Scientific Journals
of the Maritime University of Szczecin

Zeszyty Naukowe
Akademii Morskiej w Szczecinie

2017, 49 (121), 119–129
ISSN 1733-8670 (Printed)
ISSN 2392-0378 (Online)
DOI: 10.17402/209

Received: 21.11.2016
Accepted: 28.02.2017
Published: 17.03.2017

Data mining models to predict ocean wave energy flux in the absence of wave records

Kumars Mahmoodi, Hassan Ghassemi✉, Hashem Nowruzi

Amirkabir University of Technology, Department of Maritime Engineering
Hafez Ave, No 424, P.O. Box 15875-4413, Tehran, Iran
✉ corresponding author: e-mail: gasemi@aut.ac.ir

Key words: ocean wave energy, meteorological parameters, GEP, LDBOD, DMM, modeling

Abstract

Ocean wave energy is known as a renewable energy resource with high power potential and without negative environmental impacts. Wave energy has a direct relationship with the ocean's meteorological parameters. The aim of the current study is to investigate the dependency between ocean wave energy flux and meteorological parameters by using data mining methods (DMMs). For this purpose, a feed-forward neural network (FFNN), a cascade-forward neural network (CFNN), and gene expression programming (GEP) are implemented as different DMMs. The modeling is based on historical meteorological and wave data taken from the National Data Buoy Center (NDBC). In all models, wind speed, air temperature, and sea temperature are input parameters. In addition, the output is the wave energy flux which is obtained from the classical wave energy flux equation. It is notable that, initially, outliers in the data sets were removed by the local distribution based outlier detector (LDBOD) method to obtain the best and most accurate results. To evaluate the performance and accuracy of the proposed models, two statistical measures, root mean square error (RMSE) and regression coefficient (R), were used. From the results obtained, it was found that, in general, the FFNN and CFNN models gave a more accurate prediction of wave energy from meteorological parameters in the absence of wave records than the GEP method.

Nomenclature

DDMs	Data Mining Methods
FFNN	Feed-Forward Neural Network
CFNN	Cascade-Forward Neural Network
GEP	Gene Expression Programming
NDBC	National Data Buoy Center
LDBOD	Local Distribution Based Outlier Detector
OFV	Outlier Feature Vector
RMSE	Root Mean Square Error
n	The total number of instances in the data set
R	Regression Coefficient
p, q, o	Some data points in the data set
p^*	The image point of p
$d(p, q)$	The distance between points p and q
k	Number of neighbors
$N_k(p)$	The k-distance neighborhood of p
$\|N(p)\|$	The number of instances located in $N_k(p)$
PSD(p,o)	The point symmetry distance between object p and o
σ	The tuning parameter
T_i	Measured wave energy corresponding to instance i
O_i	Predicted wave energy corresponding to instance i
$\overline{T}$	Average of measured wave energies corresponding to all instances
$\overline{O}$	Average of predicted wave energies corresponding to all instances

Introduction

Interest in renewable energy sources has seen a recent dramatic increase. This is due partly to pollution, and partly because sources of fossil energy are limited. Wave energy is one of the most interesting areas of renewable energy sources for scholars

(Ming & Aggidis, 2008; Cornejo-Bueno et al., 2016; Kamranzad et al., 2016; Minh Tri et al., 2016; Sannasiraj & Sundar, 2016).

Wave energy can be used for various purposes, such as the generation of electricity. To convert ocean energy into electricity, wave energy converters (WECs) are used (Falcao, 2010). The energy extracted from waves is sensitive to the type of WEC used and its location in the marine environment. Wave energy output depends on water density, wave group celerity, and wave height, while wave height is related to wind speed, duration of the wind, and fetch length. Group celerity depends on the wave period and water depth and the gradients of air pressure generate the wind. Given this, the most important parameter in characterizing wave energy is wave height. It is possible to make a wave energy assessment when wave measurements are lacking by using meteorological data such as wind speed and air and sea temperatures instead of wave data (Özger, 2011). Therefore, the purpose of the current paper is to investigate the relationship between ocean wave energy flux and meteorological parameters by using some well-known DMMs. These methods allow an estimation of the amount of wave energy wherever meteorological information is available. To this end, a FFNN, a CFNN, and GEP are the DMMs selected for the current study.

Data mining is the process of discovering and revealing previously unknown, hidden, meaningful, and useful patterns in databases (Fayyad, Shapiro, & Smyth, 1996). It has arisen from the intersection of machine learning, pattern recognition, statistics, database management systems, intelligent systems, and data visualization. Data mining is widely used in many scientific fields. Examples of data mining applications in renewable energy research are presented in Table 1. In our modeling, meteorological and wave data from the NDBC are used.

Different parameters may lead to outliers in studied data sets. For example, outliers may occur due to an error in the measurements. Outliers in the data sets must be detected before modeling to create models with higher accuracy. In the present study, the LDBOD method is applied to detect outliers. LDBOD is a powerful data-mining method used to detect outliers in multi-dimensional data sets.

Table. 1. Different applications of data mining in research in the field of renewable energies

