

An Integrated Long-short Term Memory Algorithm for Predicting Polar Westerlies Wave Height

Chenhua Ni ^{1*} and Xiandong Ma ²

1 National Ocean Technology Center, Tianjin 300112, China

2 Engineering Department, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK,
xiandong.ma@lancaster.ac.uk

* Correspondence: jeff_ni@126.com; Tel.: +86-13803025669

Abstract:

The improved knowledge of wave height and period conditions has considerably influenced on ocean navigation, marine fishery and engineering, especially in the polar regions. The methods of predicting ocean wave height which involve field measurements, numerical simulation, physical models and analytical solutions have been gradually developed with intelligent functions. Despite numerical wave models being dominant for recent decades, wave forecasting is still facing many challenges such as small region forecasting and large amounts of data needed. This paper presents a novel deep learning algorithm, namely Long Short Term Memory (LSTM), incorporating with Principal Component Analysis (PCA) to predict the wave height by using data from four wave buoys as deployed in the polar westerlies for two and half months. The PCA method is used to extract principal components from a set of input signals while LSTM is adopted to avoid long term independences during the forecasting. The novelty of this paper is to investigate an artificial intelligence (AI) based model in the field of time sequence forecasting in order to determine the performance of wave conditions by using AI technology. The result from this integrated method demonstrates that the LSTM model has the potential to better predict wave height in the polar condition based on time-space domain information. The PCA is proved essential for selection of input signals and for correlation analysis. For comparison, different data-driven models are applied and the results also show the purposed model achieves the highest scores in terms of R-squared value. Besides, the paper also discusses the challenges for long term and high-value prediction which needs to be optimized in the future work.

Keywords: Long Short Term Memory (LSTM); Deep learning (DL); wave height prediction; principal component analysis (PCA); artificial intelligence (AI)

1. Introduction

In the ocean condition, wind wave results from the wind blowing over an area of the sea surface. After the wind ceases, the waves still exist and are then called swells. The improved knowledge of height and period of oscillatory short waves is essential for almost any engineering activity in the ocean, including planning, design, construction and operation related to the harbor and coastal structures [1]. In the field of nautical and fishery industry, the wave

forecasting plays a significant role in enduring harsh sea conditions and keeping secure the operations. Otherwise, the shipping industry requires guidance for operational planning and tactical seakeeping purposes. The main dimensions associated with waves are wave height, wave period and wave direction, of which wave height usually takes the first priority among wave parameters from maritime civilization. However, the complexity and uncertainty of the wave generation phenomenon are such that despite considerable advances made in computational techniques, the solutions obtained are neither exact nor uniformly applicable at all sites and at all times [2].

Ocean wave characteristics are mainly determined through field measurements, numerical simulation, physical models and analytical solutions. Each method has its own advantages and disadvantages. Early forecasts of the sea state were created manually based on empirical relationships between the present state of the sea, the expected wind conditions, the fetch/duration, and the direction of the wave propagation [2]. But nowadays, numerical models emerge as one of the most powerful tools for the study of surface water waves [4]. Wind wave models are used as forecasting or hindcasting tools with use of varying data assimilation and associated methods. For example, the operational wave forecasting systems at NOAA are based on the WAVEWATCH III model which has a global domain of approximately 50 km resolution and regional domains for the northern hemisphere oceanic basins at approximately 18 km and approximately 7 km resolution [5]. The Integrated Forecast System (IFS) which integrated European Centre for Medium-Range Weather Forecasts (ECMWF) and the wave model WAM currently comprises 36 frequency bins and 36 propagation directions at an average spatial resolution of 25 km [6]. In the regional area, Texas A&M University used the SWAN model to forecast waves in the Gulf of Mexico [7]. As another example, CCHE2D-COAST is a processes-based integrated model which is capable of simulating coastal processes in different coasts with complex shorelines. The MIKE21 SW, as a new third-generation numerical model, was developed to simulate scenarios for the North Sea, parts of the Norwegian Sea and the Baltic Sea. The results were validated from wave rider buoys and found that the model was better in prediction than those models that did not use fine mesh. Later, a third-generation spectral wave model, called Simulating Waves Nearshore (SWAN) developed for small-scale, coastal regions with shallow water, (barrier) islands, tidal flats, local wind, and ambient currents, was verified in stationary mode with measurements in five real field cases [8]. Despite numerical models being performed fabulously in both accuracy and timely delivery, there are still plenty of domains where they cannot cover. For instance, the prediction of waves near wave energy converters (WECs) needs a small-region predicting approach such as 0.1 km resolution. Coastal aquaculture depends on an accurate wave and current forecasting in order to organize feeding schedules and take risk assessment. Therefore small-region wave prediction has been proven to be an uncertain and random process and hence numerical models such as SWAN and WAVEWATCH III if considered would be time-consuming and complicated to model by using deterministic equations.

As the AI bloomed, the neural network (NN) has been proven suitable for modelling since NN is primarily aimed at recognition of a random pattern in a given set of input values and use the same to predict the desired property. Generally, NNs are useful to map random inputs with the corresponding random output and their application does not require good knowledge of the underlying

physical process as a precondition. For example, Scott C. James used Multilayer Perceptron (MLP) and artificial NN to perform two different tasks: regression analysis for wave height and classification analysis for characteristic wave period. A case performed at Monterey Bay demonstrated a RMSE of 9 cm and correctly identified over 90% of the characteristic periods for the test-data sets [9]. C. Feng employed a two-layer machine learning (ML) structure model to perform with up to 130% accuracy in 1-hour-ahead wind speed forecasting [10]. Over 15-years buoys data at Italian Sea were analyzed by D.J.Peres with ANNs to extend an observed time series of significant wave heights. Results were compared to WAVEWATCH III/CFSRandERA-Interim, and shown better than latter ones, thus potentially appealing to engineering practice [11]. Researches on normal NN also showed that with the number of inputs increasing and becoming complicated, the accuracy of NN may reduce sharply because the model cannot extract enough features. Furthermore the NN will become very time consuming when dealing with complicated problems [12][13].

