Renewable and Sustainable Energy Reviews 132 (2020) 110003

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: http://www.elsevier.com/locate/rser

Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms

Mumtaz Ali [a], Ramendra Prasad [b,*], Yong Xiang [a], Ravinesh C. Deo [c]

[a] Deakin-SWU Joint Research Centre on Big Data, School of Information Technology, Deakin University, VIC 3125, Australia
[b] The University of Fiji, School of Science and Technology, Department of Science, Saweni, Lautoka, Fiji
[c] School of Sciences, Centre for Applied Climate Sciences & Centre for Sustainable Agricultural Systems, University of Southern Queensland, Springfield, QLD 4300, Australia

ARTICLE INFO

Keywords:
Wave energy
Significant wave height
MLR
CWLS
MARS
M5 tree

ABSTRACT

Globally, major emphasis is currently being put in utilization and optimization of more sustainable and renewable energy resources, to overcome the future energy demand issues and potential energy crises due to many socioeconomic factors. A near-real-time *i.e.*, half-hourly significant wave height (H_{sig}) forecast model is designed using a suite of selected model input variables where the multiple linear regression (MLR) model, considering the influence of several variables, is optimized by covariance-weighted least squares (CWLS) estimation algorithm to generate a hybridized MLR-CWLS model with a capability to forecast 30-min ahead H_{sig} values. First, a diagnostic statistical test based on the correlation coefficient is performed to determine relationships between inputs denoting historical behaviour and the target (H_{sig}) at one lag of 30-min ($t-1$) scale. Subsequently, the data are split into training and testing subsets, following a normalization process, and the MLR-CWLS hybridized model is then trained and validated on the testing dataset adopted from eastern coastal zones of Australia that has a high potential for wave energy generation. Hybridized MLR-CWLS model is benchmarked against competing modelling approaches (multivariate adaptive regression splines-MARS, M5 Model Tree, and MLR) via statistical score metrics. The results show that the hybridized MLR-CWLS model is able to generate reliable forecasts of H_{sig} relative to the counterpart comparison models. The study ascertains the practical utility of the hybridized MLR-CWLS model for H_{sig} modelling with significant implications for its potential application in wave and ocean energy generation systems, and some of the other renewable and sustainable energy resource management.

1. Introduction

The juxtaposition between warming of the natural climate system and the subsequent climate change risk arising from consequent rise in sea level driven by atmospheric carbon dioxide at alarming rates is necessitating the need for modern and adaptive measures that can help ameliorate the impacts of global warming on human societies. A key mitigating strategy for global warming and its impacts is to implement efficient renewable energy (RE) resources and environmentally friendly energy technologies that promote cleaner energy production. Such economically and socially responsible energy harnessing devices, specially designed for RE utilization, must, therefore, be actively promoted, as advocated by the United Nations Sustainable Development Goal (SDG). In particular, the SDG# 7 aims to empower many RE-rich nations to ensure its citizens' appropriate access to, and utilization of reliable, sustainable and modern energy harnessing methods that have benefits spanning across a spectrum of areas such as the eradication of poverty and mitigation of climate change impacts [1]. Since the access to clean energy can vary widely among developing and first-world nations, significant efforts are still required in many countries where the RE resources are available in abundance, but they have either been underutilized or have a large energy access deficits despite the rising energy consumption profiles. Therefore, the development of novel energy harnessing technologies that can capture a wide range of resources (*e.g.*, solar, wind, hydro, geothermal and wave energy) are paramount tools required to support UN's SDGs to help meet both the energy

* Corresponding author.
E-mail addresses: mumtaz.ali@deakin.edu.au (M. Ali), ramendrap23@gmail.com, ramendrap@unifiji.ac.fj (R. Prasad), yong.xiang@deakin.edu.au (Y. Xiang), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.rser.2020.110003
Received 24 September 2019; Received in revised form 3 June 2020; Accepted 15 June 2020
Available online 23 July 2020

demand targets and also, to provide a cleaner option that helps combat climate change impacts on global citizens.

As the knowledge and the awareness of climate change science increases, advancements in RE technologies are inevitable. Despite these significant advancements, the technologies for electricity generations from renewable ocean energy resources, which is the focus of this research paper, has not been accepted so widespread [2]. For a country such as Australia that is surrounded by a vast ocean and having a high ocean energy generation potential, the ocean electricity generation projects are likely to massively reduce the carbon emissions and enable the country to meet the agreed Nationally Determined Contributions (NDC) targets of reducing the GHG emissions up to 26–28% by the year 2030 [3]. The International Renewable Energy Agency (IRENA) also recommends more research in the area of energy optimization and other operation to develop next-generation wave energy technologies [4]. Despite the fact that wave energy resources exhibit better consistency and certainty in comparison to some of the other sources such as wind energy resources [5], the occasioning of waves from fluid pressure differences in the ocean makes the wave behaviour chaotic. Subsequently, the electricity generation from the 'never still ocean' can be erratic, and so, might require relevant wave energy forecasting techniques to be explored when emulating future wave energy generation capabilities for consistent and reliable electricity supply.

In its practical terms, the energy output from an ocean-based wave source is predominantly dependent on the density of water, wavelength, and the actual wave height (or its respective amplitude). Since the density of water is constant and the wavelength is quite predictable, the primary varying-parameter for ocean wave energy that plays a critical role in wave energy generation is the significant wave height (denoted as H_{sig}) [6]. For wave forecasting, the common practice is to use the numerical weather prediction (NWP) modelling. Additionally, the European Centre for Medium-Range Weather Forecasting (ECMWF) has also been enhanced with the incorporation of a wave modelling feature. However, the global wave field forecasts associated with the NWP modelling approaches and the ECMWF wave model are at specific grid points, which are horizontally distanced [7,8]. With that, physically-based NWP modelling is based on large-scale features such as the progression of high and low-pressure systems for wind generation, large-scale oceanic circulation, ocean currents, overturning, and the global incident solar radiation, yet the finer spatial scale processes can be left out [7]. To compound that, relevant hydro-meteorological forcing is required to construct a forecasting model [9]. Another intricacy is that NWP is affected by numerous uncertainty sources, which requires ensemble and statistical post-processing via incorporating past information verifications [8]. The data-driven models, on the other hand, provides an alternative modelling approach due to its design and usage simplicity requiring historical data for model development.

Among the other global players, Australia has been trialling wave energy conversion systems and related convertors technologies in various capacities. For instance, the Perth Wave Power Project (PWEP) has installed CETO-5 in the Garden Island of Western Australia, and the Stage 1 has generation capacity of 2 MW peak capacity launched in 2010, and this project was completed in April 2011. The second stage has a capacity of 3 MW at peak [10,11]. In another emerging project in wave energy, Carnegie Clean Energy Ltd has installed the Carnegie's 6th generation CETO wave energy unit (i.e. the CETO 6 system) that is more efficient and aims to deliver increased power generation. This project, located in Albany, Western Australia, started in June 2014, was completed in October 2015, and has an installed capacity of 1 MW [11]. In another project developed by BioPower Systems Pty Ltd, which started installation in June 2012 with a capacity of 250 kW at Port Fairy, Victoria used the bioWAVE technology to efficiently convert wave energy to mechanical energy utilizing a unique mechanical-to-electrical energy converter, and it was decommissioned after the testing ceased in 2017 [11]. Furthermore, the Oceanlinx oscillating water column (OWC) technology that has a capacity of 1 MW started in June 2012 at Port MacDonnell, South Australia also sustained damages beyond repair in 2014 [11]. Although the island-continent Australia is not constrained by the availability of wave energy resource, the issues such as the economic efficiency of energy extraction, including its transmission, environmental and social impacts, and the factors of intermittency [5] remain as key challenges. Therefore, the forecasting of short-term H_{sig} may be able to overcome some of these challenges by providing new modelling technology for these systems and help attain better economic outcomes in terms of energy extraction and intermittency issues. The values of H_{sig} are contingent upon the atmospheric and oceanographic factors including wind speed and direction, sea-surface-temperatures [12–17], and oceanographic bathymetry. The effects of these potential inputs for wave energy need to be incorporated in H_{sig} forecasting studies. The study of Mahjoobi and Etemad-Shahidi [18] in their study of hindcasting of the H_{sig} values have shown that the wind speed and its direction can be one of the main factors for wave energy generation. Previous literature shows the most common model input for wave energy was considered to be the wind-speed, followed by the wind direction [18–22] while the other less important model inputs included the fetch-length, wind duration [21], and the antecedent (or historical value of) H_{sig} [23]. Yet in order to fully grasp and understand the non-linear dynamics of the stochastic ocean waves, a number of other input variables need to be also used. Therefore, in this research study, five different input variables including the sea-surface-temperature and antecedent H_{sig} are employed as the model inputs to forecast the future value of H_{sig}.

In any prediction problem that is focussed on wave energy generation, short-term as well as long term forecasting of H_{sig} is crucial for monitoring the electricity generation process [24], so both of these methods have resulted in a number of new modelling and forecasting approaches for H_{sig}. For example, the study of Mahjoobi and Etemad-Shahidi [18] have explored the prediction of ocean wave energy parameters at 1-h intervals, using a regression tree-based learning algorithm. In addition, a soft-computing technique based on the Adaptive-Network-Based Fuzzy Inference System [20,21] was also trialled at 1-h interval ocean energy prediction. In a number of other studies, artificial neural networks (ANN) [18–20] and support vector machines [19,20] have been developed to forecast H_{sig} at 1-h [18,21] and ½ hour intervals [19], respectively. From a review of these applications of computational intelligence in wave energy forecasting, the study of Cuadra, Salcedo-Sanz [25] established that an ANN was the most commonly adopted model. In a more recent study, Ali and Prasad [23] developed a hybrid extreme learning machine algorithm that satisfactorily predicted significant wave height using historical lagged series of H_{sig} at 30-min intervals. It should be noted that previous methods have largely used black-box type models, however; the application of other approaches such as a decision tree (*e.g.*, M5 model Tree) and the more simpler method such as Multiple Linear Regression (MLR) that uses clear rules between the model inputs and target variable (H_{sig}) have not been explored to forecast H_{sig}. In addition, MLR models can easily be updated when the newer chunks of input and target data arrive, so this transparency in terms of the association between inputs and outputs and their flexibility in implementation has prompted for wider acceptability of these models in real-life applications. In spite of their usage in energy prediction problems (e.g. Ref. [26]), a traditional MLR model can only be utilized in a univariate sense, aiming to establish associations between one independent variable and one dependent variable, making it overly simplistic with very limited practical applicability. As a practical model with its candidate variables is dependent upon many interacting variables in real-life situations, the MLR model can perform poorly if the non-linear dynamics within model inputs are directly introduced without a prior pre-processing of the input data, since the model is unable to capture the complex interactions and patterns embedded within these time-series data sets. One drawback of the MLR model is a lack of optimization approach that may deter the model from capturing non-linear dynamics in input variables.

