Environmental Modelling and Software 142 (2021) 105066

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Using machine learning to derive spatial wave data: A case study for a marine energy site

Jiaxin Chen [a,*], Ajit C. Pillai [a], Lars Johanning [a,b], Ian Ashton [a]

[a] College of Engineering, Mathematics and Physical Sciences, Exeter University, Cornwall, TR10 9FE, UK
[b] College of Shipbuilding Engineering, Harbin Engineering University, Harbin, 150001, China

ARTICLE INFO

Keywords:
Nearshore wave modelling
Random forest
Machine learning
Spatial prediction
Optimal gridding

ABSTRACT

Ocean waves are widely estimated using physics-based computational models, which predict how energy is transferred from the wind, dissipated, and transferred spatially across the ocean. Machine learning methods offer an opportunity to predict these data with significantly reduced data input and computational power. This paper describes a novel surrogate model developed using the random forest method, which replicates the spatial nearshore wave data estimated by a Simulating WAves Nearshore (SWAN) numerical model. By incorporating in-situ buoy observations, outputs were found to match observations at a test location more closely than the corresponding SWAN model. Furthermore, the required computational time reduced by a factor of 100. This methodology can provide accurate spatial wave data in situations where computational power and transmission are limited, such as autonomous marine vehicles or during coastal and offshore operations in remote areas. This represents a significant supplementary service to existing physics-based wave models.

1. Introduction

Met-Ocean data play a significant role in the design and operation of offshore and coastal infrastructure. Wave conditions impact ship navigation and fuel-efficient operation (James, 1957; MEPC, 2012). In particular, the sea state is a key factor that determines vessel design and operational management strategies for autonomous marine systems (Johnston and Poole, 2017). For marine renewable energy, offshore oil and gas, and offshore aquaculture, wave conditions influence activities across the full life-cycle of the infrastructure. Cyclic wave loads impact fatigue, reliability, and performance of systems (DNV, 2014), whilst continuous wave data are key to determining the "weather windows" which govern the accessibility of renewable energy devices (Ardente et al., 2008; Balog et al., 2016; Gentry et al., 2017; Reikard et al., 2017).

Virtually all forecasts and characterisations of wave conditions are currently based on deriving time-series of spatial wave conditions using phase-resolving physics-based, computational models. A series of 3rd generation wave models such as WAM (WAve Modelling) (Günther et al., 1992; Komen et al., 1996), WAVEWATCH-III (Tolman, 2009; Tolman et al., 2002), and Simulating Waves Nearshore (SWAN) (Booij et al., 1999; Ris et al., 1999) have become universal numerical methods. These models determine wave conditions based on the energy-balance equations, considering energy input from surface winds with processes dissipating wave energy. By incorporating the propagation of waves across the model domain and modelling interaction with the bathymetry, spatial wave data sets are created. These models are widely used, providing past wave climates and wave forecasts across the world (Berrisford et al., 2011; Chawla et al., 2012; Service (C3S), 2017). The spatial resolution of these global datasets range from $0.28 \times 0.28^\circ$ (about 30 km) to $1 \times 1^\circ$ (about 111 km). SWAN was designed as a tool for coastal modelling, focusing more on wave propagation in shallow water (Booij et al., 1999). It was designed for application in coastal regions around the world and has also been widely used to quantify wave conditions for offshore renewable energy sites (e.g. Ashton et al., 2014; Liang et al., 2014; Wu et al., 2020).

Physics-based models are commonly validated and calibrated with in-situ measurements or remote-sensing data. Presently, global-scale modelling assimilates satellite-based remote sensing data e.g. the Global Data Assimilation System (GDAS) system (NOAA, 2020). For nearshore areas, waves observed by in-situ buoy measurements have been used for validation of physics-based models, including the data used in this study (Van-Nieuwkoop et al., 2013).

Combining measured time-series of wave data with physics-based models, offers possibilities for deriving spatio-temporal wave data. In

* Corresponding author.
E-mail address: jc1083@exeter.ac.uk (J. Chen).

https://doi.org/10.1016/j.envsoft.2021.105066
Accepted 20 April 2021
Available online 5 May 2021
1364-8152/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the past decade, machine learning methods have demonstrated their accuracy in predicting various environmental variables. Research has explored forecasting of wave energy flux forecasts based on in-situ measurement by machine learning algorithms and have achieved similar accuracy to physics-based models in different forecast horizons. Specifically, Sánchez et al. (2018) used an artificial neural network (ANN) to estimate the wave height at a buoy station with a mean absolute percentage error (MAPE) of 5.27%, while Pirhooshyaran and Snyder (2020) used long short-term memory (LSTM) and sequence-to-sequence networks to forecast significant wave height (H_s) and power at multiple buoy stations. Their proposed networks can predict H_s with MAPE of 18.2% which outperformed alternate networks and a random forest (RF) method, a machine learning alternative.

The spatial correlations of environmental variables can be captured by machine learning methods. Oh and Suh (2018) proposed a hybrid model combining empirical orthogonal function (EOF) analysis and wavelet analysis with neural network (EOFWNN) that can forecast wave heights for the following 24 h at multiple locations with values of normalized root mean squared error (NRMSE) between 15.5% and 26.3%. Li et al. (2011) compared the application of 23 methods, including RF, to the spatial interpolation of environmental variables. Their work confirmed both the effectiveness and sensitivity of RF to predict spatial patterns. This suggests that it is an ideal candidate for application to ocean wave data.

Some research has attempted to make "grey-box" models combining a numerical model with a data-driven approach (Ibarra-Berastegi et al., 2015; Serras et al., 2019). These systems take output from a physical model (e.g. European Centre for Medium-Range Weather Forecasts (ECMWF) and National Centres for Environmental Prediction (NCEP)) as features in a machine learning model. Nencioli and Quartly (2019) proposed a synergistic method to combine satellite and in-situ observations to map an area of wave parameters, validated by a global numerical wave model. Ibarra-Berastegi et al. (2015) applied RF with a physics-based model (WAM), to issue short-term forecasts of wave energy flux from 1h to 24 h at five buoys, with mean absolute log-differences of less than 20%–60%. Serras et al. (2019) has also combined RFs with physics-based data from ECMWF to forecast wave energy flux at the Mutriku Wave Farm up to 24-h ahead with 60% MAPE.

Considering the computational requirements for coastal models such as SWAN, surrogate models can reduce the necessary computational cost associated with modelling. For example, James et al. (2018) generated a SWAN-based machine learning framework model in which a multi-layer perceptron method was used for wave height prediction, while a SVM method was used to predict wave period. O'Donncha et al. (2018b) produced an ensemble model integrating ridge regression and exponentiated gradient algorithms as a surrogate of a SWAN model. Subsequently, their research group aggregated their models to an ensemble computationally lightweight machine learning model, applied to a site in Monterey Bay, California (O'Donncha et al., 2018a). Their surrogate model showed good agreement with a physics-based model, and with a five-thousand-fold improvement in computational speed. The RMSE of the predicted significant wave height against their SWAN model averaged 9 cm and the predicted wave period had an RMSE below 0.1 s.

The demonstrated accuracy and the low computational cost of relevant machine learning systems, when compared to conventional physics-based model outputs, demonstrates an opportunity to improve accuracy and availability of wave data for a wide variety of applications. This paper initiates that research by examining whether, given sufficient data, machine learning techniques can capture the spatial patterns derived by physics-based models within a surrogate model. Acting as an addition to the physics-based model, such a system would have the potential to provide low computational cost estimates of wave conditions and effectively assimilate measured data.

In this study, a RF algorithm was used to learn from an existing physics-based SWAN wave model output in order to produce an operational surrogate model that can provide an immediate, accurate estimate of wave conditions across a domain.

With this in mind, the work presented in this paper addressed three principle objectives:

1) Generate a surrogate model that applied machine learning method on the physics-based outputs to learn the spatial relationship between input buoy data at a few locations within the domain to the full spatially distributed wave conditions across the domain.
2) Run the surrogate model using input data from three locations within the domain.
3) Run the surrogate model using wave buoy measurements as input and validated against further buoy data measured within the domain. This represents using the surrogate model and wave measurements for now-casting wave conditions at any point in the domain without running a full numerical model such as SWAN.

