Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-025-02993-3

ORIGINAL PAPER

Robust storm surge forecasts for early warning system: a machine learning approach using Monte Carlo Bayesian model selection algorithm

E. Macdonald[1] · E. Tubaldi[1] · E. Patelli[1]

Accepted: 12 April 2025

Abstract
Machine-learning based methods are increasingly employed for the prediction of storm surges and development of early warning systems for coastal flooding. The evaluation of the quality of such methods needs to explicitly consider the uncertainty of the prediction, which may stem from the inaccuracy in the forecasted inputs to the model as well as from the uncertainty inherent to the model itself. Defining the range of validity of the prediction is essential for the correct application of such models. A methodology is proposed for building a robust model for forecasting storm surges accounting for the relevant sources of uncertainty. The model uses as inputs the mean sea level pressure and wind velocity components at 10 m above sea level. A set of Artificial Neural Networks are used in conjunction with an adaptive Bayesian model selection process to make robust storm surge forecast predictions with associated prediction intervals. The input uncertainty, characterised by comparing hindcast data and one day forecasted data, is propagated through the model via a Monte Carlo based approach. The application of the proposed methodology is illustrated by considering 24 h target forecast predictions of storm surges for Millport, in the Firth of Clyde, Scotland, UK. It is shown that the proposed approach significantly improves the predictive performance of existing machine learning based models and provides a meaningful prediction interval that characterises feature, model and forecast uncertainty. The forecast system has negligible computational time requirements and showed very good agreement with observations acccording different metrics and achieving e.g., a correlation coefficient of 0.942 for 24 h forecasted surge from 2021 to 2023. The mean absolute error was 0.06 m for all observations and only 0.10 m for observations above 0.75 m showing its accuracy for predicting extreme events.

1 Introduction

Coastal flooding is one of the most pressing effects of climate change, with a growing body of research and evidence suggesting that the problem is likely to worsen in the coming years. The Intergovernmental Panel on Climate Change has projected that global mean sea level will rise by 0.43 to 0.84 m by the end of this century, depending on future emissions scenarios (IPCC 2022). As sea levels rise, the risk of coastal flooding increases, as higher sea levels will most likely result in more frequent and severe flooding events (Nicholls et al. 2018). Climate change is also contributing to more intense and frequent storms, see e.g. (Knutson 2010). These storms can produce surges, which contribute to coastal flooding (Emanuel 2017). In addition, warmer oceans are providing more energy to fuel these storms, making them more dangerous and unpredictable (IPCC 2022). Addressing the above problems requires investment in measures to protect vulnerable coastal communities. There is also an urgent need for robust early warning tools that are capable of forecasting extreme surge events with sufficient lead time so that preventive risk-mitigation and emergency-management measures can be taken. Accurately forecasting storm surge events is difficult owing to abrupt variations in the track and intensity of storms, bathymetric changes, and data assimilation issues, to name a few (Li and Nie 2017).

✉ E. Patelli
edoardo.patelli@strath.ac.uk

E. Macdonald
euan.macdonald@strath.ac.uk

E. Tubaldi
enrico.tubaldi@strath.ac.uk

[1] Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow, Scotland

Published online: 07 May 2025

To enable stakeholders to take optimal risk-mitigation decisions, early warning systems must account for these uncertainties in the predictions.

The field of storm surge modelling and forecasting has benefitted greatly from the development of high-fidelity numerical models, providing detailed description of hydrodynamic processes to generate accurate and reliable storm surge responses, see e.g. (Taflanidis et al. 2013). For surge forecasting, an extensive review was conducted by Al Kajbaf and Bensei (2020), grouping existing hydrodynamic modelling types as either high fidelity physics-based models, often with a substantial computational cost, or low-fidelity models with an associated drop in accuracy but with computational costs that are better suited to forecasting applications. Significant attention has been given in recent years to landfall location, where a large error in the forecasted landfall location can lead to a large error in surge height prediction (Kohno et al. 2018). Numerical weather prediction models have become more popular in recent years for use as surge forecast inputs with the draw back that storm intensity is often underpredicted for storms that are either too weak or very strong (Kohno et al. 2018). Storm surges are characterised by high variability and contribute the largest source of uncertainty in predicting sea level elevation (Mel et al. 2014). Around the UK, 2 day surge forecasts are made by the National Tidal and Sea Level Facility (NTSLF) 7-km NEMO Community Ocean model (NTSLF 2019).

Some studies have explored the nature of the uncertainty associated with different surge prediction methods such as surge response functions (Taylor et al. 2015), with particular attention given to the lead time of the surge height predictions (Mel and Lionello 2016). Multi-scenario storm surge forecasts are becoming increasingly popular, considering the variability in various parameters such as size, location of landfall and intensity (Kohno et al. 2018).

The overall uncertainty in storm surge prediction is multi-faceted. From a modelling perspective, errors arise from simplifying assumptions, parameterisation, local topographic and bathymetric effects, and initial conditions. Additionally, uncertainties surrounding storm characteristics such as intensity, track, and speed of storms which come from weather models carry their own uncertainties that propagate to the final surge height estimate. Similarly, variability in atmospheric pressure, wind speed, and precipitation predictions also contribute to uncertainty in forecasting the meteorological conditions that drive storm surges. In a study into the interrelationship between the effects of combined random errors and bias in numerical weather prediction (NWP) models and in surge models, Resio et al. (2017) found that surge model bias can play a dominant role in distorting forecast probabilities.

An alternative to high fidelity numerical models is to use statistical models that rely on historical data to predict future storm surges by identifying patterns and trends, using methods like regression analysis and time series analysis. Additionally, machine learning or data-driven models use statistical techniques and algorithms to learn from data and make predictions. Hybrids methods also exist, which combine different forecasting approaches, such as statistical methods and numerical weather prediction models, to provide more accurate predictions. In recent years, machine learning forecasting systems have become increasingly popular. The National Oceanic and Atmospheric Administration identified machine learning as a key technology for advancing weather forecasting, thanks to its ability to improve the accuracy of forecasts and to reduce the time required to generate them (NOAA 2020). The World Meteorological Organization has also recognised the potential of machine learning in weather forecasting, emphasising its ability to provide rapid and accurate predictions for severe weather events (WMO 2021).

Due to their ability to model nonlinear systems, artificial neural networks (ANNs) have proven to be an extremely versatile hydrodynamic surrogate modelling tool. An abundance of literature exists describing the use of ANNs for surrogate modelling for coastal (Sztobryn 2003), pluvial (Asadi et al. 2013), and fluvial (Fazel et al. 2014) flood forecasting applications. A recent study (Kim et al. 2019) described a methodology for defining the best artificial neural network surge forecasting model by comparing the correlation coefficients between the observed and forecasted surge height of different network architectures using meteorological and storm characteristic inputs. The study trained 20 ANNs and compared their performance in predicting typhoon characteristics for storm surge forecasting with different lead times. They found that for a 5- or 12-h lead time, typhoon characteristics such as longitude, latitude, central atmospheric pressure, and highest wind speed were important. However, as the lead time was extended to 24 h, highest wind speed ceased to be influential. The study suggested different approaches for developing forecasting models for different lead times and noted that the optimal network architecture varied accordingly.

Tiggeloven et al. (2021) explored the use of four architecture of deep learning methods, namely ANN, convolutional neural network (CNN), long-short term memory (LSTM), and Convolutional LSTM to predict the surge component of sea-level variability based on local atmospheric conditions. The models were constructed using global tide station data and showed the best performance in the mid-latitudes. The study found that overall the deep learning models could be useful for predicting extreme sea levels. LSTM model generally outperformed the other models since LSTM models

are designed to process sequential data, making them well-suited for weather forecasting tasks that involve time-series data (Salman et al. 2018).

Yu and Hong (2020) discussed in detail the challenges of hyperparameter optimisation in neural networks, including problems with dimensionality and the high cost of evaluating different combinations of hyperparameters (Yu and Hong 2020). This leads to variations in performance and output accuracy, introducing uncertainty in selecting the best performing ANN. Yu and Hong (2020) also discussed various solutions, such as using Bayesian optimisation, gradient-based methods, and evolutionary algorithms for hyperparameter optimisation. Kingston et al. (2008) used Bayesian model selection for water resources modelling, finding that it was an objective method for accurately selecting the optimal complexity of an ANN model when used in conjunction with the Bayesian training procedure (Kingston et al. 2008). Additionally, despite sharing the same architecture and training process, ANNs can produce different models due to the random initialisation of their weights during training (Oparanji et al. 2017). This leads to variations in performance and output accuracy, introducing uncertainty in selecting the best performing ANN. This can be mitigated by using optimised weight initialisation methods and seeding to ensure reproducibility. Usually, multiple ANN models are trained in searching for the optimal one (Kim et al. 2019) and cross-validation tests are commonly used to select the best performing ANN (Tolo et al. 2018), whereas the other ANNs are discarded. However, this approach does not consider the potential noise and imprecision in the validation data or the ANN's performance on unseen data, and it does not fully exploit the results of the training process in developing an optimal model. Oparanji et al. (2017) proposed a Bayesian model averaging algorithm to provide an averaged prediction across a set of networks with each network weighted given its likelihood of being correct. The prediction obtained according to this approach is more accurate than that obtained with a single ANN. Moreover, the variation across the network set provides an estimate of the epistemic uncertainty in the neural network modelling approach.

This paper aims to enhance current approaches for surge forecasting by developing a robust machine learning-based forecasting method designed to take operational weather forecasts and make surge height predictions with prediction intervals derived from the uncertainty of the weather forecast (*forecast uncertainty*), the uncertainty associated with the ability of the available data to describe the variability in surge heights (*feature uncertainty*), and the uncertainty inherent to the model itself (*network uncertainty)*. Whilst Bayesian model averaging methods have previously been applied to surge ensemble forecasting (Salighehdar et al. 2017), in this paper a consistent network architecture is used across all ANNs where variations among the networks stem solely from the use of distinct random seeds. This leads to different final weights after training and variances across the set of network predictions are then used to quantify the uncertainty associated with the training of the ANN model. This methodology is then further enhanced to consider *feature uncertainty* which stems from inadequate feature selection and incomplete coverage of features across all physically possible scenarios and (weather) *forecast uncertainty* estimated by an integrated Monte Carlo based approach. The resulting Monte Carlo Average Bayesian (MCBA) model is only deployable due to the reduced computational burden of ANNs compared to high-fidelity physics-based models.

The rest of the paper is organised as follows: Sect. 2 details the inputs at the base of the proposed early surge warning system, together with pre-processing requirements and uncertainty considerations. Section 3 describes the model structure for a general case, the Adaptive Bayesian Model Selection (ABMS) algorithm for describing *network uncertainty*, and the extension to propagate the input uncertainty, along with the error metrics used to evaluate model performance. Section 4 shows the results of the validation of the model for the Firth of Clyde region of Scotland and the application of the model with uncertainty with pressure and wind forecasts. Section 5 details the conclusions of the study along with recommendations for future work.

2 Input analysis

2.1 Input selection

An important issue in developing surrogate models for storm surge prediction is the choice of the relevant set of input variables, and the characterisation and propagation of the uncertainty inherent to these. A careful choice of the input variables can simplify the modelling process and reduce the computational burden, making it more efficient and less prone to overfitting. Second, it can help to identify the most important variables and their relationships, leading to more interpretable models and better insights into the underlying system.

The behaviour of storms is influenced by Rossby and Kelvin waves**.** Rossby waves are the large-scale dynamical response of the ocean to atmospheric conditions (Chelton and Schlax 1996). Their behaviour is responsible for pressure and windspeed variations that can drive storm surges, particularly in mid-latitudes. Kelvin waves, on the other hand, are a type of gravity wave that can rapidly propagate sea level changes along coastlines, contributing to the intensity and reach of storm surges (Wang 2002). In this study,

we consider the wind velocity components in the eastward and northward directions at 10 m above the ground, referred to as *U10* and *V10* respectively, and mean sea level pressure (*MSLP*) across the input grid. Additionally, the *U10*, *V10*, and the pressure difference at the target location, along with the current surge level, are included in the analysis.

Machine learning studies based on observational data are limited due to the sparsity of in situ data and the lack of large and well- structured datasets that are suitable for training machine learning models (Qin et al. 2023). This issue can be overcome by using reanalysis data sets which have increased temporal and spatial coverage compared to in-situ measurements. The size of area over which the inputs are considered, or 'footprint', is significant (Tiggeloven et al. 2021). A larger input area can improve surge height predictions with the cost of additional computational demand for model training and run time.

The early warning surge system is trained on reanalysis data and provides real time prediction on forecast data. More specifically, reanalysis data is first used so that forecast errors are not introduced to the models during training, providing the best opportunity to find patterns between the surge response and the 'true' wind and pressure reanalysis data. It is important to clarify that reanalysis data is not free from uncertainty. In fact, reanalysis data is the product of a physics-based model and therefore subject to the errors listed in Sect. 1 (a specific example of ERA-5 wind error is discussed in Sect. 4.3). Nevertheless, reanalysis data is used as the truth, against which a (weather) *forecast error* can be quantified through comparison of historical forecast data and its concurrent reanalysis data. Then, the surge prediction, i.e. the model output, is obtained using wind speed and pressure forecasts, whilst accounting for the corresponding amount of forecast error. Through this methodology the uncertainty in the weather forecast is leveraged and added to *network uncertainty* and *feature uncertainty* to estimate the final uncertainty in the surge height prediction.

In this study, reanalysis data for model development is taken from ERA-5 with the developed early warning system driven by European Centre for Medium-Range Weather Forecasts (ECMWF) operational forecasts. ERA-5 (C3S 2017) is the fifth edition of the ECMWF atmospheric reanalysis of the global climate, providing hourly estimates of meteorological variables on a global 30 km grid. This is used for model training and validation. The operational forecasts are made by the ECMWF Integrated Forecasting System (IFS) which is a global numerical model supported by a sophisticated data assimilation system that estimates the likely evolution of the weather (ECMWF 2024).

2.2 Input dimension reduction

Input grids for surge models can cover extensive footprints and many of the most popular modelling options consider time-dependent variables (see Sect. 3.1), which can imply vast input arrays with large numbers of dimensions. Input grid dimensionality can be reduced using principal component analysis (PCA). PCA is a mathematical method used for reducing the dimensionality of large data sets by extracting the most important features that explain the variability of the data. PCA has been widely and successfully applied to help understand and interpret large, spatially extensive climate datasets (Reusch et al. 2005).

PCA is based on space transformation where the original data set is linearly transformed onto a new coordinate system represented by a set of orthogonal axes created to capture the maximum variance of the data in each subsequent dimension. The original data can then be represented in this new coordinate system using a smaller number of components able to preserve the most significant features of the original data. The result is a compressed representation of the data that captures the most important information. The principal components that explain the desired variability are kept and those deemed to represent an insufficient amount of variability are discarded. A second advantage of PCA is that principal components are independent of one another, and this removes correlation in the inputs. This is useful since highly correlated data reduces the distinctiveness of data representation and can impede ANN model training and result in models that struggle to generalise (Mohamad-Saleh and Hoyle 2008). A downside to PCA is that principal components are just mathematical constructs that represent variance in the data and do not necessarily have an inherently physical explanation, meaning that there can be issues with feature interpretability (Reusch et al. 2005).

The number of principal components to be selected for the model fitting is a trade-off between computational demand and preserved input variation. The principal components are normalised by subtracting the mean and dividing by the standard deviation to eliminate scale issues with ANN training. The outputs of principal component analysis include the principal components (also referred as the score), the amount of variation explained by each principal component (used to determine the desired number of principal components) and the principal component coefficients (used to transform the input data to the principal component domain and back).

2.3 Forecast uncertainty characterisation

As mentioned in Sect. 2.1, the early warning surge system is trained on reanalysis data and run using forecast data.