Deep learning allows computational models of multiple processing layers to learn representations of data with multiple levels of abstraction. LSTM method, as derived from the family of recurrent neural network (RNN), was considered one of the most appropriate methods to address predicting issues because it has shown potential to solve a range of problems involving sequential learning in recent years. It features sharing parameters across time series between hidden units and becomes one of the most prevalent forecasting algorithms to settle regression and prediction problems [14][15]. More and more attention has been received on the predicting ocean wave parameters using deep learning network by now. The RNN and sequence-to-sequence neural networks have been introduced to predict short and long-term significant wave height and output power of the ocean waves, and the case studies proved that the Adam and AMSGrad optimization algorithms are the most robust ones to optimize the sequence-to-sequence network based on real data obtained from NOAA buoy measurements [16]. A two-step wind-wave prediction model was explored to predict wind speed and wave height based on deep RNNs with a lower prediction error being produced when compared with shallower MLP [17]. The deep neural network has also been employed for the application of actual radar detection and inversion technology and an accuracy of 99.01% was achieved from the study as detailed in [18]. The LSTM has been investigated in the situation of short term wind power forecasting and weather forecasting in recent years. M. Zayta used LSTM to forecast 24 and 72 hours' worth of temperature, humidity and wind speed for 9 cities in Morocco and produced a better alternative to forecast general weather conditions [19]. **A LSTM method was proposed in combination with proper orthogonal decomposition to provide the corresponding induced wave conditions along the coastline. The result showed that the CPU cost decreased by three orders of magnitude and the use of this approach can provide the flood prediction in seconds [20].** An approach combining the discrete wavelet transform and LSTM networks was **also** proposed to perform wind power short-term forecasting, aiming to improve prediction accuracy using real data from three wind farms as the benchmarks located in Inner Mongolia, The Netherlands and Southern China, respectively [21].

This paper aims to predict time-varying wave heights by utilizing deep learning algorithms incorporating LSMT with PCA whilst considering harsh environments in polar regions. The remainder of the paper is organized as follows.

Section 2 depicts the integrated methodology of LSTM algorithm and PCA analysis method. Section 3 explains the data acquired from buoy measurements and describes the architecture for prediction of the purposed model. Section 4 presents the results and analyzes performance of the PCA method and LSTM network, where findings and discussion from experimental data are also given. Finally, Section 5 summarizes the conclusions presented in the study.

2. Methodology

2.1 LSTM network

RNN networks can theoretically use their feedback connections to store representations of recent input events in forms of activations, mostly by changing weights for short-term memory and long-term memory. This feature makes it potentially useful for many time series applications such as speech recognition and music composition. But it still fronts oscillating weights in some cases, resulting in an unacceptable amount of time or even not working at all [22]. For these reasons, Sepp Hochreiter and Jürgen Schmidhuber introduced a novel recurrent network architecture called LSTM to tackle these error back-flow problems in 1997. It can learn to make the connections between time indexes beyond 1000 steps and keep the specific architecture constant against exploding and vanishing gradients [23]. In 1999, Felix A. Gers, et al. pointed out the LSTM cannot solve previous learning algorithms for RNNs and therefore the authors adapted “forget gate” as a remedy that enabled an LSTM cell to learn to reset itself at appropriate times and thus releasing internal resources. All AI algorithms (including LSTM) were not able to solve continual versions of these problems, but LSTM joint with forget gates could well solve them simply and conveniently [24].

As a family member of RNN, LSTM network architecture also includes an input layer, hidden layers, and an output layer. Each hidden layer of a traditional RNN contains one short-term memory vector h . As shown in Figure 1, the left side illustrates the cell architecture of RNN which means A_t not only receive messages from A_{t-1} but also send messages to A_{t+1} [25] [26]. As shown on the right side of the figure, the key to LSTM is the cell state c , which can remember long-term information by obtaining cell state and inputs from last time series or sequence. There are three inputs of LSTM in time t , namely the output h_{t-1} in last time series, the cell stats c_{t-1} in last time series and the present input x_t . Obviously, the output of this layer includes output value h_t and cell state c_t which are used as inputs of the next hidden layer. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. As shown in Figure 2, the gates include a forget gate (f) meaning how many memories will be reserved from c_{t-1} to c_t , an output gate (o) deciding how many memories will be outputted to h_t and an input gate (i) meaning how many memories will be reserved from c'_t . The interaction (element-wise multiplication) of input and update gate performs write functions into memory. The input gate, using hard sigmoid, decides which values to write while the update gate, using hyperbolic tangent (tanh) activation, creates a vector of new cell values [27] [28].

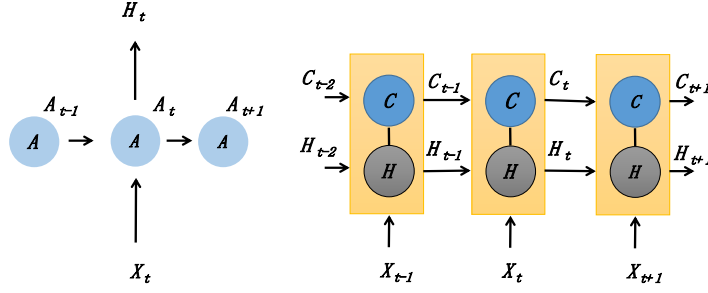


Figure 1 The improvement architecture from RNN to LSTM

The mathematical representation of LSTM can be obtained as follows [29][30].

Input gate (i_t) decides the information which will be added to the cell.

$$i_t = \sigma(W_i * X_t + U_i * H_{t-1} + b_i) \dots \dots \dots (1)$$

Forget gate (f_t) decides the information which will be abandoned from the cell.

$$f_t = \sigma(W_f * X_t + U_f * H_{t-1} + b_f) \dots \dots \dots (2)$$

Output gate (o_t) decides the information which will be exported from the cell.

$$o_t = \sigma(W_o * X_t + U_o * H_{t-1} + b_o) \dots \dots \dots (3)$$

The states in the cell are expressed.

$$c_t = f_t c_{t-1} c'_t + i_t \dots \dots \dots (4)$$

$$c'_t = \tanh(W_c * X_t + U_c * H_{t-1} + b_c) \dots \dots \dots (5)$$

For the output vector,

$$H_t = o_t \tanh(c_t) \dots \dots \dots (6)$$

From the equations above, x_t represents the present input vector, W , U , V and b denote the hyper-parameters for weights and biases. H_t is the value of the memory cell at time t . σ is the sigmoid function, and \tanh represents the hyperbolic tangent function [31][32].

If an input z is given, the output of the hyperbolic tangent function is calculated as follows.