2. Physics-based wave model data

A SWAN spectral wave model was developed for the South West UK, longitude 4°W to 7°W and latitude 49°N to 51°N (Fig. 1) and run for 23 years between 1989 and 2011, as described by Van-Nieuwkoop et al. (2013). This used 3-hourly gridded ECMWF ERA-Interim winds fields, subjected to spatio-temporal interpolation to 10 × 22 gridded data points, which drove the SWAN wave model over a 1 × 1 km^2 grid resolution, i.e. 219 × 223 cells in the grid.

This work considers significant wave height (H_s), mean wave direction (m_{Dir}), mean zero-crossing period (T_z), and peak wave period (T_p) within this region. The simulation time resolution was 1 h, however, due to storage constraints, wave parameters were only recorded every 12 h. The 12-h interval data from 1989 to 2011 were concatenated to build a single data structure in which the first three columns included time, longitude position, and latitude position, and the remaining columns contained the wave parameters. This dataset is henceforth referred to as the original dataset. It includes the training dataset, validation dataset and test dataset.

This SWAN model has previously been validated against a global WAM model (ERA-Interim) at individual grid points, but also with measurements at three buoy locations at Looe Bay, Penzance, and Perranporth. The three buoy locations and corresponding information are shown in Fig. 1 and described in Table 1. Each of the buoys is within approximately 500 m of a SWAN grid point, which is considered sufficiently close in the surrogate model. The comparison between the numerical model results and measurement data can be found in the work of Van Nieuwkoop et al. (2013), where the relative biases of H_s, energy period ($T_{m-1,0}$) at Penzance buoy, Looe Bay buoy, and Perranporth buoy

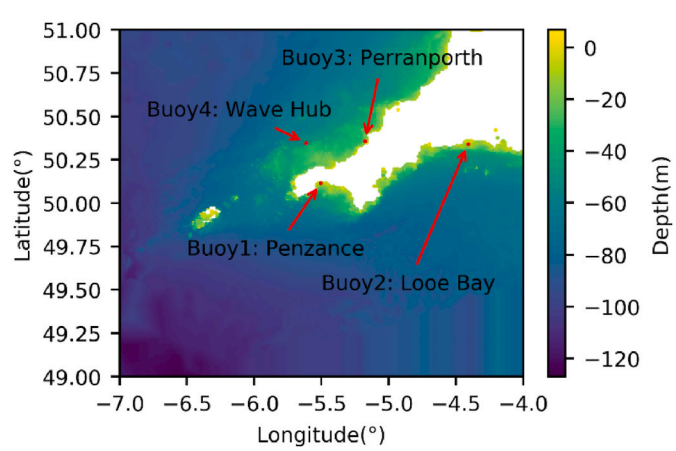

Fig. 1. Area of study (Cornwall, South West UK) and the positions of validation and test measurements from wave buoys.

Table 1
Details of the locations of wave buoys used in this study.

Buoy #	Name	Lon (°W)	Lat (°N)	Depth (m)	Nearest SWAN grid point		Distance from the nearest SWAN grid point
					Lon (°W)	Lat (°N)	(km)
Buoy 1	Penzance	5.5030	50.1144	8.84	5.5054	50.1153	0.2835
Buoy 2	Looe Bay	4.4110	50.3387	10.32	4.4086	50.3402	0.3133
Buoy 3	Perranporth	5.1750	50.3536	19.97	5.1764	50.3582	0.5330
Buoy 4[a]	Wave Hub	5.6143	50.3473	35.85	5.6152	50.3492	0.2334

[a] Buoy 4 is in a Marine Energy Test Centre Site, which was not included in surrogate model but was used for verification of surrogate model.

remained < 20%, the RMSE of mean direction remained < 40% The comparison plots between numerical results at the Perranporth buoy location are shown in Fig. 2, as an example of the validation process.

3. Methodology: machine learning techniques for surrogate regression

The high-fidelity physics-based model is governed by underlying nonlinear equations that relate the wave conditions throughout the domain. A RF approach was implemented as a multivariate surrogate model to represent the spatial patterns in the wave field predicted by the physics-based model. In addition, the RF model was benchmarked against a linear regression (LR) model (section 5.1), developed based on methods in Hutcheson, (2011).

3.1. Multivariate random forest regression

RFs are one of the most effective machine learning algorithms for predictive regression and classification purposes (Pedregosa et al., 2011). It is an ensemble machine learning algorithm proposed by Breiman (2001). As an ensemble approach, it consists of multiple, aggregated simpler machine learning constructs; the RF therefore uses multiple parallel decision tree models to train and predict sample data. Each decision tree for regression was a non-parametric supervised learning model that indicated a set of rules that were hierarchically structured to make decisions in forms of branches and to get real value consequences in forms of leaves from each node. In this research, binary decision trees were used, splitting each node at most into two. The ensemble models and random concepts in RF greatly reduce overfitting of individual tree models, increase diversity in the forest and result in more robust overall predictions (Hastie et al., 2008).

The flow chart of RF algorithm is shown in Fig. 3. Before building trees, several iterations of bootstrap resampling (random sampling with replacement) from the training dataset were applied. The bootstrapping process split each sample group into data for training trees called 'In Bag', and data not included in training trees are referred as "out-of-bag" (OOB) for evaluation. The objective of each tree model was to minimize the mean squared error (MSE) of the OOB sample. The output included ensemble results of K trees. Because each tree is independent and identically distributed, the regression result was the average of K trees.

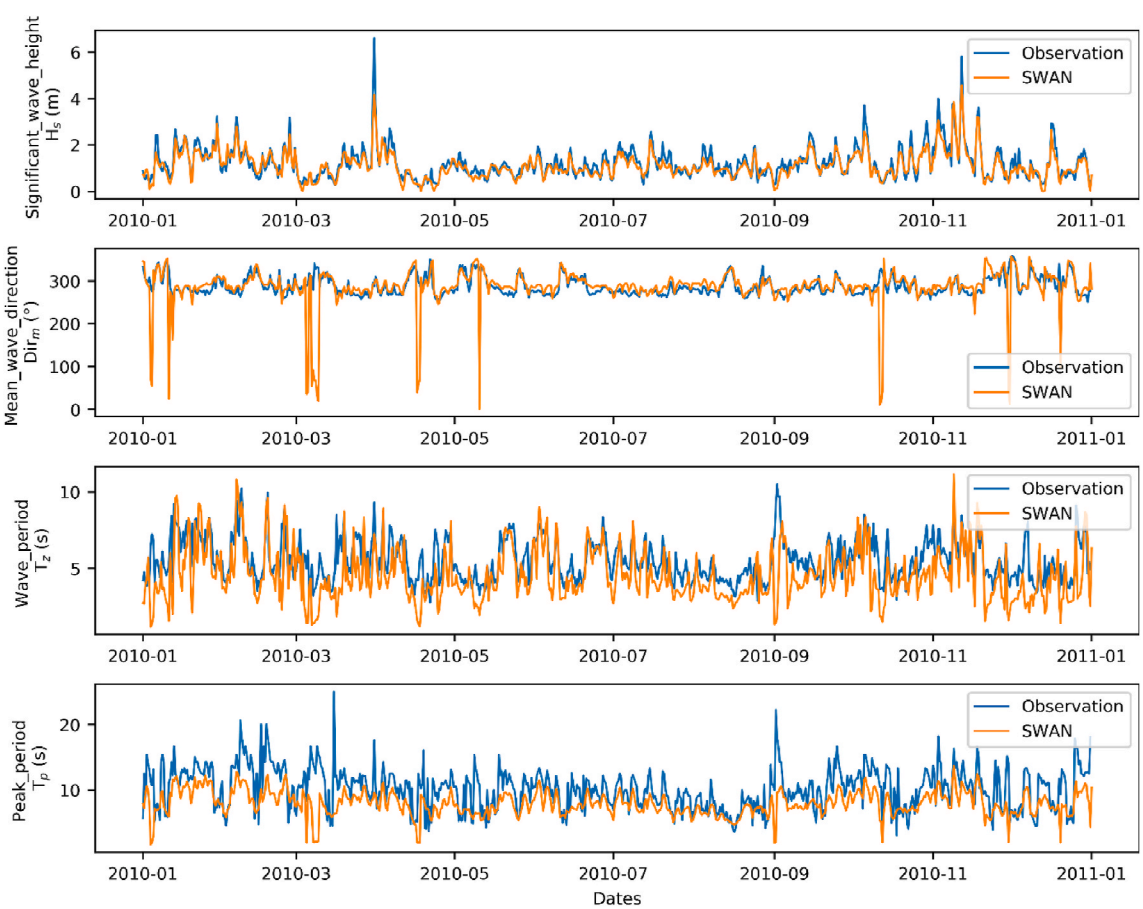

Fig. 2. Observation (blue) and SWAN simulation (orange) data at the wave buoy location close to Perranporth, Cornwall, UK.