This is to allow the expected uncertainty in the forecast to be propagated to the surge prediction uncertainty. To quantify the error between the reanalysis data and forecast data, the IFS data must be mapped to the same domain as the reanalysis data, as outlined in Sect. 2.2. This is done by normalising the IFS forecast inputs and multiplying them with the principal component coefficients for the reanalysis data. The 'true' reanalysis value of the *z*th principal component $PC_{z,R}$ can be expressed as the sum of an unbiased forecasted principal component $\widetilde{PC}_{z,forecast}$ and a zero mean forecast error $\varepsilon(0,\sigma)$:

$$PC_{z,R} = \widetilde{PC}_{z,forecast} + \varepsilon(0,\sigma) \tag{1}$$

The expected value of $PC_{z,R}, E[PC_{z,R}]$, can be expressed in terms of the expected value of the biased forecast $E[PC_{z,forecast}]$, adjusted with regression coefficients α_z and β_z:

$$E[PC_{z,R}] = \alpha_z \cdot E[PC_{z,forecast}] + \beta_z \tag{2}$$

Since the expected value of an unbiased forecast and the expected value of the true value are necessarily equal, $\widetilde{PC}_{z,forecast}$ can substituted in Eq. (1) for the biased forecast with applied regression coefficients:

$$PC_{z,R} = \alpha_z \cdot PC_{z,forecast} + \beta_z + \varepsilon(0,\sigma_z) \tag{3}$$

Hence $PC_{k,R}$ can be expressed in terms of the forecast with an error that can be propagated through the model via Monte Carlo sampling. This process is applied to Z forecast features and a set of error distributions $\epsilon = \{\varepsilon(0,\sigma_1),\ldots,\varepsilon(0,\sigma_z),\ldots,\varepsilon(0,\sigma_Z)\}$ is obtained.

3 Robust ANN-based surge forecasting system

The proposed robust ANN-based surge forecasting system (RSFS) schematic is shown in Fig. 1, describing the steps involved in converting a 24-h weather forecast to a 24-h surge height prediction. The pre-processing and input uncertainty evaluation described in Sect. 2 are written in Python as modular functions for ease of operation. The steps required for ANN Set prediction, model averaging, combining uncertainties and producing the robust surge height prediction are detailed in the next sub-sections.

3.1 Network architecture

A LSTM architecture is used in the surge forecasting framework. LSTMs are designed to remember information for prolonged periods. This capability is achieved through a sophisticated internal architecture featuring gates that regulate the flow of information (Le et al. 2019) as shown in Fig. 2. The forget, new memory, and output gates, denoted by numbers 1, 2 and 3, decide what information should be retained or discarded as the sequence progresses. As data flows through an LSTM, the new memory gate controls how much new information enters the memory cell, the forget gate determines what part of the existing memory to keep, and the output gate decides what part of the current cell state makes it to the output. The sigmoid functions used in the gates decide how much information to retain or let through by outputting values between 0 and 1. The *tanh* functions act as nonlinear transformation function to modify the information to be added to the cell state. Additionally, they process the cell state information to help determine what ultimately gets outputted from the cell, influencing both the next cell state in the sequence and the final predictions. These processes enable LSTM networks to retain long term dependencies effectively.

A Bidirectional LSTM is a configuration of LSTM where data is processed both in a forward and backward direction, effectively learning from sequences in a way that considers both past and future data. This dual-direction processing provides a more complete understanding of the sequence, enhancing performance in tasks where the context from both before and after a data point is important.

In this study a Bidirectional LSTM network is designed that uses meteorological variable principal components n_{pca} from the computational grid. This network also uses additional inputs, including the pressure difference between the target location and the maximum pressure in the domain ΔP_{local}, as well as the *U10* and *V10* wind components at the target location, $U10_{local}$ and $V10_{local}$. All inputs are normalised to have zero mean and unit standard deviation. They are considered across a 49-h window, spanning 48-h prior to the prediction hour and the prediction hour. Given that the forecast lead time is 24 h this means the previous 24 h are considered along with the present hour and the 24-h forecast. In addition, the current and previous 24 h surge heights are considered. For areas where there is a dependency between surge height and tide, it may be desirable to include tide height as well (Williams et al. 2016). This model structure is shown in Fig. 3.

Given the large number of trainable parameters in the model, the dense layer is accompanied by a dropout layer with a 0.1 dropout rate. Dropout layers uncouple a random group of weights between layers, preventing all neurons in a layer from synchronously optimizing their weights. This decorrelates the weights by preventing all the neurons from converging to the same targets (Labach et al. 2019). Initial tests also revealed that using kernel, bias and activity

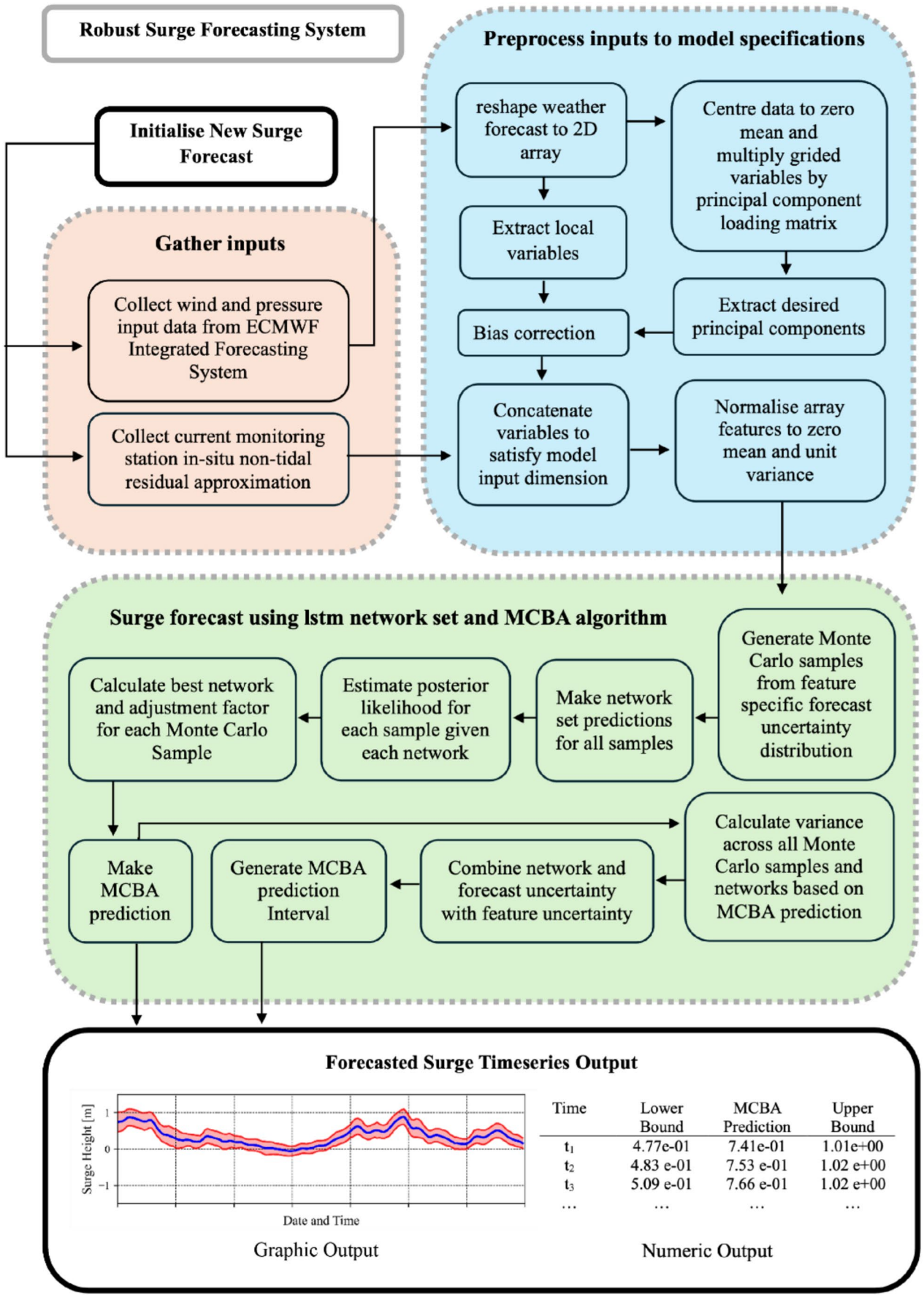
Robust Surge Forecasting System
Initialise New Surge Forecast
Gather inputs
Collect wind and pressure input data from ECMWF Integrated Forecasting System
Collect current monitoring station in-situ non-tidal residual approximation
Preprocess inputs to model specifications
reshape weather forecast to 2D array
Extract local variables
Bias correction
Concatenate variables to satisfy model input dimension
Centre data to zero mean and multiply grided variables by principal component loading matrix
Extract desired principal components
Normalise array features to zero mean and unit variance
Surge forecast using lstm network set and MCBA algorithm
Generate Monte Carlo samples from feature specific forecast uncertainty distribution
Make network set predictions for all samples
Estimate posterior likelihood for each sample given each network
Calculate best network and adjustment factor for each Monte Carlo Sample
Calculate variance across all Monte Carlo samples and networks based on MCBA prediction
Combine network and forecast uncertainty with feature uncertainty
Generate MCBA prediction Interval
Make MCBA prediction
Forecasted Surge Timeseries Output
Surge Height [m]
Date and Time
Graphic Output
Time | Lower Bound | MCBA Prediction | Upper Bound
t1 | 4.77e-01 | 7.41e-01 | 1.01e+00
t2 | 4.83 e-01 | 7.53 e-01 | 1.02 e+00
t3 | 5.09 e-01 | 7.66 e-01 | 1.02 e+00
Numeric Output

Fig. 1 Operational schematic of robust surge forecasting system

regularizes reduced overfitting and provided better a trade-off in variance and bias during the training process especially for predictions in the extreme range. The dense layer contains 100 nodes.

3.2 Adaptive Bayesian model selection algorithm (ABMS)

Typical Bayesian averaging methodologies explore the use of Bayesian model selection techniques based on approximating the posterior probability of a particular model to either identify the 'best' network in the set or to average the predictions of the individual models within the set, see for ensemble surge forecasting (Salighehdar et al. 2017). This differs from the aim of this paper in that the network architecture of the ANN is fixed but trained multiple times producing a set of different performing networks. During the training, the likelihood that the ANN weights will arrive at the global optimal is limited. Hence, instead of making predictions with a single ANN, a set of networks is used to make predictions with different models performing to different degrees for different parts of the target range. By so doing, the posterior probability for each network prediction in the set can be thought of as the degree of belief that its given prediction is 'true', and the variance across the set quantifies the level of uncertainty of the model itself (Oparanji et al. 2017). The Bayesian model selection process uses the posterior probability calculated using probability points to identify the best trained network for a given observation. The process then uses the variance of all the predictions to create a prediction interval and adjustment factor that form the robust prediction.

A set of M neural network models $N_k (k=1, 2, \ldots, M)$, each with a different seed, is trained over a dataset $D_{train}(x, y)$, where $\boldsymbol{x}$ represents the feature vector and $\boldsymbol{y}$ the target vector (here the target vector reduce to the scalar value of the surge). A second, independent dataset $D_{eval} = (x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$ is defined with the same structure over n observations. For an unseen feature vector, each of the M ANNs generates a response $y_k = N_k(x)$. Using Bayes' theory, the empirical posterior probability for the k^{th}-network predicted response y_k given the evaluation data, $P(N_k, y_k|D_{eval})$, can be expressed as:

$$P(N_k, y_k|D_{eval}) = \frac{P(D_{eval}|N_k) \cdot P(N_k, y_k)}{P(D_{eval})} \tag{4}$$

where $P(N_k, y_k)$ denotes the prior probability (i.e. assumed probability distribution before evidence) of the N_k model, $P(D_{eval}|N_k)$ denotes the likelihood term (i.e. the probability for the sample data given the k^{th}-network predicted response), and $P(D_{eval})$ is the evidence, which can be expressed as:

$$P(D_{eval}) = \sum_{k=1}^{M} P(D_{eval}|N_k) \cdot P(N_k, y_k) \tag{5}$$

Since the ANNs only differ for the seed number used to initialise the weights, there is no difference in terms of the individual ANN credibility and therefore the same prior probability is assigned to the various networks, i.e. $P(N_k) = 1/M$. For any N_k the true target value y in the evaluation dataset can be written as $y = y_k + \varepsilon_k$, with ε_k as the corresponding prediction error. It is assumed that the error ε_k follows a normal distribution with zero mean and variance σ_k^2, i.e. $\varepsilon_k = N(0, \sigma_k^2)$. To support this assumption, it is desirable to select networks with minimal bias along the length of the target range. The variance of network N_k can be estimated from the maximum likelihood estimation for n independent observations (Kleinbaum and Klein 2010):

$$\sigma_k^2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{i,k}^2 \tag{6}$$

where σ_k^2 represents the intrinsic variability in the prediction errors of N_k, and assumes that the model structure is appropriate and that all relevant variables have been included. The likelihood function for true target response y given N_k can be approximated as:

$$P(y|N_k) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot e^{\frac{-(y-y_k)^2}{2\sigma_k^2}} \tag{7}$$

Heteroscedastic variance can also be used by calculating the variance associated with different magnitudes of prediction by splitting the outputs into different bins along the target range. In this case, Eq. (7) is adapted so that all parameters become bin dependent e.g. σ_k^2 is replaced by l^{th} bin variance $\sigma_{l,k}^2$.

Denoting y_r as the prediction of the network with the highest posterior probability in the set given the evaluation data, i.e.:

$$r = \underset{k}{\operatorname{argmax}} \{P(N_k, y_k|D_{eval})\}, \quad k = 1, 2, \ldots, M \tag{8}$$

the Bayesian averaged prediction y_{abms}, is expressed as:

$$y_{abms} = y_r + \sum_{k=1}^{M} P(N_k, y_k|D_{eval}) \cdot F_k \tag{9}$$

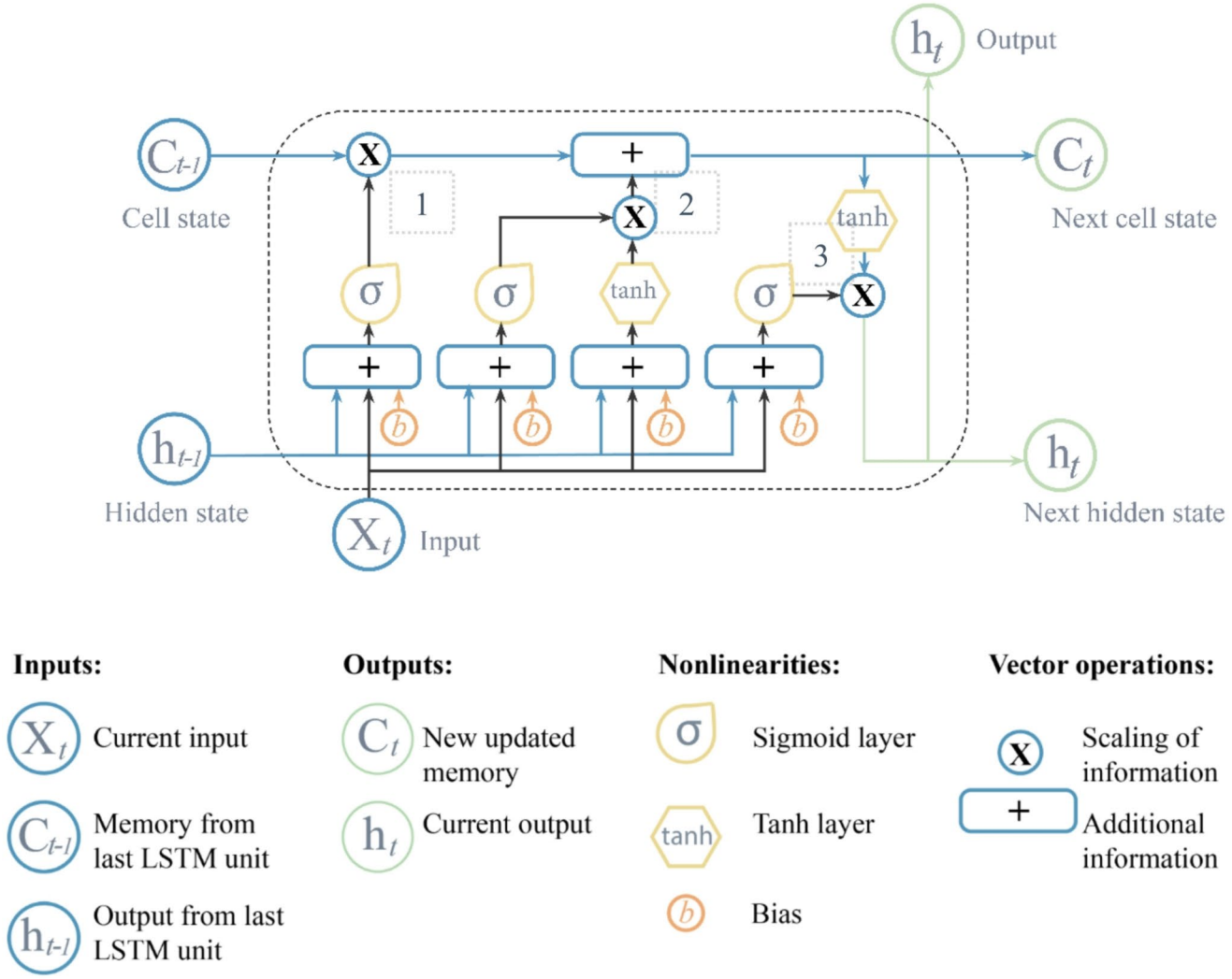

Fig. 2 The structure of the long short-term memory (LSTM) cell. The figure has been redrawn and adapted from (Yan 2016). Forget gate is denoted by 1, new memory gate by 2 and output gate by 3

