Applied Energy 359 (2024) 122624

Contents lists available at ScienceDirect

Applied Energy

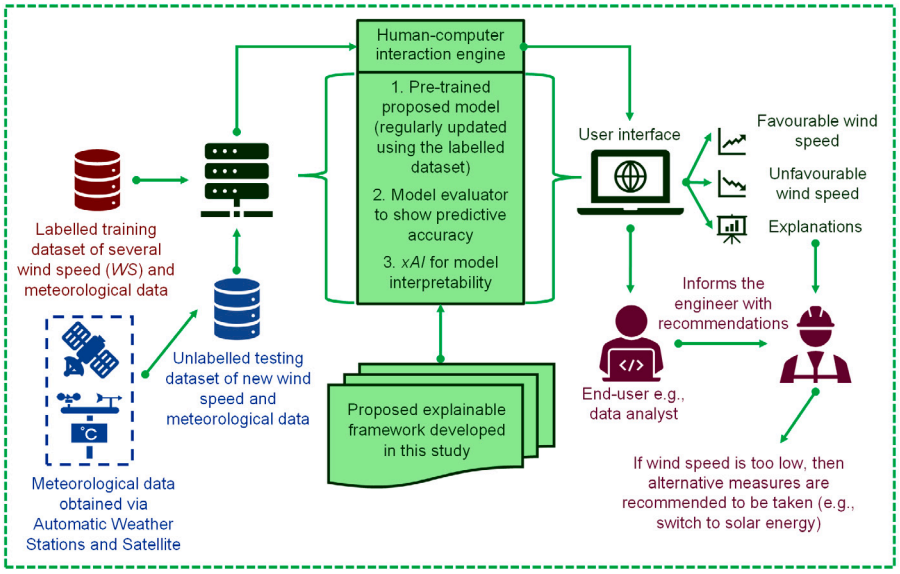
journal homepage: www.elsevier.com/locate/apenergy

Short-term wind speed forecasting using an optimized three-phase convolutional neural network fused with bidirectional long short-term memory network model

Lionel P. Joseph [a], Ravinesh C. Deo [a,b,*], David Casillas-Pérez [c], Ramendra Prasad [d], Nawin Raj [a], Sancho Salcedo-Sanz [e,a]

[a] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD, 4300, Australia
[b] Centre for Applied Climate Sciences, University of Southern Queensland, Springfield, QLD, 4300, Australia
[c] Department of Signal Processing and Communications, Universidad Rey Juan Carlos, Fuenlabrada, 28942, Madrid, Spain
[d] Department of Science, School of Science and Technology, The University of Fiji, Saweni, Lautoka, Fiji
[e] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, 28805, Madrid, Spain

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Wind speed forecasting
Convolutional neural networks
Bidirectional LSTM
Feature selection

ABSTRACT

Wind energy is an environment friendly, low-carbon, and cost-effective renewable energy source. It is, however, difficult to integrate wind energy into a mixed energy grid due to its high volatility and intermittency. For wind energy conversion systems to be reliable and efficient, accurate wind speed (*WS*) forecasting is fundamental. This study cascades a convolutional neural network (CNN) with a bidirectional long short-term memory (BiLSTM) in order to obtain a model for hourly *WS* forecasting by utilizing several meteorological variables

* Corresponding author at: School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD, 4300, Australia.
E-mail addresses: lionel.joseph@usq.edu.au (L.P. Joseph), ravinesh.deo@usq.edu.au (R.C. Deo), david.casillas@urjc.es (D. Casillas-Pérez), ramendrap@unifiji.ac.fj (R. Prasad), nawin.raj@usq.edu.au (N. Raj), sancho.salcedo@uah.es (S. Salcedo-Sanz).

https://doi.org/10.1016/j.apenergy.2024.122624
Received 11 June 2023; Received in revised form 1 December 2023; Accepted 4 January 2024
Available online 24 January 2024

eXplainable Artificial Intelligence

as model inputs to study their effects on predicted *WS*. For input selection, the mutation grey wolf optimizer (TMGWO) is used. For efficient optimization of CBiLSTM hyperparameters, a hybrid Bayesian Optimization and HyperBand (BOHB) algorithm is used. The combined usage of TMGWO, BOHB, and CBiLSTM leads to a three-phase hybrid model (i.e., 3P-CBiLSTM). The performance of 3P-CBiLSTM is benchmarked against the standalone and hybrid BiLSTMs, LSTMs, gradient boosting (GBRs), random forest (RFRs), and decision tree regressors (DTRs). The statistical analysis of forecasted *WS* reveals that the 3P-CBiLSTM is highly effective over the other benchmark forecasting methods. This objective model also registers the highest percentage of forecasted errors ($\approx$ 53.4 – 81.8%) within the smallest error range $\leq |0.25|$ ms^{-1} amongst all tested study sites. Despite the remarkable results achieved, the CBiLSTM model cannot be generally understood, so the eXplainable Artificial Intelligence (*xAI*) technique was used for explaining local and global model outputs, based on Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Both of the *xAI* methods determined that the antecedent *WS* is the most significant predictor of the short-term *WS* forecasting. Therefore, we aver that the proposed model can be employed to help wind farm operators in making quality decisions in maximizing wind power integration into the grid with reduced intermittency.

1. Introduction

Despite the global advocacy on green recovery post pandemic, the fossil fuel-based energy demand has increased. Unless appropriate and immediate action is taken, non-renewable energy sources may finally run out. Furthermore, fossil fuel combustion contributes to climate change, which primarily impacts the low-lying Pacific Small Island Developing States (PSIDS). To encourage the use of renewable energy (RE) whilst mitigating the adverse impacts, the Sustainable Development Goal (SDG) 7 was established in 2015. Wind energy is one of the most promising RE resources that has been adopted in many countries as it is an abundant, eco-friendly, cost-effective, and renewable resource [1]. For instance, despite the COVID-19 pandemic, 2020 and 2021 were the record years for global wind energy industry, where over 186 GW of wind power capacities were added globally [2]. However, the existing growth needs to quadruple by 2030 if the world is to stay on course for a net-zero carbon pathway by 2050 [2]. To achieve this, ambitious technological and policy initiatives are required to rapidly increase the uptake of global wind installations.

The wind turbine (WT) technology used to convert the kinetic energy of wind into electrical energy has significantly improved over the years. For optimal performance, the wind speed (*WS*) largely determines the amount of energy that can be generated by a WT. Hence, every WT model has a designated cut-in and cut-out speed to ensure that two criteria are met: *(a)* oncoming *WS* should be greater than the cut-in speed for maximum energy generation and *(b)* lower than the cut-out speed to avoid any mechanical or electrical damage. In the short term, wind energy generation is inherently intermittent and stochastic, which pose significant challenges for wind farm operators during generation planning and operation. [3]. Utility companies are primarily concerned with maintaining a stable grid, which requires stable energy production from every source connected to the grid. The intermittent and variable nature of wind energy further complicates the issue of generating electricity and forces utilities to either resort to diesel-based generation or refrain from incorporating wind energy. To overcome these and expedite large-scale global wind power adoption; accurate, efficient, and trustworthy *WS* forecasting is needed [4] with models that are easily understood by engineers and operators, and used for smart grid development.

The important prerequisite of forecasting is deciding the forecast horizon and the forecast model class, as these are determined based on the specific context and the purpose of the forecast. The forecast horizon for *WS* forecasting is divided into four categories [1,5]. The ultra-short-term (few seconds to 30-min) forecasting is used for real-time grid operation applications and WT regulation steps. The short-term (30-min to 6-h) range is crucial for economical planning of load dispatch (i.e., preload sharing) and operational security. The medium-term (6-h to 1-day) forecast horizon is used for energy trading and electric power system management. Lastly, the long-term (> 1-day) forecasting range is used for maintenance scheduling, operational management, and feasibility study for optimal site selection. This study is focused on short-term hourly *WS* forecasting. The research outcome will help obtain reliable 1-h ahead estimations of the expected wind power generations from the WTs since wind power is proportional to the cube of *WS* (i.e., $P \propto WS^3$).

Essentially, there are four main categories of forecasting models including physical, statistical, artificial intelligence (AI), and hybrid methods. Physical models forecast *WS* by simulating the physical laws based on meteorological conditions, geographical parameters, and boundary conditions [4]. The widely used physical models include numerical weather prediction (NWP) and weather and research forecasting (WRF). These methods can generate reliable long-term forecasting results [6]. However, physical models have high computational complexity and suffer from information latency. Hence, it is in many cases unsuitable for short-term *WS* forecasting [5,7]. Unlike physical models, statistical models forecast *WS* by using historical *WS* data and are better at dealing with short-term forecasting problems. The commonly used statistical models include autoregressive (AR), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) [8]. These models are characterized by simple structure and high stability. Although statistical models are advantageous in terms of computational complexity, they are more suited for linear time series applications [8]. However, *WS* data are stochastic with complex fluctuations and nonlinear features, which statistical models cannot capture.

