3-бөлім Раздел 3 Section 3

Информатика Информатика Computer Science

IRSTI 20.23.17; 20.23.21; 20.23.25 DOI: https://doi.org/10.26577/JMMCS.2021.v110.i2.08

Zahoor Ahmad, Madina Mansurova*
Al-Farabi Kazakh national university, Kazakhstan, Almaty
*E-mail: madina.mansurova@kaznu.kz

MACHINE LEARNING APPROACH TO PREDICT SIGNIFICANT WAVE HEIGHT

To estimate significant wave height of ocean wave, a machine learning framework is developed. Significant wave height and period can be used by supervised training of machine learning to predict ocean conditions. In this paper we proposed a method to predict significant wave height using Support vector regression (SVR). Buoy dataset taken from the Queensland government open data portal the input from which were aggregated into supervised learning test and training data sets, which were supplied to machine learning models. The SVR model replicated significant wave height with a root-mean-squared-error of 0.044 and performed on the test data with 95% accuracy. Comparing to forecasting with the physics-based model the Machine learning SVR model requires only a fraction ($< 1/1200^{\text{th}}$) of the computation time, to predict Significant wave height.

Key words: Machine learning, significant wave height, Support vector regression.

Захур Ахмад, Мадина Мансурова*
Әл-Фараби атындағы Қазақ ұлттық университеті, Қазақстан, Алматы қ.
*E-mail: madina.mansurova@kaznu.kz

Толқынның елеулі биіктігін болжауға арналған машиналық оқыту негізіндегі тәсіл

Мұхит толқынының елеулі биіктігін бағалауға арналған машиналық оқыту жүйесі құрылды. Толқынның елеулі биіктігі мен толқын периоды мұхит жағдайларын болжау үшін бақыланатын машиналық оқыту барысында пайдаланылуы мүмкін. Бұл жұмыста тірек векторы әдісі негізіндегі регрессия көмегімен (Support vector regression – SVR) толқынның елеулі биіктігін болжау әдісі ұсынылды. Буй деректер жиыны Квинсленд үкіметінің ашық деректер порталынан алынды, кіріс деректері бақыланатын оқыту мен тестілеу үшін деректер жиынтығына біріктірілді. SVR моделі толқынның елеулі биіктігін 0,044 орташа квадраттық қателікпен көрсетті және тестілеу деректерінде 95% дәлдікпен бойынша орындалды. Толқынның елеулі биіктігін физикалық модель негізінде болжаумен салыстырғанда, машиналық оқыту негізіндегі SVR моделі айтарлықтай аз есептеу уақытын ($< 1/1200$) қажет етеді.

Түйін сөздер: Машиналық оқыту, толқынның елеулі биіктігі, тірек векторы әдісі негізіндегі регрессия.

Захур Ахмад, Мадина Мансурова*
Казахский национальный университет имени аль-Фараби, Казахстан, г.Алматы
*E-mail: madina.mansurova@kaznu.kz

Подход на основе машинного обучения для прогнозирования значительной высоты волны

Разработана система машинного обучения для оценки значительной высоты океанской волны. Значительная высота и период волны могут быть использованы при контролируемом машинном обучении для прогнозирования состояния океана. В данной работе предложен метод для прогнозирования значительной высоты волны с помощью регрессии на основе метода опорных векторов (Support vector regression – SVR). Набор данных буев взят с портала открытых данных правительства Квинсленда, входные данные с которого были объединены в наборы данных для контролируемого обучения и тестирования.

Модель SVR воспроизводила значительную высоту волны со среднеквадратической ошибкой 0,044 и выполнялась на тестовых данных с точностью 95%. По сравнению с прогнозированием значительной высоты волны на основе физической модели, для модели SVR с машинным обучением требуется значительно меньше ($< 1/1200$) времени вычислений.

Ключевые слова: Машинное обучение, значимая высота волны, регрессия на основе метода опорных векторов.