Application	Models used in the study	Ref.
Prediction of significant wave height and energy flux	Genetic Algorithm – Extreme Learning Machine approach (GA-ELM)	(Cornejo-Bueno et al., 2016)
Prediction of sea wave energy	Fuzzy logic, Artificial Neural Network (ANN)	(Özger, 2011)
Prediction and optimization of wave energy converter arrays	Active learning, Genetic Algorithm (GA), Gaussian process	(Sarkar et al., 2016)
Prediction of the performance of solar chimney power plants	ANN, Adaptive Neuro Fuzzy Inference System (ANFIS)	(Amirkhani et al., 2015)
Assessment of solar energy potential	ANN, J48 algorithm	(Yadav & Chandel, 2015)
Wind power prediction	Decision trees, Support Vector Regression (SVR)	(Heinermann & Kramer, 2016)
Fault diagnosis technique for photovoltaic systems	ANN	(Chine et al., 2016)
Optimization of biodiesel engine performance	Kernel-based Extreme Learning Machine, Cuckoo search	(Wong et al., 2015)
Fault diagnosis for a wind turbine transmission system	Orthogonal Neighborhood Preserving Embedding (ONPE), Shannon wavelet support vector machine	(Tang et al., 2014)
Environmental data processing	k-means clustering	(Di Piazza et al., 2011)
Time series prediction	Artificial Wavelet Neural Network	(Doucoure et al., 2016)
Estimation of the daily global solar radiation	Linear Autoregressive Moving Average (ARMA), ANN	(Gairaa et al., 2016)
Placement of wind turbines	Gas	(Grady et al., 2005)
Wind speed prediction	Hybrid KF-ANN	(Shukur & Lee, 2015)
Energy storage management	ANN, Adaptive learning procedures based on Bayesian approach and Gaussian approximation	(Blonbou et al., 2011)
Prediction of wind turbine faults	Neural Network (NN), Neural Network Ensemble (NN Ensemble), Boosting Tree Algorithm (BTA), Support Vector Machine (SVM)	(Kusiak & Li, 2011)

The rest of the paper is organized as follows. Section 2 is an introduction to the discussed DMMs, while data sets are presented in section 3, and wave energy calculations are summarized in section 4. The modeling experiments are presented in section 5. Finally, concluding remarks are made in section 6.

Prediction and outlier detection methods

In this section, details of the DMMs used for wave energy prediction and outlier detection are introduced. Prediction methods include a FFNN, a CFNN, and GEP. Moreover, the LDBOD method is discussed as regards outlier detection in the studied data sets.

Feed-forward Neural Network (FFNN)

FFNNs are the most popular and widely-used models in many practical applications, and are known by many different names, such as "multi-layer perceptron". FFNNs can be used for any kind of input to output mapping and consist of a series of layers. Generally, these networks contain three layers: input, hidden, and output. The first layer has a connection from the network input, and each subsequent layer has a connection from the previous layer. The final layer produces the network's output. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any), and to the output nodes. The connection between the layers is made by means of processing elements called neurons (Benardos & Kaliampakos, 2004). The role of neurons in ANNs is information processing. This process is performed by a mathematical processor called an activation function. The activation function can be linear or non-linear, and is selected by the user according to problem type. If the objective is prediction, the linear function must be used in the output layer. Each neuron is connected to its neighbors with varying coefficients called weights, in which the knowledge of an ANN is stored (Maged, Khalafallah, & Hassanien, 2004). The weights are unknown values determined by training algorithm and training data.

In addition to inputs and weights, neurons include another component, called bias. Bias values accumulate with product inputs and their respective weights. The number of neurons in the input and output layers depends on the nature of the problem. The number of neurons in the input and output layer is equal to the number of input and output variables, respectively. However, the number of neurons in the hidden layer is determined in a trial-and-error approach.

The learning algorithm is a dynamic and iterative process which consists of the modification of the network's parameters in response to the received environmental signals (Moller, 1993). The goal of learning is to minimize the error between the desired output (target) and the network output (output) (Ebrahimabadi, Azimipour, & Bahreini, 2015). The learning algorithm in a FFNN is supervised. Supervised learning is a type of learning that takes place when the training instances are labeled with the correct results; in fact, the target dataset is provided and used to train the machine and obtain the desired outputs. One of the most widely-used training algorithms is back-propagation. In the back-propagation algorithm, when each entry of the sample set is presented to the network, the network examines its output response to the sample input pattern. The output response is then compared to the known and desired output and the error value is calculated. The connection weights are adjusted according to the error. The set of these sample patterns is repeatedly presented to the network until the error value is minimized (Guillermo, 1998). A FFN with one hidden layer and enough neurons in the hidden layers can fit any finite input-output mapping problem (Salari et al., 2005). In this research, therefore, this type of network is used to predict ocean wave energy.

Cascade-forward Neural Network (CFNN)

Cascade-forward neural networks (Scott, Lebiere, & Lebiere, 1990) are similar to FFNs, but include a connection from the input and every previous layer to the following layers. A CFNN can approximate any bounded continuous function with enough hidden neurons.

Gene Expression Programming (GEP)

A GA is one of the well-known adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics (Holand, 1975). In the conventional version, chromosomes were represented as a fixed length binary string. Genetic Programming (GP) (Koza, 1992) derives from the extended version of GA, where chromosomes are represented as a LISP expression translated graphically into tree structures of different sizes. LISP (Robin, Clive, & Ian, 2012) is a family of computer programming languages based on formal functional calculus. GEP is a new evolutionary Artificial Intelligence (AI)

technique developed by Ferreira (Ferreira, 2001). This technique is an extension of GP and consists of encoded individuals as linear chromosomes of fixed length, represented by a tree structure of different sizes and shapes. In fact, the linear structure of chromosomes makes genetic operators such as recombination, mutation, and duplication constantly generate accurate and reliable constructs (Keshavarz & Mehramiri, 2015). GEP is one of the best evolutionary methods for complex, non-linear modeling that automatically creates computer programs. These computer programs can take many forms, such as conventional mathematical models, neural networks, decision trees, sophisticated nonlinear regression models, logistic regression models, and so on. It combines the advantages of both GA and GP, and removes their limitations with two elements, the chromosomes and the expression trees (ETs). The chromosome is the encoder of the candidate solution, which is then translated into an ET. Linear chromosomes are composed of genes structurally organized into a head and a tail. The information is used to generate the overall GEP model stored in the head of the gene. Terminals are stored in the tail of the gene. The tail consists of information that can be used in producing subsequent GEP models (Ferreira, 2001). In GEP, the number of genes in a chromosome can be one or more.