As a result, AI-based machine learning (ML) and deep learning (DL) models are more effective in handling the volatile *WS* data with nonlinear characteristics. Common interpretable ML models used for *WS* forecasting include tree-based regressors, such as random forest (RFR), decision trees (DTR), and gradient boosting (GBR) [9]. Tree-based models are explainable and obtain better short-term predictive accuracy over physical and statistical models. However, when used with large datasets, a single tree may grow a large number of nodes, which raises model complexity and leads to overfitting [10]. Other popular non-interpretable ML models include support vector regression (SVR) [11–13] and artificial neural networks (ANN) [14–16]. SVR achieved better *WS* predictions over AR, ARMA, and ARIMA models in [13]. While SVR has a good generalization ability to efficiently reach the global solution, its scalability for larger datasets is limited. ANN can address this, owing to its powerful multivariable mapping capability as demonstrated in [17]. However, ANN often gets trapped in the local optimal solution and loses sight of the internal influence of the data [18]. This restricts further improvements of the forecasting accuracy.

On the other hand, DL models are able to overcome the limitations of ML-based models as it can handle large amounts of data and capture the complex underlying patterns in the data. DL is proposed for feature extraction and temporal dependence modelling-related tasks [10]. Feature extraction makes the model concise by reducing the number of parameters. The popular models applied for feature extraction of *WS* forecasting include stacked autoencoder (SAE), deep belief network

Table 1
Literature review of recent hybrid LSTM, BiLSTM, and CBiLSTM-related forecasting tools used to forecast short-term wind speed (*WS*). Key: ★ represents univariate data, ★★ denotes multivariate data, ▲ indicates the advantage(s) of the proposed model, and ▼ states the drawback(s) of the proposed approach.

Data type	Hybrid model acronym	Model description	Model key features (i.e., pros and cons)	Reference
1-h ★	WTD-FS-CSA-LSTM	WTD is used to decompose data, MI is used for FS, all sub-layers are forecasted by LSTM with CSA used to optimize the HPs.	▲ WTD eliminates the stochastic behaviour of *WS*. Good accuracy improvement over benchmark models. ▼ Only learning rate and batch size HPs optimized. CSA needs initialization. MI is a filter FS method, which ignores interaction with the predictive model.	[19]
1-h ★	ICEEMDAN-LSTM-GWO	ICEEMDAN is used for decomposition, LSTM predicts the individual IMFs, and GWO optimizes the weighted coefficients of each IMF to combine the results.	▲ Signal processing and weight optimization improves forecasting accuracy. ▼ Forecasting each IMF raises computational complexity. HPs not optimized.	[20]
1-h ★	FWA-LSTM	FWA is used to optimize the HPs of LSTM.	▲ FWA optimization gives better results compared to PSO and is superior in terms of convergence speed. ▼ Had a *MAPE* of 30.05% indicating that the accuracy could have been further improved. FWA needs initialization.	[21]
10-min ★	EWT-BiLSTM	EWT decomposes data, BiLSTM individually predicts the low and high-frequency sub-series and the results are aggregated.	▲ EWT is a self-adaptive decomposition method with self-adjusted parameters, which combines the advantages of WT and EMD. ▼ BiLSTM HPs are manually tuned.	[22]
10-min & 1-h ★	ED-HGNDO-BiLSTM	ED is used to decompose the data, BiLSTM is used to forecast the decomposed components, and HGNDO is used to tune BiLSTM HPs.	▲ ED efficiently extracts the features of the nonlinear *WS* data. HGNDO is useful for solving computationally expensive optimization problems. ▼ Complex "black-box" model, but no *xAI* used. HGNDO needs initialization.	[18]
10-min ★★	FS-BO-BiLSTM	Hybrid FS (i.e., stage 1 = PACF & CCF, stage 2 = RReliefF, and stage 3 = Boruta-RF) used for dimensionality reduction, BiLSTM used for forecasting, and BO used to optimize BiLSTM.	▲ Combination of filter and wrapper-based FS allowed robust dimensionality reduction. BO efficiently optimized the HPs. ▼ Although multivariate data is used, the way these features interact with output to generate predictions is not interpretable.	[10]
10-min & 1-h ★★	SSD-MEMD-AMCBiLSTM	SSD is used to denoise of the original multivariate data, MEMD is used to decompose the denoised series, and CNN with attention mechanism (AM) and BiLSTM is used to forecast the individual signals.	▲ SSD successfully extracted the trend. MEMD simultaneously decomposed the multivariate data into IMFs and residuals. AM enhanced the nonlinear spatial feature extraction ability of CNN. ▼ No FS used. HPs selected manually. Not fully explainable.	[23]
10-min ★	TVFEMD-RF-CBiLSTM-ISCA	TVFEMD is used to decompose the data, RF is used to analyse the importance of each decomposed component, CBiLSTM is used to forecast the *WS*, and ISCA is used to optimize the HPs of BiLSTM.	▲ TVFEMD is an improved EMD, which overcomes the problem of model aliasing. RF eliminated the redundant data. Convergence of ISCA > SCA. ▼ Forecasting each IMF increases computational complexity. ISCA is used only to optimize BiLSTM and not CNN. ISCA needs initialization.	[24]
1-h ★	EEMD-GA-CBiLSTM	EEMD is used to decompose the data, CBiLSTM optimized by GA is used to forecast the individual components.	▲ EEMD is an improvement over EMD. GA is a popular optimization tool that helped achieve good results. ▼ EEMD not benchmarked against ICEEMDAN. Optimizing individual components of EEMD increased complexity (e.g., took ≈ 8-h to train the model). Needs parallel computing. GA needs initialization.	[25]

(DBN), and convolutional neural network (CNN) [26]. One-dimensional (1-D) CNN has shown remarkable results for time series-related feature extraction via its convolutional kernels to autonomously mine pertinent information from the data [27]. For instance, CNN used for short-term *WS* forecasting in [28] outperformed the benchmark SVR and kernel ridge regression (KRR) models. Recurrent neural network (RNN) is widely used for temporal dependence forecasting [26], which helps in predicting future events using past information. Unfortunately, RNN suffers from the vanishing and exploding gradient issue, which makes learning of long data sequences difficult [10]. Long short-term memory network (LSTM) overcomes this issue but is limited to processing the information in a single direction; hence, can miss out on pertinent information [29]. An improvement over LSTM is the bidirectional LSTM (BiLSTM), which can process the information in both forward and backward directions [30]. This dual information flow characteristic facilitated efficient learning of long-term dependencies in [10], registering better performance over LSTM and RNN. Moreover, the single DL models discussed for feature extraction (e.g., CNN) and temporal modelling (e.g., BiLSTM) are quite powerful on their own. However, integration of these methods would help maximize the forecasting accuracy. Therefore, CNN and BiLSTM (i.e., CBiLSTM) are combined in this study to take advantage of both methods.

With the continuous development of *WS* forecasting methods, relying solely on standalone models may not suffice. Hybrid models are required, which combine the strengths of different techniques. An overview of recent hybrid models are furnished in Table 1. Literature shows that integration of various methods enhances the predictive accuracy. For instance, the hybridized LSTM in Table 1 had 91.35% [19], 70.38% [20], and 21.81% [21] decrease in mean absolute percentage error (*MAPE*) over standalone LSTM. The BiLSTM-based studies also revealed the superiority of hybridized models, where as much as 41.84% [22], 16.44% [18], and 10.99% [10] improvements were established over the standalone BiLSTM. Additionally, CBiLSTM-studies also favoured the hybrid variants. The hybrid CBiLSTM models in Table 1 had 61.79% [23] and 61.94% [24] decrease in *MAPE* over standalone BiLSTM, and 50.84% [25] improvement in *MAPE* over standalone CBiLSTM. It is evident from the results that the researchers have used a variety of techniques to achieve accurate results. Data decomposition is a commonly used tool in these studies, where researchers applied different methods to extract hidden information from chaotic data.

These included wavelet transform decomposition (WTD) [19], empirical WT (EWT) [22], improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) [20], multivariate EMD (MEMD) [23], time-varying filter based EMD (TVFEMD) [24], evolutionary decomposition (ED) [18], and singular spectrum decomposition (SSD) [23]. However, data decomposition increases computational complexity for short-term forecasting as numerous decomposed series are forecasted individually and later combined [31]. Hence, data decomposition is not tested in this current study. Other hybridization tools employed by researchers in Table 1 include hyperparameter optimization (HPO) and feature selection (FS). Both hybridization tools are crucial for improving model performances. Thus, after identifying the necessary gaps in the literature, robust HPO and FS methods are tested in this study.

Model hyperparameters (HPs) need to be optimized for ideal predictive performance. Manual tuning through trial-and-error is highly inefficient as observed in [20,22,23] (Table 1). Studies have also explored *meta*-heuristic (MH) optimization techniques like the crow search algorithm (CSA) [19] and fireworks algorithm (FWA) [21] to tune LSTM, where both outperformed particle swarm optimization (PSO). Literature on the application of other MHs is summarized in Table 1 including hybrid generalized normal distribution optimizer (HGNDO), improved sine and cosine algorithm (ISCA), and genetic algorithm (GA). However, these MH algorithms require initialization. For instance, GA requires the initialization of crossover rate, mutation rate, and population size before HPO [32]. This is time-consuming, and improper parameter adjustment gives poor solutions during HPO. The popular grid search (GS) [33] and random search (RS) [34] do not require initialization. GS exhaustively evaluates all HP combinations, whereas RS randomly selects predefined HP combinations in the search space. Unfortunately, for both GS and RS, each evaluation in their iterations is independent of previous ones [34], which prompts time wastage in exploring suboptimal areas. The state-of-the-art HyperBand (HB) [35] and Bayesian Optimization (BO) [36,37] methods help overcome the drawbacks of GA, GS, and RS (Table B.1). HB makes HPO efficient by allocating more budget to promising HP configurations. It mimics an early-stoppage strategy, where an unpromising learning curve eliminates the poor HP configuration [38]. A limitation of HB is that it assumes that all HP configuration points are independent when it is generally smooth [38]. BO overcomes this as it is based on the smoothness assumption [36]. When sampling the next trial point, BO balances between exploration and exploitation, which allowed effective optimization of BiLSTM for *WS* forecasting in [10]. However, BO uniformly allocates the computational budget, causing efficiency issues [38]. For an efficient and accurate HPO, studies [39,40] recommend combining Bayesian Optimization and HyperBand (BOHB) [39] to integrate their benefits and eliminate their drawbacks. Therefore, a robust BOHB is used in this study to optimize CBiLSTM.