1 Introduction

Many people are unaware of a single climate factor that can have a profound effect on the living conditions and health of coastal people. Wave weather is the distribution of wave signals measured at a given time and place, just as atmospheric weather is defined as the "intermediate weather" of a given time and place. About 10% of the world's population lives within 20 kilometers of coastline and less than 20 meters above sea level (Kummu et al. 2016). For these people, hot weather can affect their daily lives like atmospheric weather. Big waves can disrupt harbors and make boats dangerous, keeping fishermen and boats afloat while their businesses suffer.

Surfers aside, there are basic reasons why information on wave conditions over the next few days is important. For example, delivery routes can be made by avoiding rough seas and thus reducing shipping times. Another industry that benefits from wave information is the $ 160 B (2014) [1] marine fishery, which can improve harvesting activities accordingly. Awareness of critical situations is critical to military and navy operations by Navy and Marine Corps teams. Also, predicting energy production from renewable energy sources is important in maintaining a stable electricity grid because more renewable energy sources (e.g. sun, wind, waves, wave, etc.) are in between. In the deep penetration of the renewable energy market, a combination of increasing energy conservation and improved speculation of energy prediction will be required.

Waves can be defined by three distinct elements: wavelength, wave duration, and direction of wave. The higher the tide, the more dangerous the boat conditions and the greater the potential for the wave to form or erode beaches and coastal cliffs. The direction of the wave is the way in which the wave comes to the observer.

In practice, it is difficult to measure these variables because the waves of different wavelengths, heights and directions can mix and produce very confusing wave patterns. Scientists and engineers use sophisticated calculations to solve the parts of the waves and produce three common summarization calculations: critical wavelengths (H_s), wavelength (T_p), and wave direction (θ_m). These three figures are then used to describe the weather of the waves, just as temperature, rain, wind speed and direction can be used to describe the local climate. Commercialization and distribution of wave energy technology will require not only addressing positive and regulatory issues, but also overcome technological challenges, one of which can provide accurate predictions of energy production. The need for any prediction is that the model that is properly represented is developed, measured and validated. In addition, the model must be able to run fast and include the correct prediction details in its predictions. A mechanical framework for this skill is developed here.

Because wave models can be awfully expensive, a new method of machine learning [2, 3, 4] is being developed here. The purpose of this approach is to train machine learning models in the more realistic model of wave-based physics forced by atmospheric and ocean history conditions to accurately represent wave conditions (in particular, significant wavelengths and feat). Computer costs are often a major limitation of real-time forecasting systems [6, 7]. Here, we use machine learning techniques to predict significant wave height by taking the predictor and predict and variable into account from the dataset. While machine learning were used to predict wave conditions [8, 9, 10, 11, 12, 13], it has not been used in the context of a surrogate model which can obtained highest accuracy with lowest root mean squared error as defined below.

2 Wave modeling

2.1 Numerical Model

The Simulating WAves Nearshore (SWAN) code FORTRAN is a standard industrial tool developed at Delft University of Technology that incorporates wave fields in coastal waters forced by wave conditions at natural boundaries, oceans, and winds [14]. SWAN mimics the energy contained in the waves as they travel in the ocean and disperse ashore. Specifically, data on the surface of the ocean contains a wave-variance spectrum, or energy density $E(\sigma, \theta)$, and these wavelengths are still distributed over wavelengths (as seen in the unused frame of the current speed reference) with distribution directions common to rotate the stems of each spectral object.

The bulk of the action is defined as $N = E/\sigma$, which is saved during the distribution along the wave element before the current one. The appearance of $N(x, y, t; \sigma, \theta)$ in space, x, y, and time, t, is governed by the action balance equation [15, 16]:

$$\frac{\partial N}{\partial t} + \left(\frac{\partial c_x N}{\partial x} + \frac{\partial c_y N}{\partial y}\right) + \left(\frac{\partial c_\sigma N}{\partial \sigma} + \frac{\partial c_\theta N}{\partial \theta}\right) = \frac{S_{tot}}{\sigma} \tag{1}$$

The left side represents the kinematic part of the equation. The second term (parent) describes an increase in the wavelength of a wave in the opposite direction of the Cartesian space where the c is wave wave. The third term represents the effect of a change in radian frequency due to differences in water depth and current mean. The fourth term presents a deeper reflection and current practice. Maximum c_σ and c_θ distribution speed in the spectral space (σ, θ). The right-hand side represents the dynamic sources of space and the sinking of all body processes that produce, disperse, or disperse the wave energy (i.e., wave growth through air, offline power transmission through three or four wave interactions, and wave decay due to white extinguishing, collision, and depth).