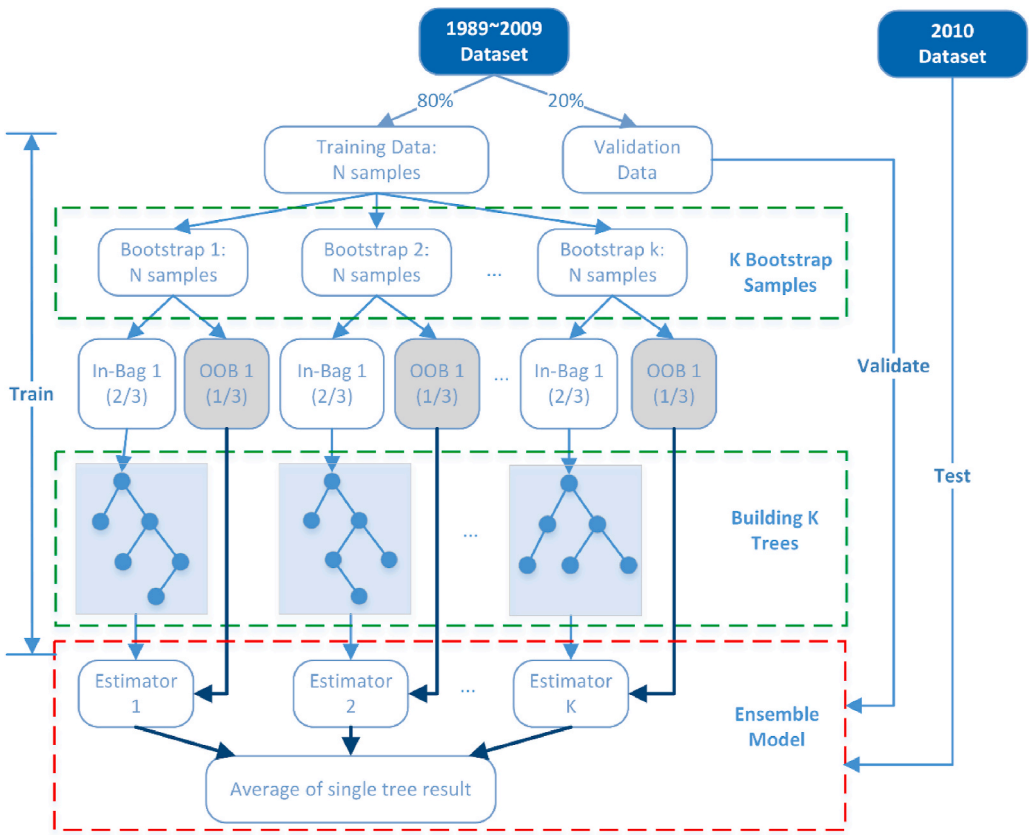

Fig. 3. A flowchart of RF for regression, where "in-bag" means data for training trees, "OOB", short for "out-of-bag", refers to data not for training trees. Adapted from Guo et al. (2011).

Hastie et al. (2008) provide a more detailed discussion of the mathematical aspects of RF.

Tsoumakas and Katakis (2007) categorized the solutions to multi-output or multivariate problems in two ways: (1) problem transformation methods, which transform the problem into several regular single-output problems, and (2) algorithm adaption methods, which directly adapt algorithm into handling multiple outputs. The multivariate RF (MRF) method can be treated as either a series of single-output regression trees, or as a multivariate model (Segal and Xiao, 2011), and the prediction scores of the two methods are similar. In this paper, the MRF regression used for each wave parameter prediction contains Y different outputs, each representing one of Y different features (grid points).

3.2. Training datasets

To implement the machine learning techniques, the original numerical results were formulated into a supervised learning framework. This required the data to be structured as feature-label pairs with a corresponding time index.

3.2.1. Input

Considering the training data set to be a two-dimensional N x M matrix, input features were represented by columns and time was represented by rows. The input feature matrix was generated using 21 years of historical data (January 01, 1989 to December 31, 2009) at the selected locations. The historical data therefore consisted of 15,340 time samples at 12-h intervals. The SWAN model in question was validated against three buoy locations; Looe Bay, Penzance, and Perranporth (Van Nieuwkoop et al., 2013). Correspondingly, wave parameters at these three locations were used as input features to the surrogate model. For each selected location, time series of the four features of interest: H_s (m), m_{Dir} (°), T_z (s), and T_p (s), are considered. To train and validate the surrogate model, the SWAN model is used exclusively, with model results nearest the buoy locations used to represent synthetic buoy data inputs. During the test phase of the model development, however, the synthetic data are substituted for real buoy measurements, demonstrating how the surrogate model can initially be built in the absence of in-situ data which can then be used in operation.

3.2.2. Correlation analysis

Prior to performing the regression analysis, the correlations between feature variables were analysed using a heat map of the Spearman's rank correlation coefficient matrix of the input feature variables, which in this case were the wave parameters at the three buoy locations (Fig. 4). For each wave parameter, correlation between locations was observed. The coefficient between the Penzance buoy and the Looe Bay buoy was moderately higher. These are along the same section of coastline and therefore more spatially correlated.

3.2.3. Output

Evgeniou and Pontil (2004) suggested that for multivariate regression, training a model on related features simultaneously rather than independently can improve predictive performance. On the other hand, if the output features are dissimilar, training separate models independently for each feature can be more time efficient than taking a multi-output approach (Faddoul et al., 2010). In this case, the model data showed spatial correlation across the measurement buoys (Fig. 4), which indicated that multivariate regression would be of value. Correlation between the same parameters at different locations was greater than that between different parameters at the same location. As such, each state variable (H_s, m_{Dir}, T_z, T_p) was modelled separately, while each spatially distributed variable was predicted simultaneously. Therefore, the outputs for each wave parameter were also in the form of a