Alongside HPO, relevant features must be considered for optimal model performance. Numerous studies reviewed in Table 1 (e.g., [18, 20,24,25]) relied only on antecedent *WS* data for *WS* prediction (i.e., univariate modelling). However, meteorological variables are important *WS* predictors also. In [10], pressure, temperature, and humidity were highly influential predictors. This is evidence-based as changes in air pressure affect *WS*, temperature difference between two locations causes a pressure gradient, and air moisture affects *WS* by changing the air density. Considering the importance of weather variables, this study uses several ground-level and satellite-based meteorological predictors. To remove extraneous inputs, FS is required, which is grouped into filter, wrapper, and embedded categories [41]. For *WS* forecasting, a few examples include filter-based mutual information (MI) [19], wrapper-based Boruta-RF [10], and embedded-based RF [42]. These methods have their pros and cons. However, wrapper-based methods ensure better predictive accuracy when fused with population-based optimization algorithms (POAs) [43]. This is because POAs have powerful search capabilities. Several POAs have been tested for FS, where a few important ones are summarized in Table B.2. The popularly used PSO [44] is simple and easy to parallelize but often stagnates before finding a globally optimum solution. Sine and cosine algorithm (SCA) [45] has good local search ability but fails to transition smoothly from exploration to exploitation. Salp swarm algorithm (SSA) [46] is less reliant on initial solutions but suffers from poor population diversity. Whale optimization algorithm (WOA) [47] has a strong neighbourhood search ability but needs more iterations to find the global optimum solution. Grey wolf optimizer (GWO) [48] is another advanced POA. It has three search agents guiding the direction towards a near-optimal solution. GWO retains prior solutions obtained over the course of an iterative process. However, GWO has poor exploitation with complex applications. To overcome this, GWO is integrated with a two-phase mutation (TMGWO) [49]. The mutation process enhances the exploitation capability and helps the search agents find the global optimum solution. TMGWO is tested in [49], where it outperformed several optimizers including GWO, WOA, and PSO. Evaluation of improved GWO in [50] also shows improvements over GWO, PSO, SSA, SCA, WOA, and several other POAs. Given its exceptional characteristics over other POAs, TMGWO is used in this study for FS.

Moreover, integration of TMGWO, BOHB, and CBiLSTM results in a three-phase hybrid model (i.e., 3P-CBiLSTM), which offers excellent predictive performance. However, this hybrid architecture is a non-interpretable "black-box" model, which requires eXplainable Artificial Intelligence (*xAI*) to increase model transparency and make the results trustworthy. Explainable models are required in the wind energy sector to avoid potential biases in decision-making. The stochastic nature of wind can have adverse consequences on the reliability and cost of energy supply as well as loss in confidence in AI forecasting systems. Hence, *xAI* can aid in making the model interpretable. Different types of AI interpretability methods are available, which vary depending on the nature of the problem. A taxonomy of AI interpretability methods are presented in [51] (e.g., Fig. 1). Based on this taxonomy, a model-agnostic explainer is required to interpret the results of CBiLSTM model. Model-agnostic explainers can be applied to any "black-box" model, where it is employed after the model training step (i.e., post hoc) without affecting its performance [52]. Popular model-agnostic explainers include Local Interpretable Model-Agnostic Explanations (LIME) [53] and SHapley Additive exPlanations (SHAP) [54,55]. LIME trains a sparse linear interpretable model locally around the prediction of interest to explain it in terms of the features used [56]. LIME is advantageous as it can provide explanations that are tailored to a specific data point (i.e., local explanation) [53]. This is useful when the model is highly nonlinear and difficult to understand or predict. On the other hand, SHAP is based on cooperative game theory and uses Shapley values to assign each input feature a score that represents its contribution to the model's output (i.e., global explanation) [54,55]. SHAP considers all possible predictions of an instance using all possible feature subsets, which makes it computationally complex. However, SHAP explanations are consistent, additive, and locally more accurate over LIME [56]. In this study, LIME and SHAP are used for local and global interpretability of the objective model.

The novel contributions of this paper are as follows:

(i) A hybrid DL model combining the benefits of CNN and BiLSTM has been applied to a short-term hourly *WS* forecasting problem. Most *WS* studies rely only on historical wind data to perform forecasts. To address this, numerous ground-level and satellite-based weather variables are integrated as CBiLSTM inputs.

(ii) The computational complexity of CBiLSTM has been reduced through a robust dimensionality reduction technique based on a grey wolf optimization algorithm integrated with a two-phase mutation (TMGWO). This improved method has enhanced exploitation capability, which achieves optimal convergence to filter out irrelevant inputs.

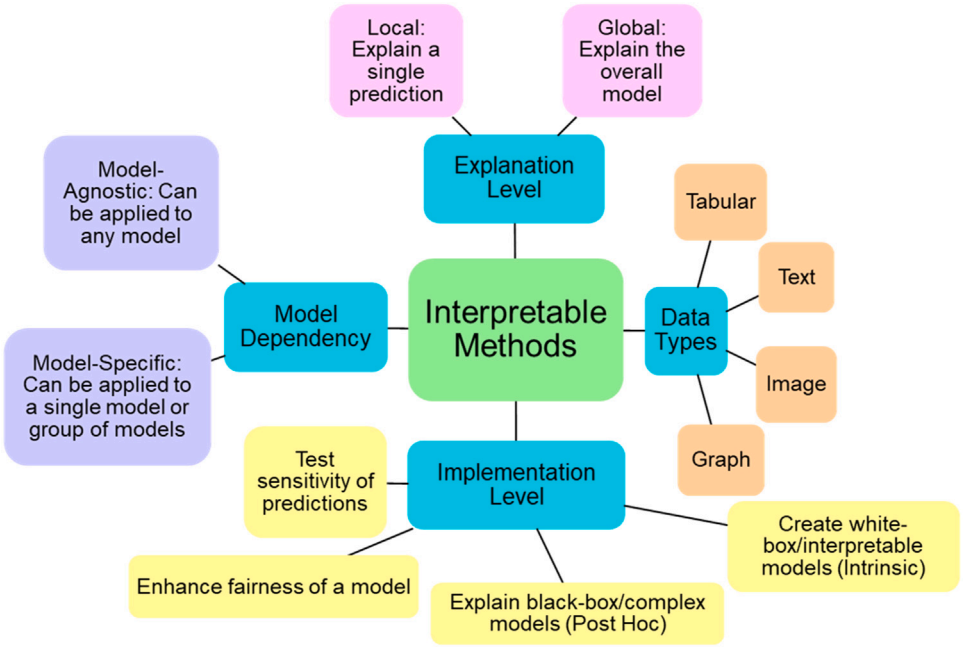


Fig. 1. Taxonomy of eXplainable Artificial Intelligence (*xAI*) interpretability techniques.

(iii) The learning competence of CBiLSTM has been improved by efficient hyperparameter tuning using a hybrid Bayesian Optimization and HyperBand (BOHB) algorithm. BOHB integrates the benefits and eliminates the drawbacks of state-of-the-art BO and HB techniques during hyperparameter optimization. This helps achieve exceptional results when tuning complex "black-box" models.

(iv) Most *WS* studies using complex "black-box" models find it challenging to interpret how specific predictions are made. This lack of transparency poses a significant drawback. To overcome this, an explainable AI *xAI* approach has been explored through model-agnostic *xAI* techniques to interpret the local model prediction results using LIME and global model results using SHAP.

(v) This research introduces a novel explainable three-phase hybrid CBiLSTM (i.e., 3P-CBiLSTM) model that enhances accuracy and reliability in *WS* forecasting. This significant contribution can help grid operators counteract fluctuations in wind power generation to balance the supply and demand of electricity and prevent blackouts.

The remainder of this paper is structured as follows: Section 2 presents the theoretical overview of the methods and algorithms used in this paper. Section 3 describes the methodology proposed. Section 4 presents and discusses the results obtained. Finally, Section 5 summarizes the conclusion and suggests future research options.

2. Theoretical overview

2.1. Hybrid CBiLSTM architecture

CBiLSTM is a combination of 1-D CNN and BiLSTM models. 1-D CNN is beneficial for time series data as it allows local connection and weight sharing, which reduces the number of parameters and improves the learning efficiency [57]. CNN includes convolutional, pooling, and fully connected (FC) layers [58]. Each convolutional layer has several convolutional kernels to extract hidden features and form a feature map, which goes through a nonlinear activation function $f(\cdot)$ to form the output c_i of the i^{th} input as follows:

$$c_i = f\left(w_i * x_i + b_i\right) \tag{1}$$

where w_i, x_i, and b_i are the weight matrix, input, and bias vector, respectively.