The second term is a weighted adjustment factor, with F_k denoting the difference between network response y_k, and best network response y_r:

$$F_k = y_k - y_r \tag{10}$$

The variance of the robust response is then evaluated as:

$$V\left(y_{abms}\right) = \sum_{k=1}^{M} P\left(N_k, y_k | D_{eval}\right) \cdot \left(y_k - y_{abms}\right)^2 \tag{11}$$

The main idea presented here is that the importance of each network's information is proportional to the quality of the prediction which varies across the range of target response values. The upper and lower bounds of the, e.g., 95% prediction interval $[\underline{y}_{abms}, \overline{y}_{abms}]$, complete the robust prediction:

$$\overline{y}_{abms} = y_{abms} + 1.96\sqrt{V\left(y_{abms}\right)} \tag{12}$$

$$\underline{y}_{abms} = y_{abms} - 1.96\sqrt{V\left(y_{abms}\right)} \tag{13}$$

As stated above, the posterior probability of each network in the set is thought of as the degree of belief that its given prediction is 'true', thereby quantifying the level of uncertainty of the ensemble (Oparanji et al. 2017). This paper will refer to this ensemble variance as *network uncertainty*. Unlike the approach by Oparanji et al. (2017), who derive posterior probabilities directly from the training data, the methodology presented here utilises evaluation data independent from the training data. This approach ensures that the probabilities reflect the model's performance on unseen data, providing a more accurate representation of uncertainty in real operational scenarios. Additionally, this methodology allows for the integration of both *feature uncertainty*,

(i.e. inadequate feature selection and incomplete coverage of features across all physically possible scenarios) and network uncertainty. Relying solely on one type of uncertainty could lead to an underestimate in surge prediction uncertainty, leading to narrower prediction intervals than appropriate for real world, real time applications. By taking this alternative approach and deriving feature uncertainty and network uncertainty from distinct sources, the assumption of independence can be maintained when combining these uncertainties. By assuming that *feature uncertainty* is independent of *network uncertainty*, Eq. (11) is adjusted to include the feature uncertainty for the k^{th} network, a_k:

$$V\left(y_{abms}\right)=\sum_{k=1}^{M} P\left(N_k, y_k | D_{eval}\right) \cdot\left\{\left(y_k-y_{abms}\right)^2+a_k\right\} \tag{14}$$

In this model, a_k represents the error variance of the trained model when evaluating its own training data. This error is assessed in percentile bins across the target range, anticipating more feature error at the extremes where the data is sparser, and the targets are more challenging to predict. a_k indicates the strength of relationship between the features and targets. Specifically, when a_k is lower, it suggests a stronger and more predictable connection between the input features and the output targets. This indicates that the model is effectively learning from the training data and is capable of accurately predicting or estimating the targets based on the inputs it receives. In practical terms, the assumption of independence means that systematic error captured by the training errors from D_{train} does not influence the random variability of the prediction errors captured by σ_k^2 from D_{eval}.

3.3 Monte Carlo Bayesian averaging algorithm (MCBA)

While the ABMS method is used to quantify the epistemic uncertainty raised from the ANN models, it does not account for the *forecast uncertainty*. Here the ABMS algorithm is developed to include a Monte Carlo based approach. Monte Carlo sampling is a powerful tool for simulation and estimating the behaviour of complex systems and propagating aleatoric uncertainty from input parameters. By generating a large number of realisations of input parameters and evaluating the response of the model under investigation, this approach can provide accurate, flexible, and robust estimates of statistical quantities.

Wind and pressure forecast uncertainties are calculated using Eqs. (1)–(2), producing a set of error distributions ϵ. *N* Monte Carlo samples are produced with *j*th Monte Carlo feature vector x_j expressed as:

$$x_j = x + \epsilon_j \tag{15}$$

where ϵ_j collects the feature specific realisations from the error distribution set. For each j^{th} sample, N_k generates a response:

$$y_{j,k} = N_k\left(x_j\right) \quad \text{for } k = 1, 2, \ldots, M \tag{16}$$

A Bayesian averaged prediction $y_{abms,j}$ can be determined for each sample by applying Eq. (9). The MCBA averaged prediction y_{mcba} is then taken as the expected value of all $y_{abms,j}$ across *N* Monte Carlo samples:

$$y_{mcba} = \frac{1}{N}\sum_{j=1}^{N}\left(y_{abms,j}\right) \tag{17}$$

The variance of the MCBA robust response $V\left(y_{mcba}\right)$ is then evaluated as:

$$V\left(y_{mcba}\right) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{M} P\left(N_k, y_{j,k} | D_{eval}\right) \cdot \left(y_{j,k} - y_{mcba}\right)^2 \tag{18}$$

Similar to Eq. (14), $V\left(y_{mcba}\right)$ can be expanded to include *feature uncertainty*, a_k as:

$$V\left(y_{mcba}\right) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{M} P\left(N_k, y_{j,k} | D_{eval}\right) \cdot \left\{\left(y_{j,k} - y_{mcba}\right)^2 + a_k\right\} \tag{19}$$

Hence the *feature*, *network* and *forecast uncertainty* are propagated to the final prediction interval by applying Eq. (17), with the variance calculated in Eq. (18) (or Eq. (19)). Finally, Eqs. (20) and (21) correspond to the upper and lower bound of the 95% prediction interval:

$$\overline{y}_{mcba} = y_{mcba} + 1.96\sqrt{V\left(y_{mcba}\right)} \tag{20}$$

$$\underline{y}_{mcba} = y_{mcba} - 1.96\sqrt{V\left(y_{mcba}\right)} \tag{21}$$

3.4 Error metrics

Various metrics should be employed to form a comprehensive evaluation of a surrogate model's performance. To quantify the error in the same unit as the observations, the mean absolute error (*MAE*), specified by Eq. (22), and root mean squared error (*RMSE*), Eq. (23), are often used. *RMSE* assigns a larger weight to larger errors, whereas *MAE* gives equal weight to all errors. The *R*2 value, represented by Eq. (24), determines how well the predictions align with the *n* observations by expressing the proportion of the unexplained variance compared to the total variance.

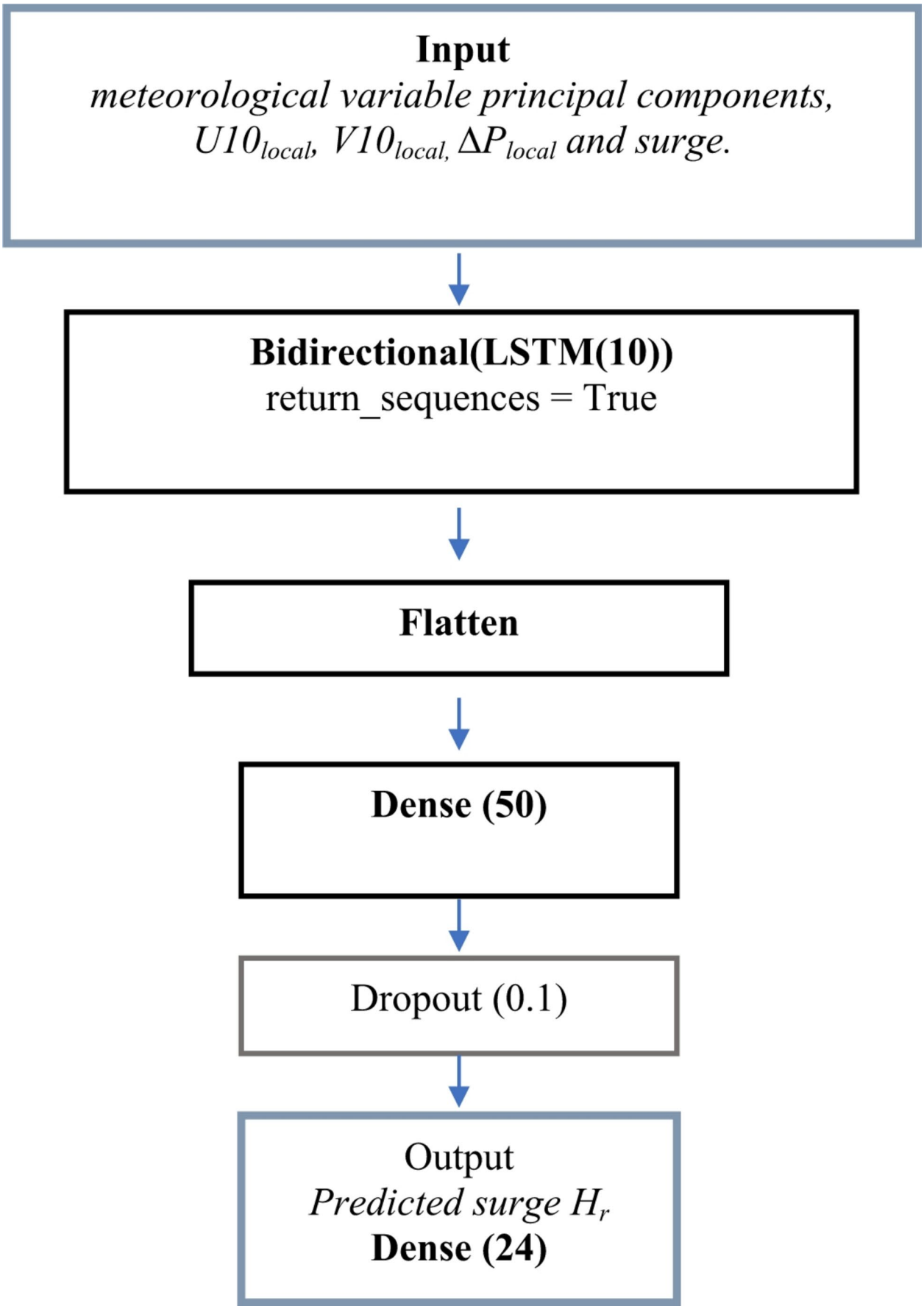
Input
meteorological variable principal components,
$U10_{local}$, $V10_{local}$, ΔP_{local} and surge.
Bidirectional(LSTM(10))
return_sequences = True
Flatten
Dense (50)
Dropout (0.1)
Output
Predicted surge H_r
Dense (24)

Fig. 3 Architecture of LSTM based surge forecasting network

Additionally, the *Bias*, represented by Eq. (25) is considered as modelled time series data frequently exhibit systematic bias (Jackson et al. 2019). The scatter index (*SI*), represented by Eq. (26), is a normalised measure of error that provides more precise and accurate information about the accuracy of a numerical simulation than *RMSE* (Mentaschi et al. 2013). Equation (27) is Pearson's correlation coefficient (CC) which is a measure of the strength and direction of a linear relationship between observed and predicted values. This provides insights into how well the predictions match the actual data.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_{p,i}|}{n} \tag{22}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_{p,i})^2}{n}} \tag{23}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f(y_{p,i}))^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{24}$$

$$Bias = \frac{\sum_{i=1}^{n} (y_{p,i} - y_i)}{n} \tag{25}$$

$$SI = \sqrt{\frac{\sum_{i=1}^{n} \left((y_i - \overline{y}) - (y_{p,i} - \overline{y}_{p,i})\right)^2}{\sum_{i=1}^{n} y_i^2}} \tag{26}$$

$$CC = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(y_{p,i} - \overline{y}_p)}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2 \sum_{i=1}^{n} (y_{p,i} - \overline{y}_p)^2}} \tag{27}$$

In each case, y_i and $y_{p,i}$ denote respectively the ith observation and prediction values, and $\overline{y}$ and $\overline{y}_p$ are the mean value of the observations and predicted values. In this study, the above metrics are used to give a complete comparison of model performance.

4 Model application

In this section, the forecasting model described in Sects. 2 and 3 is applied to and validated for the *Firth of Clyde* basin in southwest Scotland, which is prone to coastal inundation due to its complex bathymetry and exposure to Atlantic generated weather systems (Sabatino et al. 2016). Model predictions are compared to Millport surge observations in order to validate the model using ERA5 inputs. The model is then adapted to accommodate forecast inputs and its operational performance is assessed against Millport surge observations. All pre-processing tools, model building tools and operational functions are contained in the Bayesian Coastal Forecasting Toolbox (Macdonald 2024).

4.1 Case study—Firth of Clyde

The Firth of Clyde Basin was identified by the SEPA Coastal Hazard Mapping Study (SEPA 2015) as a vulnerable coastal area in Scotland. Along the east coastline of the Firth of Clyde, the A78 trunk road connects the ferry port of Stranraer to Glasgow and the rest of Scotland. This road link is critical for trade and connecting isolated towns and villages with emergency services and is prone to annual road closures due to overtopping. This makes the area suitable for an early warning surge forecasting system to inform responsible authorities of the likelihood of occurrence of extreme surges with sufficient warning time for them to take meaningful risk and impact mitigation actions. Figure 4 illustrates the area of interest along with the trunk road network and relevant locations.

The Transport Research Laboratory (TRL) report (Milne et al. 2017a, 2017b) examined surge events for Millport between 1995 and 2013 stated that flooding is likely to occur on the A78 if the flood potential value (FPV) exceeded 5.3 m Chart Datum (CD). The report defines the *FPV* as:

$$FPV = A + 2S \tag{28}$$

where A is the astronomical tide and S is storm surge. This formulation was based on the assumptions that the height of waves breaking at the coast during a storm are at least as high as the storm surge and that larger surges will have higher waves breaking at the shore. Figure 5 shows the Chart Datum tidal height against surge height at Millport for the years 1980 to 2019. In this plot, the extreme value boundary (estimated as 0.75 m) is established as the minimum surge height value required for the *FPV* value to exceed the high risk threshold.

While Eq. (28) is an adequate approximation to establish an extreme surge height threshold, more comprehensive wave height and overtopping models exist, e.g. (Pullen et al. 2018), that explicitly consider the water level, wave characteristics, and defence structure characteristics along with their associated uncertainty. It is important to mention that the proposed methodology can be easily integrated into a larger and detailed overtopping framework, or/and extended to include also wave forecasts in the output.