Haas et al. [5] define wave power consumption as a function of the critical wavelength, Hs and time wavelength, T. This information can be used to calculate wave power. Therefore, the time limit of T and, in particular, Hs because J is proportional to the wavelength, is necessary to predict the intensity of the wavelength.

3 Machine learning

3.1 Proposed method

Supervised machine learning regression models are tested to perform tasks of predicting significant wave height. Support-vector Regression (SVR) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for regression (Hs prediction), or other tasks like outlier's detection. The function used to map a lower dimensional data into a higher dimensional data though kernel. Two parallel lines drawn to the two sides of Support Vector with the error threshold value, (epsilon) are known as the boundary line. These lines create a margin between the data points.

3.2 Background

Python toolkit SciKit-Learn [34] was used to access high-level programming interfaces to machine learning libraries and to cross validate results. Machine learning have shown the greatest potential for pattern recognition in large data sets. Consider that a physics-based model acts as a non-linear function that converts input (wave signals and variable ocean currents and wind speeds) to output (spatially variable Hs). The predictor and predict and from buoy data can be collected in input vector, x, and output vector, y, respectively.

Because the purpose of this effort is to develop a framework of machine learning to effectively predict Hs from buoy data, the nonlinear function mapping inputs to the best representation of outputs, $\underline{\widehat{y}}$, is sought:

$$g(\underline{x}; \underline{\underline{\Theta}}) = \underline{\widehat{y}}. \tag{2}$$

The machine learning sufficiently trained model provides a mapping matrix, $\underline{\underline{\Theta}}$, which is a machine learning data model driven by vector-matrix functions included in (3).

The Python Toolkit SciKit-Learn [21] has been used to access high-level frameworks for cross-validation results and python machine learning libraries. Machine learning SVR model is used considering the root mean squared error, less training data for better prediction and is faster to compute output (more on this later).

3.3 Training dataset

Cairns wave monitoring of Queensland Coastal weather Observation data from Datawell 0.7 m Waverider Buoys were downloaded. Measured and derived wave parameters from data collected by a wave monitoring buoy anchored at Cairns (1 Jan 2020 to May 2020). The dataset has six fields: *Hs, Hmax, Tz, Tp, Di_ TpTrue* and *SST*. These fields are defined in table (2). There were 130 occasions when data from wave monitoring buoy at cairns were missing. Those missing values were deal using feature engineering by replacing them with the average values. These data were compiled into $\underline{\underline{X}}$ vectors. In total, the design matrix $\underline{\underline{X}}$ has 4,369 rows and 7 columns.

Table 1. Dataset fields (Attributes)

Field	**Definition**
Hs	Significant wave height, an average of the highest third of the waves in a record (26.6 minute recording period)
Hmax	The maximum wave height in the record The zero upcrossing wave period
Tz	The zero upcrossing wave period
Tp	The peak energy wave period
Dir_Tp TRUE	Direction (related to true north) from which the peak period waves are coming from
SST	Approximation of sea surface temperature
Date_Time	The Date and time of the record

For SVR algorithms, $\underline{\underline{Y}}$ is composed of the 4,369 model runs (rows), each of which contains 7 attributes (columns) defining the Hs field.

Note that in practice, data on design matrices is pre-processed. Specifically, $\underline{\underline{X}}$ undergoes a generalized global variable (e.g., all existing members are measured so that their total distribution is Gaussian with zero mean and unit variance). Here, no pre-processing of SVR's $\underline{\underline{Y}}$ is required.