GEP carries out the following stages of solving a problem: 1) the process initiates (Ferreira, 2001) in tree form; and the fitness of each individual is evaluated; 3) a check is made as to whether the termination condition is satisfied or not. If it is satisfied, then the evolution stops and the program terminates with the current population displaying the favorable solution; if not, the best present population is retained; 4) the other population is chosen based on its performance, 5) certain modifications (mutation, recombination, and duplication) are made on the selected population so as to produce new children; and 6) after some of the above-mentioned operations have been applied, a new population is generated. This process is repeated for a certain number of generations or until the required accuracy is achieved (Ferreira, 2001). In the GEP system, the operators used for the genetic modification of chromosomes are explained in (Ferreira, 2006).

Local Distribution Based Outlier Detector (LDBOD)

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. Scholars have proposed many definitions for an outlier but there is seemingly no universally accepted one. In this paper, we will take the definition of Grubbs (Grubbs, 1969), quoted in Barnett & Lewis (Barnett & Lewis, 1994): an outlier observation is one that appears to deviate markedly from other members of the data set in which it occurs.

LDBOD (Zhang, Yang & Wang, 2008) is a powerful outlier detection algorithm. It detects local outliers from the viewpoint of local distribution, which is characterized through three proposed measurements: local-average-distance, local-density, and local-asymmetry-degree. Details of LDBOD are given below.

At the outset, it is necessary to construct a neighborhood diagram among all the data points. Here a kNN diagram (Lee, 1982) is used. Local distribution needs to be quantified with some specific measurements. p and q are some data points in the data set. Also, $d(p, q)$ is representative of the distance between points and q (in this research, Euclidean distance). Moreover, $N_k(p)$ denotes the kNN neighborhood (Breunig et al., 2000). In this regard, certain definitions are presented below:

Definition 1. The local-average-distance of is defined as

$$\frac{1}{|N(p)|}\sum_{q\in N_k(p)} d(p,q).$$

Definition 2. Given the local neighborhood $N_k(p)$ of an object p, the local-density of p is defined as the distance between p and its k-th neighbor, that is, local-density $(p) = \max_{q\in Nk(p)} d\,(p, q)$.

Definition 3. The point symmetry distance between object p and o is defined as the distance between p^* and the nearest neighbor of p^* in $N_k(o)$, where p^* is the image point of p with respect to object o, i.e., $\mathrm{PSD}(p, o) = \min_{q\in Nk(o)} d\,(p^*,q)$.

Definition 4. The local asymmetry-degree of p can be defined as the weighted average of the point symmetry distances between the neighbors of p in $N_k(p)$ and p, that is, local-asymmetry-degree

$$(p) = \frac{1}{|N_k(p)|}\sum_{q\in N_k(p)} w(q) * \mathrm{PSD}(q,p).$$

Also, $w(q) = e^{d(p,q)/\sigma}$, where σ is a predefined tuning parameter. There is no general guideline for the selection of this parameter. In this research, based on the nature of the data, we set it at 100 intuitively.

Definition 5. An object p is an outlier if it is labeled as an outlier through the 2-class clustering analysis performed over the Outlier Feature Vectors (OFVs) of the data set. Every object p can be represented as

a 3-dimensional feature vector of local-average-distance (p), local-density (p), local-asymmetry-degree (p). We refer to this feature vector as an OFV. The different clustering algorithms can be used for OFV clustering. We consider utilizing Fuzzy C-means (FCM) (Chiu, 1994) as our clustering algorithm due to its efficient computation and small storage requirement.

Performance evaluation

To evaluate the performance and accuracy of our intended models in terms of the measured and predicted values, RMSE and R were employed, according to the following equations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - O_i\right)^2} \tag{1}$$

$$R = \frac{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2 \sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}} \tag{2}$$

Here, n represents the total number of instances, while T_i and O_i are representative of experimental and predicted values using models, respectively. Moreover, $\overline{T}$ and $\overline{O}$ are the average of the mentioned data.

Data sets

The historical meteorological and wave data were taken from the NDBC (http://www.ndbc.noaa.gov). Two standard meteorological data stations were used for our modeling. Table 2 shows the main characteristics of the two buoys considered and dataset locations are illustrated in Figure 1. All historical data were collected in the year 2015, and a subset in the year 2014. In order to apply the DMMs, it was necessary to divide the data into training and testing sets. Herein, data for one complete year from station 44009 was selected for the training set, and a random subset data from station 42058 was used for the test set. Before modeling, missing values and outliers must be detected and then removed from data sets. This process increases the accuracy of the created models. Missing values are easily detectable, but it is difficult to detect outliers. In this research, first the missing values were removed, and then outliers were detected and removed. To implement the LDBOD method, we had to determine the value of parameter k (number of neighbors). However, how to select this parameter depends on the nature of the data and can be determined through trial and error. In the current study, in general and for a good interval confidence, $k = 200$ was selected for all the data sets. Table 3 shows the number of detected outliers after removing missing values in the studied stations. After cleaning the data sets, the total number of training and test data obtained were 3884 and 2488, respectively. The details of the predictive variables in all the studied stations are displayed in Table 4.

Table 2. Geographic coordinates and buoy description (NDBC site[1])

Characteristics	Station 44009 (38°27'40" N 74°42'9" W)	Station 42058 (14°55'23" N 74°55'4" W)
Site elevation	sea level	sea level
Air temperature height	4 m above site elevation	4 m above site elevation
Anemometer height	5 m above site elevation	5 m above site elevation
Barometer elevation	sea level	sea level
Sea temperature depth	0.6 m below water line	0.6 m below water line
Water depth	30.5 m	4161 m
Watch circle radius	63.1 m	4344.3 m

Figure 1. Buoys considered in this study

Table 3. Detected outliers in the studied stations

Station ID	Number of samples	Number of detected outliers
44009	3934	50
42058	2509	21