The convolutional layer output is reduced by the pooling layer, which mitigates overfitting. The reduced feature map is transferred to the FC layer, which in this study is coupled with BiLSTM for *WS* forecasting.

BiLSTM is an improved LSTM [30]. LSTM addresses the vanishing and exploding gradient problem of RNN [29]. LSTM uses memory cells to remember long-term past information and regulates it through a gate mechanism. It has three gates: input i_t, forget f_t, and output o_t. Information from the current input data x_t and the outputs h_{t-1} of the memory cells at the previous time-step $(t-1)$ are transitioned by the sigmoid function σ. Thus, f_t is computed using:

$$f_t = \sigma\left(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f\right) \tag{2}$$

where W and b are the weight matrices and bias terms of any gate unit, respectively.

Then the model selects new information that needs to be kept in the cell states c_t. To calculate new c_t, two additional prior computations are needed to obtain the value of i_t at time-step t and the new candidate value $\widetilde{c}_t$. These are expressed as:

$$i_t = \sigma\left(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i\right) \tag{3}$$

$$\widetilde{c}_t = \tanh\left(W_{\widetilde{c},x}x_t + W_{\widetilde{c},h}h_{t-1} + b_{\widetilde{c}}\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t, \tag{5}$$

where $\odot$ represents the Hadamard product (i.e., element-wise product). Finally, the o_t selects the output h_t of the memory cells using:

$$o_t = \sigma\left(W_{o,x}x_t + W_{o,h}h_{t-1} + b_o\right) \tag{6}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{7}$$

LSTM processes information in forward direction only. BiLSTM has both forward and backward LSTM layers. The forward LSTM processes the past data information of the input sequence and backward LSTM uses the future information. Forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$ hidden states are combined as follows:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{8}$$

where $\oplus$ denotes summation. Since BiLSTM uses both preceding and subsequent information, it produces better learning efficiency over LSTM [30]. The theoretical details of gradient boosting (GBR), random forest (RFR), and decision tree regressor (DTR) used for comparison are described elsewhere [59–61].

2.2. Bayesian Optimization and HyperBand (BOHB)

BOHB [39] is a powerful HPO algorithm, which locates optimal solution with fewer evaluations over other optimizers. It is suitable for complex optimization problems as it combines the advantages of BO and HB.

BO [36] uses a probabilistic surrogate model for modelling the objective function and an acquisition function to explore new areas in sample space and exploit areas already known for better results. It applies to model the objective function $f(x)$ using a probabilistic model $P(f|D)$ based on the observed data $D = \{(x_0, y_0), \ldots, (x_{i-1}, y_{i-1})\}$. This study uses an Expected Improvement (EI) criterion [40] as an acquisition function, which in the best observed location x is given as:

$$EI_{y(x)} = \int_{-\infty}^{f_{\min}} \max\left\{f_{\min} - f, 0\right\} dP(f|D) \tag{9}$$

where $f_{\min} = \min\left\{f_0, f_1, \ldots, f_n\right\}$. To model the objective function, tree-structured Parzen estimator (TPE) is used as a surrogate model for efficiency [62]. TPE uses a kernel density estimator (KDE) to model the probability densities over the input configurations, given as:

$$l(x) = P(y < y(x)|x, D) \tag{10}$$

$$g(x) = P(y \geq y(x)|x, D) \tag{11}$$

HB [35] uses successive halving (SH) [63] to find the best out of n randomly-sampled configurations. SH evaluates the n configurations with a small budget, removes the worst half and doubles the budget. The process is continued until the best configuration is remaining. However, SH suffers from budget (B) vs. number of configuration (n) issue. HB solves this by balancing highly complex evaluations with many n on the smallest B.

In BOHB, at the beginning of each iteration, HB-based random sampling is applied to get the required model configuration via continuous SH operation. BOHB uses the same sampling strategy as HB, but BO is used to select new configurations based on prior trials. The surrogate model of BO used in BOHB resembles TPE, which is a single multidimensional KDE. Minimal number of configurations $N_{\min}$ is needed to fit a suitable KDE, where Eqs. (12) and (13) represent the model density of the best and worst configurations, respectively.

$$N_{B,g} = \max(N_{\min}, N_B - N_B, l) \tag{12}$$

$$N_{B,l} = \max(N_{\min}, q \cdot N_B) \tag{13}$$

where N_B is the number of samples for B and q is the percentile for N_B.

2.3. Grey Wolf Optimizer algorithm integrated with a Two-phase Mutation (TMGWO)

GWO [48] simulates the hunting behaviour of grey wolves, where the pack leader (alpha – α) guides other members (beta – β, delta – δ, and omega – ω) to look for prey. It uses a population of potential solutions to a problem, and iteratively updates it to find the best one. For instance, α, β, δ, and ω represent the first, second, third-best, and remaining candidate solutions, respectively.

The wolves encircle their prey during hunting, which is mathematically represented as:

$$\vec{D} = \left|\vec{C} \cdot \vec{X_p} - \vec{X}(t)\right| \tag{14}$$

$$\vec{X}(t+1) = \left|\vec{X_p}(t) - \vec{A} \cdot \vec{D}\right| \tag{15}$$

where $\vec{X_p}$ and $\vec{X}$ are the prey and grey wolf positions, respectively at iteration t. $\vec{A}$ and $\vec{C}$ are coefficient vectors, given as:

$$\vec{A} = \left|2\vec{a} \cdot \vec{U}_1 - \vec{a}\right| \tag{16}$$

$$\vec{C} = 2 \cdot \vec{U}_2 \tag{17}$$

where $\vec{U}_1$ and $\vec{U}_2$ are uniform random vectors between 0 to 1, and $\vec{a}$ is a linearly decreasing number from 2 to 0 at each iteration i:

$$\vec{a} = 2 - t\left(\frac{2}{i}\right) \tag{18}$$

The grey wolf position (X, Y) can be updated based on the prey position (X^*, Y^*), where adjusting $\vec{A}$ and $\vec{C}$ give the best position. The wolves update their positions in form of three best solutions (i.e., $\vec{X_1}$, $\vec{X_2}$, and $\vec{X_3}$) as follows:

$$\vec{X}(t+1) = \frac{\left(\vec{X_1} + \vec{X_2} + \vec{X_3}\right)}{3} \tag{19}$$

$$\vec{X_1} = \vec{X_\alpha} - \vec{A_1} \cdot \left(\vec{D_\alpha}\right), \vec{D_\alpha} = \left|\vec{C_1} \cdot \vec{X_\alpha} - \vec{X}\right| \tag{20}$$

$$\vec{X_2} = \vec{X_\beta} - \vec{A_2} \cdot \left(\vec{D_\beta}\right), \vec{D_\beta} = \left|\vec{C_2} \cdot \vec{X_\beta} - \vec{X}\right| \tag{21}$$

$$\vec{X_3} = \vec{X_\delta} - \vec{A_3} \cdot \left(\vec{D_\delta}\right), \vec{D_\delta} = \left|\vec{C_3} \cdot \vec{X_\delta} - \vec{X}\right| \tag{22}$$

The additional steps in TMGWO [49] include initialization, evaluation, transformation function, and two-phase mutation as follows:

- **Initialization:** This phase randomly generates a population of N wolves (i.e., search agents). Each search agent has a dimension d equivalent to the number of features in the dataset. Each agent could be a possible solution and is assigned a binary number of 0 (rejected feature) or 1 (selected feature).
- **Evaluation:** To achieve balance between selecting the least number of features and maximizing the accuracy, the fitness function for evaluating the solutions is given as:

$$fitness = \alpha\gamma_r(D) + \beta\frac{|s|}{|d|} \tag{23}$$

where $\gamma_r(D)$ is the error rate of attribute r relative to decision D calculated using KNN due to its simple implementation [64]. $|s|$ and $|d|$ represent the cardinality of selected feature subset and all features in the dataset, respectively. α and β are weight parameters. Before evaluation, the dataset is spit into training and testing data. Each sample in the test set locates its nearest K neighbours from the train set using Euclidean distance as:

$$Euc_D = \sqrt{\sum_{h=1}^{d}\left(train_{f_h} - test_{f_h}\right)} \tag{24}$$

where $train_{f_h}$ and $test_{f_h}$ represent feature h in a sample of d features from training and testing partitions. Using only training and testing data leads to model overfitting. Hence, a time series split (TSS) 5-fold cross-validation (CV) is used.
- **Transformation function:** The continuous search space of GWO is mapped into a binary one using sigmoid [65] and tanh [66] functions as:

$$X_{S_i} = \frac{1}{1+e^{-X_i}}, X_{binary} = \begin{cases} 0 & \text{if } U < X_{S_i} \\ 1 & \text{if } U \geq X_{S_i} \end{cases} \tag{25}$$

$$X_{v_i} = |\tanh(x)|, X_{binary} = \begin{cases} 0 & \text{if } U < X_{v_i} \\ 1 & \text{if } U \geq X_{v_i} \end{cases} \tag{26}$$

where X_{S_i} and X_{v_i} are continuous feature values for the respective functions, $i = 1, \ldots, d$, and X_{binary} is 0 or 1 based on the uniform random sample (U).