Data $\underline{\underline{X}}$ and $\underline{\underline{Y}}$ were randomly divided into two groups to form a training data set consisting of 90% of 4,369 rows of data and test data sets the remaining 10%. The mapping matrix is calculated using training data and then used in the test data set and RMSE between the vector of the test data, $\underline{y}$, and its machine learning representation, $\underline{\widehat{y}}$ is calculated.

The SVR algorithm needs to be supplied only by $\underline{\underline{X}}$ and the vector of the Hs value column compiled as y. Data were further subdivided into two groups with 90% of $\underline{x}$ vector randomly assembled in training dataset and reserved the remaining for testing. The SVR model returns three files; the first describes the normal change applied to $\underline{x}$, the dot product taken with the mapping matrix $\underline{\underline{\Theta}}$ described in the second file, and the third file is used to convert $\underline{\widehat{y}}$ back to the characteristic Hs.

3.4 Support vector regression model

In training data set Xn is a multivariate set of N observations with Yn response value observed. To find the linear function (4) and make sure it is as flat as possible, find $f(x)$ having minimal norm value

$$f(x) = x'\beta + b \tag{3}$$

(β', β). This is constructed as a convex optimization problem to minimize (5) subject to all residuals

$$J(\beta) = \frac{1}{2}\beta\beta' \tag{4}$$

having a value less than ε; or, in equation form (6):

$$\forall n : |y_n - (x_n'\beta + b)| \leq \varepsilon. \tag{5}$$

It is possible that no such function $f(x)$ exists to satisfy the constraints of all points. To deal with impossible obstacles, enter the slink ξn and $\xi * n$ variables for each point. This approach is similar to the concept of "soft margin" in SVM segmentation, because the flexible flexibility allows regression errors to exist until the ξn and $\xi * n$ values, but still satisfy the required conditions.

The inclusion of slack variables leads to the primal formula, also known as the objective function [25]:

$$J(\beta) = \frac{1}{2}\beta\beta' + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \tag{6}$$

Subject to:

$$\forall n : y_n - (x'_n\beta + b) \leq \varepsilon + \xi_n \tag{7}$$

$$\forall n : (x'_n\beta + b) - y_n \leq \varepsilon + \xi_n^* \tag{8}$$

$$\forall n : \xi_n^* \geq 0 \tag{9}$$

$$\forall n : \xi_n \geq 0 \tag{10}$$

Constant C is the limit of the box, a positive numerical value that controls the penalty placed on the observation which lies outside the epsilon margin (ε) and helps prevent overfitting. This value determines the trade-off between $f(x)$ fatness and the value until the deviation greater than ε is tolerated.

The linear loss function of ε-insensitivity ignores errors that are in ranges of ε distance of the observed values by considering them as equal to zero. Loss is measured based on the distance between the observed value y and the ε boundary. This is described by

$$L_\varepsilon = \begin{cases} 0 \quad \text{If} \quad |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon \quad \text{Otherwise} \end{cases} \tag{11}$$

In most of the linear regression models, the objective is to minimize the sum of squared errors. For example, take Ordinary Least Squares (OLS). For OLS with one predictor (Maximum wave height) the objective function is as follows:

$$MIN\sum_{i=1}^{n}(y_i - w_i x_i)^2 \tag{12}$$

Where y_i is the target, w_i is the coefficient, and x? is the predictor (Maximum wave Height).

Ridge, Lasso and ElasticNet are all extensions of this simple equation, with an additional penalty parameter, Correlation-based Feature Selection (CFS), that aims to minimize complexity and reduce the number of features used in the final model. The aim is to reduce the error of the test set.

In contrast to OLS, the SVR's objective function is to reduce coefficients – in particular, the $l2$-norm of the vector coefficient – not the squared error. The term error is rather handled in constraints, where we set the absolute error below or equal to the specified margin, called the maximum error, ϵ (epsilon). We can tune the epsilon to get for our model the desired accuracy. The new objective function and constraints for our model are as follows:

$$\begin{aligned} &\text{Minimize } \ \frac{1}{2}\|w\|^2, \\ &\text{Subject to } \ |y_i - \langle w, x_i \rangle - b| \leq \varepsilon \end{aligned} \tag{13}$$

Where x_i is a training sample with target value y_i.