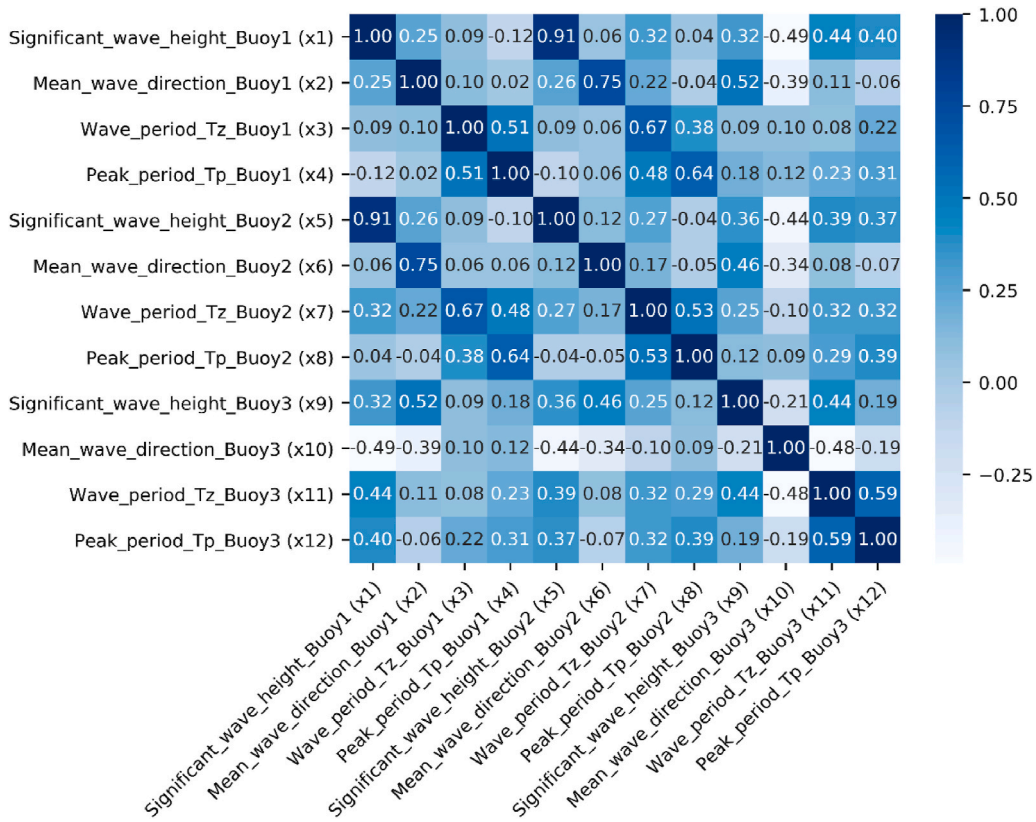

Fig. 4. Input features correlation coefficient matrix, where ($|R| \leq 0.3$ indicated a negligible correlation; $0.3 < |R| \leq 0.5$ was a weak correlation; $0.5 < |R| \leq 0.7$ represented a moderate correlation; $0.7 < |R| \leq 0.9$ denoted a strong correlation; $0.9 < |R| \leq 1$ was fully correlated (Hinkle et al., 2003).

two-dimensional matrix, defined as the Y matrix. In the Y matrix, the rows represent samples at different times, while each column represents the result of a spatial grid point individually.

4. Model setup and application

The application of the surrogate model, including data processing and implementation, used Python 3.6, including the Python toolkit SciKit-Learn (Pedregosa et al., 2011).

4.1. Pre-processing

SWAN model results were available in the Network Common Data Form (netCDF). During pre-processing, these were transformed to a matrix with three indexes; time, longitude, and latitude. These became the first three columns of the matrix and wave parameters (H_s, m_{Dir}, T_z, T_p) at a specific time and location corresponded to the remaining columns of the data matrix. Secondly, invalid samples were removed from the data. Invalid samples in this model included grid points within the computational area that correspond to land. These appeared as "NaN" (not a number) values and were removed before generating the cleaned dataset.

The cleaned dataset for 21 years from 1989 to 2009 was randomly segmented into a training dataset (80%) and validation dataset (20%) (Fig. 3). The data simulated in the year 2010 were held separately and processed as the test dataset. Normally, machine learning models require cross-validation to ensure a robust algorithm. However, for the RF algorithm, the accuracy was evaluated on each OOB sample, which was equivalent to N-fold cross-validation and the results were obtained directly from the model.

In the RF model, there is no requirement for feature engineering or transformation and normalisation of input features. As an interpretable machine learning algorithm, tree-based algorithms can always compare prediction with "what-if"-scenarios which makes them work equally well with any monotonic transformation of a feature (Molnar, 2020).

4.2. Evaluation criteria

In this paper, the accuracy of the surrogate model was quantified using the coefficient of determination (R^2), RMSE and NRMSE as assessments of the uncertainty as well as the mean proportional differences to evaluate bias.

$$\text{Coefficient of Determination: } R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=1}^{N}(\overline{y} - y_i)^2} \tag{1}$$

$$\text{Root Mean Square Error: RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{y}_i - y_i\right)^2} \tag{2}$$

$$\text{Normalized Root Mean Square Error: NRMSE} = \frac{RMSE}{\overline{y}} \tag{3}$$

$$\text{Mean proportional difference (\%): (\%)D} = \frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{y}_i - y_i}{y_i} \tag{4}$$

where N denotes the number of fitted samples, $\widehat{y}$, y and $\overline{y}$ represent the predicted value by the surrogate model, actual value, and mean of actual values respectively. In the initial assessment of surrogate model performance over the study period based on the SWAN model, the actual value refers to SWAN results, while in the later model validation with measured data, the actual value refers to the buoy observations.

4.3. Optimal gridding selection

Ideally, the surrogate model would represent the same resolution as the SWAN model (219 × 223 points in this case), representing each grid point as an output feature. However, the original SWAN resolution contained 42,500 valid points in the domain, i.e. 42,500 output variables. Using this full resolution, led to a requirement of over 2 TB of addressable memory for an 8-year training set. To enable the approach using the full 21-years of data, a dimension reduction process was implemented, to reduce the spatial resolution.

The statistics of waves were considered stationary within each simulation and homogeneous area over the domain. The surrogate model, therefore required an evenly scaled resolution representing a smoothed version of the original model that captured the spatial distribution with a good agreement. An effective resolution adjustment method was to use bilinear interpolation techniques (Accadia et al., 2003) to transform the SWAN data to different grid resolutions.

In order to find an optimal gridding resolution for the surrogate model, different scales of horizontal and vertical resolutions were assessed and compared. The optimal resolution was affected by several factors, including the computational cost and overall accuracy of the scaled resolution to represent the high-resolution dataset.

To assess whether the low-resolution data after the dimension reduction (DR) were accurately representing the original high-resolution dataset, dimension ascension (DA) was applied to the adjusted, low-dimensional data using the same bilinear interpolation method. The combined assessment, which went through the DR-DA process took the following into consideration (Table 2).

1) The minimum acceptable spatial resolution was set at 0.125 ° × 0.125 °, which resulted in 425 (25 segments longitudinal x 17 segments latitudinal) grid points in the area of study.
2) The ratio of valid grid points (non-NaN values) after the DR-DA process. Selecting different resolutions generates a different number of NaN values at the edges. The ratio of non-NaN values was considered an important factor to the subsequent modelling and was used to evaluate the resolution selected.
3) The mapping error after the DR-DA process. The NRMSEs of four wave variables associated with each set of interpolated values were computed as the average of RMSE at each valid point over the mean value of the wave parameter. To avoid seasonal trends of wave parameter distributions, 100 timestamps from the 21 years were sampled to execute the mapping NRMSE assessment (Fig. 5).
4) The training time for the surrogate model. In this comparison, the time taken to train using 21 years of significant wave height with different resolutions was quantified.
5) The accuracy of the surrogate model, which was evaluated by the R^2 value in the test dataset.

The 100-sample-averaged NRMSE between interpolation from low resolution and the original data did not vary significantly (Fig. 5). When varied between 1/2 and 1/8, the NRMSE remained less than 2% of spatial average value. For computational efficiency, the scales from 1/4 to 1/8 were processed for training the surrogate model, with results listed in Table 2. The accuracy (R^2) of the surrogate model against the scaled SWAN model remained stable around 0.957, but the training time dropped from 30 min to 6 min. The combined evaluation considered the factors including Non-NaN ratio, mapping NRMSE, and training accuracy. This indicated that the scale of 1/5 performed best. The training time for a 1/5th scale surrogate model was around 17 min. As a result, the 1/5th scale was used in this work, which transformed the original SWAN data to a grid with 43-longitude segments and 44-latitude segments. The resolution change along the transformation is illustrated in Fig. 6.

4.4. Model hyper-parameter setup

The RF algorithm contains several hyper-parameters including the number of estimators, maximum tree depth, maximum features at each split and maximum samples. When looking for the best split, all features were taken into consideration. In each estimator, the maximum tree depth was set to be default, which means the nodes of each estimator were expanded until all leaves are pure; the maximum sample was the length of the training dataset. Therefore, the number of estimators was the key hyper-parameter requiring tuning during model development. Generally, the prediction accuracy improved with increasing numbers of estimators. However, increasing the number of estimators resulted in increased training time. A parameter study of training the H_s surrogate model showed diminishing returns when increasing the number of estimators (Fig. 7). With greater than 200 estimators, the R^2 curve and the RMSE curve flattened and converged for both the training and test sets, while the required computational time continued to increase linearly. Based on this with respect to both accuracy and training efficiency, the number of estimators used in the present surrogate models was set to 200.

5. Results

5.1. Accuracy of the surrogate model relative to SWAN model

The surrogate outputs were compared to the equivalent SWAN estimates. For each of the wave parameters studied, R^2 values exceeded 0.9, with the exception of the mean wave direction (Table 3) and scatter plots demonstrate this strong correlation (Fig. 8). The relative RMSE of H_s and T_p are below 10% of each wave parameter's average value. Low RMSE values within both the validation and test datasets indicated high confidence in the model's ability to replicate the SWAN results. The prediction of peak period performed best among the four wave

Table 2
Parameters used for assessment of optimal gridding process. Column (10) which takes all the factors into considerations determines the optimal scale. The row (resolution scale of 1/5) with bold values means the optimal resolution scale selected for the surrogate model.