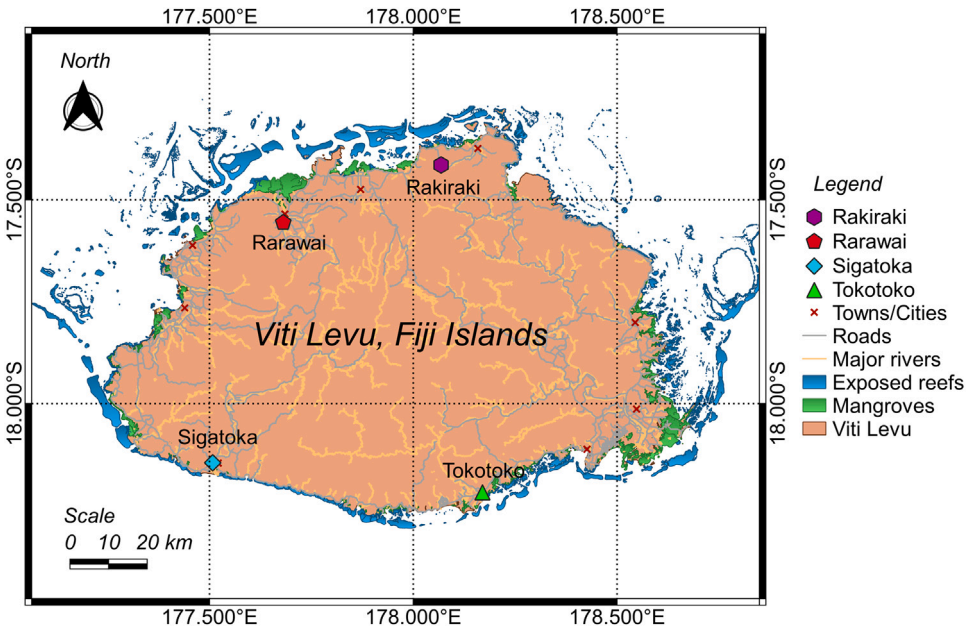


Fig. 2. Location of four selected study sites in Viti Levu, Fiji, where the proposed model has been implemented to forecast hourly wind speed (*WS*).

- **Two-phase mutation:** Mutation phase 1 is used to reduce the number of selected features while maintaining a high accuracy. Phase 2 aims to add more relevant features to further improve accuracy. The two-phase mutation is done using a probability M_p to reduce computational complexity.

2.4. eXplainable Artificial Intelligence (xAI)

The proposed DL-based CBiLSTM is a non-interpretable "black-box" model with a complex architecture. This makes it difficult to establish clear relationship between the inner workings of the model and the output. Hence, *xAI* [67] is used to interpret the underlying model behaviour. LIME [53] and SHAP [54,55] are two popular model-agnostic explainers that provide highly reliable interpretations. Therefore, LIME and SHAP are respectively used for local and global interpretations of the proposed model, which are described as follows:

2.4.1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME explains the objective model by locally approximating it using a surrogate sparse linear interpretable model as follows [53]:

- The dataset to be explained is perturbed n times to generate replicated data.
- The "black-box" model performs prediction on the perturbed data.
- The distance from each perturbed instance to the original observation is converted to a similarity index.
- From the perturbed data, features that best describe the "black-box" model predictions are selected.
- A surrogate interpretable model is trained using the features of the perturbed data.
- The feature weights obtained by the surrogate model is used to explain the "black-box" models' local behaviour.

To approximate the "black-box" model f, LIME minimizes the following objective function:

$$\xi(x) = \arg\min_{g \epsilon G} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g) \tag{27}$$

where g is the interpretable model, x represents original observation, π_x is the proximity measure from all permutations to the original observation, $\mathcal{L}\left(f, g, \pi_x\right)$ is a measure of unfaithfulness of g in approximating f in the locality defined by π, and $\Omega(g)$ is a model complexity measure.

2.4.2. Shapley additive explanations (SHAP)

SHAP is used to identify the global feature influences, dependencies, and interactions on prediction outcomes [54,55]. This tool is derived from cooperative game theory, which assigns each input an importance score for a respective prediction. In game theory, *players* have a set of *strategies* and a *reward* associated with each *strategy*. Shapley values are used to determine the contribution of each player to the outcome of the game. For explaining the model, the *strategies* correspond to the results of the procedures, the *players* correspond to the features, and the *reward* is the quality of the results obtained. Using this concept, in SHAP, Shapley values indicate the contribution of a given feature to the overall prediction.

The SHAP value can then be computed as the weighted average of the marginal contributions over all possible coalitions $|F|!$ Using [54]:

$$\phi_i(f) = \sum_{\{S \subseteq F\} \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \cdot \left[f\left(x_{S \cup \{i\}}\right) - f\left(x_S\right)\right] \tag{28}$$

where $\phi_i(f)$ is the weighted average Shapley value that feature i provides in the context of all coalitions that exclude i, F is the total number of features, S is the subset (i.e., coalition) of F, $f(x_{S \cup \{i\}})$ represents the model prediction using feature i, and $f\left(x_S\right)$ denotes the model prediction without i.

3. Material and methods

3.1. Description of dataset

This study is based on a PSIDS–Fiji, which lies in the South-West Pacific in between latitudes 15.5°S – 19.5°S and longitudes 177°E – 179°W. Fiji has a tropical marine climate with wet season from November to April and a dry season from May to October. The nation lies in the region of South-Easterly trade winds, which powers the existing 10 MW wind farm in Butoni, Sigatoka. Fiji has 332 islands, where Viti Levu and Vanua Levu take up around 87% of the total 18,333 km^2 land area. Viti Levu is larger and more populated with higher energy demand, which can be met by a future wind farm. Thus, four sites in Viti Levu: Rakiraki (RK), Sigatoka (SG), Rarawai (RW), and Tokotoko (TK) (Fig. 2) were selected with their geographical information presented in Table 2. RK site was selected due to its high *WS*, making it a viable site for future wind farm commissioning. SG site

Table 2
General description of four selected wind speed (*WS*) datasets obtained from Fiji Meteorological Services (FMS). (Note: The hourly *WS* data are recorded from 01-01-2017 to 31-12-2019.)

Site no.	Site name	Site acronym	Latitude	Longitude	Elevation (m)	Expected data points	Data missing (%)	Training data split (%)	Testing data split (%)
1	Rakiraki	RK	17.34°S	178.22°E	8.1	26,280	3.34	66.67	33.33
2	Sigatoka	SG	18.14°S	177.50°E	6.7		7.95		
3	Tokotoko	TK	18.22°S	178.17°E	4.9		1.84		
4	Rarawai	RW	17.56°S	177.68°E	9.3		9.83		

is very close to the existing Butoni wind farm, which made it a good benchmark site. The remaining two sites were selected based on data availability (Table 2) as most monitored sites (excluded in this study) had over 20% of missing data.

The predictive models developed in this study were based on the historical data with 1-h temporal resolution. The ground-level *WS* and meteorological data were provided by Fiji Meteorological Services (FMS), which were recorded at a height of 10 m above ground level (AGL).

A total of seventeen satellite-based climate variables (Table 3) were acquired from NASA's online public database managed by the Prediction of Worldwide Energy Resource (POWER) project http://power.larc.nasa.gov/. The solar-related inputs (e.g., *ASWDiR – AUVI*) were based on NASA's Clouds and the Earth's Radiant Energy System Synoptic (CERES SYN1deg Edition 4.1) product [68]. The remaining meteorological inputs (e.g., *T2M – PS*) were based on NASA's Modern Era Retrospective-Analysis for Research and Applications (MERRA-2) assimilation model from Goddard's Global Modelling and Assimilation Office (GMAO) [68]. These reanalysis data are averaged over 0.5° latitude by 0.625° longitude geographical grids. The POWER project team processes the data daily to provide low-latency products. The reliability of NASA's reanalysis-based global solar radiation data was tested in [69], which found a high correlation (r = 0.60 – 0.94) with ground-based observed data. In this study, only sites TK and RW had availability of ground-level *Radn* (Table 3). The correlation analysis showed that *Radn* at both sites had a high correlation with NASA's solar-related inputs (Figure B.1). Thus, satellite data can be useful in the absence of ground-level data.

The statistical mean and standard deviation (SD) of all ground-level and satellite-based attributes are furnished in Table 3, where *WS* is the target variable to forecast and the remaining variables are used as model inputs. A large number of climate variables were used in this study to evaluate their effect on *WS* forecasting. It is important to consider these variables as they are directly linked to the atmospheric conditions that drive wind patterns.

Other than the statistical mean and the SD value, Figure B.2 also shows the Weibull distribution of *WS*. Weibull distribution is characterized by scale (λ; ms^{-1}) and shape (k; dimensionless) parameters [70]. The λ parameter is directly related to the mean *WS*. The k parameter determines the shape of the *WS* distribution. A higher value leads to a more peaked distribution concentrated around the mean. A lower value gives a flatter distribution with a longer tail, indicating that *WS* is variable. The k values for the respective sites include 2.33 (RK), 1.73 (SG), 1.77 (TK), and 1.65 (RW) (Figure B.2). A smaller k range (1.65 – 1.77) show a high probability of stochastic *WS*; hence, there is more need for *WS* forecasting at these sites. Alongside Weibull curves, Figure B.3 shows the wind rose plots for all four sites, which share a common prevailing South-East wind direction.