The inner product plus intercept $\langle w, x_i \rangle - b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold. The Kernel applied here is RBF(Radial basis function) due to non-linearity in the data set.

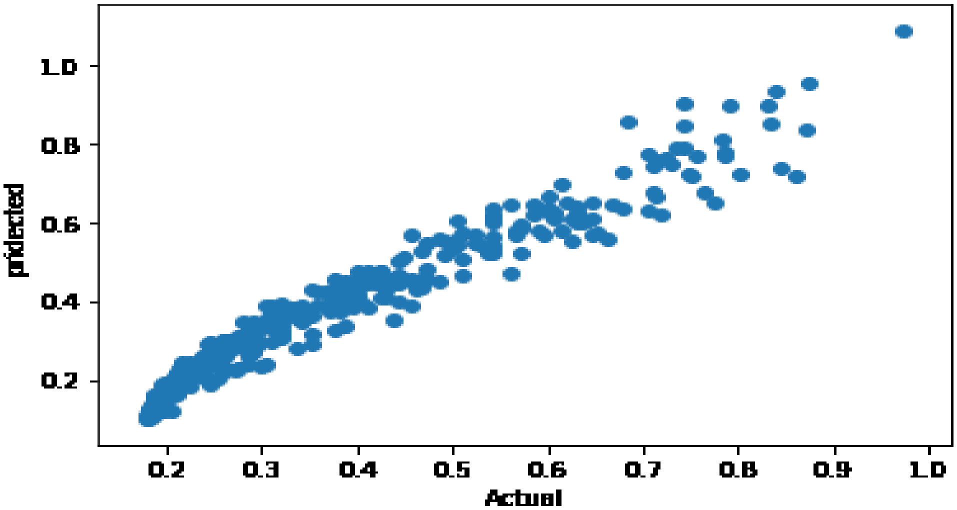

Figure 1: Scatter plot showing the actual and predicted values for Hs. The Swan model values horizontally on X-axis and SVR predicted values are represented by Y-axis vertically.

The SVR model effectiveness was evaluated according to the accuracy percentage in predicting Hs value. In the SVR results no bias was observed and 95.7% of the time correctly predicted the characteristic Hs in the test data set. The scattered plot in Figure 2 visualize the characteristic Hs from SWAN and the SVR representation which revel that there is no outlier found with the final model.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers based on a training sample. To analyze the performance of SVR, the model was trained on 90% of the dataset and the remaining 10% was allocated as test data. The accuracy of SVR was 95% on test dataset with a root mean squared error of 0.044.