Resol. scale	Long. seg. (1)	Lat. seg. (2)	Resol. (3) = (1) x (2)	Valid points after DA (4)	Non-NaN ratio (5) = (4)/ 48,837	Avg. NRMSE of 100 samples (6)	Comb. eval. (7) = (6)/(5)	Train time (min) (8)	R^2 (Test) of surrogate model (9)	Comb. eval. (10) = (6)/(5)/ (9)
Origin	219	223	48,837	42,500	0.8702	0.00%	0.0000			
1/2	110	112	12,320	42,306	0.8663	0.88%	1.02%			
1/3	73	74	5402	42,156	0.8632	0.87%	1.01%			
1/4	55	56	3080	42,004	0.8601	1.15%	1.34%	30.8	0.9577	1.40%
1/5	**43**	**44**	**1892**	**41,830**	**0.8565**	**1.09%**	**1.27%**	**17.6**	**0.9575**	**1.33%**
1/6	37	37	1369	41,697	0.8538	1.31%	1.53%	12.8	0.9574	1.60%
1/7	31	32	992	41,358	0.8469	1.23%	1.45%	9.1	0.9571	1.52%
1/8	27	28	756	41,374	0.8472	1.92%	2.27%	6.6	0.9566	2.37%
1/9	24	25	600	41,041	0.8404	2.12%	2.53%			
1/10	22	22	484	41,075	0.8411	1.65%	1.96%			

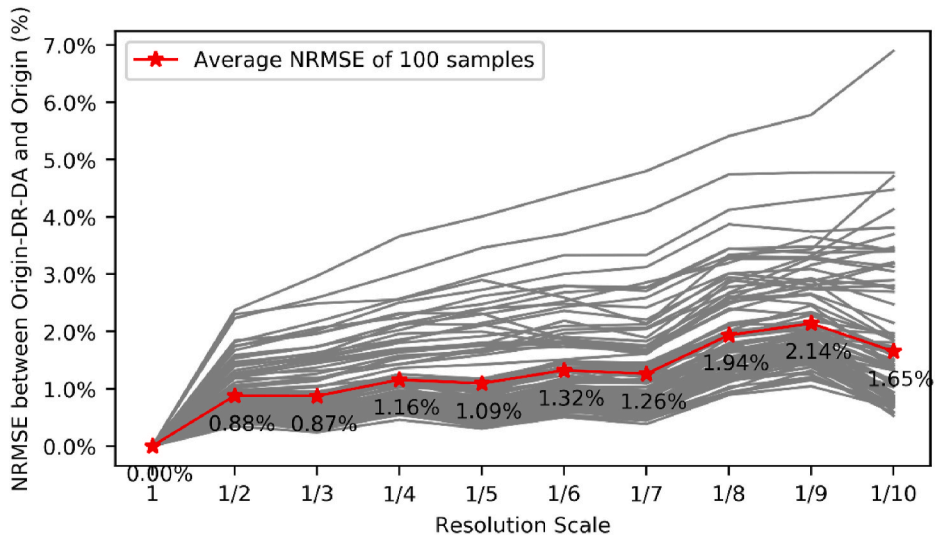

Fig. 5. NRMSE comparison of different resolution scales.

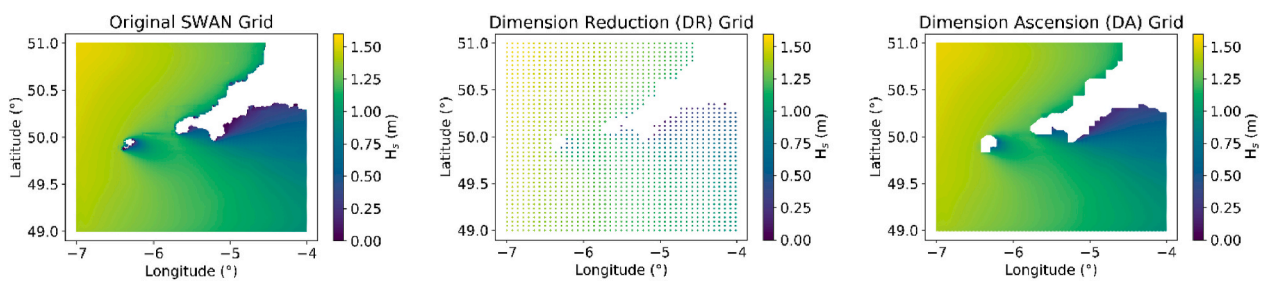

Fig. 6. Plot comparison among original, dimension reduction and dimension ascension.

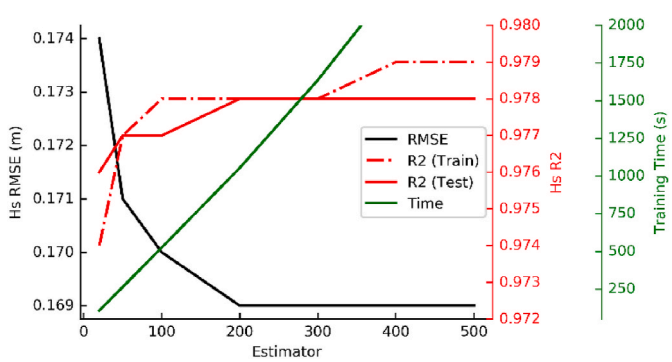

Fig. 7. Parameter study of surrogate model training H_s: Relation between RMSE and estimator number (black); relation between R^2 and estimator number (red); relation between training time and estimator number (green).

parameters with NRMSE values 5.3% of the average peak period over the test data set among the target area. The surrogate prediction results of mean wave direction (m_{Dir}) had relatively low prediction accuracy with the R^2 value below 0.9 in all of the datasets, the NRMSE accounted for 14.31% of the average direction. The RF model outperformed the benchmark LR model, reducing the NRMSE by a factor of approximately 1.5 for all wave parameters (Table 3), indicating the efficacy of the RF algorithm.

The spatial distribution of the differences between the surrogate and SWAN models are illustrated in Fig. 9. The annual averaged proportional differences between the surrogate and SWAN model in 2010 were less than 1.8% of SWAN values. In the northern area, the surrogate model over-estimated H_s, while in the southern area, it underestimated. The largest mean proportional difference within the domain was, $D(H_s) = -2\%$ for the area close to the east of the Isles of Scilly. However, the annual averaged value from SWAN at that point was 0.66 m, which resulted in a small actual difference of 0.0198 m. Areas closer to input positions were more likely to have lower RMSE between the original SWAN simulation and the surrogate model in 2010.

Table 3
Surrogate model performance parameters.

Input location	Train set period	RF config.	Wave para.	Para. avg	R^2 (train)	R^2 (val)	RMSE (val)	R^2 (test)	RMSE (test)	NRMSE (test)	LR NRMSE (test)
Buoy 1 Penzance	21 year (1989-01-01	Estimators =	H_s	1.8086	0.9783	0.9782	0.1680	0.9575	0.1724	9.53%	13.27%
Buoy 2 Looe	00:00–2009-12-31 12:00)	200	T_z	5.1897	0.9007	0.9015	0.5137	0.8800	0.5423	10.45%	14.06%
Bay			T_p	7.9756	0.9636	0.9594	0.3705	0.9398	0.4238	5.31%	8.42%
Buoy 3			m_{Dir}	231.01	0.8302	0.8279	27.242	0.8171	33.058	14.31%	23.79%
Perranporth											