3.2. Proposed WS forecasting model development

This study designs an explainable three-phase hybrid modelling framework that integrated TMGWO for feature selection (FS) and BOHB for hyperparameter optimization (HPO), and the outcomes were channelled into the CBiLSTM for *WS* forecasting leading to 3P-CBiLSTM model. The proposed 3P-CBiLSTM model schematic is depicted in Fig. 3.

All experiments were implemented in Python under the Google Colaboratory environment (Intel Xeon CPU @2.20 GHz, 13 GB RAM). The ML and DL models were developed using Sklearn [71] and Tensorflow [72] libraries, respectively. Optuna library [73] was used for BOHB. *xAI* libraries LIME [53] and SHAP [54] were employed for model interpretability. The step-by-step procedure to develop the model is as follows:

Step 1: The data were partitioned into two components (Table 2), where 2017–2018 data were allocated for FS, HPO, model training and validation; and 2019 data were assigned for model testing. Data pre-processing was done to check for missing values and confirm data stationarity. The satellite data had no missing values. However, the ground-level data had few missing values (Table 2). These were backfilled using calendar-averaged values [74]. All extreme outliers were replaced with the median values for better model learning.

To test stationarity of the data, all of the attributes summarized in Table 3 were screened using the augmented Dickey–Fuller (ADF) test [75]. For the data to be stationary, the null hypothesis (H_0) of this test needs to be rejected. The rejection of H_0 is stronger when the ADF statistic test value is more negative than the critical value. The ADF statistic test results for the partitioned data are furnished in Table B.3. The critical values at 1% and 5% significance levels were −3.43 and −2.86, respectively. The results confirmed that all attributes were stationary at a 1% significance level, except for the test attribute *AAlb* (RK site), which was stationary at a 5% significance level. Hence, there was no complication regarding non-stationarity in this study.

Step 2: Partial auto-correlation function (PACF) and cross-correlation function (CCF) statistical assessment was performed to extract the significant lagged inputs. PACF was utilized to ascertain the best lag of antecedent *WS*. CCF measures the correlation of target *WS* with the lags of antecedent climate variables; hence, it was employed to determine the best lags of ground-level and satellite-based meteorological variables. Only 20 lags (i.e., past 20-h) were examined since the correlation coefficient (r) for PACF and cross-correlation coefficient (r_{cross}) for CCF decreased with antecedent lags > 20. This is because wind gust and wake effect are short-lived random events, and it is difficult to study the relationship of *WS* with other climate indices at longer antecedent lags [10]. For both PACF and CCF, the attributes with lags exceeding the 95% confidence band were deemed significant. For each input, only the most significant lag (out of 20) with the highest r for PACF and r_{cross} for CCF were selected (Table 4). These significant inputs were then fed to the TMGWO system.

The TMGWO approach was applied on the training dataset to select optimal features from the pool of 24 variables for RK, 23 for SG, and 25 for TK and RW. For robust dimensionality reduction, optimal TMGWO parameters were used. The weight parameters α = 0.99 and β = 0.01, and mutation probability M_p = 0.5 were recommended in [49]. Too large (i.e., 0.9) or too small (i.e., 0.1) M_p values respectively result in high computational complexity and poor precision; hence, are suboptimal [49]. The number of independent runs (*NRuns*) was 10 to avoid randomness [50]. Optimal allocation of number of iterations (*Iter*) and population size (N) is critical. For instance, use of small *Iter* may cause an optimizer to converge to local minima solutions and stagnate, while a large *Iter* has high computational burden. Fortunately, TMGWO does

Table 3
Descriptive mean and standard deviation (SD) of hourly wind speed (*WS*) and other ground-level and satellite-based meteorological variables.

Attribute name	Attribute acronym	Unit	RK		SG		TK		RW	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fiji MET Data (Ground-level)										
Wind speed	*WS*	ms^{-1}	5.86	2.80	1.96	1.17	2.95	1.65	2.15	1.18
Wind direction	*WD*	deg	155.81	54.10	126.00	66.73	137.71	79.56	165.07	72.50
Maximum temperature	*Tmax*	°C	26.22	2.42	24.92	3.26	25.17	2.79	25.77	3.99
Minimum temperature	*Tmin*	°C	25.68	2.36	24.31	3.17	24.60	2.73	25.12	3.91
Relative Humidity	*RH*	%	80.09	8.73	84.19	11.51	83.19	10.11	82.00	15.13
Mean sea-level pressure	*Pmsl*	hPa	1010.98	3.54	NS	NS	1011.19	6.06	1010.65	3.60
Total rainfall	*Rain*	mm	0.04	0.26	0.02	0.22	0.06	0.34	0.03	0.26
Total solar radiation	*Radn*	kWm^{-2}	NA	NA	NA	NA	0.09	0.15	0.11	0.16
NASA Power Data (Satellite-based)										
All Sky Surface Shortwave Downward Irradiance	*ASWDiR*	Whm^{-2}	204.67	279.43	214.00	290.24	202.53	277.52	223.34	300.11
Clear Sky Surface Shortwave Downward Irradiance	*CSWDiR*	Whm^{-2}	282.37	359.54	281.09	358.01	280.95	357.88	282.59	359.79
All Sky Insolation Clearness Index	*ACI*	dimensionless	0.24	0.26	0.26	0.27	0.24	0.26	0.27	0.28
All Sky Surface Albedo	*AAlb*	dimensionless	0.03	0.03	0.03	0.04	0.03	0.04	0.04	0.04
Solar Zenith Angle	*SZA*	deg	27.57	31.21	27.27	31.04	27.57	31.21	27.17	30.98
All Sky Surface Photosynthetically Active Radiation Total	*APARtot*	Wm^{-2}	97.91	133.97	102.27	138.89	97.30	133.42	105.82	142.60
Clear Sky Surface Photosynthetically Active Radiation Total	*CPARtot*	Wm^{-2}	132.23	168.40	131.49	167.44	131.50	167.49	132.09	168.22
All Sky Surface Ultraviolet A Irradiance	*AUVA*	Wm^{-2}	13.44	18.56	13.95	19.14	13.38	18.49	14.25	19.47
All Sky Surface Ultraviolet B Irradiance	*AUVB*	Wm^{-2}	0.39	0.63	0.40	0.64	0.39	0.62	0.41	0.65
All Sky Surface Ultraviolet Index	*AUVI*	dimensionless	2.06	3.33	2.09	3.38	2.02	3.28	2.16	3.46
Temperature at 2 Meters	*T2M*	°C	25.02	2.14	25.81	2.09	24.61	2.32	26.41	1.84
Dew/Frost Point at 2 Meters	*T2Mdew*	°C	21.43	2.19	21.78	2.19	21.18	2.26	22.15	2.08
Wet Bulb Temperature at 2 Meters	*T2Mwet*	°C	23.22	1.89	23.80	1.91	22.89	2.04	24.28	1.73
Specific Humidity at 2 Meters	*QV2M*	gkg^{-1}	16.37	2.13	16.51	2.18	16.12	2.19	16.85	2.09
Relative Humidity at 2 Meters	*RH2M*	%	81.04	9.64	78.88	8.74	81.76	9.79	77.83	8.24
Precipitation Corrected	*PCNcorr*	$mmhour^{-1}$	0.24	0.58	0.20	0.50	0.27	0.66	0.21	0.53
Surface Pressure	*PS*	kPa	98.74	0.34	100.01	0.35	98.71	0.35	100.18	0.34

not require a very large *Iter* given its excellent exploitation capability, which facilitates faster convergence to a global optimum solution. Hence, the FS process was examined using $Iter \in \{50, 80, 100\}$ [48,76,77], where the global optimum solution was achieved within 80 iterations and stagnation point was reached with $Iter > 80$. Similarly, a smaller N reduces population diversity within the search agents, which traps an optimizer in local minima [46]. A large N has a higher computational cost, but it prevents the search agents from stagnating in the FS process. Consequently, N was evaluated on a range of values {10, 20, 50, 80, 100, 200, 300, 500} [76–78] to obtain the best solution with the lowest fitness value (*FV*) (i.e., *RMSE*). For *FV* computation, a wrapper-based KNN regressor with K = 5 [49] and a time series split (TSS) 5-fold cross-validation (*CV*) was used to prevent overfitting. The *FV* convergence plots comparing TMGWO and GWO with different N is shown in Fig. 4 for RK. It is observed that the *FV* decreases as the N increases for both TMGWO and GWO. The comparison of both optimizers reveals that the improved variant has better convergence to global optima at all N. Hence, the best solution was obtained with the smallest *FV* of 0.5321 ms^{-1} at $N = 500$ and $Iter = 49$. Nonetheless, the time required for FS with TMGWO increased with larger N compared to GWO. This can be resolved in future by using a parallel version of TMGWO [49]. The selected features are furnished in Table 5.

Step 3: The new training dataset with optimal predictors were fed to the CBiLSTM model for HPO using BOHB. Similar to FS, HPO was done using a TSS 5-fold *CV*. This *CV* strategy involved the sequential division of data into multiple subsets (Figure B.4). In the first split, one subset was used for training and one was reserved for validation, simulating future unseen data. In the second split, the first two subsets were combined for training and the subsequent subset was allocated for validation. This process was repeated until the fifth split, which had five subsets for training and one for validation. The results were then averaged over the five splits to obtain ideal HPs. This *CV* approach helped preserve the temporal structure of the data whilst preventing overfitting. For more stable results, the maximum number of BOHB iterations was tested between 30, 50, and 80, where 50 iterations gave the best results with a low compute time. Fig. 5 illustrates the selection of HPs at each iteration for the proposed CBiLSTM model for RK site. The horizontal line represents the selected HP, whereas the vertical line shows the optimal iteration. The selected HPs for the proposed and benchmark models are summarized in Table 6. Additionally, Adaptive Moment Estimator (Adam) optimizer was used for all DL-based models to minimize the loss function during model development.