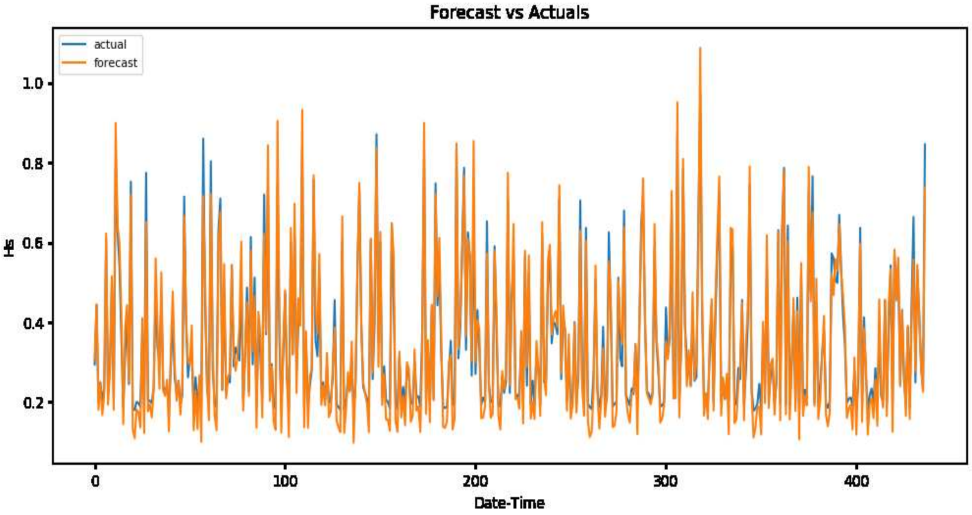

Figure 2: Plot visualizing the actual and forecasted values for Hs against time series. Blue indicate the actual Swan model Hs while the orange represents SVR predicted Hs.

4 Discussion

Advanced machine learning models have been developed here to create improved mapping matrix (or vector) and pre- and post-processor functions, to predict significant wave height. Instead of the historical data used to create an input vector, x, now the weather data can be used. In order to work with the weather mode, the data from buoy are used, both the predictor and predicant from the same data, to train the machine learning algorithm to form Hs field. Such data is part of the Marine Information System which is the state of WAVEWATCH III-predictable waves conditions available for the next 10 days. Also, forecasts for ROMS-simulated ocean-currents and wind forecast are available for the next 48 hours from CeNCOOS and The Weather Company [19], respectively. For the Cairns wave the historical data is available on Queensland Coastal weather Observation [24] for data taken from Datawell 0.7m Waverider Buoys.

The execution of machine learning models quickly produces the Hs field. Computationally, this only need a multiplication of the $L+1$ matrix. In fact, for a 24-hour forecast, on a single-core processor the machine-learningSVR took 0.044 s to calculate the Hs field | well over three orders of magnitude (485,833%) faster than the running the full physics based models. In fact, performance that requires a lot of wall clock time loads metrics files for memory.

The machine learning models presented here are specific to the Queensland coast cairns region and will need to be re-training to apply to other locations. Of course, using a physics-based model on a new site requires the creation of a grid and the integration of all the boundary and conditions of coercion with all the efforts of the server. However, the important thing is that the framework needed to develop this technology is introduced for the first time for wave modeling in 2017. As expected [22], the data-centric modeling machine learning approaches has grown increasingly common in last few years and are expecting to grow in near future rapidly.

5 Conclusion

An improved version of machine learning models has been developed with new approach, as computationally efficient to predict Hs fields. From supervised training of machine learning models determined appropriately trained mapping matrices, give in representations of similarly accurate Hs in the domain of interest. Thus, this approach of machine learning models can contribute to a fast and efficient wave-condition forecast system. The power-generation potential of WECs or surf conditions can be estimated using these forecasted wave conditions. Ultimately, it is envisioned that such improved version of machine learning models which don't required many parameters for accurate prediction could be installed locally on a WEC thereby making their own forecast system. In addition, the buoy itself can collect wave-condition data that can be used to update machine learning models. As machine learning technology advances, they can be adapted to integrate the continuous distribution of real-time data collected locally with predictions available to change and improve the parameters of the machine learning model. In fact, such methods have already been widely used "online learning" [23].

The approach previously proposed by author to predict characters Hs using MLP requires a large amount of data and more calculation to form mapping matrix due to lack of an important parameter maximum wave height, which is considered in this work. This parameter when used as a predictor gives better forecasting with low calculation cost and high model efficiency. Additional efforts are currently underway using ensemble machine learning approaches to predict Hs and Wave period T. The results are expected to further improve the process of wave characteristics prediction and take into account how bathymetry effects wave heights.

References

[1] FAO, The State of the World Fisheries and Aquaculture 2016. Con tributing to the food security and nutrition for all *Technical Report, Food and Agriculture Organization of the United Nations* (2016).

[2] I. Goodfellow, Y. Bengio, A. Courville, "Deep Learning" , *MIT Press,* (2016).

[3] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning" , *Nature* 521 (2015): 436-444.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview" , *Neural Networks* 61 (2015): 85-117.