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \dots \dots \dots (7)$$

Figure 2 also illustrates how these equations work. Every rectangle represents a single LSTM cell and three σ denote three functional gates respectively. As shown in the middle rectangle, the long-term memory is calculated with c_{t-1} multiplying by f_t , which means forget gate. The short-term memory is calculated with c'_t multiplying by i_t which means the input gate. As a result, the c_t representing the new cell state is able to combine with both long-term and short-term information [33] [34] [35]. In terms of output h_t , it is decided by cell state c_t and output gate o_t . With the equations (1)–(7), the LSTM architecture can follow the steps to pass the information to the next one based on the three functional gates.

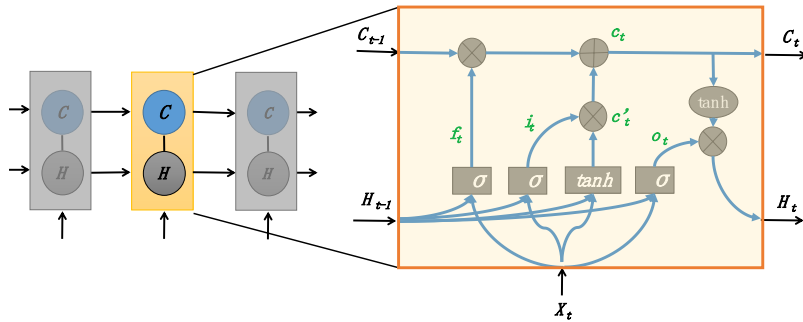


Figure 2 The repeating cell in an LSTM contains four interacting gates

2.2 PCA method

It is worth noting that prediction of wave height requires large amounts of data to be processed during the training and prediction phases. There is also some redundancy in these data, with many of the signals being highly correlated. This would lead to over-fitting of the LSTM model and subsequently poor model fit, and it is therefore necessary to employ some feature selection algorithm, such as PCA, employed in this paper.

PCA is a statistical method that uses an orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). It was invented as an analogue of the principal axis theorem in mechanics by Karl Pearson in 1901. It has been improved many times until now [36] [37]. Nowadays, PCA plays an important role in data analysis and constructing predictive networks in many domains such as population statistics, molecular dynamics, mathematical analysis and mathematical modelling. If a multivariate dataset with a high-dimension is given, the PCA is able to generate a low-dimension dataset based on the specific method to reduce the transformed data by the first few PCs. More specifically, an m -dimension dataset can be orthogonally transformed to an n -dimension dataset ($n < m$). The PCA method will reconstruct the n -dimension dataset (new dataset) rather than simply remove the $(m-n)$ dimension dataset from the m -dimension dataset. Consequently, adopting the PCA method contributes to reduction of the complexity of inputs and hence the processing time notably [38].

If F_i represents the i^{th} PC, then variance-covariance matrix is expressed as follows. The bigger the $var(F_i)$ is, the more information F_i contains from the original dataset.

$$\text{var}(F_i) = \Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{pmatrix} \dots\dots\dots (8)$$

Consider the linear combinations,

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \dots\dots\dots (9)$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \dots\dots\dots (10)$$

...

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \dots\dots\dots (11)$$

where a_{1i} , a_{2i} , \dots , a_{pi} ($i=1, \dots, m$) denote the feature vectors of variance-covariance matrix, and X_1 , X_2 , \dots , X_p represent the normalized input variables [39] [40].

The eigenvalues of the PCs are sorted as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$, and the number of the PCs selected mostly depend on the value of the cumulative variance contribution rate. Generally speaking, it is considered that enough information is obtained if the cumulative variance contribution rate is above 75%. The variance contribution rate is described as follows.

$$\eta_i = \frac{\lambda_i}{\sum_{i=1}^m \lambda_i} \times 100\% \dots\dots\dots (12)$$

2.3 Performance evaluation index

Here we choose three mainstream performance evaluation indexes to measure the accuracy of prediction, which are the coefficient of determination (R^2), RMSE and the mean absolute error (MAE). R^2 value provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model [41] [42]. RMSE is more sensitive to a large deviation between the predicted values and the actual outcomes while the MAE measures the absolute difference between the forecasts and the actual values. These measures are calculated by the following equations.

$$R_T^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} \dots\dots\dots (13)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (e_{t+k/t})^2} \dots\dots\dots (14)$$

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |e_{t+k/t}| \dots\dots\dots (15)$$

where σ_e^2 is the sample variance of the residual, σ_y^2 is the variance of the actual output, N is the number of samples, and e is the residual between actual and predicted values.

3. The LSTM integrated prediction architecture

3.1 Data acquisition and preparation

The Antarctica continent is considered as abundant natural resource such as coal, gold, copper and hydrocarbons which remain untapped [43]. However, it is not easy to voyage to Antarctica by sea because of the barrage of the Westerlies

which dominate with cool air, low pressure, strong westerly wind and cyclones [44]. The data used in this article were collected by buoys deployed in the westerlies in the northern hemisphere, as shown in Figure 3.

The four buoys from B1 to B4 were classified into two categories based on the deployed methods: mooring and floating. B1 to B3 use floating technology by drifting along with the force of current and wind. B4 uses mooring technology which means the buoy is anchored at the seabed. Table I. The type, date and observing information of deployed buoys. Observing parameters were logged within the data collection system inside the buoys and sent to the laboratory hourly by satellites.

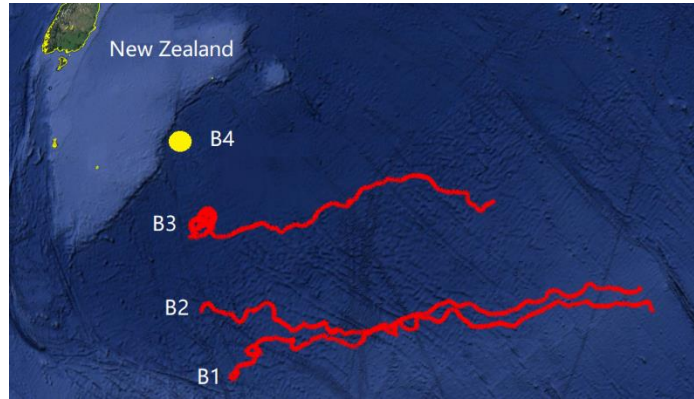


Figure 3 The deployed positions of westerlies buoys

More than 7,000 group data were collected and recorded in the database. A data preprocessing was inevitably used to process abnormal and constant data which may exist due to malfunction condition and harsh environment. The gap between data may exist due to communication faults or sensor malfunction (always logged as 999.99) that should be replaced using Mean/Mode Completer [46], hot deck imputation [45] or K-means clustering [47]. It is found some signals are correlated and hence the PCA method is adopted to extract the PCs and get rid of the redundant data. It is worth mentioning that the date, time and location of buoys are strongly related to the wave prediction results because the meteorological condition and time-domain information change from each other. Here, the day of year and the time of day are represented as time domain information and the distance between the historical location and predicted destination is calculated respectively based on the equations (17)-(18) below. The direction parameters are also represented using a pair of sine and cosine values in an effort to capture their periodic nature.