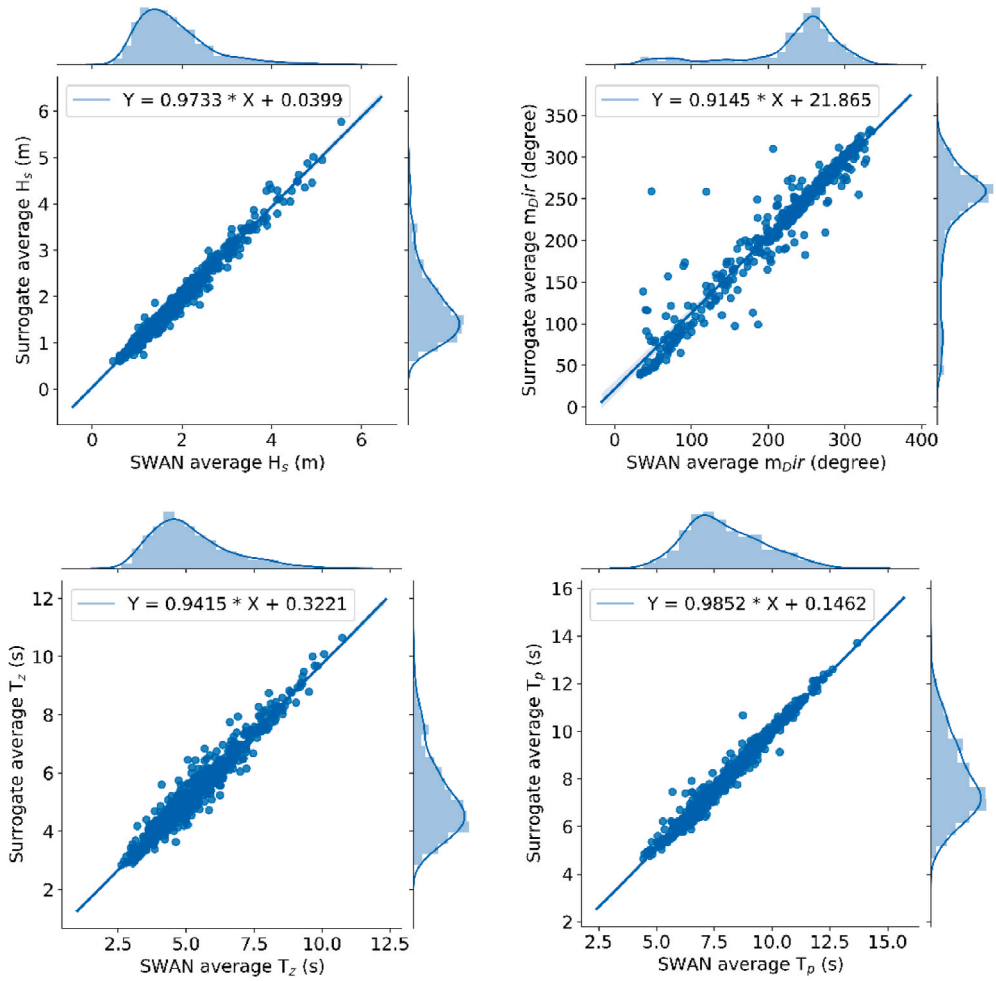

Fig. 8. The spatially averaged H_s, m_{Dir}, T_z, and T_p, from the surrogate model estimates in the test data set, compared with SWAN model runs. Blue lines and corresponding equations represent a fitted trend line of scatters.

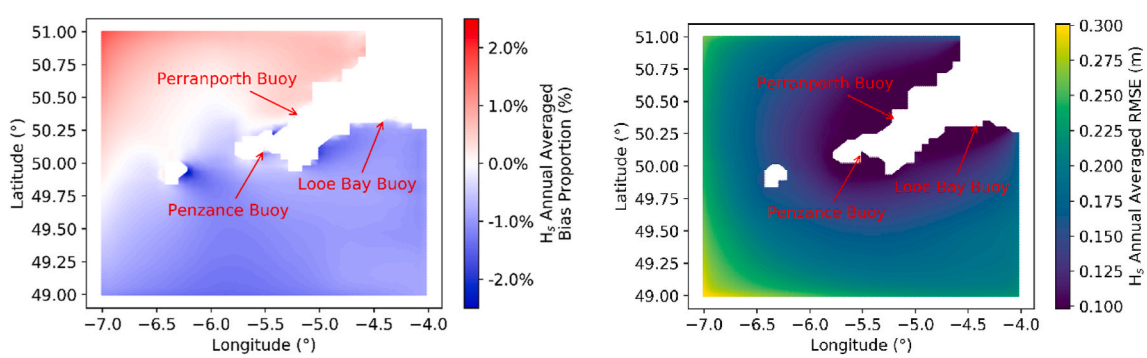

Fig. 9. Annual averaged proportional difference, D (left) and annual averaged RMSE distribution (right) of the surrogate estimation after the Dimension Ascension (DA) process, compared with the SWAN simulation over the test dataset in 2010.

5.2. Performance of the surrogate model

To test the feasibility of the surrogate model for potential deployment applications, a validation stage using measured data both as input and output was necessary. The observation data of the three buoys from Penzance, Looe Bay, and Perranporth were used as input and the result at another buoy location was used for output validation. This validation buoy, close to the Wave Hub Marine Energy Test Centre Site, is representative of an offshore site with a valuable resource for marine renewable energy development (Ashton, 2012; Saulnier et al., 2012). It is the primary resource for any real-time decisions made for marine operations at this offshore renewable energy test site. To compare the result from a more accurate location, the DA process was undertaken to generate a high resolution spatial data set. The wave buoy data were then compared to the nearest high-resolution grid point of the surrogate model.

For all four wave parameters studied, the surrogate model consistently matched the real data better than the SWAN model (Table 4 & Fig. 10). All R^2 and RMSE values comparing the real data and surrogate model output were smaller than the equivalent statistics between real

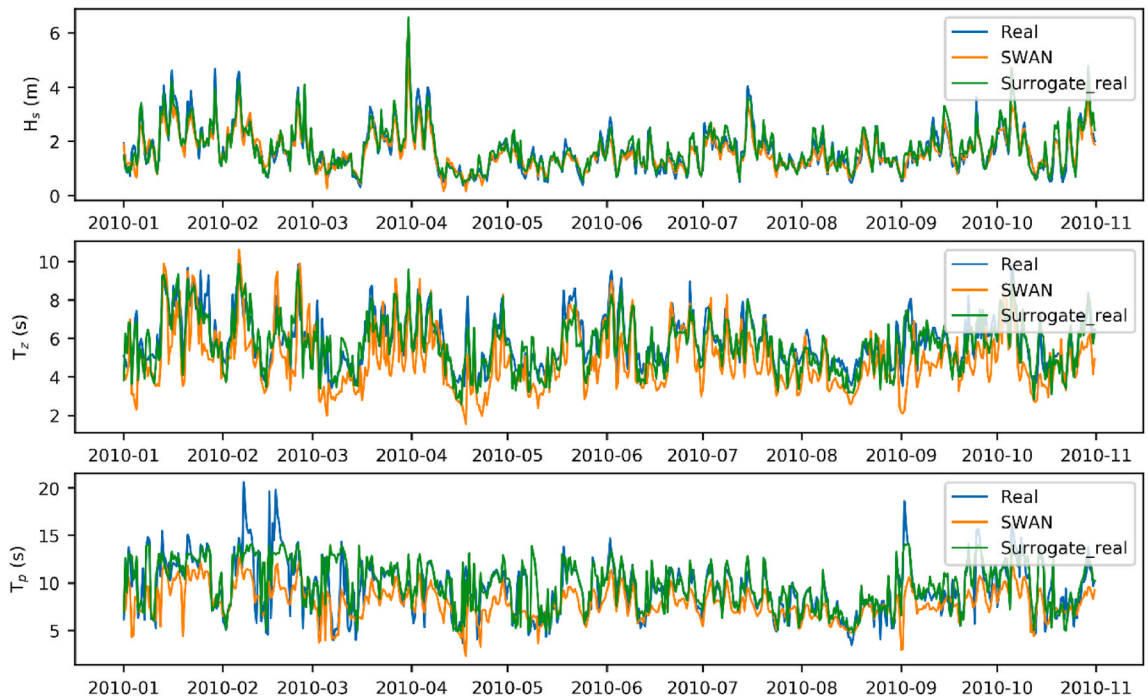

Fig. 10. Comparison of four wave parameters in the year 2010 between SWAN outputs, the surrogate model outputs with real data inputs, and buoy observations at Wave Hub. Blue curves represent real buoy data, the orange and green curves represent SWAN and surrogate model outputs, respectively. The time-step for both sets of output was consistent with data plotted every 12 h.