[5] K. Haas, S. Ahn, V. S. Neary, S. Bredin, "Development of a wave energy resource classification system, in: Waterpower Week" , *METS, Washington, DC,* 1-5.

[6] P. M. DeVries, T. B. Thompson, B. J. Meade, "Enabling large-scale viscoelastic calculations via neural network acceleration" , *Geophysical Research Letters* 44 (2017): 2662-2669.

[7] V. Mallet, G. Stoltz, B. Mauricette, "Ozone ensemble forecast with machine learning algorithms" , *Journal of Geophysical Research: Atmospheres* 114 (2009).

[8] D. Peres, C. Iuppa, L. Cavallaro, A. Cancelliere, E. Foti, "Significant wave height record extension by neural networks and reanalysis wind data" , *Ocean Modelling* 94 (2015): 128-140.

[9] O. Makarynskyy, "Improving wave predictions with artificial neural networks" , *Ocean Engineering* 31 (2004): 709-724.

[10] A. Etemad-Shahidi, J. Mahjoobi, "Comparison between M5' model tree and neural networks for prediction of significant wave height in lake superior" , *Ocean Engineering* 36 (2009): 1175-1181.

[11] J. Mahjoobi, A. Etemad-Shahidi, "An alternative approach for the prediction of significant wave heights based on classification and regression trees" , *Applied Ocean Research* 30 (2008): 172-177.

[12] M. Browne, D. Strauss, B. Castelle, M. Blumenstein, R. Tomlinson, C. Lane, "Empirical estimation of nearshore waves from a global deep-water wave model" , *IEEE Geoscience and Remote Sensing Letters* 3 (2006): 462-466.

[13] M. Browne, B. Castelle, D. Strauss, R. Tomlinson, M. Blumenstein, C. Lane, "Near-shore swell estimation from a global wind-wave model: Spectral process, linear, and artificial neural network models" , *Coastal Engineering* 54 (2007): 445-460.

[14] The SWAN Team, SWAN Scientific and Technical Documentation *Technical Report SWAN Cycle III version 40.51, Delft University of Technology* (2006).

[15] G. J. Komen, L. Cavaleri, M. Donelan, "Dynamics and Modelling of Ocean Waves" , *Cambridge University Press* (1996).

[16] C. C. Mei, M. Stiassnie, D. K.-P. Yue, "Theory and Applications of Ocean Surface Waves: Part 1: Linear Aspects. Part 2: Nonlinear Aspects" , *World Scientific* (1989).

[17] Y. Song, D. Haidvogel, "A semi-implicit ocean circulation model using a generalized topography-following coordinate system" , *Journal of Computational Physics* 115 (1994): 228-244.

[18] J. Patterson, J. Thomas, L. Rosenfeld, J. Newton, L. Hazard, J. Scianna, R. Kudela, E. Mayorga, C. Cohen, M. Cook, et al., "Addressing ocean and coastal issues at the west coast scale through regional ocean observing system collaboration, in: Oceans'12" , *IEEE* 1-8.

[19] The Weather Company *The Weather Company* (2017).

[20] G. Chang, K. Ruehl, C. Jones, C. C. Roberts, J. D.and Chartrand, "Numerical modeling of the effects of wave energy converter characteristics on nearshore wave conditions" , *Renewable Energy* 89 (2016): 636-648.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "SciKit-Learn: Machine learning in Python" , *Journal of Machine Learning Research* 12 (2011): 2825-2830.

[22] Scott C James, Yushan Zhang, Fearghal O'Donncha, "See discussions, A Machine Learning Framework to Forecast Wave Conditions" , *Coastal Engineering* (2017): 46556-5637.

[23] N. Cesa-Bianchi, A. Conconi, C. Gentile, "On the generalization ability of on-line learning algorithms" , *IEEE Transactions on Information Theory* 50 (2004): 2050-2057.

[24] Queensland Government Coastal Data System - Near real time wave data, open data portal https://www.data.qld.gov.au/dataset/coastal-data-system-near-real-time-wave-data

[25] Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, New York, 1995).