To address these issues, the aim of this paper is:

1. To design a novel multiple linear regression model (MLR) to forecast H_{sig} using multiple inputs (five different input predictors and antecedent H_{sig}) at near-real-time *i.e.*, half-hourly, forecast horizons.
2. A robust optimization algorithm based on the Covariance-Weighted Least Squares (CWLS) estimation method is ardently applied to optimize the prescribed MLR model leading to the development of the MLR-CWLS model. The CWLS algorithm can easily decrease the objective function with a fixed covariance diagonal matrix [27] where the covariance diagonal matrix consists of relative weights. Moreover, the CWLS computes the covariance of coefficients using the initial covariance [28]. It has the power to detect the number of parameters and covariance matrix which makes it more stringent than other optimizing algorithms [27].
3. The results of improved MLR-CWLS are benchmarked with other competing models that have a lucid structure including the standalone MLR, M5tree, and MARS models.

The newly designed hybridized MLR-CWLS model (which aims to generate improved H_{sig} forecasts) and is benchmarked with a number of comparison models, is validated at three high wave energy potential study sites in Queensland, Australia, located at the off the shores of Gladstone, Townsville and Tweed Head. The next section of this paper presents the theoretical background of the models (Section 2) followed by the Case Study Description and Data (Section 3), Results (Section 4), Further Discussions (Section 5) and Conclusions (Section 6).

2. Theoretical review

To construct and evaluate a near real-time, half-hourly significant wave height forecasting, we adopt the covariance-weighted least squares estimation algorithm (as described in the next section).

2.1. Covariance-weighted least squares (CWLS) algorithm

An identity error covariance matrix in a multivariate problem generates biased standard error estimates [29]. Let **C0** be an invertible diagonal matrix with dimension d to approximate CWLS utilizing the name-value pair argument [28]. The weights of **C0** contains the inverse matrix with each dimension.

The CWLS can be solved by the following equation for a given $\mathbf{C}_0$:

$$\sum_{\varepsilon=1}^{n} (\imath_\varepsilon - \Psi_\varepsilon \delta)^T C_0 (\imath_\varepsilon - \Psi_\varepsilon \delta) \tag{1}$$

In the above Eq., $\imath$indicate the $nd \times 1$ vector of stacked d-dimensional responses and Ψ denote the $nd \times K$ matrix. The term δ is the solution is the vector. The $K \times 1$ vector of CWLS regression coefficient estimates is the first mvregress algorithm output given as:

$$\delta_{CWLS} = \left(\Psi^T (I_n \otimes C_0)^{-1}\right)^{-1} \Psi^T (I_n \otimes C_0)^{-1} \delta \tag{2}$$

The term $\Psi^T$is the transpose of Ψ matrix while I_n ordinary least squares (OLS) estimates. If $\sum = C_0$, the CWLS becomes a generalized least square method. The square root of diagonal this matrix gives the standard errors of CWLS regression coefficients.

In this study, we adopt the CWLS algorithm, which has the ability to reduce the objective function of input data with a fixed covariance diagonal matrix [27]. Additionally, the inverse of the covariance diagonal matrix consists of relative weights of every single input variable. The CWLS algorithm estimates the covariance of coefficients using the initial covariance [28]. Another advantage of utilizing the CWLS in this study is to identify the number of parameters and covariance matrix since the CWLS algorithm is essentially more robust than some of the other optimizing algorithms [27].

2.2. Multiple linear regression (MLR)

The MLR is a linear regression model that capitalizes on the causativeness among target and input variables utilizing the extreme deviations of the data to estimate their analogous regression coefficients. Another major advantage of MLR is that it has the capability to minimize the variations due to the unexplained "noise". Mathematically, MLR is expressed [30,31] with n being the number of observations of k input variables:

$$\Pi = \theta + \varpi_1 \nu_1 + \varpi_2 \nu_2 + \ldots + \varpi_k \nu_k \tag{3}$$

Where the forecasted H_{sig} is represented by $\Pi(n+1)$, $\nu(n \times k)$ denotes the input vectors, θ and ϖ_i denotes the y-intercept and coefficient respectively [32,33]. The magnitude ofϖ is calculated by the least-squares (*e.g.* Ref. [34,35]). During the training process, a set of Π and Δ matrix are fitted following Eq. (4) to model the causality. The coefficients and the y-intercept in the validation period give the forecast in terms of fitted MLR.

2.3. M5 tree (M5tree) model

The M5tree model [36] is mainly constructed on a binary decision structure. The linear regression establishes an association among predictors and response variables [37] where these variables are partitioned into subgroups at further stages [38]. An N-dimensional sampling matrix of training data of input predictors (H_{max}, T_z, T_p, Dirr, and SST) with respect to the response data (H_{sig}) is constructed by M5tree [39] algorithm. The whole process is recursively constructed on divide-and-conquer rule splitting the N data samples by creating similar subsets to test using standard deviation and the reduction of error $\pi_{\Re}$ [39,40] as in the following:

$$\pi_{\Re} = \pi(\Phi) - \sum \left(\frac{\Phi_k}{\Phi} \pi(\Phi_k)\right) \tag{4}$$

Here π indicates the set of examples and Φ_k is the respective*jth* result. The best split of data sample containing patterns and attributes acquired to optimize the M5tree model with a smaller$\pi_{\Re}$until all the sample data reaches a stationary node. The pruning is established due to a complex network structure formed as a result of splitting the input data. Any simple procedure is required to avoid the unexpected breaks arise in the minor training data among the neighbouring linear models at the leaves of the pruned tree [39,40]. This process improves the accuracy to update the linear equation of the model [36].

2.4. Multivariate auto-regressive spline (MARS)

The MARS model non-parametrically consolidates a set of simplified linear functions in an additive style [41]. Fundamentally, the MARS divides the training set into numerous splines over identical breaks, and then further split these into many subclasses separated by knots. At respective knots, a pair of basis functions (F_B) depicting the correlations between the input variables and the predictant are generated leading to continuous models with continuous derivatives [41]. The output (O) at the knot of the ith sub-group positioned at κ can mathematically be found as:

$$\left.\begin{array}{l} O = F_{B_i}(x) = \max imum(0,\ x - \kappa) \\ \quad \textit{and the respective mirror} \\ O = F_{B_i}(x) = \max imum(0,\ \kappa - x) \end{array}\right\} \tag{5}$$

The final modelled output, $\boldsymbol{Y}$', is the summation of a series of F_Bs that generates a large number of potential knots building a bulky model that generally over-fits the training data as:

$$\boldsymbol{Y}' = \lambda + \sum_{i}^{I} (\tau_i \times F_{B\,i}) \tag{6}$$

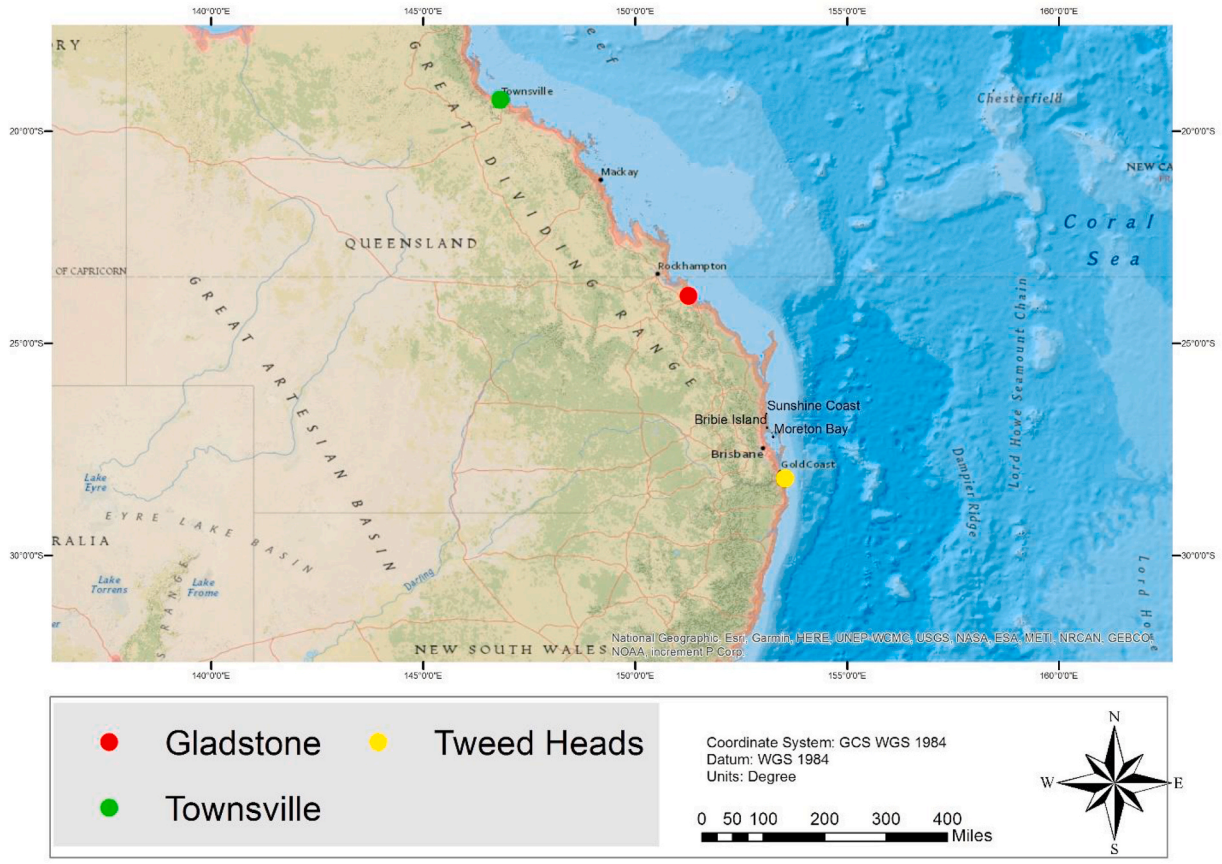

Fig. 1. Map of the selected wave modelling coastal study stations in Queensland.

Where λ is a general constantis a unique constant for respective F_B and I represent the maximum number of subgroups.

Then, a backward deletion phase is iteratively engaged to prune those F_Bs that make the least contributions with respect to model fitting based on minimum values of Generalized Cross-Validation (*GCV*) and finally, the optimum model with smallest *GCV* values are selected.

3. Case Study Description and Data

3.1. Study locations

Australia is the largest island-continent with a vast and diverse coastal region and approximately 85% of the populations living within 50 km of these coasts. Australia has great potential for wave energy derived from coastal waves. In this paper, the coastal stations are picked cautiously which are representative of the diverse geophysical and bathymetric conditions. The Gladstone coastal station located in Queensland, Australia where the wave data is gathered by oceanographic wave measuring buoys anchored. The Buoy Type is Datawell WR Mk3 0.9 Cu and the water-depth is 16 m. The highest wave, as recorded at this site, was 6.1 m at 08:00 p.m. in February 2010 [42]. The second station is Townsville located in Queensland, which is vulnerable to tropical cyclones developing out in the Coral Sea. The data is also collected by anchored oceanographic wave measuring buoys [42]. The Buoy Water Depth is 16 m and during the study period the highest wave observed was 10.1 m in February 2011. The Tweed head coastal station is located in New South Wales next to the border of the Gold Coast. Here the Buoy type is Datawell WR Mk4 0.9 Cu with water depth 24 m. In May 1996, the highest recorded wave was 13.1 m. Fig. 1 displays the map of the coastal sites.