$$h \rightarrow \left(\sin \frac{h}{24}, \cos \frac{h}{24} \right) \dots \dots \dots (16)$$

The distance between two geological locations A and B is calculated by following equations.

$$C = \sin A_{lat} \sin B_{lat} \cos(A_{lon} - B_{lon}) + \cos A_{lat} \cos B_{lat} \dots \dots \dots (17)$$

$$d = R * \cos^{-1}(C * Pi/180) \dots \dots \dots (18)$$

where A_{lat} , A_{lon} , B_{lat} and B_{lon} represent the latitude and longitude of location A and B. The east longitude represents positive while west longitude represents

negative. In terms of latitude, the north latitude takes $90 - \text{latitude}$ and south latitude takes $90 + \text{latitude}$. The parameter R is a constant representing the radius of earth and the value is 6371.004 km. The parameter d is the distance between A and B.

Table I. The type, date and observing information of deployed buoys

Buoy No.	Type	Date	Observing Parameters
B1-B3	Floating	From 9 Jan 2019 to 24 Mar 2019	Date, location, wave height, wave period, wave direction
B4	Mooring	From 9 Jan 2019 to 24 Mar 2019	Date, location, wind speed, wind direction, air temp, humidity, air pressure, water temp, salinity, wave height, wave period, wave direction

3.2 Prediction architecture

The purposed LSTM prediction model can be briefly classified into two steps, as shown in Figure 4, comprising the PCA process and the LSTM model. The PCA process takes the responsibility to obtain useful signals and get rid of abundant information from the original signals while the LSTM model predicts wave parameters through the LSTM block and fully connected layers. The proposed LSTM network sets out with a multi-variant time-series input layer followed by an LSTM layer. The LSTM layer learns long-term dependencies between time steps in time-series and sequence data. Then, the fully connected layer connects every neuron in LSTM layer to every neuron in the next layer, and finally ends with a regression output layer. This diagram illustrates the stream of time-series S with D dimensions of length S as input signals from the wave buoys. The PCA method then extracts the useful P signals from the input of D dimensions, which means P features are finally treated as the flow of inputs to the LSTM layer.

“0” represents the output (also known as the hidden state) of the network. In the LSTM layer, the first block takes the initial state of the network and the first time-series component from the feature extraction section. Also, it calculates the result h_1 and cell state c_1 (as defined in Figure 2). At time step t , the t block receives the current state of c_{t-1} and h_{t-1} , as well as the input signals at time t . After processing, the state c_t and result h_t will be updated and the computing is completed. At each step, the output state (h) and the cell state (c) are included in the state of layer. The output state h contains the output information while the cell state c contains information learned from the previous time steps. The gates which represent the specific feature of LSTM network control the information to add to or remove from the cell state.

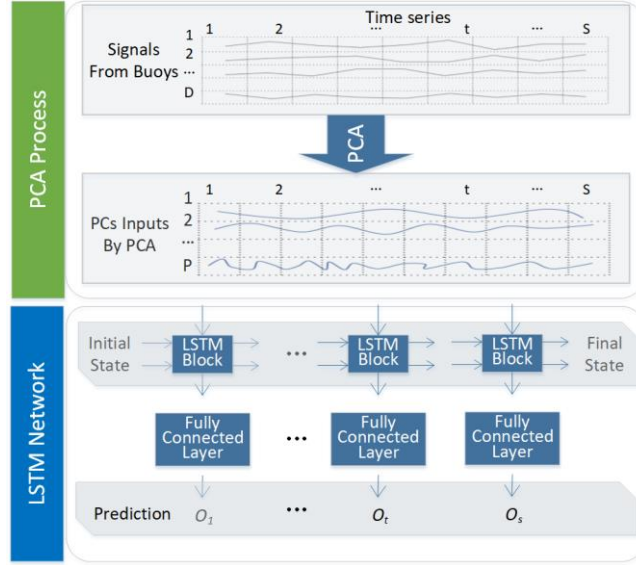


Figure 4. The LSTM forecasting model integrated with the PCA process

The activation function for both input layers and hidden layers are set to ReLU function because it is very practical in regression. It is not limited to outputs between 0 and 1 and thus improves the disparity of the solution in the hidden neurons. Besides, it is able to tackle the vanishing gradient problem which prevailing in RNN. The expression for the activation function for the l -th layer of a neural network with N layers is shown below.

$$\sigma_l(x) = \begin{cases} \max(0, x), & \text{if } l \neq N \\ x, & \text{if } l = N \end{cases} \dots \dots \dots (19)$$

where x indicates the input of the ReLU function.

4. Results and discussions

4.1 PCA result

The PCA process is normally classified into several steps such as data normalization, covariance calculation, eigenvalue deduction and calculation of contribution rates. For this case, the time-series observed data mentioned in Table I from B1 and B4 buoys were processed as inputs of the PCA, including wave height, wave period, wind speed and direction, day of year and time of day. The total dataset includes 1800 group of data and each group has 18 inputs. After the processing, the result of the PCA process based on B1 and B4 data is listed in Table II. The PC1 to PC10 represent the PCs associated with inputs 1 to 10 whereas the eigenvalue and the contribution rate indicate how important the PC dominates the overall inputs. Obviously, the cumulated rate reaches 88.3% (between 85% and 95% by empiricism) by the PC8 column, meaning PC1 to PC8 contain most of the information and the PCs after PC9 considered as redundant data. Furthermore, the PC can be expressed by the following equation,