Wave energy flux can be obtained using the following deep-water expression (Fernández et al., 2015; Cornejo-Bueno et al., 2016; Sierra et al., 2016):

$$P = \frac{\rho g^2}{4\pi}\int_0^\infty \frac{S(f)}{f}\mathrm{d}f = \frac{\rho g^2}{4\pi} m_{-1} =$$

$$= \frac{\rho g^2}{4\pi} H_s^2 T_e = 0.491 H_s^2 T_e \tag{3}$$

Table 4. Predictive variables statistics corresponding to all studied stations

Predictive variable	Unit	Max		Min		Mean		Std	
		ST. 44009	ST. 42058	ST. 44009	ST. 42058	ST. 44009	ST. 42058	ST. 44009	ST. 42058
Wind speed (WSPD)	[m/s]	18.70	15.10	0.00	0.30	6.26	8.12	3.30	2.02
Significant wave height (WVHT)	[m]	6.11	3.92	0.27	0.43	1.21	1.46	0.80	0.48
Atmospheric pressure (PRES)	[hPa]	1042.4	1017.70	1000.4	1006.70	1018.46	1012.12	6.90	1.91
Air temperature (ATMP)	[°C]	27.70	29.70	2.90	24.70	18.64	27.91	5.70	1.03
Water temperature (WTMP)	[°C]	27.90	30.30	12.00	26.00	19.73	28.20	5.07	0.98

where, P is the wave energy flux (or power density per meter of wave crest) in kW/m. H_s is the significant wave height (i.e., defined as the average of the highest one-third of waves). The sea is composed of many random waves of different lengths and heights. It is not possible to consider all these waves at the same time for design or research purposes. For this reason, a wave that represents all of them will be considered, this being the significant wave height. T_e is the energy period. The parameters ρ and g are density of seawater, which is assumed to be 1025 kg/m^3, and gravitational acceleration, respectively. As suggested by (Boronowski et al., 2010), a conservative value of $T_e = 0.9T_p$ was used to assess the wave energy resource, where T_p is peak period.

Results and discussion

In this section the results of FFNN, CFNN, and GEP DMMs on the data sets is presented. All methods were implemented in MATLAB software, with the exception of the GEP method, which was modeled on *GeneXproTools* 5.0 software (Ferreira, 2001). In all models, input parameters are wind speed, atmospheric pressure, air temperature, and water temperature with output wave energy flux (see Eq. (3)).

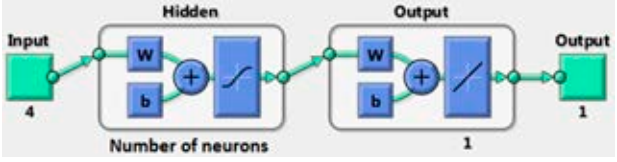
Feed-forward (FFN) and Cascade Neural Network (CNN)

Since, according to Bishop's (Bishop, 1995) study, more than one hidden layer is often unnecessary, our architectures have only one hidden layer. All the used networks are trained using a back-propagation algorithm with gradient descent and momentum terms. A neural network must be learned by network parameters before utilization. The characteristics of the ANNs employed in this study are presented in Table 5, while Figure 2 represents the schematic of defined FFNs and CNNs. To avoid over-fitting, each dataset was randomly split into three sets: 70% for model training (to compute the gradient and updating of the network parameters, such as weights and biases); 15% for model testing; and 15% for validating. The model weights were randomly initialized.

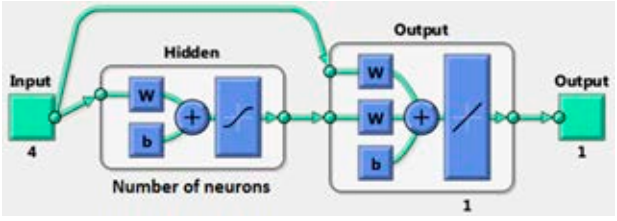
A neural network is a random process and in each run, may produce different results under the same conditions. Therefore, different networks were created to achieve the best model. To create different networks, the numbers of hidden neurons varied from five to 15, and other conditions are considered as being the same. For each neuron, 30 networks, with a total of 330 networks, were created, and the best result, with the least RMSE and maximum R^2, is shown in dark gray in Table 6. For both networks, the best result was obtained for 15 neurons.

Table 5. The characteristics of selected neural networks

Training subset	70% of dataset
Validation subset	15% of dataset
Test subset	15% of dataset
Number of input layer neurons	4
Number of output layer neurons	1
Number of hidden layer neurons	Varied from 5 to 15
Hidden layer activation function	Hyperbolic tangent
Output layer activation function	Linear
Training algorithm	Levenberg-Marquardt
Maximum number of training epochs	1000



FFNN



CFNN

Figure 2. Schematic of defined neural networks of a) FFNN and b) CFNN

Table 6. Results of neural network implementation

Number of neurons	FFNN		CFNN	
	RMSE (kW/m)	R^2	RMSE (kW/m)	R^2
5	5.30	0.89	5.34	0.89
6	5.01	0.90	5.03	0.90
7	4.90	0.91	5.16	0.90
8	4.64	0.92	4.86	0.91
9	4.67	0.92	4.90	0.91
10	4.68	0.92	4.47	0.92
11	4.48	0.92	4.70	0.91
12	4.53	0.92	4.62	0.92
13	4.29	0.93	4.43	0.92
14	4.47	0.92	4.41	0.92
15	4.24	0.93	4.37	0.93

In general, for training data, the FFNN performance was better than that of the CFNN, although the difference is negligible. It was also observed that, with an increase in the number of neurons, the accuracy of the models did not significantly increase. The regression plot and error histogram for the best obtained models by a FFNN and CFNN are presented in Figures 3 and 4, respectively. As is clear from Figures 3 and 4, the accuracy of the resulting models is acceptable, because in Figure 3 most data are distributed around the bisector exact model line. This means that the created networks were able to estimate the nonlinear relationship between the meteorological and wave energy flux with reasonable accuracy.