3.2. Data collection and preparation

The coastal waves are periodic in nature where crests and troughs

Table 1
Basic statistics of the inputs (or predictor variables) and the target variable (i.e., significant wave height) including the geographic coordinates of the study locations.

Sites	Gladstone					Townsville					Tweed Head				
Long. (°S)	23.53°					19.09°					28.10°				
Lat. (°E)	151.30°					147.03°					153.34°				
Elevation (m)	6.0 m					16.3 m					1 m				
Input predictors	Mean	Max	Min	Std.	Skew	Mean	Max	Min	Std.	Skew	Mean	Max	Min	Std.	Skew
H_{max}	1.2	6.8	0.1	0.5	1.1	1.1	10.1	0.1	0.6	1.3	2.0	11.8	0.4	0.8	1.8
Tz	4.79	17.2	1.6	1.7	2.3	3.9	14.7	1.7	1.0	2.0	6.8	21.2	3.0	2.0	1.5
T_p	6.2	18.9	1.6	2.6	0.9	4.7	25.6	1.7	1.6	1.6	8.6	21.2	2.6	2.7	0.3
Dir	88.8	359.6	−90.3	65.9	3.2	88.6	365.6	−92.4	54.7	3.1	94.9	358.0	1.0	23.2	−0.7
SST	22.9	32.2	0.03	3.5	−0.7	25.5	33.5	0.0	4.0	−2.6	21.9	28.0	5.0	3.1	−0.6
H_{sig}	0.70	3.15	0.12	0.34	1.0	0.6	5.5	0.1	0.3	1.3	1.2	6.7	0.3	0.5	2.0

persistently alternate and repeat during the propagation. The data consisted of the following objective and input variables:

Significant wave height (H_{sig}): is the objective variable at a 30-min interval and it is gauged as the average of the highest one-third of wave heights in a wave record. This is the approximated wave height that someone can see. The H_{sig} is proportional to the square root of total wave energy that determines how much bulk wave energy is available at that specific position.

Maximum wave height (H_{max}): is one of the important predictor variables which have a significantly largest statistical relationship with the objective variable (Fig. 3). It is characterized as the height of the highest single-wave recorded.

Zero Up-Crossing Wave Period (T_z): is another predictor which reasonably affects the H_{sig}. It is the mean value of the zero up crossing wave periods in a wave record.

Peak energy wave period (T_p): is the wave period that carries the maximum amount of energy and is characterized as the time accompanying the most energetic waves in the total wave spectrum at a specific point.

Direction (Dir): is the direction that the peak waves are coming from. The statistical relationship shows this predictor variable has a reasonable influence on H_{sig}.

Sea surface temperature (SST): is the approximated sea surface temperatures at the respective wave monitoring buoys measured in degrees Celsius.

Table 1 presents a summary and some descriptive analysis of the data used in this paper.

3.3. The performance assessment Metrics

The forecasting ability of the MLR-CWLS *vs.* other benchmark models were examined on the basis of statistical metrics based on earlier approaches of [43–51].

I. Correlation coefficient (R):

$$R = \left(\frac{\sum_{\upsilon=1}^{N} (Hsig^{Obs,\upsilon} - \overline{Hsig^{Obs,\upsilon}})(Hsig^{For,\upsilon} - \overline{Hsig^{For,\upsilon}})}{\sqrt{\sum_{\upsilon=1}^{N} (Hsig^{Obs,\upsilon} - \overline{Hsig^{Obs,\upsilon}})^2} \sqrt{\sum_{\upsilon=1}^{N} (Hsig^{For,\upsilon} - \overline{Hsig^{For,\upsilon}})^2}} \right) \tag{7}$$

II. Willmott's Index (E_{WI}) formulated as:

$$E_{WI} = 1 - \left[\frac{\sum_{\upsilon=1}^{N} (Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon})^2}{\sum_{\upsilon=1}^{N} (|Hsig^{For,\upsilon} - \overline{Hsig^{Obs,\upsilon}}| + |Hsig^{Obs,\upsilon} - \overline{Hsig^{Obs,\upsilon}}|)^2} \right], \; 0 \le E_{WI} \le 1 \tag{8}$$

III. Nash-Sutcliffe efficiency (E_{NS}) metric:

$$E_{NS} = 1 - \left[\frac{\sum_{\upsilon=1}^{N} (Hsig^{Obs,\upsilon} - Hsig^{For,\upsilon})^2}{\sum_{\upsilon=1}^{N} (\overline{Hsig^{Obs,\upsilon}} - \overline{Hsig^{For,\upsilon}})^2} \right], 0 \le E_{NS} \le 1 \tag{9}$$

IV. Root mean square error (*RMSE*) is mathematically derived as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{\upsilon=1}^{N} (Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon})^2} \tag{10}$$

V Mean absolute error (*MAE*)

$$MAE = \frac{1}{N} \sum_{\upsilon=1}^{N} \left| (Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon}) \right| \tag{11}$$

VI. Legates and McCabe's (E_{LM}):

$$E_{LM} = 1 - \left[\frac{\sum_{\upsilon=1}^{N} |Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon}|}{\sum_{\upsilon=1}^{N} |Hsig^{Obs,\upsilon} - \overline{Hsig^{Obs,\upsilon}}|} \right], 0 \le E_{LM} \le 1 \tag{12}$$

VII. Root mean squared percentage error (*RMSPE*; %), is stated as

$$RMSPE = \frac{1}{N} \sum_{\upsilon=1}^{N} \left| \frac{(Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon})}{Hsig^{Obs,\upsilon}} \right| \times 100 \tag{13}$$

VIII. Mean absolute percentage error (*MAPE*; %):

$$MAPE = \frac{1}{N} \sum_{\upsilon=1}^{N} \left| \frac{(Hsig^{For,\upsilon} - Hsig^{Obs,\upsilon})}{Hsig^{Obs,\upsilon}} \right| \times 100 \tag{14}$$

In equations (7)-(14), $Hsig^{Obs,\upsilon}$ is observed and $Hsig^{For,\upsilon}$ is the forecasted υ^{th} magnitudes of significant wave height *Hsig*, while the observed and forecasted mean *Hsig* values are represented by $\overline{Hsig^{Obs,\upsilon}}$ and $\overline{Hsig^{For,\upsilon}}$ respectively, where N is the number of the entire testing dataset.

Firstly, the magnitude of correlation coefficient (R) is ranging between −1 and +1, which exhibits the relationships in terms of the proportion of variance in between the observed and forecasted *Hsig* from the machine learning model [46]. A value of +1 illustrates that the observed and forecasted data are highly correlated with the least variances. The Willmott's Index (E_{WI}) assessment metric varies between 0 and 1. The E_{WI} handles the insensitivity problems as the differences between the observed and forecasted data are not squared [52] where the differences are substituted by the ratio of the mean squared error in calculations [49–51,53,54]. The next metric, *i.e.*, Nash-Sutcliffe efficiency (E_{NS}) metric fundamentally equates the variance of forecasted and observed H_{sig} [55] with the ideal magnitude from 1 to negative infinity. To handle the restrictions and weaknesses of E_{WI} and E_{NS} [48] correlation-based metrics, the measures of Legates-McCabe's (E_{LM}) is introduced ranges from 0 to 1 and The *RMSE* and *MAE* errors assessment are centered on the accumulation of residuals between observed and forecasted H_{sig} [56]. The larger H_{sig} are mainly captured by the *RMSE* while the *MAE* likewise evaluates variations entirely irrespective of negative or positive values, however, equally, they range from 0 (ideal value) to positive infinity.

Nevertheless, all the aforementioned measures should not principally be utilized to compare model accuracy at geographically different locations [57]. The relative root mean squared error (*RRMSE*) and mean absolute error (*RMAE*) were employed [46,53] in this regard. The relative magnitudes are in percentages and categorise the models to be excellent when the (*RRMSE, RMAE*) < 10%, good when the range is 10% < (*RRMSE, RMAE*) < 20%, while the model is considered to be fair if 20% < (*RRMSE, RMAE*) < 30% and the model is supposed to have a poor accuracy if the (*RRMSE, RMAE*) > 30% [58,59].

3.4. Modelling strategy

The hybridized MLR-CWLS model has been constructed in the MATLAB R2018b programming environment (The Math Works Inc. USA). All these simulations were attained on 2.93 GHz dual-core PC with Pentium 4 operating system. The construction of MLR-CWLS consists of the following steps:

Step 1. Determining statistically significant lagged inputs

A set of correlograms based on the correlation coefficient (R_{Cross}) was

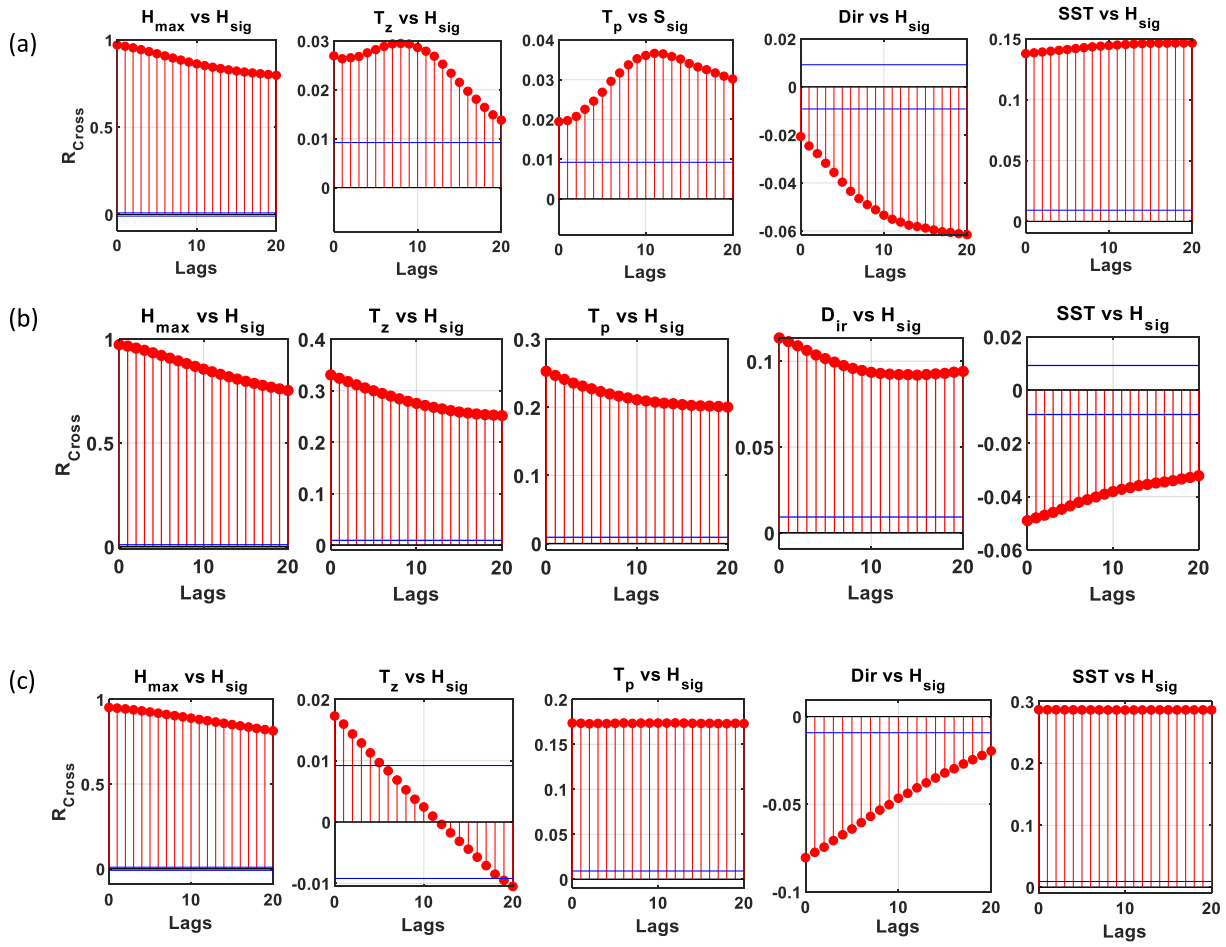

Fig. 2. Correlograms based the correlation coefficient (R_{Cross}) performed on each of the study sites examining covariance between H_{sig} and the predictor variables for (a) Gladstone, (b) Townsville; and (c) Tweed Head. Blue lines indicate the significance of R at the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

performed to examine the covariance between H_{sig} and the predictor variables (H_{max}, T_z, T_p, Dirr, and SST). Fig. 2 shows these statistically significant lagged data of each coastal station. The time-lagged inputs at $(t-1)$ were used to develop the hybridized MLR-CWLS model to forecast the half-hour lead-time ahead Hsig values.

