L. S.: A value for n-person games, 1953.

Smith, A., Sampson, C., and Bates, P.: Regional flood frequency analysis at the global scale, Water Resources Research, 51, 539-553, https://doi.org/10.1002/2014WR015814, 2015.

Spearman, C.: The proof and measurement of association between two things, 1961.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research, 15, 1929-1958, 2014.

Tadesse, M., Wahl, T., and Cid, A.: Data-Driven Modeling of Global Storm Surges, Frontiers in Marine Science, 7, 2020.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F.: The computational limits of deep learning, arXiv preprint arXiv:2007.05558, 10, 2, 2020.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., and Ward, P. J.: Exploring deep learning capabilities for surge predictions in coastal areas, Scientific Reports, 11, 17224, 10.1038/s41598-021-96674-0, 2021.

Tiwari, P., Rao, A. D., Pandey, S., and Pant, V.: Investigation of the impact of complex coastline geometry on the evolution of storm surges along the eastern coast of India: a sensitivity study using a numerical model, Ocean Sci., 21, 381-399, 10.5194/os-21-381-2025, 2025.

Watson, P. A. G.: Machine learning applications for weather and climate need greater focus on extremes, Environmental Research Letters, 17, 111004, 10.1088/1748-9326/ac9d4e, 2022.

Webster, P. J., Holland, G. J., Curry, J. A., and Chang, H. R.: Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment, Science, 309, 1844-1846, 10.1126/science.1116448, 2005.

Westfeld, P., Richter, K., Maas, H.-G., and Weiß, R.: Analysis of the effect of wave patterns on refraction in airborne lidar bathymetry, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 41, 133-139, 2016.

Wood, M., Haigh, I. D., Le, Q. Q., Nguyen, H. N., Tran, H. B., Darby, S. E., Marsh, R., Skliris, N., Hirschi, J. J. M., Nicholls, R. J., and Bloemendaal, N.: Climate-induced storminess forces major increases in future storm surge hazard in the South China Sea region, Nat. Hazards Earth Syst. Sci., 23, 2475-2504, 10.5194/nhess-23-2475-2023, 2023.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S.: A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems, 32, 4-24, 2020.

Yeo, I. K. and Johnson, R. A.: A new family of power transformations to improve normality or symmetry, Biometrika, 87, 954-959, 10.1093/biomet/87.4.954, 2000.

Zhang, Y. J., Ye, F., Stanev, E. V., and Grashorn, S.: Seamless cross-scale modeling with SCHISM, Ocean Modelling, 102, 64-81, https://doi.org/10.1016/j.ocemod.2016.05.002, 2016.