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \dots + a_{pi}X_p, i = 1, \dots, p \dots \dots \dots (20)$$

where F_i represents the i^{th} PC, X_p means p -th input signal. In terms of PC1 in Table II, the expression will be as follows.

$$F_1 = 0.4677X_1 - 0.4662X_2 + 0.0141X_3 - 0.0134X_4 + 0.0886X_5 - 0.052X_6 + 0.0724X_7 - 0.0146X_8 + 0.4677X_9 - 0.4662X_{10} + 0.0141X_{11} - 0.0134X_{12} - 0.2111X_{13} - 0.1563X_{14} - 0.0819X_{15} + 0.0732X_{16} - 0.1661X_{17} + 0.0496X_{18} \dots \quad (21)$$

The values highlighted with yellow indicate they are much larger than others, meaning the corresponding input variables dominate this PC. In other words, we choose the most representative inputs based on those values and remove other redundant inputs. After checked the values from PC1 to PC8 (as PC1 to PC8 contain most information of the original dataset as stated above), the wave period and wave direction from B1 and wave period from B4 seem as redundant inputs and therefore removed, and other 15 inputs are remained.

Table II. The result of the PCA process based on B1 and B4 observing signals

Inputs	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
B1 Day of year(sin)	0.4677	0.0157	0.0506	-0.0002	0.1038	0.0324	0.0112	0.0002	-0.0536	-0.0785
B1 Day of year(cos)	-0.4662	-0.0150	-0.0714	-0.0085	-0.0309	-0.0628	-0.0044	-0.0141	0.0427	0.1373
B1 Time of day(sin)	0.0141	-0.4992	-0.0015	0.0083	0.0330	0.0024	-0.0118	-0.0267	0.0026	0.0125
B1 Time of day(cos)	-0.0134	0.4996	-0.0024	0.0116	0.0017	0.0019	0.0020	-0.0224	0.0048	0.0246
B1 Distance	0.0886	0.0000	0.2522	-0.3138	-0.1628	0.5994	-0.3171	0.0675	-0.0925	0.4843
B1 Wave height	-0.0520	0.0113	0.5418	0.0895	0.0246	0.0093	0.0447	0.1773	0.1956	-0.4308
B1 Wave period	0.0724	-0.0120	0.3950	0.3124	-0.1579	-0.0910	-0.0411	-0.1333	0.6882	0.3361
B1 Wave direction	-0.0146	-0.0157	0.0074	-0.1848	-0.0843	0.3425	0.9060	-0.0616	0.1090	0.0611
B4 Day of year(sin)	0.4677	0.0157	0.0506	-0.0002	0.1038	0.0324	0.0112	0.0002	-0.0536	-0.0785
B4 Day of year(cos)	-0.4662	-0.0150	-0.0714	-0.0085	-0.0309	-0.0628	-0.0044	-0.0141	0.0427	0.1373
B4 Time of day(sin)	0.0141	-0.4992	-0.0015	0.0083	0.0330	0.0024	-0.0118	-0.0267	0.0026	0.0125
B4 Time of day(cos)	-0.0134	0.4996	-0.0024	0.0116	0.0017	0.0019	0.0020	-0.0224	0.0048	0.0246
B4 Distance	-0.2111	-0.0073	0.1502	0.2129	-0.2364	0.4853	-0.1546	-0.4656	-0.1696	-0.4691
B4 Wind speed	-0.1563	0.0158	0.2224	-0.2088	0.6672	0.1162	-0.0434	-0.0972	-0.0605	0.1599
B4 Wind direction	-0.0819	-0.0074	-0.0933	0.5730	0.0823	0.3530	0.0414	0.6874	-0.1042	0.0570
B4 Air temperature	0.0732	0.0177	-0.2114	0.5032	0.4464	0.1567	0.0527	-0.4538	0.0296	0.1799
B4 Wave height	-0.1661	0.0015	0.4774	-0.0432	0.3484	-0.1378	0.1006	0.1057	-0.2071	-0.0780
B4 Wave period	0.0496	-0.0119	0.3437	0.2980	-0.2945	-0.2964	0.1842	-0.1538	-0.6092	0.3525
Eigenvalue	4.4042	3.9076	2.2149	1.3308	1.2697	1.0476	0.9830	0.7398	0.7142	0.5922
Contribution Rate (%)	24.4678	21.7088	12.3053	7.3935	7.0538	5.8197	5.4609	4.1101	3.9677	3.2902
Cumulative Rate (%)	24.4678	46.1766	58.4819	65.8754	72.9292	78.7489	84.2098	88.3199	92.2876	95.5778

The bars in Figure 5 also illustrate the same trend, where six PCs from PC1 to PC6 are considered and the height of each bar represents the contribution from each input. Normally, the input is of great importance and treated as a

dominant input if its absolute value is higher than 0.4. After the total number of dominant inputs is identified, the results of the PCA can be decided. Let us take the first three PCs as an example, the date, time and wave height contribute to more than half of the total inputs and therefore the results can be primarily concerned with these inputs.

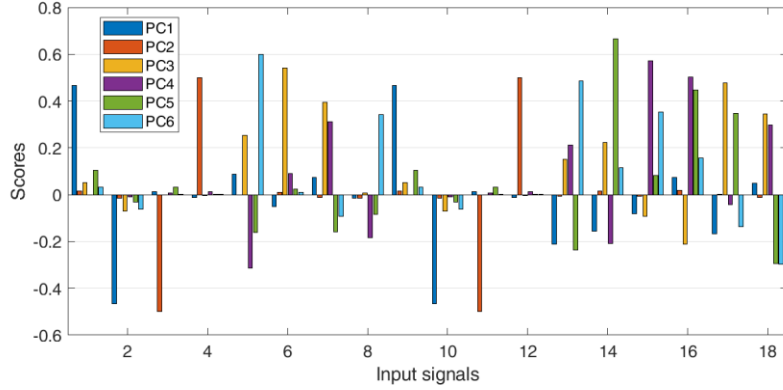


Figure 5. The contribution of all inputs from PC1 to PC6

4.2 LSTM model result

In order to obtain better performance and prediction results, the outcome of the PCA method is used in combination with LSTM model. The PCs can be set as the updated input signals for the time-series LSTM network not only to keep accuracy of the model output but also to reduce computing-time consumption. As mentioned in Table II, the yellow cells to represent the PCs in this case are rearranged as new inputs for the LSTM model. The 27 time-series signals include date, time, distance, and mean wave height values of B1 to B2 and date, time, distance, air temperature, wind speed, wind direction, and mean wave height values of B4 are selected as the inputs of the model. In terms of output, the mean wave height of each buoy is selected respectively. Each time-series sequence data has 1800 records with each record being one-hour long interval. Thus approximately two and half months' time long records are used with 80% data being used for training and other 20% data used for testing.