For Gladstone site, the input predictor H_{max} shows a high correlation (R_{Cross}) value of above 0.90 at - time-lagged $(t-1)$, followed by SST with a magnitude of $R_{Cross} = 0.14$, $T_z = 0.03$, $T_P = 0.2$ and Dirr $= -0.02$. Similarly, all input predictors H_{max}, T_z, T_p, Dirr appear to be positively correlated with Hsig except SST where H_{max} indicates a high correlation for the study location Townsville. Again, the H_{max} predictor acquired a significant amount of correlation for Tweed Head. Fig. 2 exhibits the significance of all the input predictors in forecasting near real-time H_{sig}.

Step 2. Data partitioning

The datasets from 1st January 2011 to 31st October 2018 were subdivided straightly into 70% - training and 30% - testing phases, before the model development phase following [60] as it is the most common approach for data partitioning [61]. The total number of data records were 137328 at 30-min time intervals. The purpose of data partitioning is to validate the hybridized MLR-CWLS model independently from training. Moreover, the cross-validation technique or any random sampling procedure cannot be used as time-series data by definition occur in a temporal order/sequence and this order or sequence must be preserved in order to keep the structure of the series intact [62].

Step 3. Normalization procedure

The normalization was performed via. Eq. (15) to smooth the data [63] whereby all data were restrained between [0, 1] and the results are not affected due to the invertible nature of this process [63].

$$\Lambda_{NORM} = \frac{(\Lambda - \Lambda_{MIN})}{(\Lambda_{MAX} - \Lambda_{MIN})} \tag{15}$$

In Eq. (15), Λ denotes the input/output, Λ_{MIN} = the lowest, Λ_{MAX} = the biggest and Λ_{NORM} = the required normalized value.

Step 4. Construction of Hybridized MLR-CWLS Model

In the final stages, the hybridized MLR-CWLS model is developed to forecast half-hour ahead significant wave height values. After combining the statistical significant lags at $(t-1)$ lags into the hybridized MLR-CWLS model, the CWLS algorithm tries to minimize the objective function of predictor variables (H_{max}, T_z, T_p, Dirr, and SST) with a fixed covariance diagonal matrix. Furthermore, the inverse of this covariance diagonal matrix consists of relative weights of each input series. The CWLS algorithm approximates the covariance of coefficients using the initial covariance of the input data. Additionally, the CWLS also determines the number of parameters and covariance matrix of the MLR model. CWLS is essentially more stringent than some of the other optimizing algorithms [27].

In other words, the CWLS to perform least squares (optionally conditionally weighted by an input covariance matrix). The parameter "covtype" has either 'full' (default) to allow a full covariance matrix or 'diagonal' to restrict it to be a diagonal matrix. In our case, covtype is a diagonal matrix. During the development stage, the MLR model utilizes

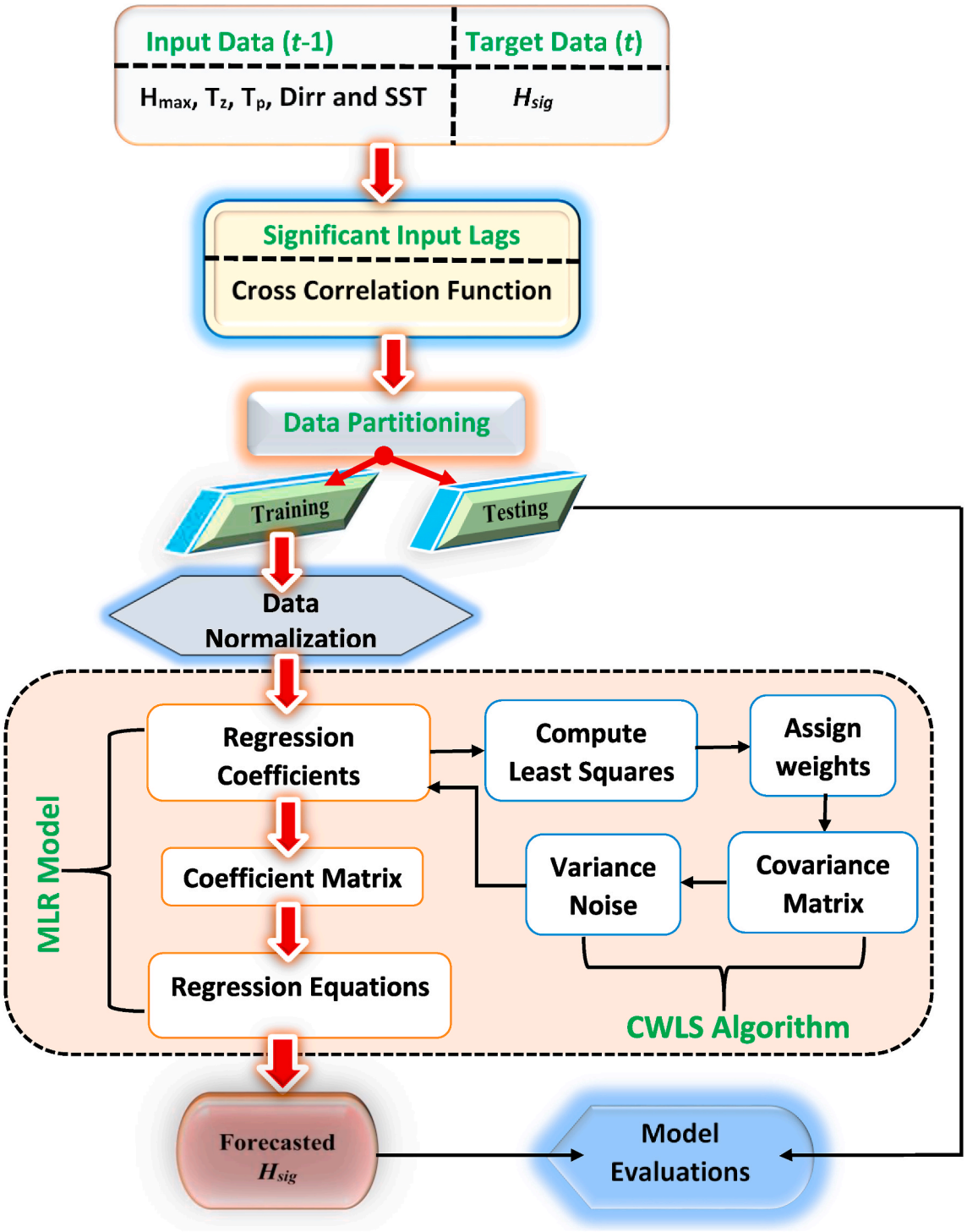

Fig. 3. Schematic view of the modelling strategy using hybridized MLR-CWLS model.

several types of aforementioned parameters including regression coefficients, F-value and covariance matrix. The values of these parameters and coefficients are: F-value = [154072.70 0] Gladstone, [175269.41 0] (Townsville), [94548.37 0] (Tweed Head). The coefficients utilized in the designing of hybrid MLR-CWLS are c = 0.559 (Gladstone), 0.545 (Townsville), 0.549 (Tweed Head); α_1 = 0.001 (Gladstone), 0.006 (Townsville), 0.008 (Tweed Head); α_2 = 0.001 (Gladstone), 0.003 (Townsville), 0.002 (Tweed Head); α_3 = −4.52 (Gladstone), 9.588 (Townsville), 0.00 (Tweed Head); α_4 = 0.001 (Gladstone), 0.00 (Townsville), 0.002 (Tweed Head). After training, the MLR-CWLS model is validated independently on testing data. Fig. 3 shows the schematic view of modelling strategy to develop the hybridized MLR-CWLS model in this paper.

Benchmark comparing Models: The performance of the hybridized MLR-CWLS model is benchmarked with M5tree, MLR, and MARS models. While constructing the M5tree model, the attention was paid to the parameters including model Tree = 1, the minimum number of training data cases = 4, pruning = 1, smoothing = 0, smoothing coefficient = 15 and split threshold = 0.05. On the other hand, the MLR model utilizes several types of parameters such as regression

Table 2
Design parameters and coefficients of the standalone MLR model.

Sites	C	b_1	b_2	b_3	b_4	b_5
Gladstone	0.002	1.217	0.006	0.005	−0.007	0.003
Townsville	−0.007	1.096	0.042	0.053	0.004	1.426
Tweed Head	0.010	0.972	0.027	0.007	−0.029	0.008

Table 3
Testing performance of hybridized MLR-CWLS *vs.* benchmark models using correlation coefficient (*R*), root mean square error (*RMSE*), and mean absolute error (*MAE*).

	R	***RMSE*** **(m)**	***MAE*** **(m)**
Gladstone			
MARS	0.970	0.33	0.26
M5tree	0.929	0.15	0.09
MLR	0.971	0.08	0.06
MLR-CWLS	**0.971**	**0.08**	**0.06**
Townsville			
MARS	0.967	0.06	13.35
M5tree	0.930	0.08	20.50
MLR	0.975	0.06	13.23
MLR-CWLS	**0.975**	**0.05**	**11.28**
Tweed Head			
MARS	0.960	0.24	24.79
M5tree	0.927	0.14	15.55
MLR	0.960	0.11	12.16
MLR-CWLS	**0.960**	**0.10**	**11.66**

Table 4
Testing performance of hybridized MLR-CWLS *vs.* benchmark models using Willmott's Index (E_{WI}), Nash-Sutcliffe efficiency (E_{NS}) and Legates and McCabe's (E_{LM}).