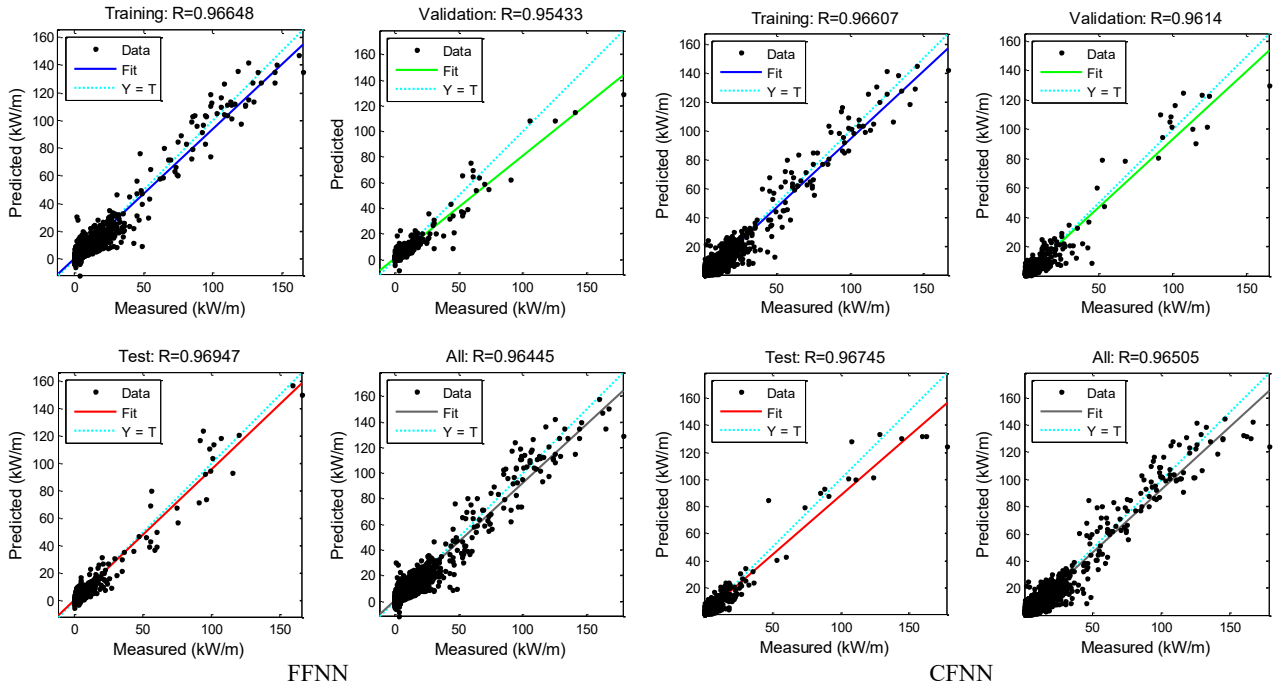


Figure 3. The measured wave energy flux versus predicted values for the best obtained model by FFNN and CFNN

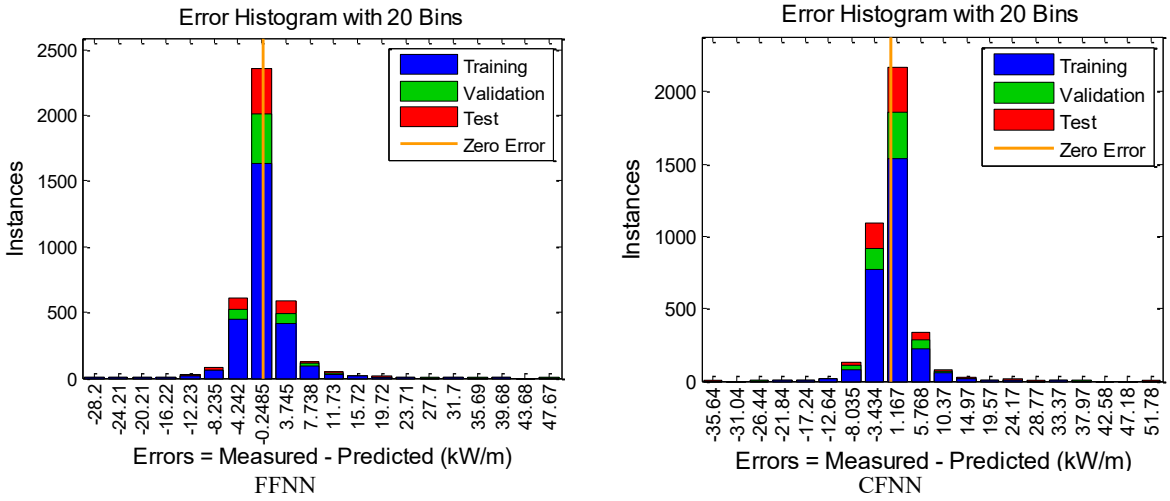


Figure 4. Error histogram for the best obtained model by FFNN and CFNN

Gene Expression Programming (GEP)

In GEP modeling, parameters should be defined. These include fitness function, the set of terminals *T*, and the set of functions *F* to create the chromosomes, chromosomal architecture, i.e., the length of the head and the number of genes and chromosomes, linking function, set of genetic operators, and their rates. There is no complete information regarding how to choose appropriate GEP parameters. Parameter values are usually determined by trial and error. The parameters used in the GEP models are given in Table 7; other parameters were set to default values in *GeneXproTools* 5.0 software. In our modeling, 70% of the training data set were implemented for the training phase, and the rest of the data were used for the test phase. The results of the GEP implementation on the training data are provided in Table 8. The accuracy of all models is almost the same overall, but their results may be different as regards the test data. In general, the performance of model 1 is better than other models. The regression plot and error histogram of training data for model 1 is presented in Figures 5 and 6.