The LSTM model runs for the time-series forecasting based on multiple signal sequences to forecast the values of future time series sequence that mimics an existing one. The model trains features at each time step along with the input signals and learns to predict the values of the next time step. When the value of each time step is predicted, it is necessary to update network state with the observed values instead of the predicted values. The procedures are given below.

- a) Initialize the network state.
- b) Reset the network state to prevent previous predictions from affecting the predictions on the new data.
- c) Reset the network state and then initialize the network state by predicting on the training data.
- d) Predict at each time step. At each prediction, the next time step value is predicted using the observed value of the previous time step.

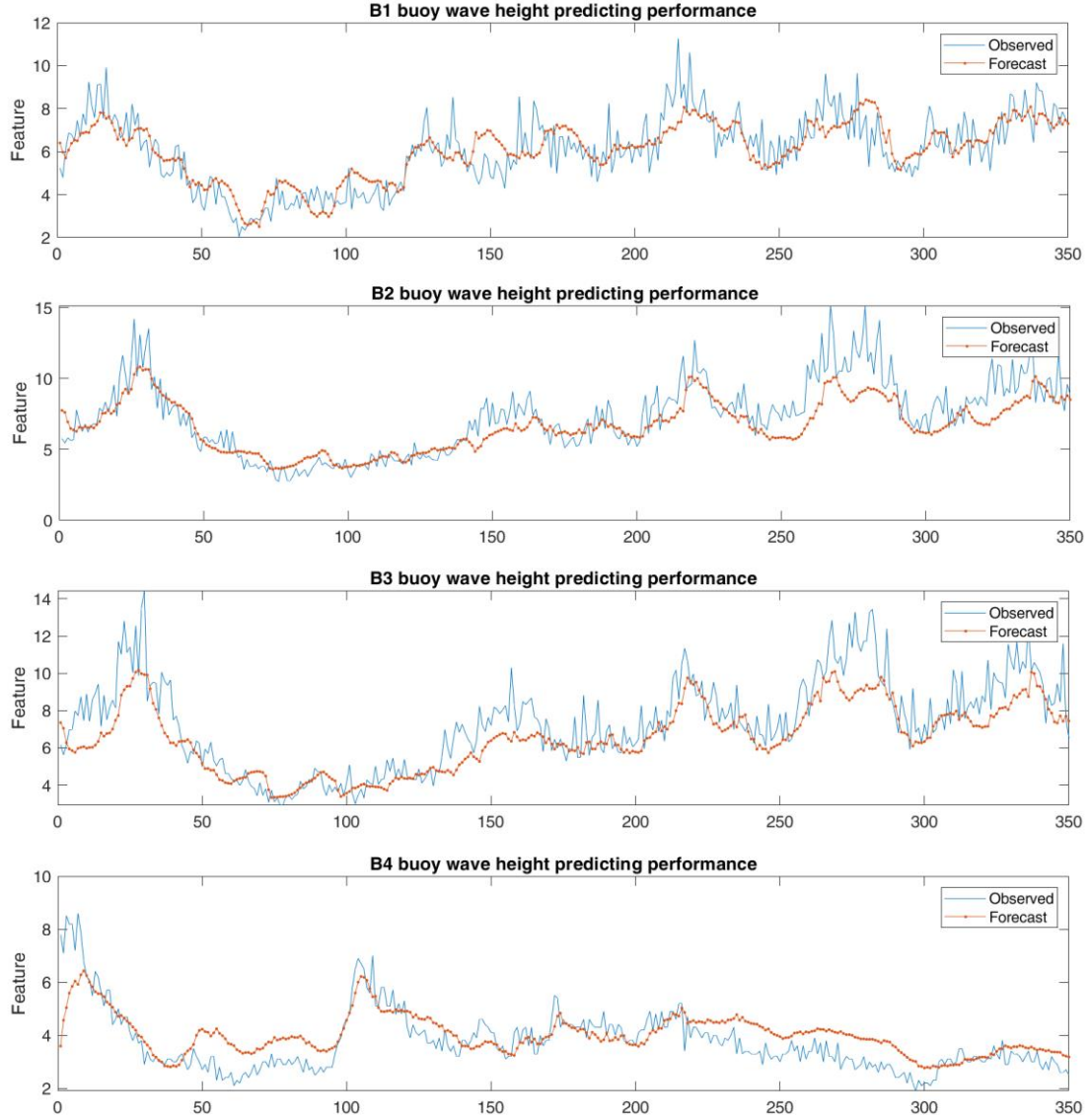


Figure 6. The predicted result of wave height form four buoys (B1-B4)

The result of the LSTM model is shown in Figure 6 to show the wave height predicted at all the buoy stations. Each subplot shows the comparison of wave height value resulting from prediction model and values actually acquired from buoy stations from B1 to B4. Generally, the predictions can be considered acceptable in terms of trend while the wave height oscillates around low values. The predicted curves for each station seem accordant with the observed curves in most time stamps. The red curves (predicted values) keep a smooth change, while, on the other hand, the actual values experience jagged spikes during whole time period. When compared among the four curves, the difference between the predicted and observed values show there are relatively larger errors when the wave height is higher. It is speculated that the worst performance in B2 station is because of not only the largest wave height recorded but also the station being the most high-density wave energy among the four stations. The values in Table III demonstrate the similar situation in terms of these performance measures. The B4 station performs a best predicted result among the four buoys while the other three show the comparable indexes in terms of R^2 and RMSE values. This can be attributed to the factors that the average wave height for B4 is less when

compared to other stations and the additional input like air temperature available at B4 is used for prediction.