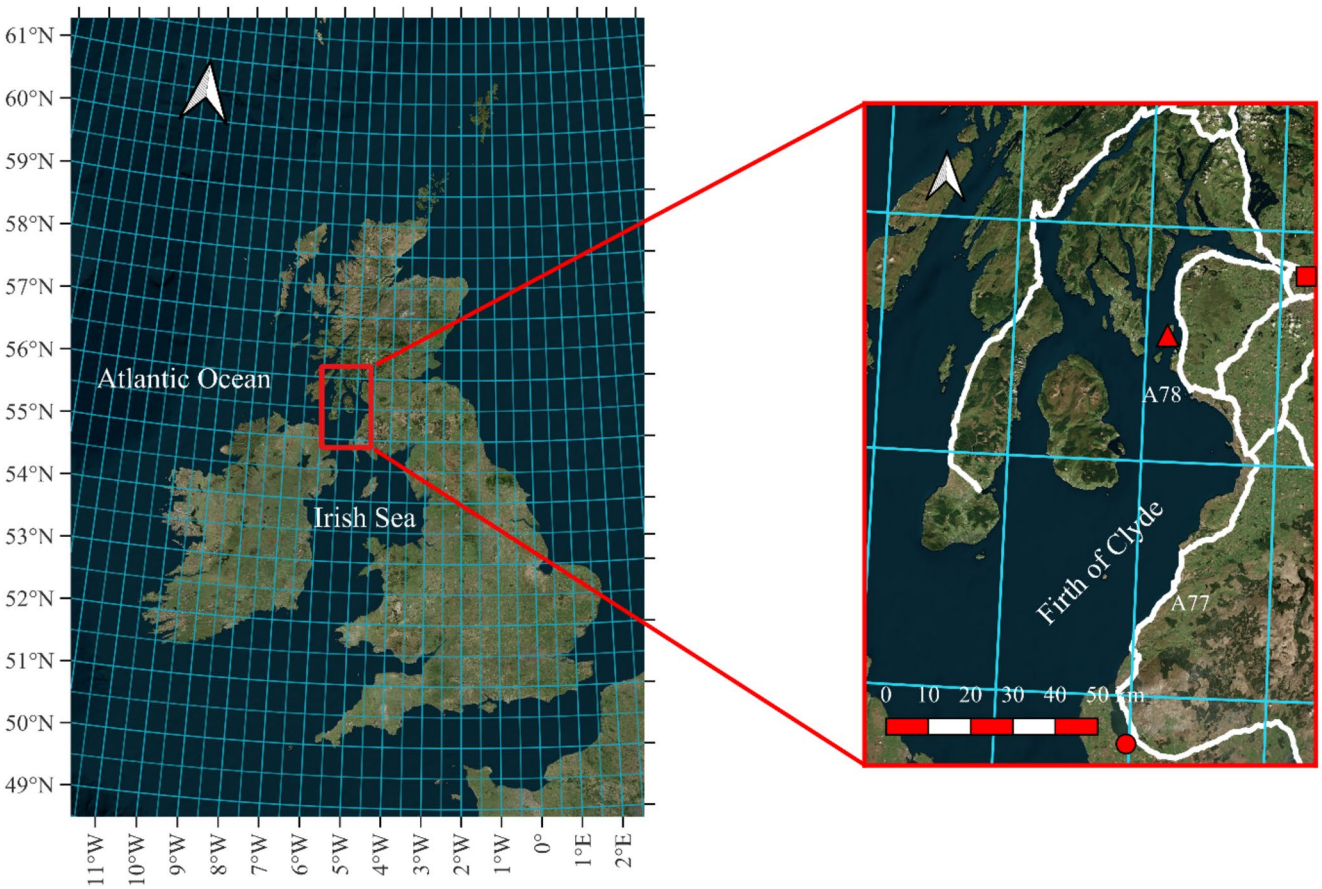

Fig. 4 map of the UK (on the left) and case study location (on the right) of the Firth of Clyde showing Millport (triangle), Stranraer (circle), Glasgow (square) and trunk road network (white). The A77 and A78 link have been specifically highlighted

4.2 Model definition

This study uses a 1000 km footprint around the target location of Millport in the Firth of Clyde as shown in Fig. 6. To reduce operating computational demand, the spatial resolution of the domain reduces with distance to the target location: 400–1000 km has 1 degree resolution, 200–400 km has 0.5 degree resolution and < 200 km has 0.25 degree resolution. Locations that lie east are removed as they do not affect the target location. There are 434 locations over water in the domain. Considering mean sea level pressure, *U10* and *V10* wind speed components, this generates 1302 inputs for every timestep. ERA-5 data was obtained for the years 1980 to 2020.

If 49 timesteps (the previous 24-h + present hour + 24-h forecast at 1 h interval) are used to make each 24-h prediction, this increases the number of inputs to 63,798 (1302 inputs at 49 timesteps). This leads to unmanageable network training times especially since multiple networks needs to be trained for the ABMS network set. To facilitate faster training by reducing the dimensionality of the problem and to de-correlate the input data which prevents overfitting, the inputs are subjected to principal component analysis. The cumulative percentage of variation explained for the principal components is shown in Fig. 7.

Selecting the desired number of principal components is a trade-off between the quantity of preserved input information and the amount of computational time required for training and running the model. 18 principal components that described 95% of the feature variation were used in this case study.

In addition to the 18 principal components, 3 local meteorological variables are added: the difference between the mean sea level pressure at Millport and the maximum pressure in the domain, *U10* and *V10*. The location of these variables is − 5° Latitude, 55.5° Longitude.

Non-tidal residual data was obtained for Millport for the years 1980 to 2023 from the British Oceanographic Data Centre National Tidal and Sea Level Facility (NTSLF) (BODC 1980-2023). For this series, any values missing or deemed questionable by NTSLF and were removed from the analysis.

Hence the 18 principal components, 3 local variables and surge level at Millport combine to make the total number of

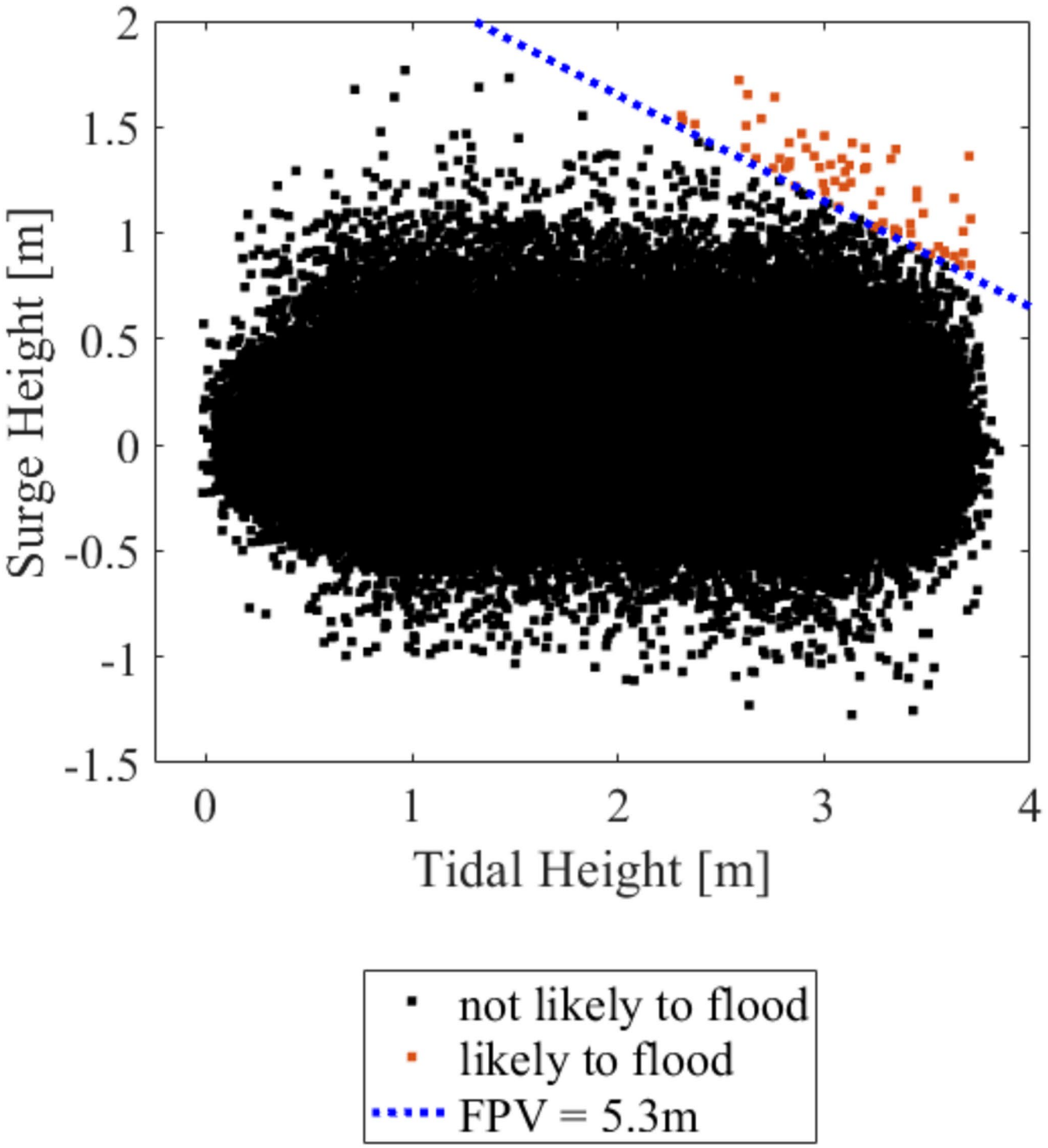

Fig. 5 Plot of tidal height over the Chart Datum against surge height for Millport for 1980 to 2019 showing flood potential value (FPV) threshold and flood likely observations

features 22. The inputs are considered with a 1 h timestep spanning a 49 h window, 24 h either side of the present hour. The output is a 24 h forecast with 1 h resolution. To preserve the input shape for the model, surge values for hours 25–49 are padded with zeros.

240,000 inputs were divided for training, validation and testing with a 60/20/20% split. The split was made chronologically to ensure independence for the groups. These percentage splits do not align precisely with the years spanning 1980 to 2020 owing to gaps in the Millport surge timeseries.

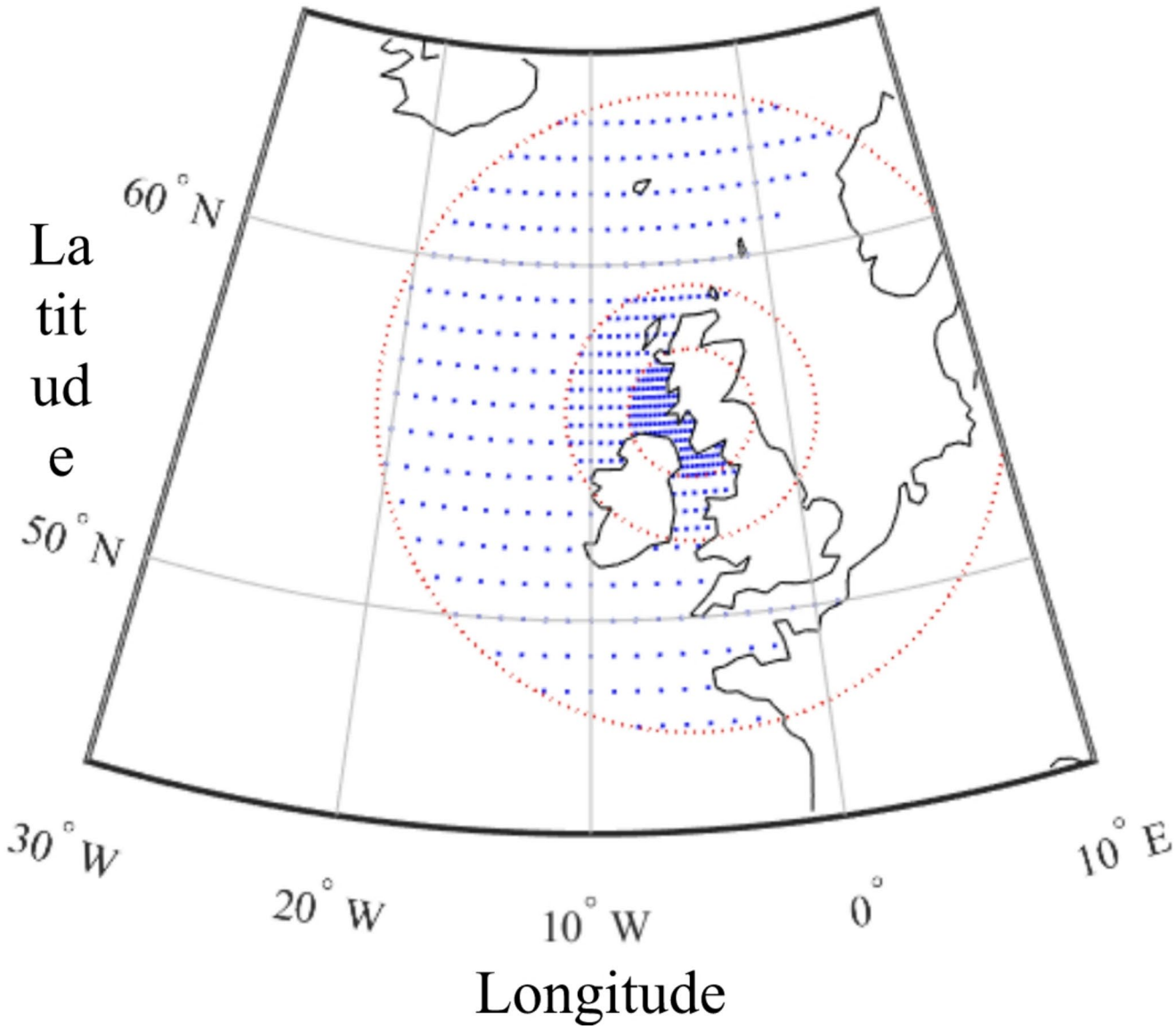

Fig. 6 Map of the UK computational domain centred around the Firth of Clyde, ERA-5 variable locations (blue), and 200 km, 400 km and 1000 km boundary (red)

Due to the low frequency of extreme events and high frequency of non-extreme events, the target distribution is highly imbalanced. To address this imbalance, the 24 h lead time normalised surge heights are binned. 126 unique values are registered when a bin width of 0.1 is used, with counts between 1 (for the most extreme values) and 10,000 (for the most common residual value, ~ 0 m). 1000 values are selected for each unique normalised surge height: When the bin count is greater than 1000, 1000 values are randomly sampled; when the bin count is less than 1000, the values are repeated to bring the bin count to 1000. Balancing the dataset prevents the high frequency, non-threatening surge heights from dominating the training process.

The Adam algorithm was selected for network training given its suitability for problems with large numbers of parameters and sparse data (Kingma and Ba 2014). Additionally, it is easily implemented and computationally efficient. The loss function and metric are both *MAE* (Eq. (22)). Networks were continuously trained with various random seeds until 8 networks, each with an absolute bias of less than 0.1 m for values above 0.75 m, were produced, with the architecture shown in Fig. 3. Approximately 1 in 4 trained models met the requisite bias criteria. Each network was trained with a learning rate of 0.001 and early stopping criteria of 3 epochs was applied to validation loss to prevent overfitting. Each model training epoch took approximately 90 s using a CPU, with models taking roughly 25 epochs

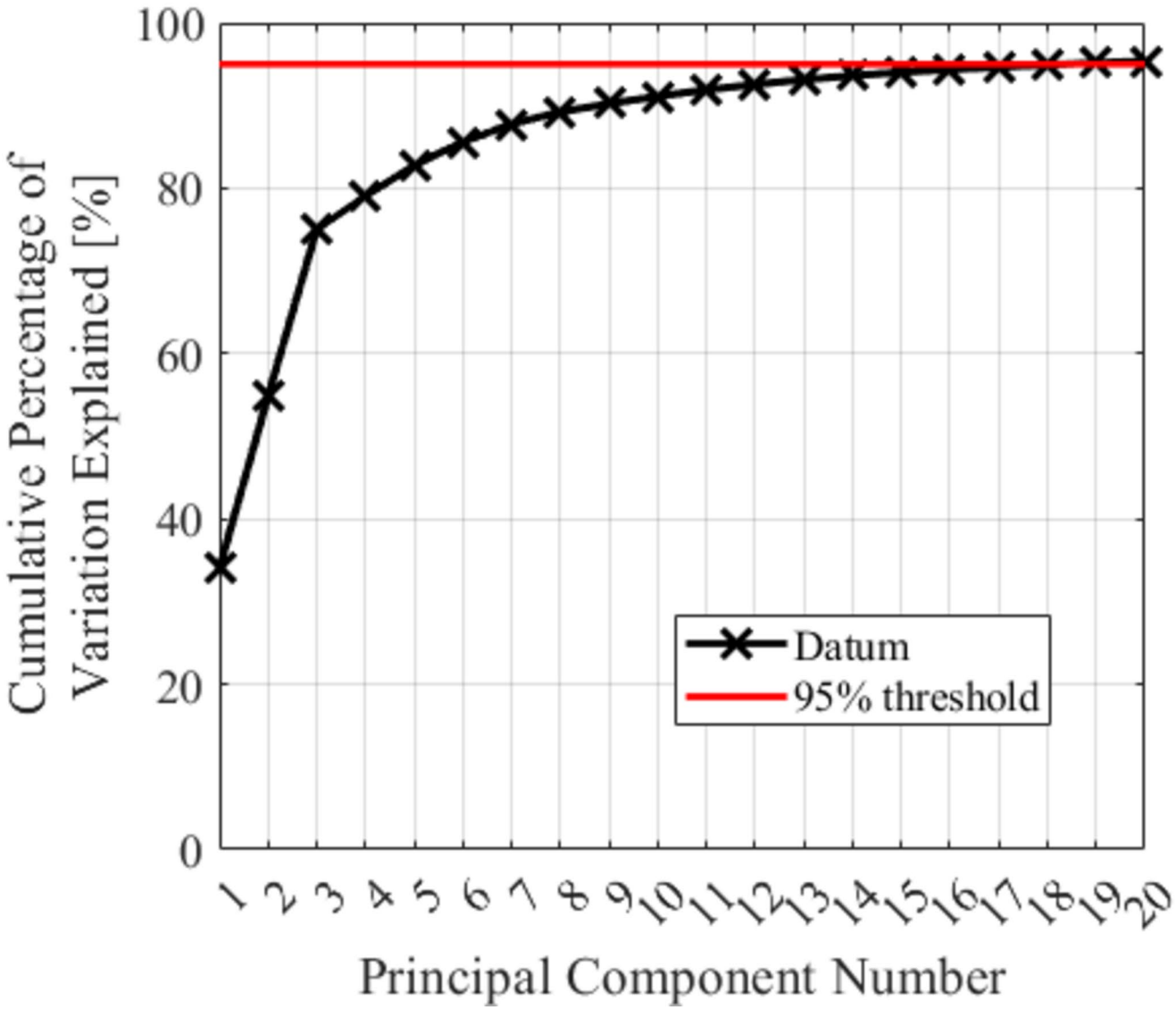

Fig. 7 Plots of PCA component number against the cumulative sum of variation explained by the first principal components

to train before early stopping criteria was met. Individual model's training error were evaluated in 0.25 m bins across the full target range.