Table 4
Model performance between the surrogate model and SWAN in 2010.

		R^2	RMSE	NRMSE
H_S	SWAN	0.8521	0.3218	19.01%
	Surrogate	0.9067	0.2556	15.10%
T_Z	SWAN	−0.0257	1.3903	23.71%
	Surrogate	0.7205	0.7258	12.38%
T_P	SWAN	0.2263	2.4852	26.26%
	Surrogate	0.5558	1.8831	19.89%

data and SWAN estimates. In particular, the surrogate model's prediction of zero-crossing wave period had a dramatic improvement in accuracy compared to SWAN, with an R^2 of 0.7205 and an RMSE value half of the corresponding SWAN value (Table 4). The surrogate model estimated significant wave height R^2 value was 0.9067, with an RMSE of 0.2556 m and NRMSE of around 15%. The NRMSEs of the surrogate model against measured data were also below 20% for both T_z and T_p. The observations did not provide mean wave direction, so the comparison of wave direction was possible.

5.3. Computational time and requirements

Training the model for each wave parameter for 21 years was completed in 17 min on a laptop with 16.0 GB RAM and an i7-8550U processor, using Python 3.6 program in a Windows environment, including the steps for data processing. Feeding the surrogate model with one set of observation data yielding equivalent estimates of all wave parameters for the whole domain took less than 1 s on the same machine. As the model architecture for the surrogate produced separate independent models for each wave parameter, these can be run in parallel to make full use of available resources.

6. Discussion

This paper has described a method for developing a surrogate wave model based on existing phase-averaged spectral wave model output. The surrogate model worked on the assumption that the spatial distribution of wave conditions created by the physical modelling process (in this case SWAN) was well defined and provides an additional service to immediately estimate wave conditions across the model domain from limited input values.

These results indicated that the RF based surrogate model represents an efficient and accurate method for predicting the spatial wave field. When using solely physics-based models, forecasts and associated spatial estimates of current conditions rely on models updating every 6–12 h, which require significant computing resources. This system offers a low-cost alternative to estimate spatial model outputs based on very limited input data and with significantly improved speed. When deployed using data from 3 in-situ wave buoys within the domain, the outputs were more accurate than physical modelling equivalents. Furthermore, the computational requirement was reduced by approximately a factor of 100.

In the surrogate spatial estimation, the most time-consuming task was loading the machine learning model, which took around 5 min for each wave parameter. "Edge computing" (Shi and Dustdar, 2016) technology would enable the model to be pre-loaded into memory and give rise to nearly instantaneous spatial wave estimation. As such, this system that would potentially be accessible using a PC, mobile phone, vessel navigation system, or autonomous vessel.

Machine learning algorithms have been verified in the literature to solve the spatial regression problem. For example, James et al. (2018) and O'Donncha et al. (2018b, 2018a) used all wave boundary data to replicate SWAN outputs. The implementation demonstrated in this paper, is different to previous studies, reducing the input data to three points within the domain and focusing on providing accurate now-casting from measurement assets. This makes direct comparison with previous studies difficult. Instead, the work used benchmarking with a LR model and validation with measured data to assess the surrogate model. Benchmarking demonstrated that the extra complexity of RF algorithm improves accuracy. Verification showed that when in-situ data were used to drive the surrogate model, the results were improved

when compared to physics-based model estimates. This is an important outcome as it showed that the combined surrogate model and in-situ data have the potential to provide the most accurate description of current conditions across the model domain. This makes the system highly suited to real-time management for autonomous vehicles or marine operations for offshore infrastructure.

The surrogate modelling process methods demonstrated are additional to physics-based models. This system relies on an accurate spatial description of the wave conditions from which the surrogate model can learn the spatial distribution of conditions across the area. As such, accurate physical modelling remains central to this process. This work shows that a surrogate model can offer real-time estimates for wave conditions across a model domain. The computational requirements and operation from limited real-time measurements mean that higher resolution model output could now be implemented as a service, but this could be considered as an additional service within the physical-modelling architecture.

For some applications, such a system based on global model data will be advantageous, particularly where forecast data are available. The spatial correlations quantified using this method could similarly be implemented to convert forecast from global models into higher-resolution output, allowing a forecast product to be defined with similar savings on data input and computational requirements. Further development in this area should consider how to effectively combine in-situ measurements with this forecast activity. Taking advantage of the improved accuracy shown in this work has the potential to create an augmented forecast using a surrogate model procedure.

This work has demonstrated the surrogate system using wave buoy data. These data were particularly suited to the application, with accurate, long-term data sets. However, the buoys were not deployed for the purpose of this study and the system showed excellent results from measurements that were not optimally placed within the domain. This highlights the potential for the surrogate approach to incorporate imperfect data. Further work should establish how such a system can work with other data sets. Satellite remote sensing offers global coverage, while installed infrastructure, vessels, or autonomous systems all have the potential to gather data. Establishing how data that may be inaccurate or contain bias can be incorporated in this system has the potential to open up significant opportunity for revolutionising met-ocean data provision and making real-time access to accurate spatial data common in a marine setting.

7. Conclusion

In this paper, a novel method was proposed to derive an accurate spatial wave data set using in-situ measured wave data from point locations, using a machine learning approach. Based on a physics-based wave model (SWAN), this approach used a RF algorithm to evaluate the spatial correlation of wave parameters within the computational domain. This created a surrogate model, which was an efficient method to replicate physical modelling, without the undertaking computationally expensive calculations. When using observations within the domain as inputs, the surrogate model was more accurate than the corresponding estimates from the SWAN model drawing on global model data at the computational domain boundary.

This method supplements a combination of physics-based modelling and in-situ observations that form the most common approach to met-ocean monitoring. It combines the real-time availability of in-situ data with the spatial capabilities of physics-based models and it is easily implemented with existing systems. Once developed, the system required little computational power for implementation and as such, it has the potential to provide real-time spatial data coverage even in situations where data transmission or computational resources are limited. This access to accurate real-time spatial data has the potential to fundamentally change the way that met-ocean data are used for the management and operation of marine infrastructure.

The system developed was highly flexible and has potential for implementation with other marine environmental parameters. Continued development will allow combined analysis of a range of in-situ monitoring devices and can incorporate measurements of opportunity to create highly detailed and accurate data sets. This will include mobile measurements from autonomous vehicles and a built in suitability to direct such measurements to improve accuracy and relevance of data sets for specific operations. Establishing this system with other data sets will create significant opportunity for making real-time access to accurate spatial data the new normal in a marine setting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Source data used in the preparation of this paper has been supplied by the Channel Coastal Observatory, UK, funded under the regional coastal monitoring programmes by DEFRA. Ongoing work is supported by the EPSRC Supergen Offshore Renewable Energy Hub (http://doi.org/10.13039/501100000266) [grant no: EP/S000747/1]. Flexible Fund support for MaLCOM and collaboration with Dr. Ed Steele at the UK Met Office.

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., Speranza, A., 2003. Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. Weather Forecast. 18, 918–932. https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Ardente, F., Beccali, M., Cellura, M., Lo Brano, V., 2008. Energy performances and life cycle assessment of an Italian wind farm. Renew. Sustain. Energy Rev. 12, 200–217. https://doi.org/10.1016/j.rser.2006.05.013.

Ashton, I., 2012. Spatial Variability of Wave Fields over the Scale of a Wave Energy Test Site. D. Phil. University of Exeter, United Kingdom.

Ashton, I., Van-Nieuwkoop, McCall, J.C.C., Smith, H.C.M., Johanning, L., 2014. Spatial variability of waves within a marine energy site using in-situ measurements and a high resolution spectral wave model. Energy 66, 699–710. https://doi.org/10.1016/j.energy.2013.12.065.

Balog, I., Ruti, P.M., Tobin, I., Armenio, V., Vautard, R., 2016. A numerical approach for planning offshore wind farms from regional to local scales over the Mediterranean. Renew. Energy 85, 395–405. https://doi.org/10.1016/j.renene.2015.06.038.

Berrisford, P., Dee, D., Poli, P., Brugge, R., Fielding, K., Fuentes, M., Kallberg, P., Kobayashi, S., Uppala, S., Simmons, A., 2011. The ERA-Interim Archive, Version 2.0.