Comparison between models

After creating models, their ability to estimate wave energy flux must be measured in other than training data. A model providing a more accurate estimation of new data has more functionality. For this purpose,

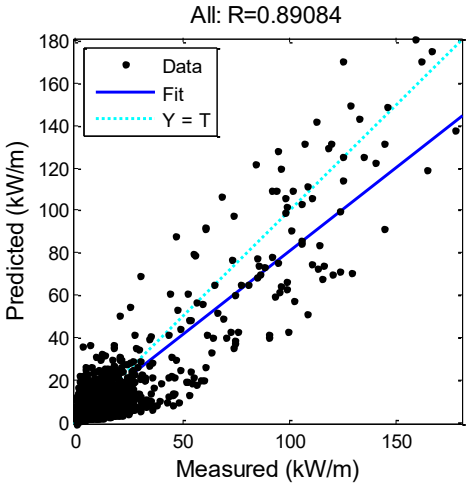


Figure 5. Measured wave energy flux versus predicted values for GEP model 1

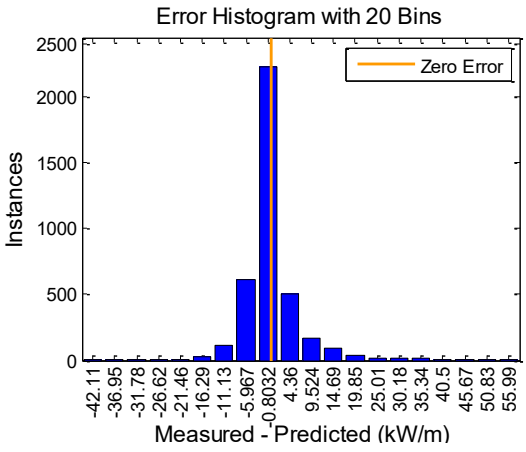


Figure 6. Error histogram for the GEP model 1

Table 7. Selected parameters for the GEP models

GEP parameters	Model 1	Model 2	Model 3	Model 4
Function set	+, _, *, /, ^, Exp, Ln, Log, Sqrt	+, _, *, /, ^	+, _, *, /, Exp, Ln, Sin, Cos	+, _, *, /, ^, Exp, Ln, Log, Sqrt, Sin, Cos,
Terminal set	WSPD, PRES, ATMP, WTMP	WSPD, PRES, ATMP, WTMP	WSPD, PRES, ATMP, WTMP	WSPD, PRES, ATMP, WTMP
Number of chromosomes	25	25	30	28
Number of genes	6	6	7	5
Head size	8	8	8	6
Linking function	Addition	Addition	Addition	Addition
Mutation rate	0.00138	0.00138	0.02	0.1
Gene recombination rate	0.00277	0.00277	0.003	0.1
One-point recombination rate	0.00277	0.00277	0.30	0.2
Two-point recombination rate	0.00277	0.00277	0.30	0.2
Gene transposition rate	0.00277	0.00277	0.00277	0.00277
Inversion rate	0.00546	0.00546	0.00546	0.00546
IS transportation rate	0.00546	0.00546	0.00546	0.00546
RIS transportation rate	0.00546	0.00546	0.00546	0.00546
Fitness function error type	RMSE	RMSE	RMSE	RMSE

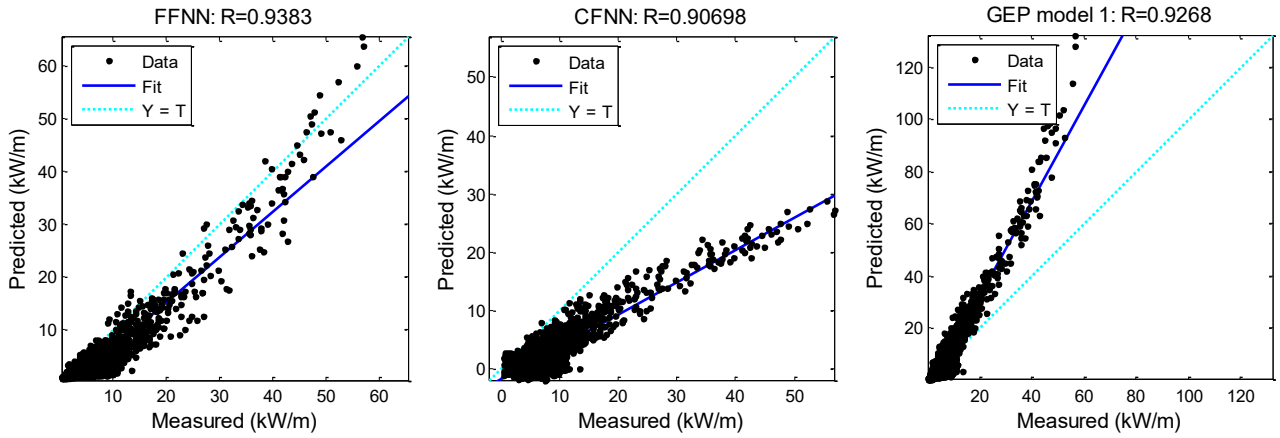


Figure 7. The measured wave energy flux versus predicted values on the test data for FFNN, CFNN, and GEP model 1

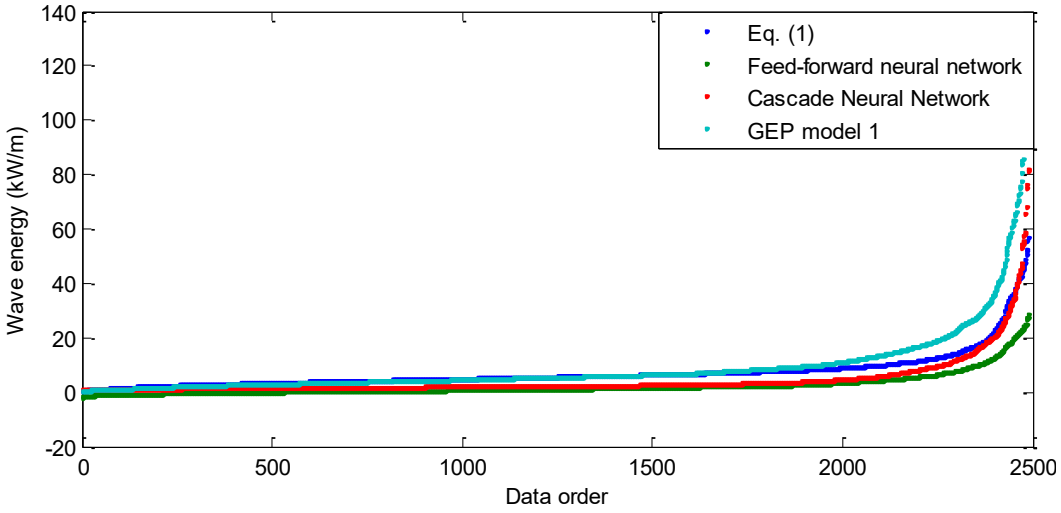


Figure 8. Sorted plot of wave energy flux estimation comparison between all models for the test data