The posterior probabilities of the unseen data are inferred from a gaussian mixture model (GMM) fitted to the posteriors of the 40,000 point validation dataset using Eqs. (4)–(5). An example for the 1st network is given in Fig. 8. The surge height targets were divided into 10 equally populated bins to ensure a reasonable minimum population for fitting each mixture model.

4.3 ABMS validation with ERA-5 inputs

A comparison between 40,000 observations from the test dataset and 24-h predictions at Millport has been used to validate the ABMS surge forecast model. Box plots for the test data for each of these trained networks are shown in Fig. 9 for the full range of values and in Fig. 10 for the most extreme values, i.e. surge > 0.75 m. The results establish good agreement for both the full range and the extreme values, demonstrating that the model inputs selected are able to describe the surge level to a good degree.

The benefits of using the ABMS algorithm are apparent as for the full series predictions, the ABMS averaged prediction outperforms every individual network in the set across all metrics (i.e. highest values for R^2, CC, lower values for MAE and SI and Bias values around 0). Similarly, for extreme predictions the ABMS averaged prediction outperforms every individual network in the set across all metrics, except for bias where a single network has a better bias value by 0.02 m but with huge variability. For both the full series and the extreme values, the best individual network

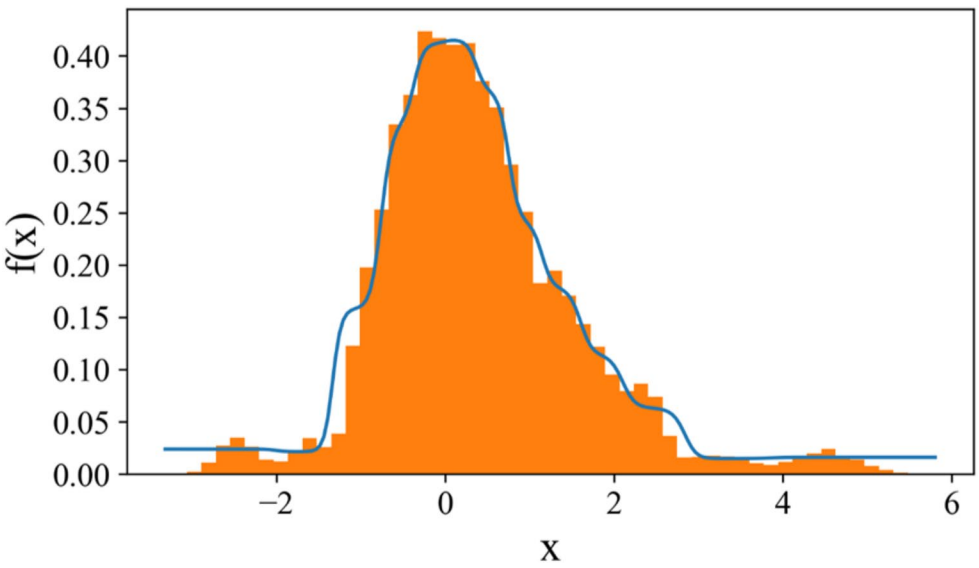

Fig. 8 Gaussian Mixture Model for the prediction from the first neural network. X is surge height normalised by subtracting the mean and dividing by the standard deviation [dimensionless], f(x) is probability density

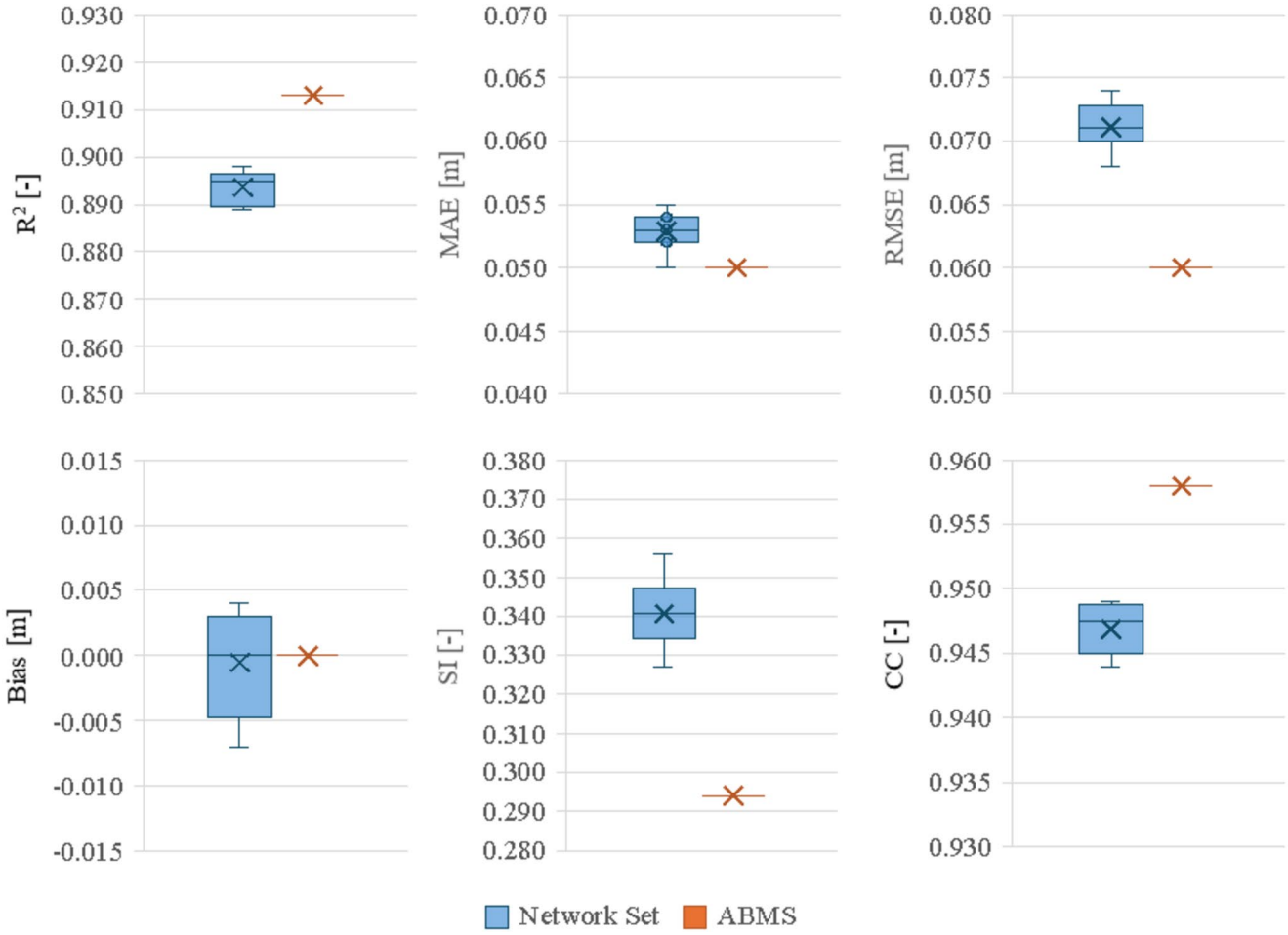

Fig. 9 Box plots of network set performance compared with ABMS result for the full test data set

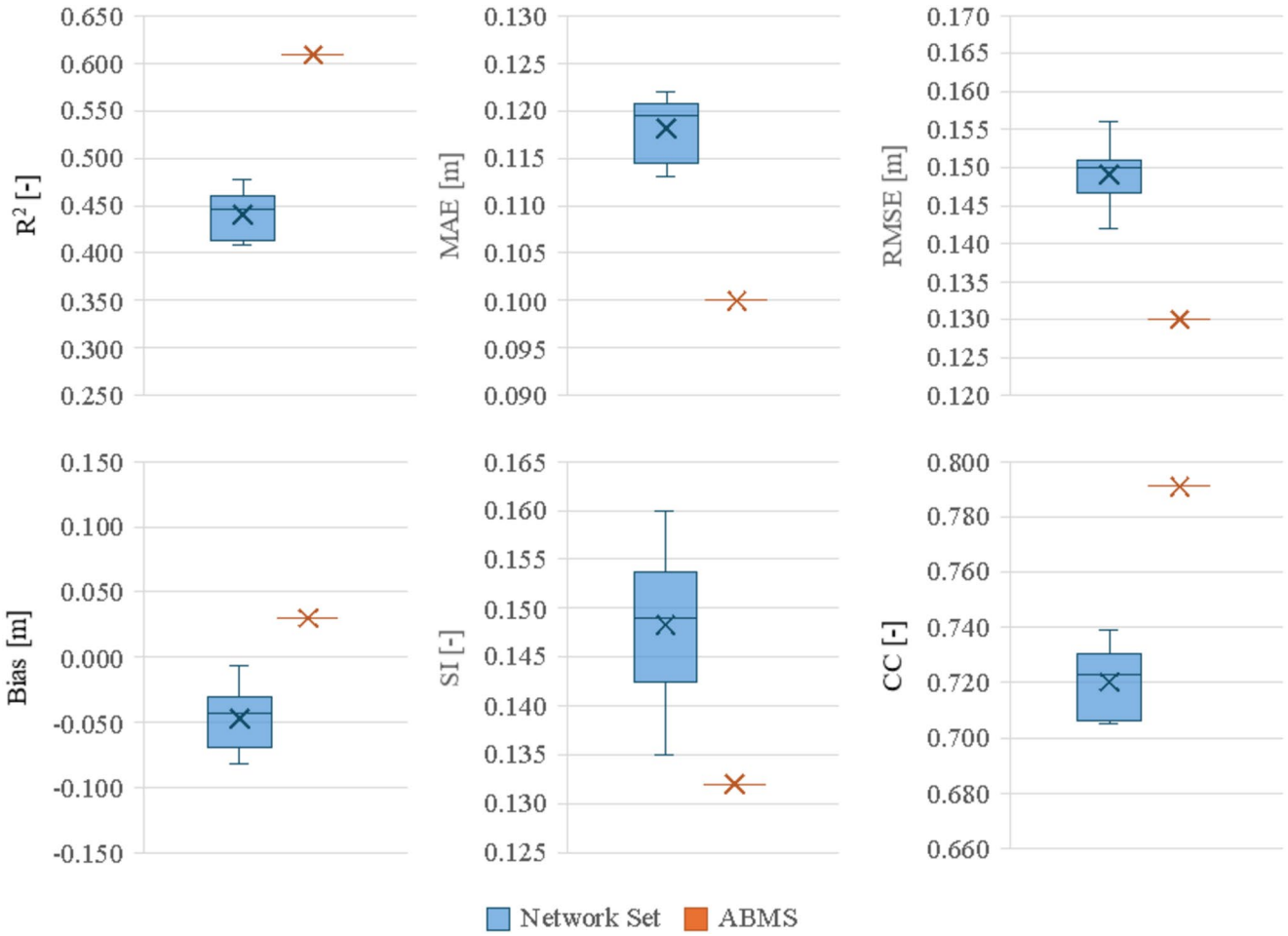

Fig. 10 Box plots of network set performance compared with ABMS result for the extreme surge observations in test data set

Table 1 24 h forecast metric comparison of ABMS averaged prediction and best net prediction within the set

24 hour lead time surge prediction method metric comparison					
Metric	Units	Best network within the set		ABMS	
		Full series	Extreme surge (>0.75 m)	Full series	Extreme surge (>0.75 m)
R^2	–	0.898	0.478	0.913	0.609
MAE	m	0.05	0.11	0.05	0.10
RMSE	m	0.07	0.14	0.06	0.13
BIAS	m	0.00	0.01	0.00	0.03
SI	–	0.327	0.135	0.294	0.132
CC	–	0.949	0.739	0.958	0.791

result for each metric is compared to the ABMS averaged prediction in Table 1.

The increase in R^2 recorded in Table 1 shows that the ABMS method enhances the amount of observed variation explained by the model, compared with any individual ANN model within the set, both for the full data set and for extreme values. Likewise, the increase in CC shows that the

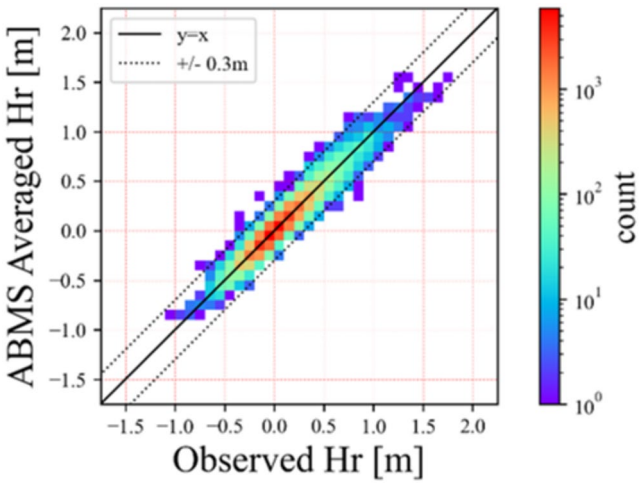

Fig. 11 ABMS Predicted surge height against observed for millport

ABMS surge predictions have a stronger relationship with the observed surge data than any individual model within the set. Figure 11 shows the performance of the 24-h forecast surge height for Millport using the ABMS model, and

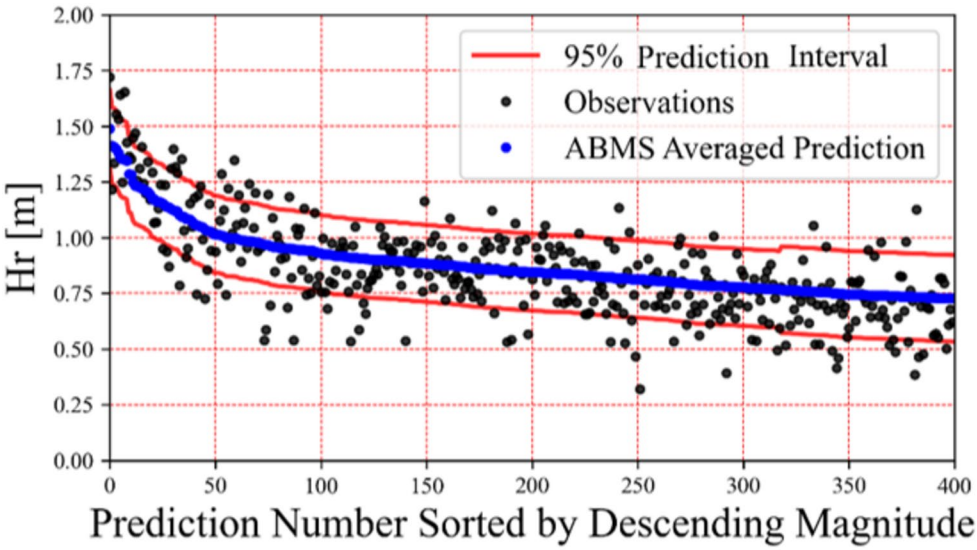

Fig. 12 Single ANN predictions including *feature uncertainty* for 400 largest predictions sorted by descending magnitude

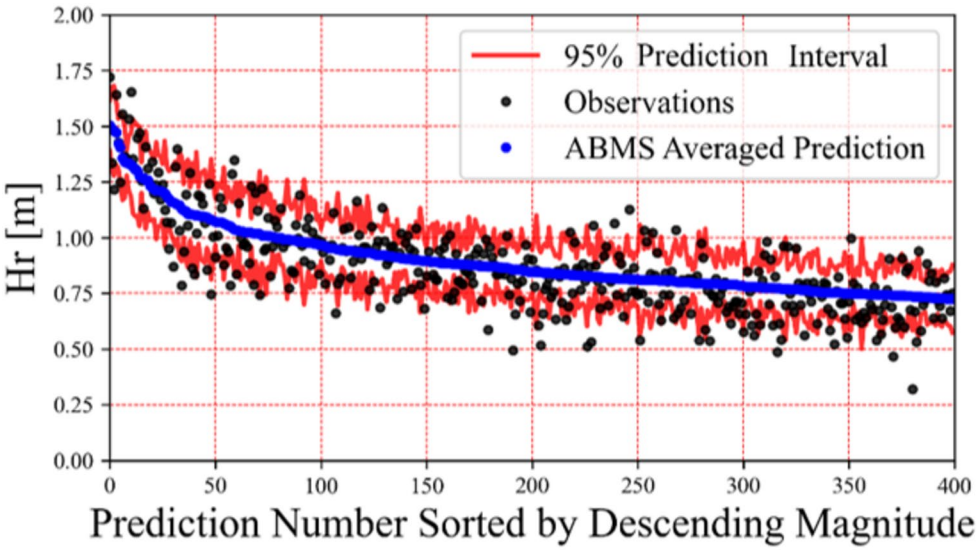

Fig. 13 ABMS predictions excluding *feature uncertainty* applied for 400 largest predictions sorted by descending magnitude

a good agreement between the model predictions and the observed surge height can be seen. In Fig. 11, the high density regions track line of y=x which aligns with the bias metric of 0.00 m for the full range and 0.03 m for extreme values.