Table III. Model performance measures of the buoys

Buoy No.	RMSE	R ²	MAE
B1	0.8630	0.7269	0.6778
B2	1.2516	0.5953	1.1606
B3	1.1352	0.7423	0.9050
B4	0.4689	0.8755	0.3038

4.3 Discussions

In terms of validations and accuracy using data-driven methods, different models are applied to measure the result of B4 station as an example, including linear regression (LR), regression tree (RT), support vector machine (SVM) and Gaussian Process Regression (GPR). The results are given in Table IV. The calculations were implemented in Matlab R2018b in an Inter® Core™ i7-8550U CPU at 1.80 GHz workstation with 16 GB RAM and a Radeon Pro WX3100 GPU. The R² values seem little higher by applying LSTM model than other popular approaches. The GPR model can predict the closest values when compared to LSTM model. For the time consuming, the LSTM seems much quicker than SVM and GPR because the average time consuming for running an LSTM model takes approximately 20 seconds. The SVM and GPR would take much longer time if more data are required. It worth mention that the process of PCA only costs 0.029 s. However the training time of LSTM model may reduce approximately 124 seconds if using PCs resulted from PCA method. In other words, the PCA analysis can reduce the training time about 104 s in this case.

Table IV. The model results from different supervised modelling approaches

Methods	RMSE	R ²	MAE	Time (s)
LR	0.6182	0.62	0.4720	3.72
RT	0.5292	0.72	0.3492	0.63
SVM	0.4530	0.79	0.3323	40.99
GPR	0.3723	0.86	0.2796	119.22
LSTM	0.4083	0.89	0.3014	18.53

It is found from Figure 6 that the proposed LSTM model is more sensitive in high values than in low values. At the time steps from 0 to 50 and from 250 to 300 at B1 to B3 stations, the predicted curves deviate far from these observed curves, meaning a poor prediction. At the higher wave heights, the smooth fitting curve may not be able to oscillate frequently and strongly. This phenomenon can also be partially explained by the nature of RNN. The RNN, from which the LSTM is derived, has the ability of keeping the old value in its memory cells. When the old value becomes larger, they may affect the prediction results. If the old value appears nonsense to the new prediction, it is necessary to remove them and keep the useful value in the model. However, although the PCA method is able to decide which information could be retained, it is still needed to process the selection further before being applied to the LSTM model. This implies different inputs to the model could evidently affect the result.

Besides, the wave height has plenty of physical attributes in both time-

series and space domain. In order to predict the wave height more precisely, it is necessary to consider information such as distance and location from the destination. For example, at a time step, the wave condition can behave similarly to what happens in a kilometer-scale space. The wave caused by wind may consider the direction of the wind whilst the wave direction should be as same as wind. That is why the wind-wave models could provide a more accurate forecasting. It can also be found from the results that the performance of each buoy station varies from each other in the R^2 value ranging from 0.60 to 0.88. This phenomenon might partially be because of the distribution and weather condition of the buoys. The buoy distribution will affect the prediction result and experience from numerical wave model shows that a grid distribution of the buoys may present better results.

5. Conclusions

There are a number of ways to forecast wave height in next hours or even days. With the third generation numerical wave models, the results could reach the accuracy standards for ocean engineering as well as maritime activities in recent decades. However, in the polar region, the shortage of data caused partly by scarce human activities and harsh ocean conditions leads to rough prediction performance even by those numerical models. The novel LSTM method presented in this paper not only provides a good and accurate prediction but also requires fewer data inputs and time consuming. The results in this case study show the LSTM network has the potential to perform prediction at a small-scale polar region by using a few amounts of data. This demonstrates that the proposed approach can be used in the maritime forecasting, aquaculture and wave energy utilization.

The model results provide acceptable values in term of performance measure indexes of R^2 and RMSE values of 0.89 and 0.4083 respectively. After comparison, the results perform much better in terms of both performance metrics and time consuming when compared with LR, RT, SVM and GPR. The PCA method is used to extract PCs from input signals, by which the LSTM adopts to avoid long term independences during the forecasting. The different performance from each buoy might be partly because of the distribution and weather condition of the data station from the buoys.

Finally, discussions on the shortages of high-value wave prediction and the curve-fitting accuracy are made. The study shows that the number of the features extracted from inputs directly dominate the prediction results and therefore future work will focus on wave height feature extraction. Furthermore, the long-terms (7 days to 15 days) prediction method has been widely needed in specific fields which will also be studied in the future work.

Acknowledgements: The first author also acknowledges the support received by SFMRE (Special Funds for Marine Renewable Energy) from MNR (Grant number GHME2019ZC01). We also gratefully acknowledge the colleagues in National Ocean Technology Center for the valuable comments and suggestions that have helped improve the manuscript.

References

- [1] Deepthi. I. Gopinath, G. S. Dwarakish, 2015. Wave prediction using neural networks at New Mangalore Port along west coast of India. In: International Conference on Water Resources, Coastal and Ocean Engineering. Aquatic Procedia. 4, 143 - 150.
- [2] Deo, M. C., Abhay Jha, A. S. Chaphekar and K. Ravikant, 2001. Neural networks for wave forecasting, *Ocean Engineering*. 28(7), 889-898.
- [3] Wittmann, Paul and Mike Clancy, 2004. Thirty Years of Operational Ocean Wave Forecasting at Fleet Numerical Meteorology and Oceanography Center. In: Symposium on the 50th Anniversary of Operational Numerical Weather Prediction, pp. 14-17 June 2004, University of Maryland.
- [4] J. Wolf, 2008. Coupled wave and surge modelling and implications for coastal flooding. *Adv. Geosci.* 17, 19 - 22.
- [5] Tolman, 2002g: User manual and system documentation of WAVEWATCH-III version 2.22. NOAA / NWS / NCEP / MMAB Technical Note 222, 133 pp.
- [6] Komen, GJ and Cavaleri, L. and Donelan, M. and Hasselmann, K. and Hasselmann, S. and Janssen, P. et al, 1994. Dynamics and Modelling of Ocean Waves, Cambridge, 534 pp.
- [7] Surf's Up: Professor Using Models To Predict Huge Waves, ScienceDaily, Feb. 23, 2005.
- [8] Justin Thomas T, G S Dwarakish, 2015. Numerical wave modelling - A review. *Aquatic Procedia*. 4, 443 - 448.
- [9] Scott C., Zhang Y., Fearghal O'Donncha, 2018. A Machine Learning Framework to Forecast Wave Conditions. *Coastal Engineering*. 137, 1-10.
- [10] C. Feng, M. Cui, B. Hodge, J. Zhang, 2017. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy*. 190, 1245 - 1257.
- [11] D. J. Peres, C. Iuppa, L. Cavallaro, A. Cancelliere, E. Foti, 2015. Significant wave height record extension by neural networks and reanalysis wind data. *Ocean Modelling*. 94, 128 - 140.
- [12] E. Alexandre, L. Cuadra, J. C. Nieto-Borge, G. Candil-García, M. delPino, S. Salcedo-Sanz, 2015. A hybrid genetic algorithm—extreme learning machine approach for accurate significant wave height reconstruction. *Ocean Modelling*. 92, 115 - 123.
- [13] Y. LeCun, Y. Bengio, G. Hinton, 2015. Deep learning. *NATURE*. 521.
- [14] N. Krishna kumar, R. Savitha, Abdullah Al Mamun, 2017. Regional ocean wave height prediction using sequential learning neural networks. *Ocean Engineering*. 129, 605-612.
- [15] Shikhar, S., Stefan, L. A., 2018. Comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data. *Solar Energy*. 162, 232-247.
- [16] Mohammad Pirhooshyaran, Lawrence V. Snyder, 2019. Multivariate, Multistep Forecasting, Reconstruction and Feature Selection of Ocean Waves via Recurrent and Sequence-to-Sequence Networks. arXiv preprint arXiv:1906.00195.
- [17] Chih-Chiang Wei, Ju-Yueh Cheng, 2020. Nearshore two-step typhoon wind-wave prediction using deep recurrent neural networks. *Journal of Hydroinformatics*. 22(2), 346-367.
- [18] T. Wu, Y. Cao, et al., 2019. Deep learning for inversion of significant wave height based on actual sea surface backscattering coefficient model. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-019-07967-6>