Table 8. Overall comparison of different methods

Method	Training Data		Test Data	
	RMSR (kW/m)	R^2	RMSR (kW/m)	R^2
FFNN	4.25	0.93	3.79	0.88
CFNN	4.31	0.93	6.02	0.82
GEP Model 1	7.30	0.79	6.94	0.85
GEP Model 2	7.68	0.77	36.35	0.34
GEP Model 3	0.48	0.77	1.49	0.13
GEP Model 4	0.53	0.75	4.30	0.66

in Table 8, the results of the models' implementation on the test data are presented. In general, the results of all three methods are appropriate as regards the test data. Models 2 and 3 using the GEP method performed poorly in estimating the wave energy flux on the test data, while FFNN performed best of all the four created models. The regression plots of the FFNN, CFNN, and GEP model 1 on the test data are presented in Figure 7. Also as regards the test data, the sorted plot of wave energy flux estimation comparison between neural networks and GEP for the best obtained models is shown in Figure 8. According to this figure, in general the total amount of energy estimated using FFNs and CNNs is lower than the actual values, and the energy estimated using GPE model 1 is higher than the actual values. The overall results for all three methods used in this study are presented in Table 8, from which it can be seen that the FFNN method performs better than the other methods.

Conclusions

For wave energy calculation, spectral wave measurements are required. In some cases, it is not possible to measure these values due to lack of laboratory equipment, financial resources, or other items. In the absence of spectral wave measurements, the current research studied wave energy flux estimation by using historical meteorological data. There

are a variety of DMMs for the prediction of problems. We used three different well-known methods of FFNN, CFNN, and GEP to estimate wave energy flux by using meteorological data. In all created models, wind speed, air temperature, and sea temperature were considered as input parameters. Wave energy flux was also selected as an output parameter. The accuracy of the mentioned methods was examined using the performance evaluation criteria. As a result of this study, it can be said that the performance of all discussed DMMs is satisfactory, but that among them, FFNN could estimate wave energy flux with a more acceptable accuracy than other methods. The main aim of this paper was to find the relationship between wave energy flux and the meteorological parameters, and the results of the present work have shown that there is a good correlation between these variables. In fact, it is possible to estimate wave energy flux in the absence of wave records in the different areas. It should also be noted that there are other useful methods, for example numerical models, to assess wave energy. Finally, it is recommended that future studies of wave energy prediction should consider the combination of DMMs and numerical models to achieve more efficient results.

Acknowledgments

The data mining computations presented here were performed on the parallel machines of the High Performance Computing Research Center (HPCRC) of Amirkabir University of Technology; their support is gratefully acknowledged. The authors would also like to thank NDBC for offering the necessary datasets.

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. Amirkhani, S., Nasirivatan, Sh., Kasaeian, A.B. & Hajinezhad, A. (2015) ANN and ANFIS models to predict the performance of solar chimney power plants. *Renewable Energy* 83. pp. 597–607.
2. Barnett, V. & Lewis, T. (1994) *Outliers in Statistical Data*. John Wiley & Sons, 3rd edition.
3. Benardos, AG. & Kaliampakos, D.C. (2004) Modelling TBM performance with artificial neural networks. *Tunnelling and Underground Space Technology* 19 (6). pp. 597–605.
4. Bishop, C. (1995) *Neural Networks for Pattern Recognition.* New York: Oxford University Press.
5. Blonbou, R., Monjoly, S. & Dorville, JF. (2011) An adaptive short-term prediction scheme for wind energy storage management. *Energy Conversion and Management* 52 (6). pp. 2412-2416.
6. Boronowski, S., Wild, P., Rowe, A. & Van Kooten, G.C. (2010) Integration of wave power in HaidaGwaii. *Renewable Energy* 35. pp. 2415–3242.
7. Breunig, M.M., Kriegel, H.P., Ng, R.T. & Sander, J. (2000) LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACMSIGMOD International Conference Management of Data (SIGMOD'00)*, Dallas, Texas.
8. Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G.A. & Pavan, M. (2016) A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renewable Energy* 90. pp. 501–512.
9. Chiu, S. (1994) Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent& Fuzzy Systems* 2 (3).
10. Cornejo-Bueno, L., Nieto-Borge, J.C., García-Díaz, P., Rodríguez, G. & Salcedo-Sanz, S. (2016) Significant wave height and energy flux prediction for marine energy applications: A grouping genetic algorithm – Extreme Learning Machine approach. *Renewable Energy* 97. pp. 380–389.
11. Di Piazza, A.D., Di Piazza, M.C., Ragusa, A. & Vitale, G. (2011) Environmental data processing by clustering methods for energy forecast and planning. *Renewable Energy* 36 (3). pp. 1063–1074.
12. Doucoure, B., Agbossou, K. & Cardenas, A. (2016) Time series prediction using artificial wavelet neural network and multi-resolution analysis: Application to wind speed data. *Renewable Energy* 92. pp. 202–211.
13. Ebrahimabadi, A., Azimipour, M. & Bahreini, A. (2015) Prediction of roadheaders' performance using artificial neural network approaches (MLP and KOSFM). *Journal of Rock Mechanics and Geotechnical Engineering* 7. pp. 573–583.
14. Falcao, A.F. (2010) Wave energy utilization: a review of the technologies. *Renewable and Sustainable Energy Reviews* 14 (3). pp. 899–918.
15. Fayyad, U., Shapiro, G.P. & Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39 (11).pp. 27–34.
16. Fernández, J.C., Salcedo-Sanz, S., Gutiérrez, P.A., Alexandre, E. & Hervás-Martínez, C. (2015) Significant wave height and energy flux range forecast with machine learning classifiers. *Engineering Applications of Artificial Intelligence* 43. pp. 44–53.
17. Ferreira, C. (2001) *Gene expression programming in problem solving*. In: Roy, R., Koeppen, M., Ovaska, S., Furuhashi, T. & Hoffmann, F. (Eds.), *Soft Computing and Industry*. Springer, UK, pp. 635–653.
18. Ferreira, C. (2006) *Gene-expression programming: mathematical modeling by an artificial intelligence*. Berlin, Germany: Springer.
19. Ferreria C. (2001) Gene expression programming: a new adaptive algorithm for solving problems. *Complex Syst* 13 (2). pp. 87–129.
20. Gairaa, K., Khellaf, A., Messlem, Y. & Chellali, F. (2016) Estimation of the daily global solar radiation based on Box–Jenkins and ANN models: A combined approach. *Renewable and Sustainable Energy Reviews* 57. pp. 238–249.
21. Grady, S.A., Hussaini, M.Y. & Abdullah M.M. (2005) Placement of wind turbines using genetic algorithms. *Renewable Energy* 30 (2). pp. 259–270.