To evaluate the contribution of different errors to the prediction interval, three test cases are considered: a single network with its own *feature uncertainty*, ABMS ensemble excluding *feature uncertainty* Eq. (11) and ABMS ensemble including *feature uncertainty* Eq. (14). The robust predictions (expected value with 95% prediction interval) for these three cases are shown in Figs. 12, 13 and 14 respectively.

In each case, the predictions run through the centre of the test data which is reflected in the low bias described in Table 1. The percentage of observations that fall inside each of these prediction intervals for both the full series and the extreme surges is recorded in Table 2. The ABMS with feature uncertainty prediction (Fig. 14) interval is less erratic than the ABMS model omitting feature uncertainty (Fig. 13). A steady prediction interval provides consistent and reliable measures of uncertainty across different estimates of similar magnitudes, indicating that the level of variability in the data or model predictions is relatively constant.

Table 2 explores the effect that different errors have on the size of the prediction interval. The single model case with feature uncertainty falls 2.4% short of the desirable 95% of points for the full series, while falling 18.8% short for extreme surges. The ABMS prediction excluding feature uncertainty falls 30.9% and 26.3% short of the desired 95% for full series and extreme surges. This means that for the network structure described in this paper, *network uncertainty* alone is not sufficient to build a usable prediction interval. The ABMS prediction including feature

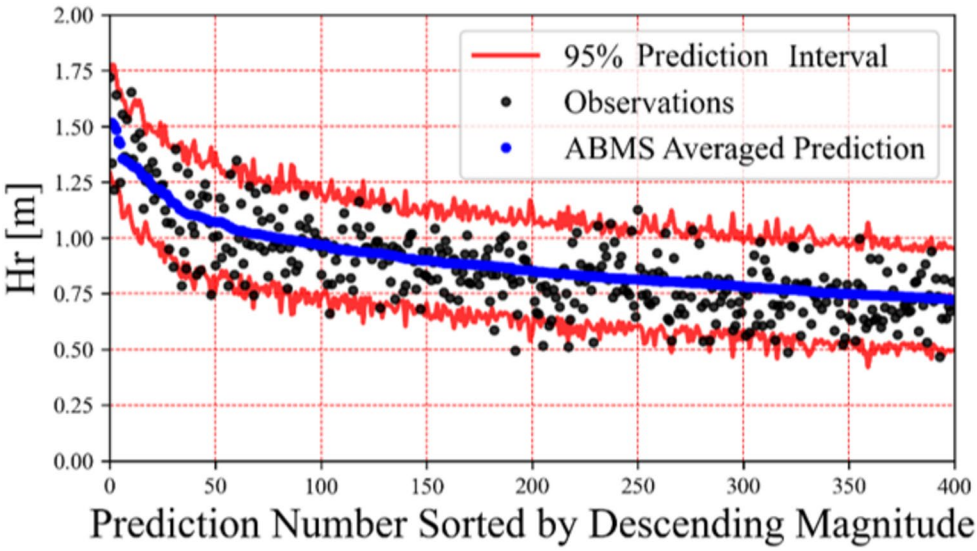

Fig. 14 ABMS predictions including *feature uncertainty* for 400 largest predictions sorted by descending magnitude

Table 2 Prediction interval results and average width for three test cases

24 hour leadtime surge prediction method interval comparison						
Error type	Percentage of values in 95% prediction Interval (%)			Average width of 95% Prediction interval (m)		
	Single Model Inc. feature uncertainty	ABMS Exc. feature uncertainty	ABMS Inc. feature uncertainty	Single Model Inc. feature uncertainty	ABMS Exc. feature uncertainty	ABMS Inc. feature uncertainty
Full series	92.6	64.1	97.3	0.25	0.11	0.28
Extreme surge (>0.75 m)	76.2	68.7	92.9	0.35	0.28	0.48

uncertainty meets the desired 95% of observations captured by the prediction interval and falls 2.1% short for extreme surges. Comparing the ABMS with feature uncertainty to a single network with feature uncertainty, the biggest difference comes for extreme surges.

There are several factors that would contribute to this. Firstly, the selected features and model architecture do not adequately describe the target space giving rise to significant feature uncertainty, and this is more significant for extreme surges, given the increased MAE and RMSE for extreme predictions compared with the full series (Figs. 9 and 10). Extreme surges are inherently rare and thus underrepresented in data sets. This scarcity of data complicates the ability of machine learning algorithms to discern patterns between features and targets. Additionally, extreme surges are influenced by complex atmospheric and oceanographic phenomena that can vary greatly in intensity and behaviour, making them more challenging to model accurately. Moreover, since the models are training on ERA5 data, its ability to describe the target space is constrained by the limitations of ERA5. ERA5 wind speed errors in the Atlantic Ocean tend to increase with wind speed intensity percentiles (Campos et al. 2022). Furthermore, each model's size and structure are likely suboptimal for capturing the intricate patterns that drive surge responses, partly due to the compromise between model size and computational demands and because a fully comprehensive optimisation study was beyond the scope of this investigation.

The improvement in prediction interval for extreme surges vs full series surges, using the ABMS with feature uncertainty method compared to a single model with feature uncertainty, indicates that there is significantly greater variance in predictions across the set for extreme predictions than for non-extreme predictions. This aligns with the challenges of limited data availability for extreme surges and the complexities of machine learning modelling mentioned earlier. It highlights the benefits of using an ensemble of machine learning models that incorporate *network uncertainty* over a single model that does not, particularly in the context of forecasting extreme surges. As discussed earlier, *network uncertainty* by itself does not provide a robust basis for constructing a reliable prediction interval. However, when considered with feature uncertainty, it becomes sufficiently robust to be usable.

The improved prediction interval of the ABMS method when applied with feature uncertainty, alongside the clear improvements in ABMS predictions across multiple metrics demonstrated by Figs. 9, 10 and Table 1: 24 h forecast metric comparison of ABMS averaged prediction and best net prediction within the set., make a very strong case for the

adoption of the ABMS method for surge prediction compared to using a single machine learning model, especially when faced with data constraints. The performance of the ABMS method with feature uncertainty across the full range of test values is shown in Fig. 15.

4.4 Robust surge forecasting system validation with IFS forecasts

To validate the surge prediction in realistic situations, one day IFS weather forecast data was obtained for the period January 2020–December 2023. To use the MCBA algorithm, the *forecast uncertainty* must be quantified and any possible bias between the IFS and ERA-5 removed. Therefore, the data for 2020 was used to quantify the uncertainty associated with the forecast by comparing them against the ERA-5 measurements for the same period. The forecasts for wind speed and direction were converted into *U10* and *V10* components. Thereafter, the wind and pressure inputs were transformed to the same domain as the reanalysis principal components by applying the loading matrix from the PCA process used on the original ERA-5 dataset. These transformed inputs were then normalised. To remove the potential bias, a linear regression model was applied to the principal components of the IFS forecast and its equivalent ERA-5 component using Eqs. (1)–(2). This process is shown for the first principal in Fig. 16.

The unbiased error histograms follow a logistic distribution as shown by the first principal component bias corrected error histogram in Fig. 17, with heavier tails indicating that larger errors are more frequent compared to those in a normal distribution. The magnitude of the forecast error varies with the magnitude of the principal components and increases from the centre to both ends of the feature range. This relationship is demonstrated by the binned logistic fit standard deviations shown in Fig. 18. As such, errors are assessed across 10 evenly distributed bins that span the range of each feature, creating a more versatile error structure. To guarantee that the errors applied within each bin are unbiased, the mean error for each bin is calculated and incorporated into the predictions. 250 Monte Carlo samples are used for each prediction.

Two cases are evaluated here, ABMS with *feature uncertainty* and MCBA with *feature uncertainty*. For clarity, the MCBA method is the development of ABMS that considers *forecast uncertainty* and so a comparison of the two methods will highlight the significance of *forecast uncertainty* to the final prediction interval. IFS forecasts are obtained for years 2021 to 2023. The years selected here lie outside the training, validation, and previous testing period (Table 3). The results of these cases for are recorded in Table 4. The volume of data available for extreme surges is low and the spread of predictions is high relative to the range of values observed. As such the calculated R^2 value for this range is not meaningful and has been omitted from the table.

The metrics recorded in Table 4 show strong agreement between the model predictions and in situ observations. The minimal differences in average predictions between the two methods are expected, given that the main difference between them is the inclusion of unbiased forecast error. The impact of this difference on the prediction interval is demonstrated in Table 4.

The prediction intervals for both methods cover the expected range of data through the full series. Since this holds true for both the ABMS and MCBA methods, it suggests that most of the prediction uncertainty arises from *feature uncertainty*, and *network uncertainty*, with *forecast uncertainty* playing a smaller role. This may be due to the way the models are set up; for example, the simplification inherent in principal component analysis can introduce

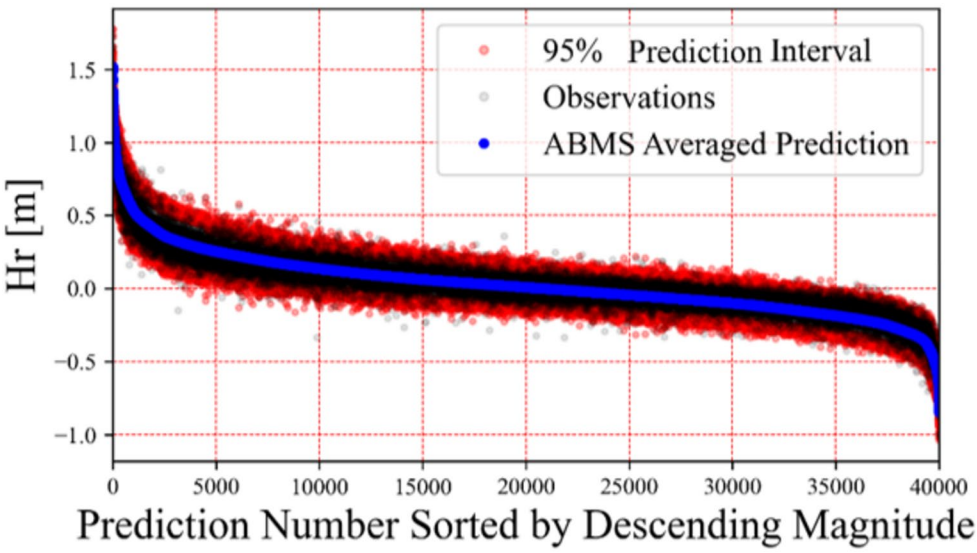

Fig. 15 ABMS predictions including model error applied for all 40,000 test predictions sorted by descending magnitude

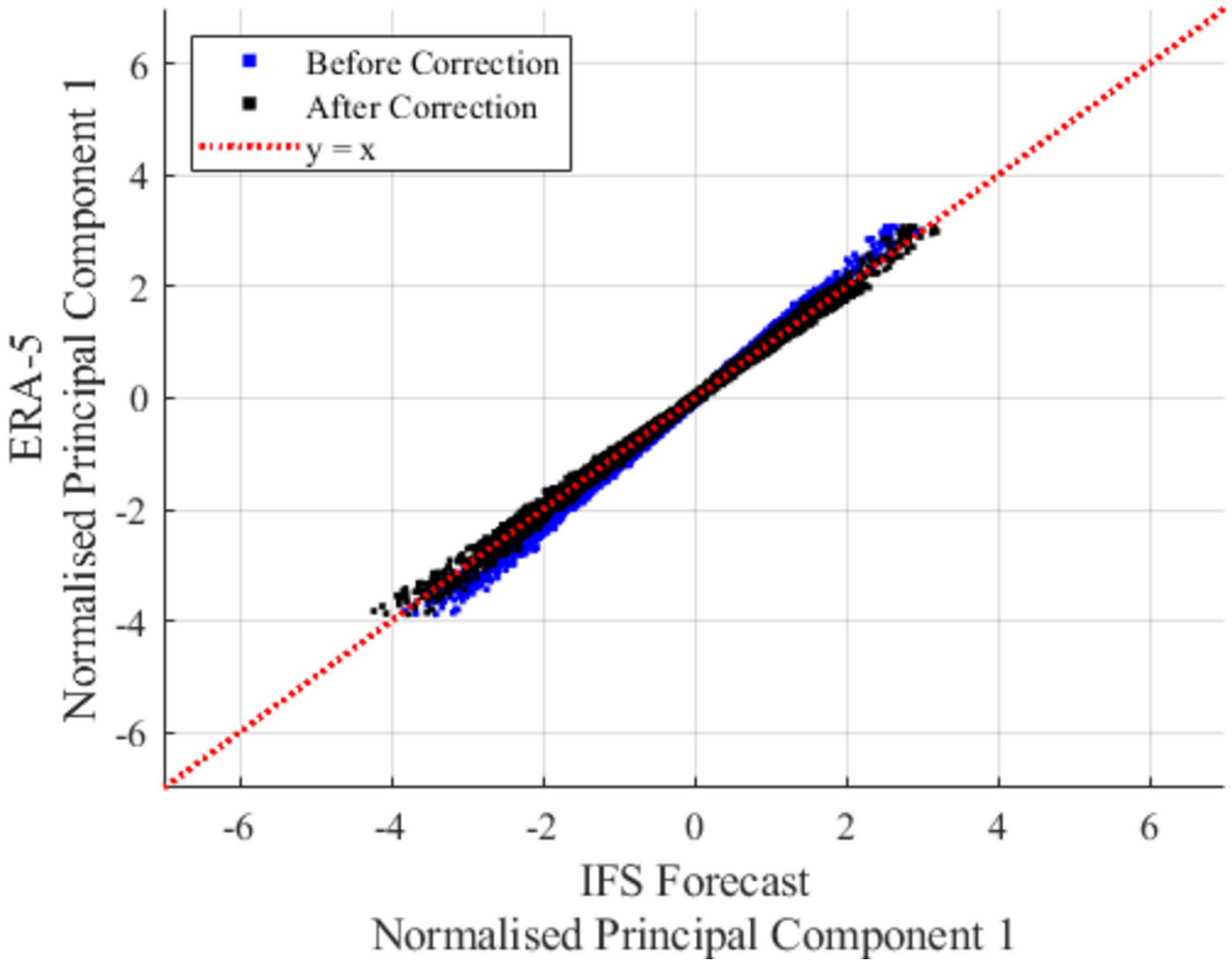

Fig. 16 Bias correction for IFS Variable Principal Component 1 before (blue) and after (black) correction

errors by reducing complex patterns to simpler forms during training. Uncertainty that might latterly be introduced as forecasting error is instead embedded during training, manifesting as *network uncertainty* and potentially explaining variation between models in the set. Additionally, forecasting errors might be less noticeable because they are averaged over a large input domain before pre-processing. However, the ABMS method falls short by 4.5% of the targeted quantity for extreme surge events, whereas the MCBA method is only 0.4% off. Although forecasting error appears to be unnecessary for low-risk surge events, it is significant in constructing satisfactory prediction intervals for extreme surges. The MCBA with feature uncertainty robust predictions for extreme surges are shown in Fig. 19, with the full series shown in Fig. 20.

The MCBA with feature uncertainty has prediction intervals that typically span 48 cm. Figure 19 illustrates that values lying outside the prediction interval generally differ by only a few centimetres. The efficacy of the model ensemble and the prediction interval is deemed satisfactory. This supports the conclusion that the model assumptions are sufficient for characterising surge *forecast uncertainty*.

Figure 21 shows the MCBA with feature uncertainty is able to predict surge height with a good agreement with observed data from Millport with data from 2021 till 2023**.** The high density regions track the identity line ($y=x$) which aligns with the bias metric of − 0.02 m for the full range and 0.03 m for extreme values recorded in Table 4. The CC for this time period is 0.942 with a RMSE of 0.07 m.