Step 4: Diverse performance metrics were used to compare the proposed model against the benchmark hybrid (three-phase; 3P and two-phase; 2P) and standalone (one-phase; 1P) models. The comparative models included: BiLSTM, LSTM, GBR, RFR, and DTR.

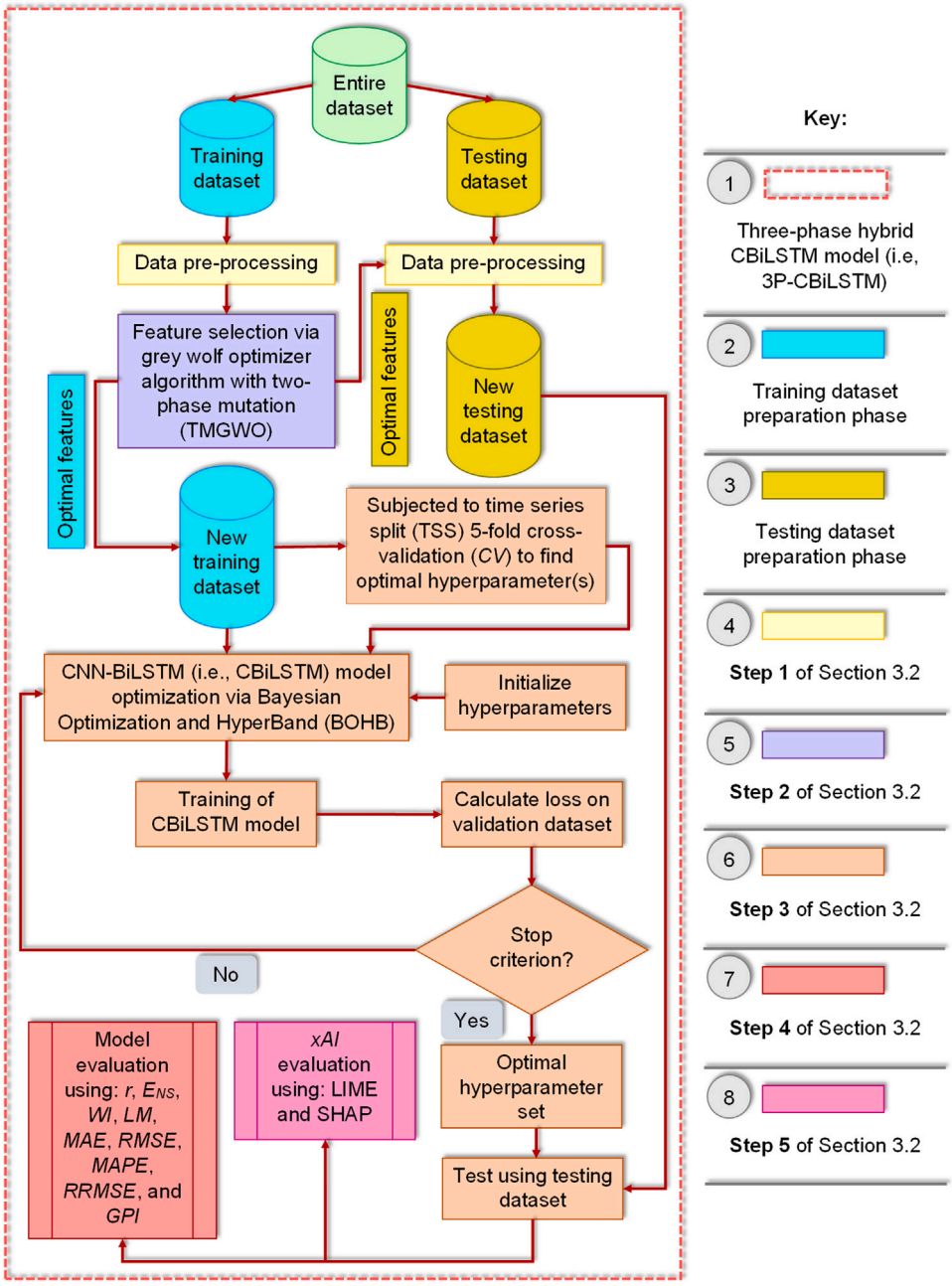


Fig. 3. Overall architecture of the objective three-phase CBiLSTM (i.e., 3P-CBiLSTM) model employed for 1-hour ahead wind speed (*WS*) (ms^{-1}) forecasting. (Note: The key to the right-hand side of the flowchart summarizes the model development steps.)

Step 5: *xAI* was used to interpret the predictions, where the proposed "black-box"-based 3P-CBiLSTM model was explained locally and globally using LIME and SHAP, respectively.

3.3. Model evaluation criteria

To compare the proposed 3P-CBiLSTM model against competing models, two classes (i.e., Class A; ideal value = 1 and Class B; ideal value = 0) of evaluation metrics were used. Several metrics were used to limit the drawbacks and utilize the benefits of different indicators [79]. The Class A (Pearson's correlation coefficient (r), Nash–Sutcliffe Efficiency (E_{NS}), Willmott's Index of agreement (*WI*), Legates and McCabe Index (*LM*)) and Class B (mean absolute error (*MAE*; ms^{-1}), root mean square error (*RMSE*; ms^{-1}), mean absolute percentage error (*MAPE*; %), and relative root mean square error (*RRMSE*; %)) indicators used are computed as follows:

$$r = \frac{\sum_{i=1}^{N}\left(WS_i^O - \overline{WS}^O\right)\left(WS_i^F - \overline{WS}^F\right)}{\sqrt{\sum_{i=1}^{N}\left(WS_i^O - \overline{WS}^O\right)^2}\sqrt{\sum_{i=1}^{N}\left(WS_i^F - \overline{WS}^F\right)^2}} \tag{29}$$

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N}\left(WS_i^O - WS_i^F\right)^2}{\sum_{i=1}^{N}\left(WS_i^O - \overline{WS}^O\right)^2}\right] \tag{30}$$

Table 4
Lag summary of output and input data. The output data is 1-hour ahead wind speed (i.e., *WSt*) and the input data are the most significant lags retrieved via partial auto-correlation function (PACF) and cross-correlation function (CCF) analysis. (Key: Ld*/Lg is Lead*/Lag, NS is not significant, NA is not available, and *WSa* is antecedent wind speed.)

Attributes	RK		SG		TK		RW	
	Ld*/Lg	r^*/r_{cross}	Ld*/Lg	r^*/r_{cross}	Ld*/Lg	r^*/r_{cross}	Ld*/Lg	r^*/r_{cross}
Target								
WSt	$t_{L+1}{}^*$		$t_{L+1}{}^*$		$t_{L+1}{}^*$		$t_{L+1}{}^*$	
Predictors								
WSa	t_{L-1}	0.939	t_{L-1}	0.905	t_{L-1}	0.921	t_{L-1}	0.800
WD	t_{L-1}	−0.426	t_{L-5}	0.057	t_{L-3}	−0.240	t_{L-1}	0.277
Tmax	t_{L-5}	−0.183	t_{L-1}	0.485	t_{L-1}	0.329	t_{L-1}	0.568
Tmin	t_{L-5}	−0.182	t_{L-1}	0.470	t_{L-1}	0.322	t_{L-1}	0.562
RH	t_{L-1}	−0.388	t_{L-1}	−0.697	t_{L-1}	−0.580	t_{L-1}	−0.690
Pmsl	t_{L-4}	0.132	NS	NS	t_{L-5}	0.047	t_{L-2}	−0.017
Rain	t_{L-1}	−0.062	t_{L-3}	0.020	t_{L-5}	0.041	t_{L-1}	0.026
Radn	NA	NA	NA	NA	t_{L-1}	0.369	t_{L-1}	0.551
ASWDiR	t_{L-1}	0.214	t_{L-1}	0.605	t_{L-1}	0.367	t_{L-1}	0.511
CSWDiR	t_{L-1}	0.215	t_{L-1}	0.636	t_{L-1}	0.395	t_{L-1}	0.551
ACI	t_{L-1}	0.227	t_{L-1}	0.586	t_{L-1}	0.340	t_{L-1}	0.494
AAlb	t_{L-1}	0.202	t_{L-1}	0.372	t_{L-1}	0.218	t_{L-1}	0.492
SZA	t_{L-1}	0.129	t_{L-1}	0.304	t_{L-1}	0.151	t_{L-1}	0.279
APARtot	t_{L-1}	0.211	t_{L-1}	0.606	t_{L-1}	0.370	t_{L-1}	0.513
CPARtot	t_{L-1}	0.213	t_{L-1}	0.635	t_{L-1}	0.393	t_{L-1}	0.550
AUVA	t_{L-1}	0.208	t_{L-1}	0.607	t_{L-1}	0.372	t_{L-1}	0.515
AUVB	t_{L-5}	0.182	t_{L-1}	0.564	t_{L-1}	0.355	t_{L-1}	0.481
AUVI	t_{L-5}	0.178	t_{L-1}	0.558	t_{L-1}	0.351	t_{L-1}	0.477
T2M	t_{L-5}	−0.126	t_{L-1}	0.408	t_{L-1}	0.244	t_{L-1}	0.390
T2Mdew	t_{L-1}	−0.269	t_{L-1}	−0.121	t_{L-1}	−0.131	t_{L-1}	−0.131
T2Mwet	t_{L-5}	−0.213	t_{L-1}	0.153	t_{L-5}	−0.091	t_{L-1}	0.127
QV2M	t_{L-1}	−0.271	t_{L-1}	−0.120	t_{L-1}	−0.134	t_{L-1}	−0.125
RH2M	t_{L-1}	−0.358	t_{L-1}	−0.568	t_{L-1}	−0.410	t_{L-1}	−0.521
PCNcorr	t_{L-5}	−0.108	t_{L-1}	−0.043	t_{L-5}	−0.040	t_{L-1}	−0.038
PS	t_{L-5}	0.221	t_{L-2}	−0.037	t_{L-5}	0.114	t_{L-1}	−0.107