Gladstone			
Models	E_{WI}	E_{NS}	E_{LM}
MARS	0.833	−0.029	−0.028
M5tree	0.872	0.798	0.644
MLR	0.948	0.932	0.763
MLR-CWLS	**0.959**	**0.944**	**0.785**
Townsville			
MARS	0.968	0.930	0.762
M5tree	0.891	0.834	0.691
MLR	0.945	0.931	0.753
MLR-CWLS	**0.964**	**0.950**	**0.797**
Tweed Head			
MARS	0.913	0.635	0.378
M5tree	0.885	0.856	0.635
MLR	0.927	0.912	0.709
MLR-CWLS	**0.936**	**0.919**	**0.725**

coefficients, F-value and covariance matrix as illustrated in Table 2. In formulating the MARS model, ARESLab toolbox (version 1.13.0) [64], was utilized whereby a piecewise cubic MARS model was constructed following a two-phase strategy, *i.e.*, forward selection and backward deletion. The iterative addition of basis function pairs to the initial intercept term is carried out during the forward selection phase, to minimize the objective function (*i.e.*, *MSE*) which eventually creates a large model. Therefore a backward deletion phase is executed to reduce the model bulkiness and overfitting, whereby the model is pruned from backward and the best-performing model is the one with least *GCV* value.

4. Results

The hybridized MLR-CWLS model is compared against a MARS, M5tree, and MLR model in forecasting H_{sig} at three coastal sites in Queensland, Australia. The performance accuracy of MLR-CWLS is evaluated with respect to the benchmark models on the basis of statistical metrics and diagnostic plots and the results are presented in this section.

The hybridized MLR-CWLS and the MLR models applied for forecasting H_{sig} values for Gladstone, produced reasonable accuracy in terms *R*, *RMSE* and *MAE* values ($R \approx 0.971$, $RMSE \approx 0.08$ m, $MAE \approx 0.06$ m) followed by M5tree ($R \approx 0.929$, $RMSE \approx 0.15$ m, $MAE \approx 0.09$ m) and MARS ($R \approx 0.970$, $RMSE \approx 0.33$ m, $MAE \approx 0.26$ m). Again, the hybridized MLR-CWLS and MLR models achieved the same performance to forecast H_{sig} for Townsville but the hybridized MLR-CWLS is slightly better ($R \approx 0.10$, $RMSE \approx 0.10$ m, $MAE \approx 11.66$ m) than MLR and other benchmark models for Tweed Head station. Overall, the preciseness of the hybridized MLR-CWLS model is reasonably good for all three tested sites (Table 3).

According to Table 3, the performance of the hybridized MLR-CWLS and the MLR model is somewhat similar based on *R*, *RMSE*, and *MAE* and we cannot distinguish which best model. Therefore, Table 4 examines the hybridized MLR-CWLS model's efficiency in predicting half-hourly H_{sig} using E_{WI}, E_{NS}, and E_{LM} criterion. The metrics for Gladstone are the hybridized MLR-CWLS model ($E_{WI} \approx 0.959$, $E_{NS} \approx 0.944$ and $E_{LM} \approx 785$), MLR ($E_{WI} \approx 0.948$, $E_{NS} \approx 0.932$ and $E_{LM} \approx 0.763$), M5tree ($E_{WI} \approx 0.872$, $E_{NS} \approx 0.798$ and $E_{LM} \approx 0.644$) and MARS ($E_{WI} \approx 0.833$, $E_{NS} \approx -0.029$ and $E_{LM} \approx -0.028$). The outcomes of the hybridized MLR-CWLS model and other benchmark predictive models (Table 3) for Townsville and Tweed Head confirms the superiority of the hybridized MLR-CWLS for all study sites in this paper (Table 4). Hence, based on the above criteria, the proposed hybridized MLR-CWLS model could be regarded as 'satisfactory' for H_{sig} forecasting at these coastal stations in Australia. It must be noted that based on E_{NS}, models can be identified as 'unsatisfactory' ($E_{NS} < 0.800$); 'fairly good' ($0.800 \leq E_{NS} \leq 0.900$), or 'very satisfactory' for E_{NS} values greater than 0.900 [65].

Furthermore, Fig. 4 demonstrates a scatterplot with the coefficient of determination, r^2 between forecasted, and observed H_{sig}. The hybridized MLR-CWLS model evidently generates better forecasts than the comparative models by achieving higher r^2 values as follows: (MLR-CWLS≈0.944, MLR≈0.944, MARS≈0.942 and M5tree≈0.870) for Gladstone. Likewise, for other sites Townsville, and Tweed Head, the hybridized MLR-CWLS model affirms better performance, followed by MLR, MARS, and M5tree models (Fig. 4).

In accordance with Tables 3 and 4, Fig. 5 showing the empirical cumulative distribution function (ECDF) also confirms that the hybridized MLR-CWLS model at all three sites registered similar forecast of H_{sig} as compared to observed/actual, revealing its better and reliable forecasting capability. The baseline MLR is seen to generate reasonably similar forecast, followed by M5tree and MARS models (Fig. 5), yet the optimized MLR-CWLS outperformed the comparative models. In addition, the box-plots made a clear distinction of performances since the distribution of forecasted and observed H_{sig} generated from benchmark comparison models (*i.e.*, MLR, MARS, and M5tree) for all three stations were fairly scattered with a large number of outlier points. Hence, the box-plots (Fig. 6) together with the ECDF plots (Fig. 5) for all stations further ascertain the better accuracy of the hybridized MLR-CWLS model in forecasting half-hourly near real-time H_{sig} compared to the competing models.

Moreover, for a more detailed appraisal of the prescribed model performances, the magnitudes of *R* (showing a comparison of observed and forecasted H_{sig}) are plotted in the form of a Taylor diagram [66]. The Taylor diagram (Fig. 7) portrays a more tangible and convincing statistical relationship between the forecasted and observed H_{sig} depending on *R* with respect to standard deviations. It is seen that the MARS and M5tree, except for the hybridized MLR-CWLS and MLR models, are not appropriate as the *R* to standard deviation points were highly parted from the ideal observed point. The hybridized MLR-CWLS and the MLR models appear to lie close to the ideal observed point which

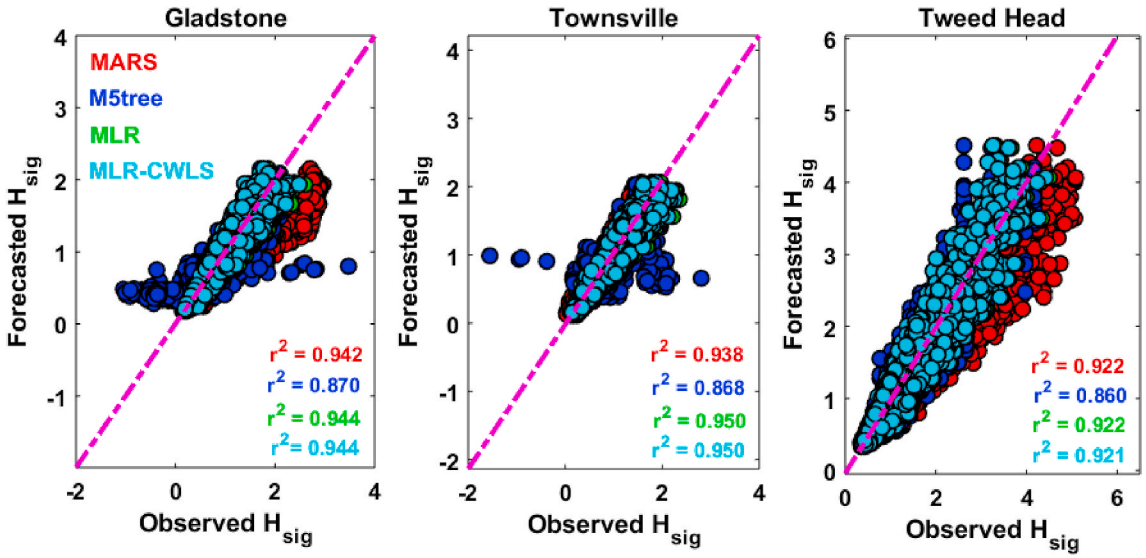

Fig. 4. Scatterplots of forecasted *vs*. observed H_{sig} for the 3 tested sites in Australia's wave energy region. A least-squares regression line and coefficient of determination (r^2) are also inserted in each sub-panel.

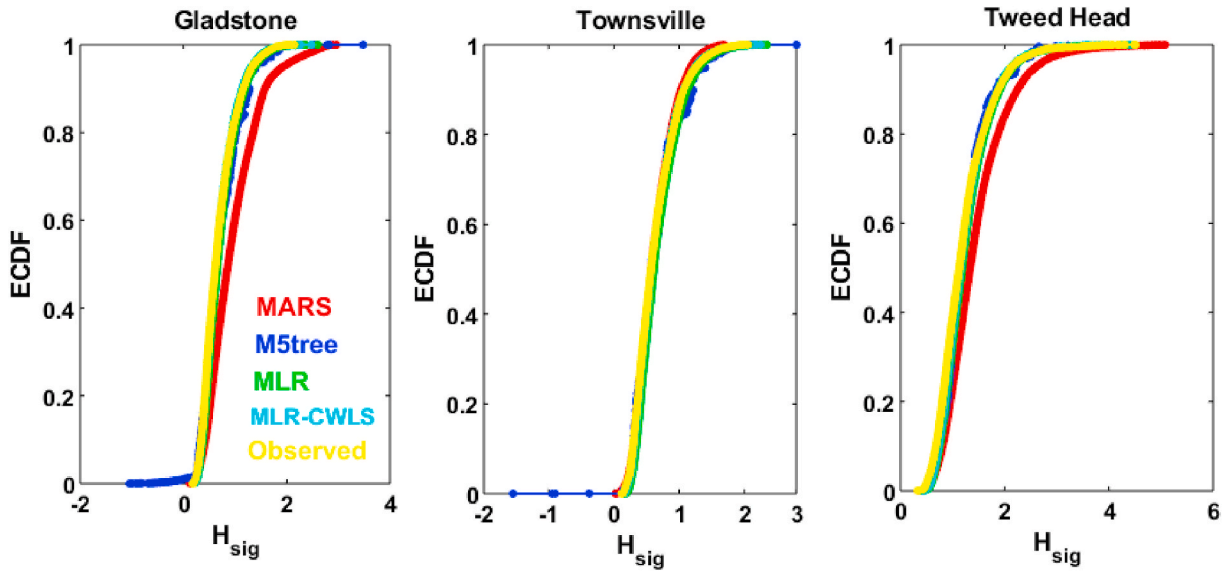

Fig. 5. Empirical cumulative distribution function (ECDF) of the forecasted and observed H_{sig} generated by the hybridized MLR-CWLS *vs.* the standalone models (MARS, M5tree, and MLR) applied at the 3 tested study sites.

confirms the forecasting accuracy was significantly better at all three stations.