For comparison, the NTSLF NEMO tide-surge model provides forecasts ranging from 0 to 6 h, covering the period from 2020 to 2024, with the data organised in monthly files. Data from 2021 to 2023 was specifically analysed for Millport. Monthly values consolidated into a composite 3-year value for RMSE and CC, with each month weighted according to its length. For the 0–6 h forecasts, the physics-based model NTSLF achieves a CC of 0.927 and an RMSE of 0.09 m. Comparatively, the MCBA surge forecasting

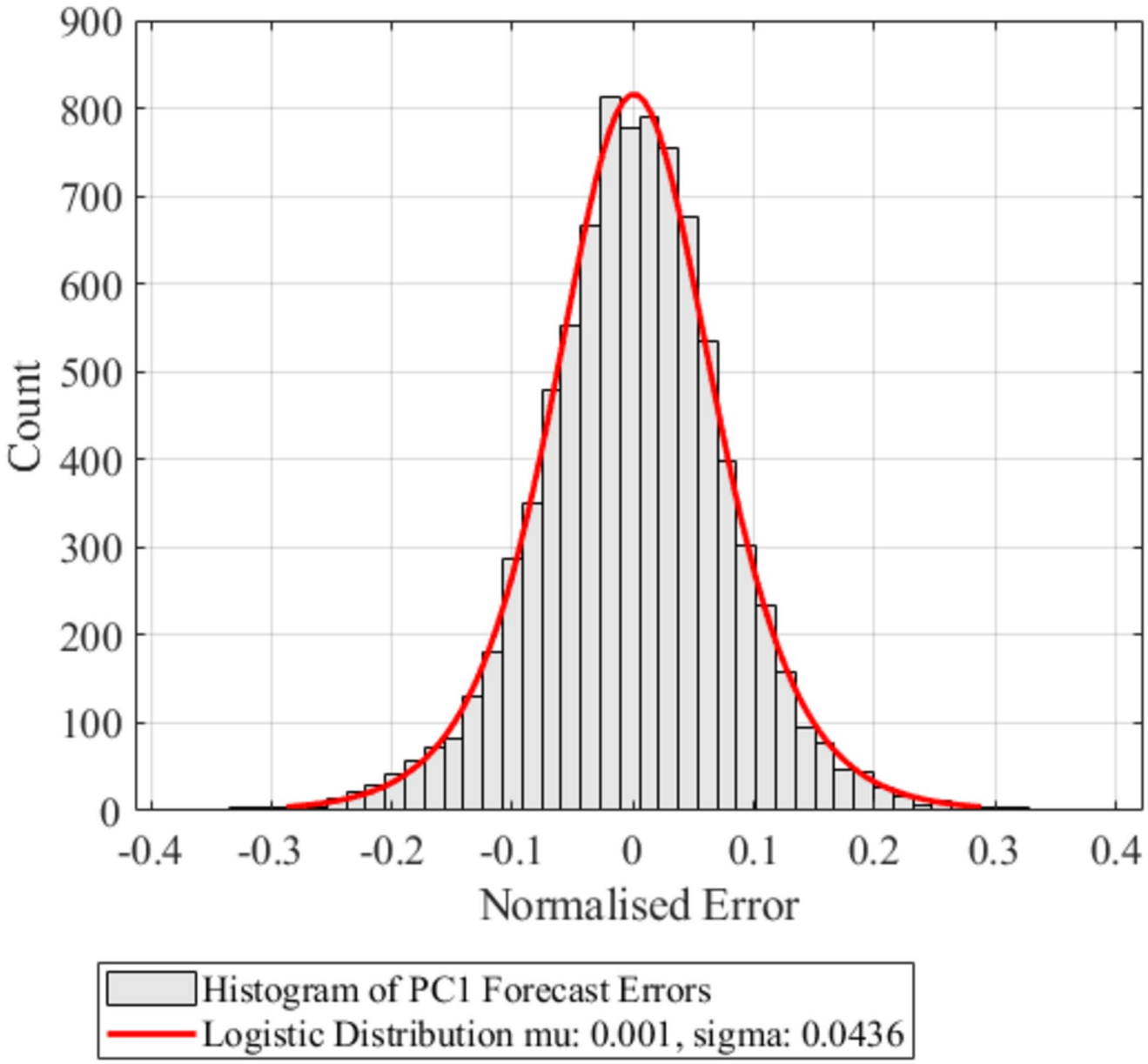

Fig. 17 Forecast error histogram and logistic distribution fit for first principal component

framework shows better CC and lower RMSE for its 24-h lead time forecasts. The MCBA with feature uncertainty method offers a robust prediction interval that takes into account all the uncertainties (from data, model and forecast). It is structured modularly, allowing for straightforward updates to the model as network design, data availability, or forecast accuracy improves. These updates can then be reflected in both the accuracy and the prediction interval of the model.

4.5 Application of the robust surge forecasting system (RSFS)

The robust surge forecasting system uses the MCBA method with feature uncertainty. The operational capability of the model is shown by applying the method to 24-h lead time surge height predictions between the 27th October and 4th November 2022 and between 18 and 25th September 2023 as shown in Figs. 22 and 23, respectively. Shown here are 2 weeks of 24-h forecasts. In practise 24 predictions are made, one for each hour up to and including 24 h. The RSFS expected value (EV) prediction tracks the observed surge height, and the prediction interval is both time periods

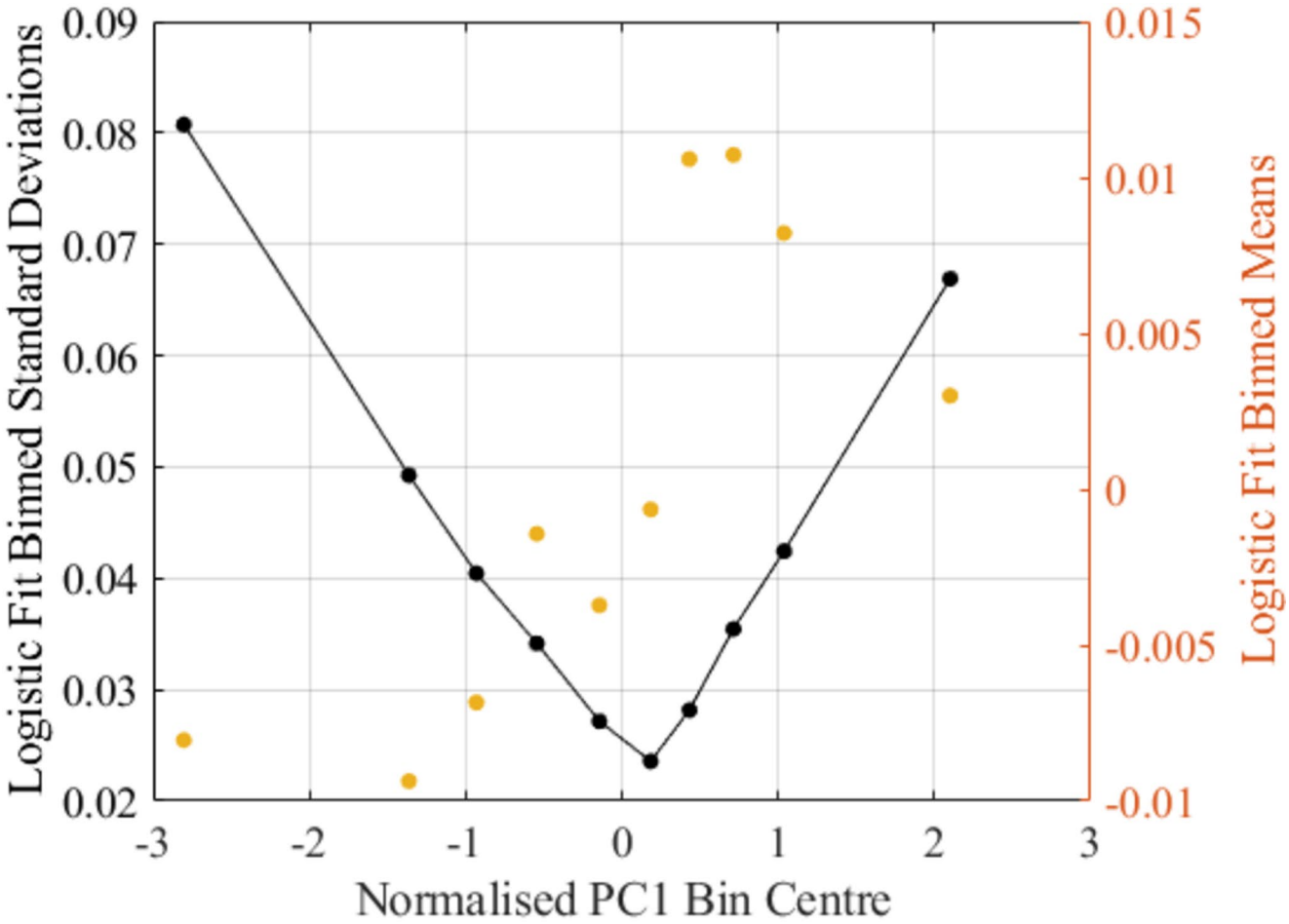

Fig. 18 Logistic fit standard deviations (left y-axis) and means (right y-axis) for 10 equally populated bins along the range of the first principal component

Table 3 24 hour surge prediction comparison for the results 2021–2023 of the MCBA and ABMS algorithm with feature uncertainty

24 hour lead time operational surge prediction method metric comparison					
Metric	Units	ABMS		MCBA	
		Full series	Extreme surge (> 0.75 m)	Full series	Extreme surge (> 0.75 m)
R^2	–	0.882	-	0.878	
MAE	m	0.05	0.10	0.06	0.10
RMSE	m	0.07	0.13	0.07	0.12
BIAS	m	− 0.02	0.04	-0.02	0.03
SI	–	0.332	0.148	0.337	0.146
CC	–	0.943	0.170	0.942	0.173

Table 4 Prediction interval results and average width comparison for ABMS with feature uncertainty and MCBA with feature uncertainty

24 hour lead time operational prediction interval comparison				
Error type	Percentage of values in 95% prediction interval (%)		Average width of 95% prediction interval (m)	
	ABMS Inc. feature uncertainty	MCBA Inc. feature uncertainty	ABMS Inc. feature uncertainty	MCBA Inc. feature uncertainty
Full series	95.6	96.6	0.29	0.30
Extreme surge (> 0.75 m)	90.5	94.6	0.46	0.48

meaningful and usable. The computational time required to obtain one prediction is in the order of seconds and therefore allowing real-time predictions.

4.6 Limitation and potential improvements

While the purpose of the paper was to demonstrate the viability of the MCBA process for surge height predictions,

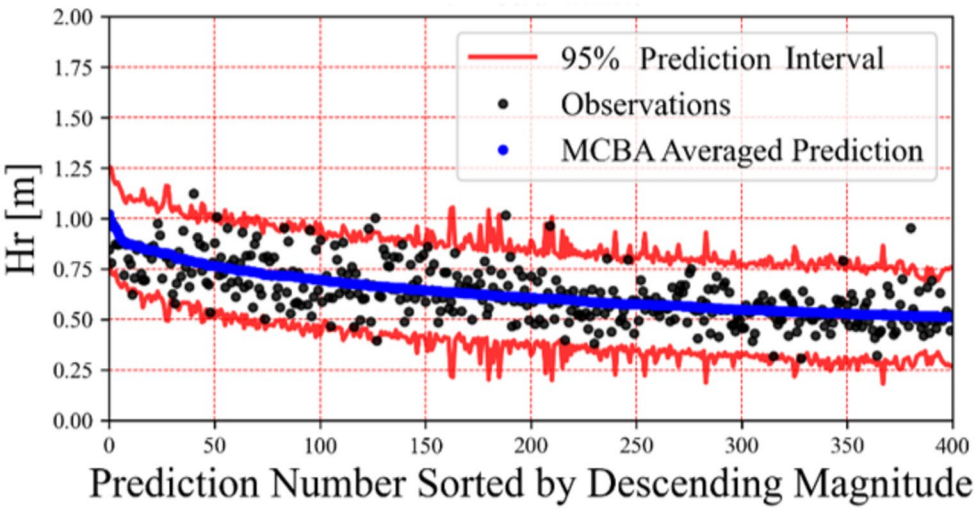

Fig. 19 2021–2023 24 h lead time surge predictions. Predictions made using MCBA algorithm with *feature uncertainty* and IFS forecasts showing 400 largest predictions sorted by descending magnitude

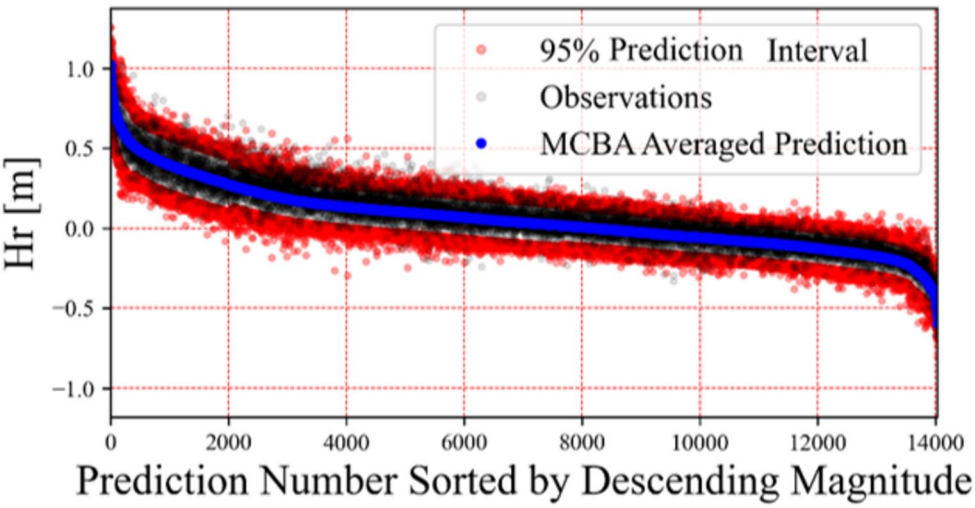

Fig. 20 2021–2023 24 h lead time surge predictions. Predictions made using MCBA algorithm with *feature uncertainty* and IFS forecasts showing all predictions sorted by descending magnitude

several improvements to the model structure can be made. Further analysis on the architecture selection of the ANN and hyper parameter tuning may lead to an improvement in the individual models within the set and an improvement in the MCBA prediction accuracy. Additionally, since the MCBA method can be applied to models with differing architectures that use the same inputs, the unique abilities of different sized architectures to capture specific surge patterns can and should be exploited to improve the averaged prediction. The predicative capability of the surge can be further improved by increasing the input grid beyond the dimensions used in this study. A larger input domain would better capture the atmospheric processes that generate surges. Investigation should be conducted to evaluate the amount of uncertainty that arises from changing the number of selected principal components.

These changes directly impact the computational cost of the analysis during the training of the model, but it might offer the opportunity for a longer time prediction. For instance, the medium-range IFS forecast extends up to two-weeks, and the methodology can readily be extended to accommodate this prediction length albeit with further consideration given to the propagation of uncertainty since an extension of the forecast lead time would strongly depend on the reliability of the atmospheric forcing. An interesting area of further work would be to investigate the spatial and temporal dependencies of meteorological forecast errors and assess their impact on surge predictions. In addition, specific storm characteristic inputs such as radius of maximum wind speed or land fall location can be easily incorporated into the model framework to improve the prediction capability of the model, or the proposed methodology can be applied to existing neural network-based models or ensemble models that have uncertain inputs.