22. Grubbs, F.E. (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11. pp. 1–21.
23. Guillermo, V. (1998) A Distributed Approach to a Neural Network Simulation Program. *Master's thesis*, The University of Texas at El Paso, El Paso, TX.
24. Heinermann, J. & Kramer, O. (2016) Machine learning ensembles for wind power prediction. *Renewable Energy* 89. pp. 671–679.
25. Holand, J.H. (1975) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The University of Michigan Press, USA.
26. Kamranzad, B., Chegini, V. & Etemad-Shahidi, A. (2016) Temporal-spatial variation of wave energy and nearshore hotspots in the Gulf of Oman based on locally generated wind waves. *Renewable Energy* 94 pp. 341–352.
27. Keshavarz, A. & Mehramiri, M. (2015) New Gene Expression Programming models for normalized shear modulus and damping ratio of sands. *Engineering Applications of Artificial Intelligence* 45. pp. 464–472.
28. Koza, J.R. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT press, USA.
29. Kusiak, A. & Li, W. (2011) The prediction and diagnosis of wind turbine faults. *Renewable Energy* 36 (1). pp. 16–23.
30. Lee, D. (1982) On k-Nearest Neighbor Voronoi Diagrams in the Plane. *IEEE transactions on computers* 31 (6).
31. Maged, M.H., Khalafallah, M.G. & Hassanien, E.A. (2004) Prediction of wastewater treatment plant performance using artificial neural networks. *Environmental Modelling and Software* 19 (10). pp. 919–928.
32. Ming, H. & Aggidis, G.A. (2008) Developments, expectations of wave energy converters and mooring anchors in the UK. *Journal of Ocean University of China* 7 (1). pp. 10–16.
33. Minh Tri, N., Truong, D.Q., Thinh, D.H., Binh, P.C., Dung, D.T, Lee, S., Park, H.G. & Ahn, K.K. (2016) A novel control method to maximize the energy-harvesting capability of an adjustable slope angle wave energy converter. *Renewable Energy* 97. pp. 518–531.
34. Moller, M.F. (1993) A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* 6 (4). pp. 525–533.
35. Özger, M. (2011) Prediction of ocean wave energy from meteorological variables by fuzzy logic modeling. *Expert Systems with Applications* 38. pp. 6269–6274.
36. Robin, J., Clive, M. & Ian, S. (2012) *The Art of Lisp Programming*. Springer Science & Business Media. p. 2. ISBN 9781447117193.
37. Salari, D., Daneshvar, N., Aghazadeh, F. & Khataee, A.R. (2005) Application of artificial neural networks for modeling of the treatment of wastewater contaminated with methyl tert-butyl ether (MTBE) by UV/H2O2 process. *Journal of Hazardous Materials* 125 (1–3). pp. 205–210.
38. Sannasiraj, S.A. & Sundar, V. (2016) Assessment of wave energy potential and its harvesting approach along the Indian coast. *Renewable Energy* 99. pp. 398–409.
39. Sarkar, D., Contal, E., Vayatis, N. & Dias, F. (2016) Prediction and optimization of wave energy converter arrays using a machine learning approach. *Renewable Energy* 97. pp. 504–517.
40. Scott, E., Lebiere, F. & Christian, L. (1990) The Cascade Correlation Learning Architecture. *Advances in Neural Information Processing Systems*. pp. 524–532.
41. Shukur, O.B. & Lee, H.M. (2015) Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA. *Renewable Energy* 76. pp. 637–647.
42. Sierra, J.P., Martín, C., Mösso, C., Mestres, M. & Jebbad, R. (2016) Wave energy potential along the Atlantic coast of Morocco. *Renewable Energy* 96. Part A, pp. 20–32.
43. Tang, B., Song, T., Li, F. & Deng, L. (2014) Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine. *Renewable Energy* 62. pp. 1–9.
44. Wong, P.K., Wong, K.I., Vong, C.M. & d anCheung, C.S. (2015) Modeling and optimization of biodiesel engine performance using kernel-based extreme learning machine and cuckoo search. *Renewable Energy* 74. pp. 640–647.
45. Yadav, A.K. & Chandel, S.S. (2015) Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. *Renewable Energy* 75. pp. 675–693.
46. Zhang, Y., Yang, S. & Wang, Y. (2008) LDBOD: A novel local distribution based outlier detector. *Pattern Recognition Letters* 29. pp. 967–976.