From the outcomes of these numerical metrics (Tables 3 and 4), a difference in performances of models at the three distinctive stations is apparent. The key limitation of these metrics is their inability to compare models at geophysically and bathymetrically disparate sites. The relative performance (Table 5) suitably discovered that the hybridized MLR-CWLS model exhibit the lowest values of *RMSPE* and *MAPE* in comparison with benchmark models. More accurately, the *RMSPE* and *MAPE* magnitudes when comparing MLR with the worst performing model MARS, in the combination [MLR-CWLS: MARS] were as follows: Gladstone: [11.11%, 8.06%: 47.51%, 35.48%]; Tweed Head: [11.66%, 8.96%: 24.79%, 19.64%], while M5tree was the worst model Townsville station: [11.28%, 8.45% (MLR-CWLS): 20.50%, 12.11% (M5tree)]. Consequently, the *RMSPE* and *MAPE* exhibited that the hybridized MLR-CWLS model accomplished the best accuracy for Gladstone, followed by Townsville, and Tweed Head.

Fig. 8 shows the forecasted H_{sig} with respect to correlation coefficient R by incorporating each input individually in the hybridized MLR-CWLS model. The vertical bars indicate the contribution of each input predictor in forecasting H_{sig}. It is clear that the input predictor H_{max} is the most responsive variable, contributing about 97% in H_{sig} forecast. This confirms the greater influence of H_{max} on H_{sig}. The peak wave energy period (T_P) is the second predictor which takes part up to 60%, followed by zero up-crossing wave period (T_z), direction (Dir), and sea surface temperature (SST).

To compare directly the hybridized MLR-CWLS *vs*. the MLR, M5tree, and the MARS models, Fig. 9 plots the forecasted and observed (actual) H_{sig} for the tested dataset. For all three coastal stations, the time-series graph of forecasted generated by the hybridized MLR-CWLS and MLR models appears to be more stable with observed H_{sig}. Contrary to that, the forecasts generated by M5tree and MARS are more unstable showing the high fluctuation of forecasted H_{sig} in comparison with observed H_{sig}. Overall, the hybridized MLR-CWLS performs very well for all three study

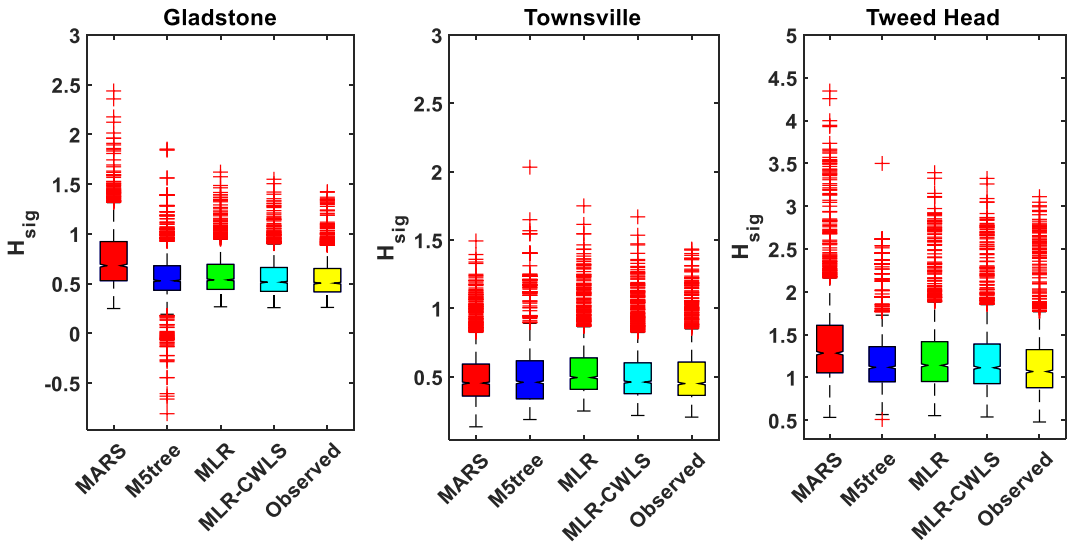

Fig. 6. Boxplots showing the observed (actual) and forecasted H_{sig} generated by the hybridized MLR-CWLS *vs.* the MARS, M5 Model Tree and MLR models in the testing phase.

regions while the baseline MLR (where the CWLS was not used) does not provide reasonably good results.

5. Further discussion

The appropriateness of the hybridized MLR-CWLS against the MLR, M5tree, and MARS models to forecast H_{sig} has been explored in this paper. The forecasting precision of the hybridized MLR-CWLS model was considerably better than the other benchmark counterpart models (Tables 3–5) for all locations illustrating that the hybridized MLR-CWLS model was a well-designed and optimized algorithm to extract pertinent features to simulate H_{sig}. The developed MLR-CWLS model was effectively appraised to generate smaller relative percentage errors in terms of RMSPE, MAPE and larger magnitudes of R, E_{WI}, E_{NS} and E_{LM} (Tables 3 and 4). The performance was high, according to the achieved assessment criterion (Eqs. (6)–(13)) used in this paper. The performance of MLR-CWLS has revealed that the CWLS algorithm was also beneficial for optimization purposes to obtain the pertinent features making the model parsimonious.

The applicability of the hybridized MLR-CWLS model in this paper is tested for several coastal waves-related predictors. The outcomes of the hybridized MLR-CWLS model proved the applicability of the CWLS algorithm to minimize the objective function of predictor variables (H_{max}, T_z, T_p, Dirr, and SST) with a fixed covariance diagonal matrix. Further, the inverse of this covariance diagonal matrix consists of relative weights of each input series. The CWLS algorithm approximates the covariance of coefficients using the initial covariance of the input data. Additionally, the CWLS determines the number of parameters and covariance matrix of the MLR model [27]. Optimization is an important drawback of MLR models and is aptly captured in this study via the CWLS optimization algorithm. The CWLS successfully optimized the MLR modelling parameters for optimum performance using the respective features. This applicability of the CWLS algorithm is a key development of this paper that in the due course increased the forecasting capacity of the MLR model. The MLR model is widely used and can be favoured by energy policy, energy researchers, practitioners, and executives because of its easily explainable regression structure. The easiness of interpretation and the overall understanding of the respective coefficients derived from the predictor variables gives a great degree of lucidity to the MLR model.

The advanced machine learning models including classification and regression trees [18]; Neural Networks [67,68] and support vector machines (SVM) [68,69] have been found to work well in Hs forecasts. These machine learning models including the ANN, Deep learning (CNN, LSTM and RNN) and RF are smart in feature extraction and modelling, yet they all use some form of regression within their architectures. Essentially, the neural networks and its variants and the deep learning models (CNN, LSTM and RNN) uses weights to perform the regression against a target variable while each regression tree in the random forest, which is an ensemble of regression trees, utilizes linear regressions in establishing association among predictors and response variables and partitions them into subgroups at each node making regression the building block of modelling. A key drawback is that ANN, Deep learning (CNN, LSTM and RNN) and RF are all black-box type models as their modelling structures are concealed (i.e., not physically visible to demonstrate the interactions between predictors and target variable) and therefore, may not be preferable for use by these practitioners. In addition, the neural nets require an iterative tuning of their neuronal weights, which are fundamentally based on weighted regressions, making the modelling process computationally exhaustive. On the other hand, the MLR-CWLS developed in this study is very fast and computationally not expansive for the very large dataset (147, 000 data points) making it suitable for half-hourly or near-real-time forecasting applications. In contrast, the deep learning models (including CNN, LSTM and RNN) are developed on the basis of multiple layers, hence have large model execution time. Additionally, the regression tree-based modelling approach, RF, has also been found to be computationally laborious and time-consuming as well [23]. These features together with predictive accuracy established that the hybridized MLR-CWLS can deliver more accurate predictions of H_{sig} together with the ability for near-real-time forecasting ability.

In addition, as discussed above, the main advantage of the CWLS algorithm is the ability to tackle regression circumstances where the dataset are of erratic quality. For example, the weights in the CWLS method are inversely proportional to every point of the explanatory variables that estimates the possible optimum parameter. Therefore, owing to the fact that the hybridized MLR-CWLS model outperformed other competing models in predicting forecasting near real-time H_{sig}, the proposed MLR-CWLS model has the potential to be effectively adopted in energy management systems for optimum significant wave height predictions to those necessities which would plummet the downtimes while increasing productivity. Furthermore, concept drifts resulting from non-stationarity processes is an important phenomenon to consider in machine learning and data science forecasting applications.

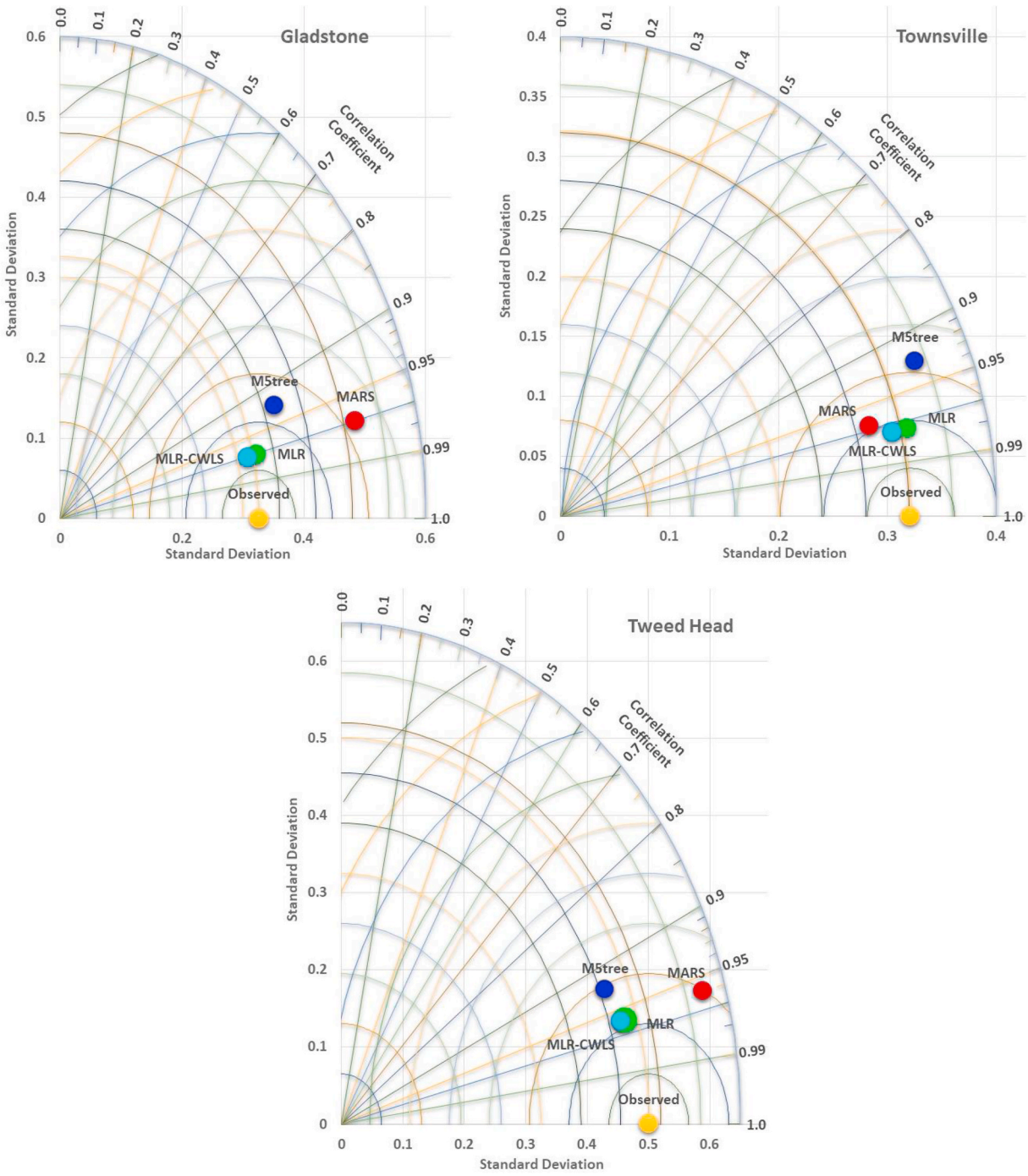

Fig. 7. Taylor diagram depicting the correlation coefficient for the hybridized MLR-CWLS *vs.* the MARS, M5 Model Tree and MLR models at all 3 tested stations in Queensland, Australia.