- [19] M. Zaytar, C. Amrani, 2016. Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks. *International Journal of Computer Applications*. 143(11).
- [20] Hu, R., Fang, F., Pain, C. C., Navon, I. M., 2019. Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method. *Hydrology*. 575: 911-920.
- [21] Y. Liu, L. Guan, C. Hou, H. Han, et al., 2019. Wind Power Short-Term Prediction Based on LSTM and Discrete Wavelet Transform. *Applied Sciences*. 9(6), 1108.
- [22] Williams, R.J., Zipser, D., 1992. Gradient-based learning algorithms for recurrent networks and their computational complexity. In *Back-propagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum,.
- [23] Sepp, H., Jürgen, S., 1997. Long short-term memory. *Neural Computation*. 9 (8): 1735 - 1780.
- [24] Felix, A., Jürgen, S., Fred, C., 2000. Learning to forget: Continual Prediction with LSTM. *Neural Computation*. 12 (10), 2451-2471.
- [25] Yan, K., Wang, X., Du, Y., Jin, N., Huang, H., Zhou, H., 2018. Multi-Step Short-Term Power Consumption Forecasting with a Hybrid Deep Learning Strategy. *Energies*. 11, 3089.
- [26] Wang, Y., Xie, D., Wang, X., Zhang, Y., 2018. Prediction of Wind Turbine-Grid Interaction Based on a Principal Component Analysis-Long Short Term Memory Model. *Energies*. 11, 3221.
- [27] Shi, X., Lei, X., Huang, Q., Huang, S., Ren, K., Hu, Y., 2018. Hourly Day-Ahead Wind Power Prediction Using the Hybrid Model of Variational Model Decomposition and Long Short-Term Memory. *Energies*. 11, 3227.
- [28] Afan, G. S., Yaya, H., Edi, A., Wayan, S., 2018. Single Layer & Multi-layer Long Short-Term Memory (LSTM) Model with Intermediate Variables for Weather Forecasting. *Procedia Computer Science*. 135, 89 - 98.
- [29] Tian, Y., Xu, Y., et al., 2018. Integration of a Parsimonious Hydrological Model with Recurrent Neural Networks for Improved Streamflow Forecasting. *Water*. 10, 1655.
- [30] Hu, C., Wu, Q. et al., 2018. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water*. 10, 1543.
- [31] Liu, J., Zhang, T. et al., 2018. TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction. *Sensors*. 18, 3797.
- [32] Md, J. H., Kim J. M., 2018. Bearing Fault Diagnosis under Variable Rotational Speeds Using Stockwell Transform-Based Vibration Imaging and Transfer Learning. *Appl. Sci*. 8, 2357.
- [33] Gers, F. A., Schmidhuber, J., 2001. LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages. *IEEE Transactions on Neural Networks*. 12 (6), 1333-1340.
- [34] Gers, F., Schraudolph, N., Schmidhuber, J., 2002. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*. 3, 115 - 143.
- [35] Klaus, G., Rupesh, K. S., Jan, K. et al., 2015. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*. 28 (10), 2222
- [36] Pearson, K., 2010. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*. 2 (11): 559 - 572.
- [37] Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 24, 417 - 441 and 498 -

520.

[38] Arkaitz, R., Alain, U., et al., 2019. MIDAS: A Benchmarking Multi-Criteria Method for the Identification of Defective Anemometers in Wind Farms. *Energies*. 12, 28.

[39] PennState Eberly College of Science. Applied Multivariate Statistical Analysis: Principal Component Analysis (PCA) Procedure. Available: <https://newonlinecourses.science.psu.edu/stat505/node/51/>

[40] Liu, H., Du, Z., 2013. Some Problems in Comprehensive Evaluation of Principle Component Analysis. *Statistical Research*. 30 (8).

[41] Steel, R. G. D., Torrie, J. H., 1960. *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*. McGraw Hill.

[42] Glantz, Stanton A., Slinker, B. K., 1990. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill.

[43] What Are the Natural Resources of Antarctica. [online] Available at: <https://www.reference.com/geography/natural-resources-antarctica-2888c72e9337cdb4>

[44] Worldatlas. What Are The Westerlies. [online] Available at: <https://www.worldatlas.com/articles/what-are-the-westerlies.html>

[45] Rebecca R. Andridge, Roderick J. A. Little, 2010. A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*. 78(1), 40 - 64.

[46] Georgiev G.Z. Mean, Median, Mode Calculator. [online] Available at: <https://www.gigacalculator.com/calculators/mean-median-mode.php> URL [Accessed Date: 24 May, 2019].

[47] Wang, Y., Ma, X. Joyce, J.J., 2016. Reducing sensor complexity for monitoring wind turbine performance using principal component analysis. *Renewable Energy*. 97(13), 444 - 456.