Booij, N., Ris, R.C., Holthuijsen, L.H., 1999. A third-generation wave model for coastal regions: 1. Model description and validation. J. Geophys. Res. 104, 7649–7666. https://doi.org/10.1029/98JC02622.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chawla, A., Spindler, D., Tolman, H., 2012. 30 Year Wave Hindcasts Using WAVEWATCH III R with CFSR Winds† Phase 23.

DNV, G., 2014. Environmental Conditions and Environmental Loads. Recommend Practice DNV-RP-C205.

Evgeniou, T., Pontil, M., 2004. Regularized multi–task learning. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04. Presented at the the 2004 ACM SIGKDD International Conference. ACM Press, Seattle, WA, USA, p. 109. https://doi.org/10.1145/1014052.1014067.

Faddoul, J.B., Chidlovskii, B., Torre, F., Gilleron, R., 2010. Boosting multi-task weak learners with applications to textual and social data. In: 2010 Ninth International Conference on Machine Learning and Applications. Presented at the 2010 International Conference on Machine Learning and Applications (ICMLA). IEEE, Washington, DC, USA, pp. 367–372. https://doi.org/10.1109/ICMLA.2010.61.

Gentry, R.R., Lester, S.E., Kappel, C.V., White, C., Bell, T.W., Stevens, J., Gaines, S.D., 2017. Offshore aquaculture: spatial planning principles for sustainable development. Ecol. Evol. 7, 733–743. https://doi.org/10.1002/ece3.2637.

Günther, H., Hasselmann, S., Janssen, P.A., 1992. The WAM Model Cycle 4. Deutsches Klimarechenzentrum (DKRZ).

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS J. Photogram. Remote Sens. 66, 56–66. https://doi.org/10.1016/j.isprsjprs.2010.08.007.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Element of Statistical Learning.

Hinkle, D.E., Wiersma, W., Jurs, S.G., 2003. Applied Statistics for the Behavioral Sciences. Houghton Mifflin College Division.

Hutcheson, Graeme D, 2011. Ordinary least-squares regression. The SAGE dictionary of quantitative management research.

Ibarra-Berastegi, G., Saénz, J., Esnaola, G., Ezcurra, A., Ulazia, A., 2015. Short-term forecasting of the wave energy flux: analogues, random forests, and physics-based models. Ocean Eng. 104, 530–539. https://doi.org/10.1016/j.oceaneng.2015.05.038.

James, Richard W, 1957. *Application of wave forecasts to marine navigation*. New York University.

James, S.C., Zhang, Y., O'Donncha, F., 2018. A machine learning framework to forecast wave conditions. Coast. Eng. 137, 1–10. https://doi.org/10.1016/j.coastaleng.2018.03.004.

Johnston, P., Poole, M., 2017. Marine surveillance capabilities of the AutoNaut wave-propelled unmanned surface vessel (USV). In: OCEANS 2017 - Aberdeen. Presented at the OCEANS 2017 - Aberdeen. IEEE, Aberdeen, United Kingdom, pp. 1–46. https://doi.org/10.1109/OCEANSE.2017.8084782.

Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., Janssen, P., 1996, 0521577810. In: Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., Janssen, P.A.E.M. (Eds.), Dynamics and Modelling of Ocean Waves. Dynamics and Modelling of Ocean Waves. Cambridge University Press, Cambridge, UK, p. 554. August 1996. 554.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.

Liang, B., Fan, F., Liu, F., Gao, S., Zuo, H., 2014. 22-Year wave energy hindcast for the China East Adjacent Seas. Renew. Energy 71, 200–207. https://doi.org/10.1016/j.renene.2014.05.027.

MEPC, R., 2012. GUIDELINES FOR THE DEVELOPMENT OF A SHIP ENERGY EFFICIENCY MANAGEMENT PLAN (SEEMP), vol. 2. International Maritime Organization, London.

Molnar, C., 2020. 4.4 decision tree | interpretable machine learning. https://christophm.github.io/interpretable-ml-book/tree.html#interpretation-2.

Nencioli, F., Quartly, G.D., 2019. Evaluation of sentinel-3A wave height observations near the coast of southwest England. Rem. Sens. 11, 2998. https://doi.org/10.3390/rs11242998.

NOAA, 2020. Global data assimilation system (GDAS) [WWW Document]. URL. https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00379. accessed 4.20.20.

O'Donncha, F., Zhang, Y., Chen, B., James, S. c, 2018a. Ensemble Model Aggregation Using a Computationally Lightweight Machine-Learning Model to Forecast Ocean Waves arXiv:1812.00511 [physics].

O'Donncha, F., Zhang, Y., Chen, B., James, S.C., 2018b. An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts. J. Mar. Syst. 186, 29–36. https://doi.org/10.1016/j.jmarsys.2018.05.006.

Oh, J., Suh, K.-D., 2018. Real-time forecasting of wave heights using EOF – wavelet – neural network hybrid model. Ocean Eng. 150, 48–59. https://doi.org/10.1016/j.oceaneng.2017.12.044.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: machine learning in Python. Machine Learning in Python, 6.

Pirhooshyaran, M., Snyder, L.V., 2020. Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks. Ocean Eng. 207, 107424. https://doi.org/10.1016/j.oceaneng.2020.107424.

Reikard, G., Robertson, B., Bidlot, J.-R., 2017. Wave energy worldwide: Simulating wave farms, forecasting, and calculating reserves. Int. J. Marine Energy 17, 156–185. https://doi.org/10.1016/j.ijome.2017.01.004.

Ris, R.C., Holthuijsen, L.H., Booij, N., 1999. A third-generation wave model for coastal regions: 2. Verification. J. Geophys. Res. 104, 7667–7681. https://doi.org/10.1029/1998JC900123.

Sánchez, A.S., Rodrigues, D.A., Fontes, R.M., Martins, M.F., Kalid, R. de A., Torres, E.A., 2018. Wave resource characterization through in-situ measurement followed by artificial neural networks' modeling. Renew. Energy 115, 1055–1066. https://doi.org/10.1016/j.renene.2017.09.032.

Saulnier, J.-B., Maisondieu, C., Ashton, I., Smith, G.H., 2012. Refined sea state analysis from an array of four identical directional buoys deployed off the Northern Cornish coast (UK). Appl. Ocean Res. 37, 1–21. https://doi.org/10.1016/j.apor.2012.02.001.

Segal, M., Xiao, Y., 2011. Multivariate random forests. In: WIREs Data Mining and Knowledge Discovery, 1, pp. 80–87. https://doi.org/10.1002/widm.12.

Serras, P., Ibarra-Berastegi, G., Sáenz, J., Ulazia, A., 2019. Combining random forests and physics-based models to forecast the electricity generated by ocean waves: a case study of the Mutriku wave farm. Ocean Eng. 189, 106314. https://doi.org/10.1016/j.oceaneng.2019.106314.

Service (C3S), C.C.C., 2017. ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate, Copernicus Climate Change Service Climate Data Store (CDS).

Shi, W., Dustdar, S., 2016. The promise of edge computing. Computer 49, 78–81. https://doi.org/10.1109/MC.2016.145.

Tolman, H.L., 2009. User Manual and System Documentation of WAVEWATCH III TM Version 3.14 220.

Tolman, H.L., Balasubramaniyan, B., Burroughs, L.D., Chalikov, D.V., Chao, Y.Y., Chen, H.S., Gerald, V.M., 2002. Development and implementation of wind-generated ocean surface wave models at NCEP. Weather Forecast. 17, 23.

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: an overview [WWW Document] Int. J. Data Warehous. Min. URL www.igi-global.com/article/multi-label-classification/1786 accessed 3.11.20.

Van Nieuwkoop, J.C.C., Smith, H.C.M., Smith, G.H., Johanning, L., 2013. Wave resource assessment along the Cornish coast (UK) from a 23-year hindcast dataset validated against buoy measurements. Renew. Energy 58, 1–14. https://doi.org/10.1016/j.renene.2013.02.033.

Wu, W., Wang, T., Yang, Z., García-Medina, G., 2020. Development and validation of a high-resolution regional wave hindcast model for U.S. West Coast wave resource characterization. Renew. Energy 152, 736–753. https://doi.org/10.1016/j.renene.2020.01.077.