Table 5
Geographic comparison of the accuracy of the hybridized MLR-CWLS *vs.* comparison models with Root mean squared percentage error (*RMSPE*), and Mean absolute percentage error (*MAPE*). Note that the best model is boldfaced (**blue**).

	Gladstone		Townsville		Tweed Head	
	RMSPE, %	MAPE, %	RMSPE, %	MAPE, %	RMSPE, %	MAPE, %
MARS	47.51	35.48	13.35	9.93	24.79	19.64
M5tree	21.04	14.19	20.50	12.11	15.55	11.80
MLR	12.20	9.28	13.23	11.58	12.16	9.77
MLR-CWLS	**11.11**	**8.06**	**11.28**	**8.45**	**11.66**	**8.96**

Particularly, for near-real-time forecasting of Hs, non-stationarity processes e.g., atmospheric and oceanographic dynamics are bound to deter the model performances over a longer time. Since the MLR-CWLS model use coefficients optimized by the CWLS estimation algorithm these coefficients can be periodically updated as newer information regarding observed data is available in the database [71]. This is an efficient method in comparison to the common periodical refitting approaches, making the MLR-CWLS model versatile for near-real-time forecasting applications. It is noted that the present study has utilized the historical coastal and sea-related datasets to successfully forecast the H_{sig} but this study does carry some limitations that can be addressed in another independent study. While the present study employing the proposed

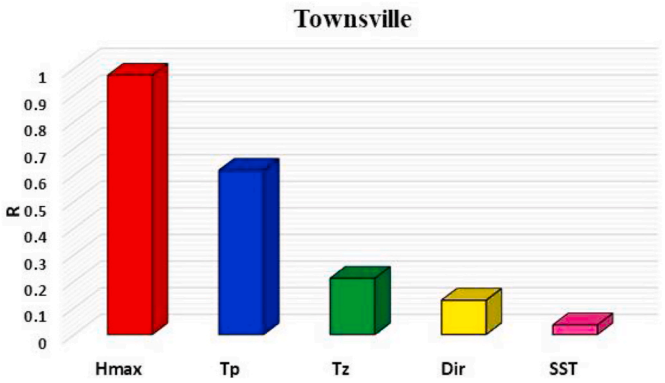

Fig. 8. Forecasting H_{sig} using each input predictor individually via hybridized MLR-CWLS model for Townsville site.

hybrid MLR-CWLS model has successfully modelled half-hourly wave height with high accuracy, the scope of this study can be further enhanced by the inclusion of other exogenous predictors derived from atmospheric sources. This is because wave energy, in fact, is a concentrated form of the wind energy driven by many factors such as solar radiation, sea surface temperature [16,24], as well as cloud cover, sunshine and air temperatures. It should be noted that the buoys at the respective marine locations in this study also recorded the H_{max}, T_z, T_p, Dirr, and SST, which has been utilized as the model predictors. Due to the unavailability of site-specific data of the other parameters, this has not been incorporated in this study and could be considered as a limiting factor. In a further study, one could employ remotely sensed predictor datasets through satellites and atmospheric simulation models (*e.g.* Ref. [72–76]) for the respective oceanic positions to model half-hourly H_{sig} values.

Furthermore, it is also of interest to this discussion that we recognize that as the process-based modelling (e.g. NWP models) can be resource-demanding and cost-prohibitive, the proposed hybridized MLR-CWLS model with many salient inputs may be seen as an alternative viable solution as this method is considerably lucid and easy to comprehend by practitioners. Moreover, medium and long term (hourly, daily, weekly, and monthly) H_{sig} can be forecasted in the follow-up study.

The hybridized MLR-CWLS model could be improved by an ensemble approach to possibly achieve more accurate results. With that new advanced optimization techniques could be utilized including Quantum-Behaved PSO and the Firefly Algorithm to select input predictors, which have been tested to hybridize with the MLR and other models (*e.g.* Refs. [77–85]). Further, empirical wavelet transform [86] and empirical mode composition [87] may be additional approaches. Copula modelling [88] can be applied in terms of statistical approaches where joint behaviour of multivariate data (*e.g.*, H_{sig} and corresponding predictors) can be modelled.

6. Conclusions

A reliable machine-learning model based on multiple linear regression and covariant-weighted least square estimation (i.e., the hybridized MLR-CWLS model) for H_{sig} modelling is designed in this paper. The designed MLR-CWLS model uses the statistically significant lags of H_{max}, T_z, T_p, Dirr, and SST at the 30-min interval to forecast half-hourly nearly real-time significant wave height values. The results of MLR-CWLS

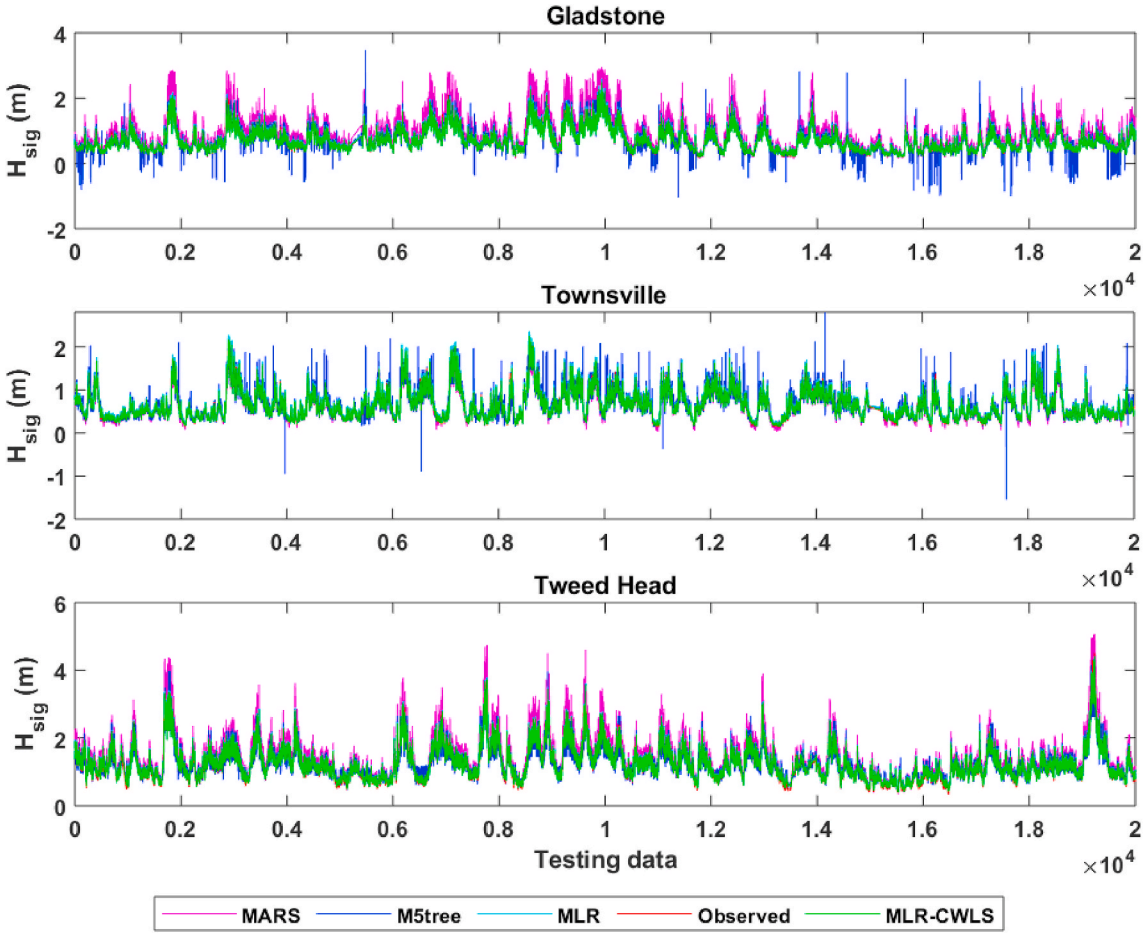

Fig. 9. Time-series plot of observed and forecasted H_{sig} using MARS, M5tree, MLR and the hybridized MLR-CWLS models for all sites.

models were benchmarked with MLR, M5tree, and MARS models using different types of assessment criteria and diagnostic plots/figures. The hybridized MLR-CWLS model is validated against benchmark models at three wave energy-rich zones in Queensland, Australia.

This proposed framework was innovative in terms of introducing the hybridized MLR-CWLS model and using several distinct types of input predictors which significantly improves the forecasting accuracy of H_{sig}. Extending the scope of the hybridized MLR-CWLS, new work can validate the model in other emerging areas of interest such as solar radiation, wind energy, rainfall patterns, drought, hydrology, agriculture crops, and energy demand, to enable the government representatives to manage the climate change scenarios, agriculture crops, and energy-related matters with clearer models.

In conclusion, the proposed MLR-CWLS based significant wave height forecasting model can cordially enable Governments and investors in the renewable and sustainable energy sector for better decision making (*e.g.*, smart grids, efficient and economic integration of wave energy and energy management systems). Moreover, the modelling strategy can be beneficial to other physical applications especially climate change scenarios where advanced artificial intelligence models can be used to make informed prior decisions.

Credit author statement

The contributions of the the respective authors are as follows: Mumtaz Ali: Conceptualization, Methodology, Software, original draft preparation, Reviewing and Editing. Ramendra Prasad: Conceptualization, Visualization, original draft preparation, Reviewing and Editing. Yong Xiang: Writing- Reviewing and Editing. Ravinesh Deo: Writing-Reviewing and Editing and, Supervision.

Declaration of competing interest

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

Acknowledgements

The authors are thankful to the Environment and Science, Queensland Government, Coastal Data System, Queensland for providing significant wave height and other predictor data.