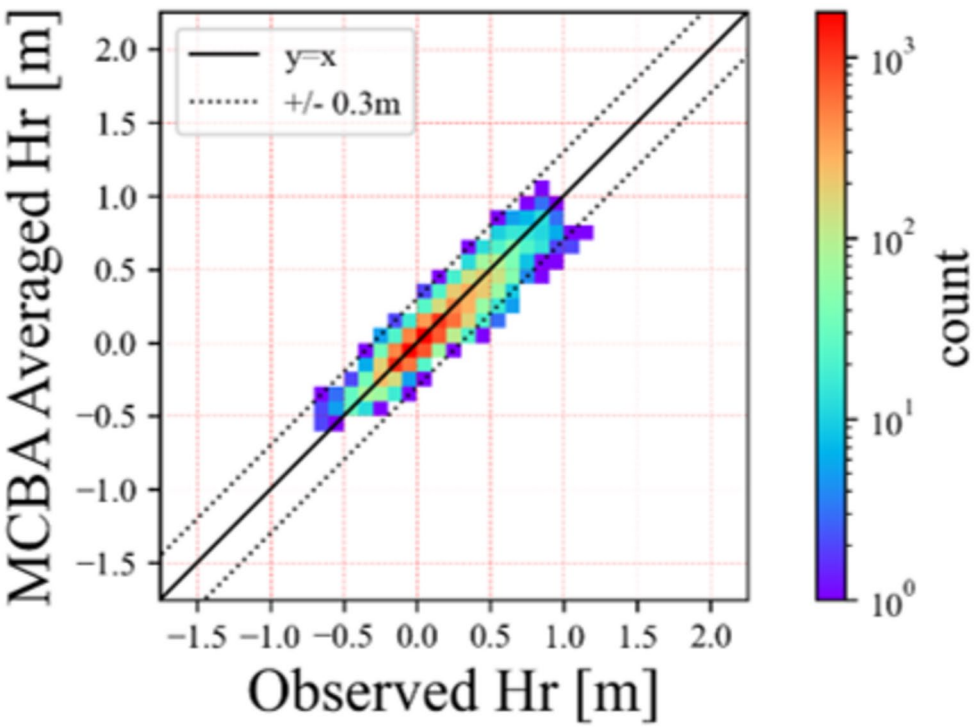

Fig. 21 Validation of the MCBA with feature uncertainty: Millport for 2021–2023

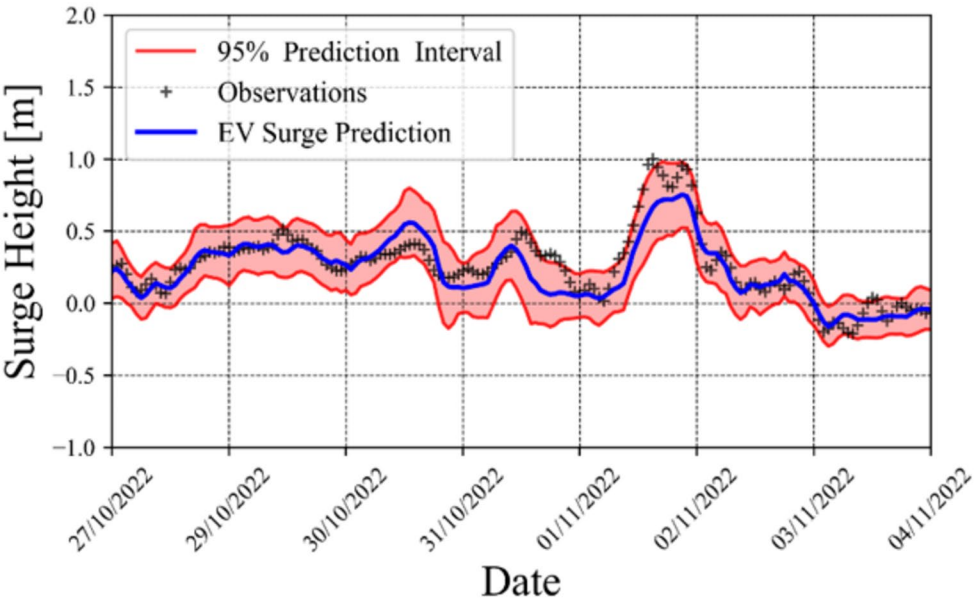

Fig. 22 24 hour RSFS 24-h predictions for the week spanning 27/10/2022 to 04/11/2022

5 Conclusions

This paper has illustrated the development and validation of a robust 24-h surge forecasting machine learning based model. This builds and improves on current methodologies. More specifically, Adaptive Bayesian Model Selection approach has been used to improve on predictions than can be achieved with a single neural network whilst providing a usable and reliable prediction interval that considers both model uncertainty and attribute uncertainty. Thereafter a Monte Carlo procedure has been integrated to include the effect of *forecast uncertainty* on the surge prediction interval. This study shows that for 24-h lead times, model, attribute and forecast uncertainties are crucial for accurately characterising surge prediction intervals, particularly for extreme surges.

The robust surge forecasting system has been successfully deployed and validated at Millport within the Clyde basin in Scotland. Operational performance has been shown through model application using weather forecast

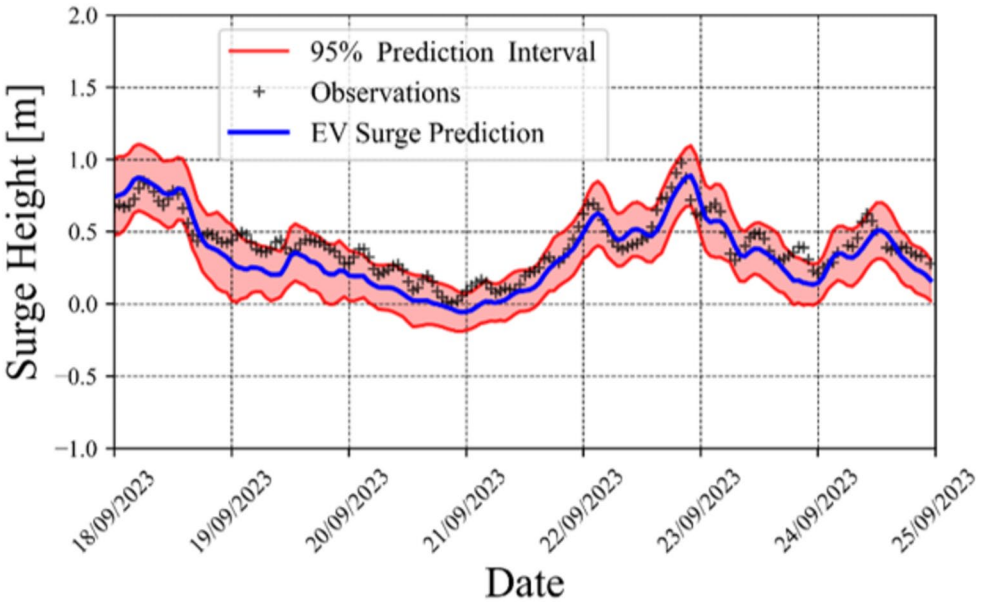

Fig. 23 24 hour RSFS 24-h predictions for the week spanning 18/09/2023 to 25/09/2023

provided by ECMWF's IFS, demonstrating that the proposed approach is able to perform robust surge predictions with an error boundary approximately of 48 cm wide which is both meaningful and usable. The methodology presented offers a straightforward and effective way to construct and train a robust storm surge model. Around the UK, the National Tidal and Sea Level Facility (NTSLF) provides a storm surge model with good spatial coverage. However, this model does not capture all the uncertainties that this approach addresses due to the computational burden of running physics-based models. Given that ERA-5 and IFS are global datasets, this methodology can, in principle, be applied to any location using historical surge height time-series or surge height reanalysis data series. While most of the coastline is not covered by tide gauges, reanalysis data sets provide a valuable alternative. This framework can be readily modified to accommodate the additional uncertainty introduced by using surge reanalysis datasets.

Acknowledgements This study uses data from The National Tidal and Sea Level Facility, provided by the British Oceanographic Data Centre; the work is supported by the Engineering and Physical Science Research Council [grant number EP/R513349/1]; results were obtained using the ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk) based at the University of Strathclyde; IFS Forecast Data was downloaded from Visual Crossing, Visual Crossing Weather, URL: https://www.visualcrossing.com/ accessed Jan 2024.

Author contributions E.M. wrote the main manuscript text, prepared all the figures, performed all the analysis, and developed the algorithms and tools. E.T. and E.P. provided funds, access to previous algorithms and supervised the research. All authors reviewed the manuscript.

Data availability All preprocessing tools along with the model algorithms are available in the Bayesian Coastal Forecasting git repository: https://github.com/emacd-domain/Bayesian_Coastal_Forecasting.

Declarations

Competing interests The authors declare no competing interests.

References

Al Kajbaf A, Bensi M (2020) Application of surrogate models in estimation of storm surge: a comparative assessment. Appl Soft Comput 91:106184

Asadi S, Shahrabi J, Abbaszadeh P, Tabanmehr S (2013) A new hybrid artificial neural networks for rainfall–runoff process modeling. Neurocomputing 121:470–480

Blöschl G, Bierkens MF, Chambel A, Cudennec C, Destouni G, Fiori A, Kirchner JW et al (2019) Twenty-three unsolved problems in hydrology (UPH)—a community perspective. Hydrol Sci J 64(10):1141–1158

BODC, British Oceanographic Data Centre. 1980-2023. Prod. National Oceanography Centre. Accessed Jan 2023. https://www.bodc.ac.uk/data/hosted_data_systems/sea_level/uk_tide_gauge_network/

C3S, Copernicus Climate Change Service (2017) *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.* Accessed Dec 2022. https://cds.climate.copernicus.eu/cdsapp#!/home.

(C3S), Copernicus Climate Change Service (2017) ERA5: fifth generations of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS).

Accessed July 2021. https://cds.climate.copernicus.eu/cdsapp#!/home.
Campos R et al (2022) Assessment and calibration of ERA5 severe winds in the atlantic ocean using satellite data. Remote Sens. https://doi.org/10.3390/rs14194918
Chelton DB, Schlax MG (1996) Global observations of oceanic Rossby waves. Science 272:234–238
Diks CG, Vrugt JA (2010) Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch Env Res Risk Assess 24(6):809–820
ECMWF, European Centre for Medium-Range Weather Forecasts (2024) Integrated forecasting system (IFS). Reading
Emanuel K (2017) Assessing the present and future probability of Hurricane Harvey's rainfall. Proc Natl Acad Sci 114(48):12681–12684
Fazel SAA, Blumenstein M, Mirfenderesk H, Tomlinson R (2014) Estuarine flood modelling using artificial neural networks. In: 2014 international joint conference on neural networks. IEEE, pp 631–637
French J, Mawdsley R, Fujiyama T, Achuthan K (2017) Combining machine learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports. Procedia IUTAM 25:28–35
IPCC, Intergovernmental Panel on Climate Change (2022) Special report on the ocean and cryosphere in a changing climate
Jackson EK, Roberts W, Nelsen B, Williams GP, Nelson EJ, Ames DP (2019) Introductory overview: error metrics for hydrologic modelling–A review of common practices and an open source library to facilitate use and adoption. Environ Model Softw 119:32–48
Kim S, Pan S, Mase H (2019) Artificial neural network-based storm surge forecast model: practical application to Sakai Minato, Japan. Appl Ocean Res 91:101871
Kingma DP, Ba J (2014) Adam: a method for stochastic optimization
Kingston GB, Maier HR, Lambert MF (2008) Bayesian model selection applied to artificial neural networks used for water resources modeling. Water Resour Res. https://doi.org/10.1029/2007WR006155
Kleinbaum DG, Klein M (2010) Maximum likelihood techniques: an overview. In: Logistic regression, pp 103–127. Springer
Knutson T (2010) Tropical cyclones and climate change. Nat Geosci 3:157–163
Kohno N, Dube SK, Entel M, Fakhruddin SHM, Greenslade D, Leroux MD, Rhome J, Thuy NB (2018) Recent progress in storm surge forecasting. Trop Cyclone Res Rev 7(2):128–139
Labach A, Salehinejad H, Valaee S (2019) Survey of dropout methods for deep neural networks. http://arxiv.org/abs/1904.13310
Le X, Ho H, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. Water 11(7):1387
Li J, Nie B (2017) Storm surge prediction: present status and future challenges. Procedia IUTAM 25:3–9. https://doi.org/10.1016/j.piutam.2017.09.002
Macdonald E (2024). Bayesian coastal forecasting toolbox. https://github.com/emacd-domain/Bayesian_Coastal_Forecasting
Mel R, Lionello P (2016) Probabilistic dressing of a storm surge prediction in the Adriatic Sea
Mel R, Viero DP, Carniello L, Defina A, D'Alpaos L (2014) Simplified methods for real-time prediction of storm surge uncertainty: the city of Venice case study. Adv Water Resour 71:177–185
Mentaschi L, Besio G, Cassola F, Mazzino A (2013) Problems in RMSE-based wave model validations. Ocean Model 72:53–58
Milne F, Winter M, Reeves S, Knappett J, Dawson S, Dawson A, Peeling D, Peeling J, Brown M (2017) Assessing the risks to infrastructure from coastal storms in a changing climate: Project Report PPR800. Transport Research Laboratory, Wokingham
Milne F, Winter MG, Reeves SJ, Knappett J, Dawson S, Dawson AG, Peeling D, Peeling J, Brown M (2017) Assessing the risks to infrastructure from coastal storms in a changing climate: project Report PPR800. Transport Research Laboratory
Moges E, Demissie Y, Larsen L, Yassin F (2021) Sources of hydrological model uncertainties and advances in their analysis. Water 13(1):28
Mohamad-Saleh J, Hoyle BS (2008) Improved neural network performance using principal component analysis on Matlab. Int J Comput Internet Manag 16(2):1–8
Nicholls RJ, Hinkel J, Lincke D, Suckall N, Tol RS (2018) Integrated assessment of global environmental change with a focus on rivers and coasts. In: Handbook of global environmental politics, pp 307–322. Edward Elgar Publishing
NOAA, National Oceanographic and Atmospheric Administration (2020) NOAA artificial intelligence strategy—analytivd for next-generation earth science
NTSLF, National Tidal and Sea Level Facility (2019) Tide-surge model. Liverpool: National Oceanography Centre. https://ntslf.org/storm-surges/surge-model
Oparanji U, Sheu RJ, Bankhead M, Austin J, Patelli E (2017) Robust artificial neural network for reliability and sensitivity analyses of complex non-linear systems. Neural Netw 96:80–90
Pullen T, Liu Y, Otinar Morillas P, Wyncoll D, Malde S, Gouldby B (2018) A generic and practical wave overtopping model that includes uncertainty. In: Proceedings of the institution of civil engineers-maritime engineering, vol 171, pp 109–120. Thomas Telford Ltd.
Qin Y, Su C, Chu D, Zhang J, Song J (2023) A review of application of machine learning in storm surge problems. J Mar Sci Eng 11(9):1729
Resio TD, Powell NJ, Cialone MA, Das HS, Westerink JJ (2017) Quantifying impacts of forecast uncertainties on predicted storm surges. Nat Hazards 88(3):1423–1449
Reusch DB, Alley R, Hewitson BC (2005) Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. Polar Geogr 29(3):188–212
Sabatino A, Murray R, Hills A, Speirs D, Heath M (2016) Modelling sea level surges in the Firth of Clyde, a fjordic embayment in south-west Scotland. Nat Hazards 84:1601–1623
Salighehdar A, Ye Z, Liu M, Ionut F, Blumberg AF (2017) Ensemble-based storm surge forecasting models. Weather Forecast 32(5):1921–1936
Salman AG, Heryadi Y, Abdurahman E, Suparta W (2018) Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. Procedia Comput Sci 135:89–98
SEPA, Scottish Environment Protection Agency (2015) Flood modelling guidance for responsible authorities Version 1.1. 59
Sztobryn M (2003) Forecast of storm surge by means of artificial neural network. J Sea Res 49(3):317–322
Taflanidis AA, Jia G, Kennedy AB, Smith JM (2013) Implementation/optimization of moving least squares response surfaces for approximation of hurricane/storm surge and wave responses. Nat Hazards 66(2):955–983
Taylor NR, Irish JL, Udoh IE, Bilskie MV, Hagen SC (2015) Development and uncertainty quantification of hurricane surge response functions for hazard assessment in coastal bays. Nat Hazards 77:1103–1123
Tiggeloven T, Couasnon A, van Straaten C, Muis S, Ward PJ (2021) Exploring deep learning capabilities for surge predictions in coastal areas. Sci Rep 1:1–15
Tolo S, Tian X, Bausch N, Becerra V, Santosh TV, Vinod G, Patelli E (2018) Robust on-line diagnosis tool for the early accident detection in nuclear power plants. Reliab Eng Syst Saf 186:110–119
Visual Crossing (2023) Visual crossing weather. https://www.visualcrossing.com/

Wang B (2002) Kelvin waves. In: Encyclopedia of atmospheric sciences, vol 1062

Williams J, Horsburgh KJ, Williams JA, Proctor RN (2016) Tide and skew surge independence: new insights for flood risk. Geophys Res Lett 43(12):6410–6417

WMO, World Meteorological Organization (2021) Future of weather and climate forecasting—WMO open consultative platform white paper #1. Public-Private Engagement Publication No. 3

Yan S (2016) Understanding LSTM and its diagrams. Accessed Dec 2024. https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714

Yu T, Hong Z (2020) Hyper-parameter optimization: a review of algorithms and applications