References

[1] UNDP UNDP. Strategic plan, 2018-2021. Special session (executive board of the united nations development programme, the united nations population fund and the united nations office for project services). New York: United Nations; 2017.
[2] Thies PR, Johanning L, Gordelier T. Component reliability testing for wave energy converters: rationale and implementation. European Wave and Tidal Energy Conference; 2013.
[3] Weiss G, Fagan E, Holt P. Australias intended nationally determined contribution to a new climate change agreement. Energetics Pty Ltd; 2016.
[4] Kempener R, Neumann F. IRENA ocean energy technology brief 4. Abu Dhabi: IRENA; 2014.
[5] CSIRO. Ocean renewable energy: 2015-2050. An analysis of ocean energy in Australia. Energy Transformed and Wealth from Oceans Flagships. CSIRO; 2012.
[6] Özger M, Altunkaynak A, Şen Z. Statistical investigation of expected wave energy and its reliability. Energy Convers Manag 2004;45:2173–85.
[7] Prasad R, Joseph L, Deo RC. Modeling and forecasting renewable energy resources for sustainable power generation: basic concepts and predictive model results 2020;68:59–79.
[8] Roulston MS, Ellepola J, Jv Hardenberg, Smith LA. Forecasting wave height probabilities with numerical weather prediction models. Ocean Eng 2005;32:1841–63.
[9] CSIRO and Bureau of Meteorology. reportClimate change in Australia information for Australia's natural resource management regions: technical report. CSIRO and Bureau of Meteorology, Australia2015.
[10] Renewables REN21. Global status report. Paris: REN21 Secretariat; 2019. 2019.
[11] Perth ARENA. Wave energy project. Australian Renewable Energy Agency; 2020.
[12] Reikard G, Robertson B, Bidlot J-R. Wave energy worldwide: simulating wave farms, forecasting, and calculating reserves. Int J Mar Energy 2017;17:156–85.
[13] Uihlein A, Magagna D. Wave and tidal current energy – a review of the current state of research beyond technology. Renew Sustain Energy Rev 2016;58:1070–81.
[14] Gunn K, Stock-Williams C. Quantifying the global wave power resource. Renew Energy 2012;44:296–304.
[15] Langhamer O, Haikonen K, Sundberg J. Wave power—sustainable energy or environmentally costly? A review with special emphasis on linear wave energy converters. Renew Sustain Energy Rev 2010;14:1329–35.
[16] Özger M. Significant wave height forecasting using wavelet fuzzy logic approach. Ocean Eng 2010;37:1443–51.
[17] Falcão AFdO. Wave energy utilization: a review of the technologies. Renew Sustain Energy Rev 2010;14:899–918.
[18] Mahjoobi J, Etemad-Shahidi A. An alternative approach for the prediction of significant wave heights based on classification and regression trees. Appl Ocean Res 2008;30:172–7.
[19] Berbić J, Ocvirk E, Carević D, Lončar G. Application of neural networks and support vector machine for significant wave height prediction. Oceanologia 2017;59:331–49.
[20] Malekmohamadi I, Bazargan-Lari MR, Kerachian R, Nikoo MR, Fallahnia M. Evaluating the efficacy of SVMs, BNs, ANNs and ANFIS in wave height prediction. Ocean Eng 2011;38:487–97.
[21] Kazeminezhad MH, Etemad-Shahidi A, Mousavi SJ. Application of fuzzy inference system in the prediction of wave parameters. Ocean Eng 2005;32:1709–25.
[22] Etemad-Shahidi A, Mahjoobi J. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Eng 2009;36:1175–81.
[23] Ali M, Prasad R. Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition. Renew Sustain Energy Rev 2019;104:281–95.
[24] Reikard G. Forecasting ocean wave energy: tests of time-series models. Ocean Eng 2009;36:348–56.
[25] Cuadra L, Salcedo-Sanz S, Nieto-Borge JC, Alexandre E, Rodríguez G. Computational intelligence in wave energy: comprehensive review and case study. Renew Sustain Energy Rev 2016;58:1223–46.
[26] Deo R, Sahin M. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renew Sustain Energy Rev 2017;72:828–48.
[27] Meng X-L, Rubin DBJB. Maximum likelihood estimation via the ECM algorithm: A general framework, 80; 1993. p. 267–78.
[28] Orsini N, Bellocco R, Greenland SJTsj. Generalized least squares for trend estimation of summarized dose–response data 2006;6:40–57.
[29] Beck N, Katz JNJAPSR. What to do (and not to do) with time-series cross-section data 1995;89:634–47.
[30] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2012.
[31] Draper N, Smith H. Applied regression analysis. New York: John Wiley; 1981. p. 709.
[32] Civelekoglu G, Yigit N, Diamadopoulos E, Kitis M. Prediction of bromate formation using multi-linear regression and artificial neural networks. Ozone Sci Eng 2007;29:353–62.
[33] Şahin M, Kaya Y, Uyar M. Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data. Adv Space Res 2013;51:891–904.
[34] Apaydın A, Kutsal A, Atakan C. Ankara. The statistics in practice. Hacettepe Pub.; 1994.
[35] Ozdamar K. The statistical data analysis with software packages. Eskis ehir: Kaan press; 2004.
[36] Quinlan JR. Learning with continuous classe. 1992. p. 343–8. s. 5th Australian joint conference on artificial intelligence: Singapore.
[37] Mitchell TM. Machine learning, ser. Computer science series singapure. McGraw-Hill Companies, Inc.; 1997.
[38] Rahimikhoob A, Asadi M, Mashal M. A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manag 2013;27:4815–26.
[39] Bhattacharya B, Solomatine DP. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing 2005;63:381–96.
[40] Kisi O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 2015;528:312–20.
[41] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19:1–141.
[42] Queensland. Environ Sci 2018.
[43] Yen BC. Discussion and closure: criteria for evaluation of watershed models. J Irrigat Drain Eng 1995;121:130–2.
[44] ASCE. Artificial neural networks in hydrology. II: hydrologic applications. J Hydrol Eng 2000;5:124–37.
[45] ASCE. Criteria for evaluation of watershed models. J Irrigat Drain Eng 1993;119:429–42.
[46] Dawson CW, Abrahart RJ, See LM. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Model Software 2007;22:1034–52.
[47] Deo RC, Wen X, Qi F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl Energy 2016;168:568–93.
[48] Legates DR, McCabe GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233–41.
[49] Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteorol Soc 1982;63:1309–13.

[50] Willmott CJ. On the validation of models. Phys Geogr 1981;2:184–94.
[51] Willmott CJ. On the evaluation of model performance in physical geography. Spatial statistics and models. Springer; 1984. p. 443–60.
[52] Legates DR, McCabe GJ. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233–41.
[53] Mohammadi K, Shamshirband S, Tong CW, Arif M, Petković D, Ch S. A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers Manag 2015;92:162–71.
[54] Willmott CJ, Robeson SM, Matsuura K. A refined index of model performance. Int J Climatol 2012;32:2088–94.
[55] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 1970;10:282–90.
[56] Nourani V, Ö Kisi, Komasi M. Two hybrid Artificial Intelligence approaches for modeling rainfall–runoff process. J Hydrol 2011;402:41–59.
[57] Hora J, Campos P. A review of performance criteria to validate simulation models. Expet Syst 2015;32:578–95.
[58] Ertekin C, Yaldiz O. Comparison of some existing models for estimating global solar radiation for Antalya (Turkey). Energy Convers Manag 2000;41:311–30.
[59] Li M-F, Tang X-P, Wu W, Liu H-B. General models for estimating daily global solar radiation for different solar radiation zones in mainland China. Energy Convers Manag 2013;70:139–48.
[60] Quilty J, Adamowski J. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. J Hydrol 2018.
[61] Cannas B, Fanni A, See L, Sias G. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. Phys Chem Earth, Parts A/B/C 2006;31:1164–71.
[62] Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci 2012;191:192–213.
[63] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2003.
[64] Jekabsons G. Adaptive regression splines toolbox for matlab/octave. Ver 1130. 2016. p. 1–33.
[65] Shamseldin AY. Application of a neural network technique to rainfall runoff. J Hydrol 1997;199:272–94.
[66] Xu Z, Hou Z, Han Y, Guo W. A diagram for evaluating multiple aspects of model performance in simulating vector fields. Geosci Model Dev (GMD) 2016;9:4365–80.
[67] Gopinath DI, Dwarakish GS. Wave prediction using neural networks at new mangalore Port along west coast of India. Aquatic Procedia 2015;4:143–50.
[68] James SC, Zhang Y, O'Donncha F. A machine learning framework to forecast wave conditions. Coast Eng 2018;137:1–10.
[69] Mahjoobi J, Adeli Mosabbeb E. Prediction of significant wave height using regressive support vector machines. Ocean Eng 2009;36:339–47.
[71] Kelly PF, Drury JC, Weston W, Devine N. Multiple linear regression as a basis for cost oriented decision support. In: Stockton D, Wainwright C, editors. Advances in manufacturing technology IX: proceedings of the 11th national conference on manufacturing research. London: Taylor & Francis Ltd; 1995.
[72] Stathakis D, Savin I, Nègre T. Neuro-fuzzy modeling for crop yield prediction. Int Arch Photogram Rem Sens Spatial Inf Sci 2006;34:p1–4.
[73] Kumar P, Gupta DK, Mishra VN, Prasad R. Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using LISS IV data. Int J Rem Sens 2015;36:1604–17.
[74] Dempewolf J, Adusei B, Becker-Reshef I, Hansen M, Potapov P, Khan A, et al. Wheat yield forecasting for Punjab Province from vegetation index time series and historic crop statistics. Rem Sens 2014;6:9653–75.
[75] Chen C, McNairn H. A neural network integrated approach for rice crop monitoring. Int J Rem Sens 2006;27:1367–93.
[76] Bauer ME. The role of remote sensing in determining the distribution and yield of crops. Adv Agron 1975;27:271–304.
[77] Hoang N-D, Pham A-D, Cao M-T. A novel time series prediction approach based on a hybridization of least squares support vector regression and swarm intelligence. Appl Comput Intell Soft Comput 2014;2014:15.
[78] Kayarvizhy N, Kanmani S, Uthariaraj R. ANN models optimized using swarm intelligence algorithms. WSEAS Trans Comput 2014;13:501–19.
[79] Kumar D, Prasad RK, Mathur S. Optimal design of an in-situ bioremediation system using support vector machine and particle swarm optimization. J Contam Hydrol 2013;151:105–16.
[80] Pal SK, Rai C, Singh AP. Comparative study of firefly algorithm and particle swarm optimization for noisy non-linear optimization problems. Int J Intell Syst Appl 2012;4:50.
[81] Sedki A, Ouazar D. Hybrid particle swarm and neural network approach for streamflow forecasting. Math Model Nat Phenom 2010;5:132–8.
[82] Taormina R, Chau K-W. Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. J Hydrol 2015.
[83] Taormina R, Chau K-W, Sivakumar B. Neural network river forecasting through baseflow separation and binary-coded swarm optimization. J Hydrol 2015;529:1788–97.
[84] Raheli B, Aalami MT, El-Shafie A, Ghorbani MA, Deo RC. Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. Environ Earth Sci 2017;76:503.
[85] Ghorbani MA, Deo Ravinesh C, Zaher Mundher Y, Mahsa HK, Babak M. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in north Iran. Theoretical and Applied Climatology; 2017 [Press].
[86] Gilles J. Empirical wavelet transform. IEEE Trans Signal Process 2013;61:3999–4010.
[87] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the royal society of london A: mathematical, physical and engineering sciences. The Royal Society; 1998. p. 903–95.
[88] Nguyen-Huy T, Deo RC, An-Vo D-A, Mushtaq S, Khan S. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. Agric Water Manag 2017;191:153–72.