Table 5
Selected and rejected predictor variables after application of two-phase mutation grey wolf optimizer (TMGWO) algorithm. (Note: Variables assigned NS and NA were not considered after application of PACF and CCF. For abbreviations, please refer to Table 3.)

Predictors	RK	SG	TK	RW
WSa	✓	✓	✓	✓
WD	✓	✗	✓	✗
Tmax	✓	✓	✓	✓
Tmin	✓	✓	✓	✓
RH	✗	✓	✗	✓
Pmsl	✗	NS	✗	✗
Rain	✗	✗	✓	✗
Radn	NA	NA	✓	✓
ASWDiR	✗	✗	✗	✓
CSWDiR	✓	✓	✓	✓
ACI	✗	✓	✓	✓
AAlb	✗	✓	✓	✓
SZA	✓	✓	✗	✗
APARtot	✗	✗	✗	✓
CPARtot	✓	✓	✓	✓
AUVA	✗	✓	✗	✓
AUVB	✓	✓	✓	✓
AUVI	✓	✓	✓	✗
T2M	✗	✓	✓	✗
T2Mdew	✗	✗	✗	✗
T2Mwet	✓	✗	✗	✗
QV2M	✓	✗	✓	✗
RH2M	✓	✗	✗	✓
PCNcorr	✗	✗	✗	✗
PS	✓	✓	✓	✓
No. of features selected	13	14	15	15

$$WI = 1 - \frac{\sum_{i=1}^{N}\left(WS_i^O - WS_i^F\right)^2}{\sum_{i=1}^{N}\left(\left|WS_i^F - \overline{WS}^O\right| + \left|WS_i^O - \overline{WS}^O\right|\right)^2} \tag{31}$$

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}\left|WS_i^F - WS_i^O\right|}{\sum_{i=1}^{N}\left|WS_i^O - \overline{WS}^O\right|}\right] \tag{32}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(WS_i^F - WS_i^O\right)\right| \tag{33}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(WS_i^F - WS_i^O\right)^2} \tag{34}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{WS_i^F - WS_i^O}{WS_i^O}\right| \times 100 \tag{35}$$

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(WS_i^F - WS_i^O\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}\left(WS_i^O\right)} \times 100 \tag{36}$$

where WS_i^O is the observed *WS*, WS_i^F is the forecasted *WS*, $\overline{WS}^O$ is the average of observed *WS*, $\overline{WS}^F$ is the average of forecasted *WS*, and N is the number of samples.

3.3.1. Global performance indicator (GPI)

The global performance indicator (*GPI*) was used to rank and establish overall model performance [80]. The *GPI* combined the results of all eight metrics used. For i^{th} model, *GPI* is given as:

$$GPI = \sum_{j=1}^{N}\alpha_j\left(\overline{y_j} - y_{ij}\right) \tag{37}$$

where a larger *GPI* is preferred for optimal models, N is the total number of metrics used (i.e., 8), $\alpha_j = -1$ for Class A metrics and +1 for Class B metrics, y_{ij} is the scaled value of metric j for model i, and $\overline{y_j}$ is the median value of scaled values of metric j.

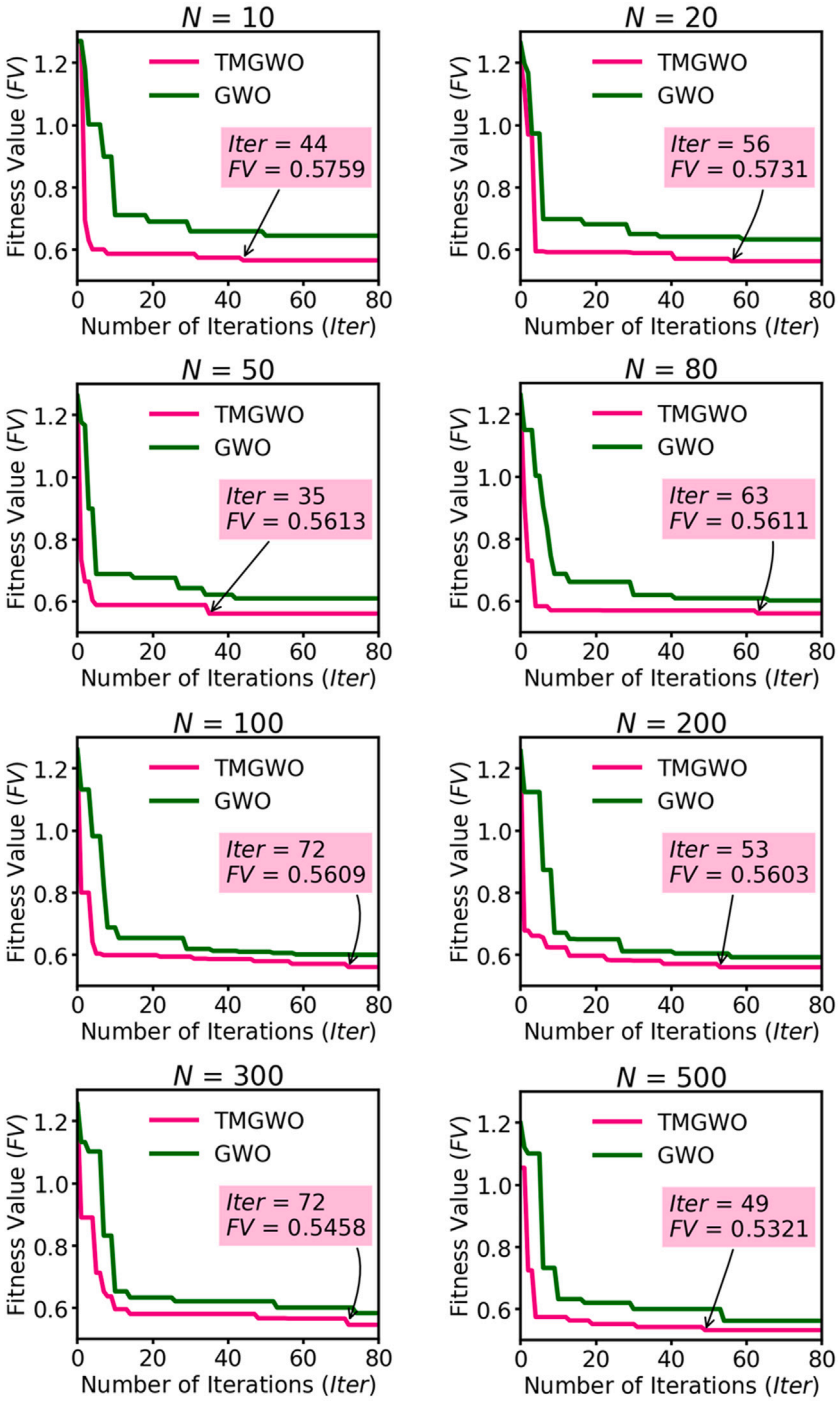


Fig. 4. Convergence curves for the proposed two-phase mutation grey wolf optimizer (TMGWO) versus the standard grey wolf optimizer (GWO) algorithm-based dimensionality reduction of predictor variables for Rakiraki (RK) study site.

3.4. Model explainability

LIME and SHAP tools were used to interpret the results of the proposed "black-box" model. First, LIME was used to explain every instance (i) of the test dataset (i.e., local explanation). For discussion, the explanations of only five instances are presented, which included the 0^{th} (first), 25^{th} (first quarter), 50^{th} (median), 75^{th} (third quarter), and 100^{th} (last) instances of the test dataset. Next, for global explanation, SHAP Kernel Explainer was used to highlight the effect of respective predictors on the entire model performance in form of SHAP summary, feature importance, and feature dependence visual plots.

4. Results and discussion

4.1. Model predictive performance

This section compares the predictive performance of the proposed 3P-CBiLSTM with the benchmark models. The initial assessment was done with *r*, *MAE*, and *RMSE*.

The value of *r* aims to describe the variance and the extent of agreement between the observed WS^O and forecasted WS^F, whereas the error indicators *MAE* and *RMSE* reveal model biases. In *WS* forecasting, large biases are undesirable; hence, *RMSE* helps evaluate model performance by registering high values to large errors. Conversely, *MAE*

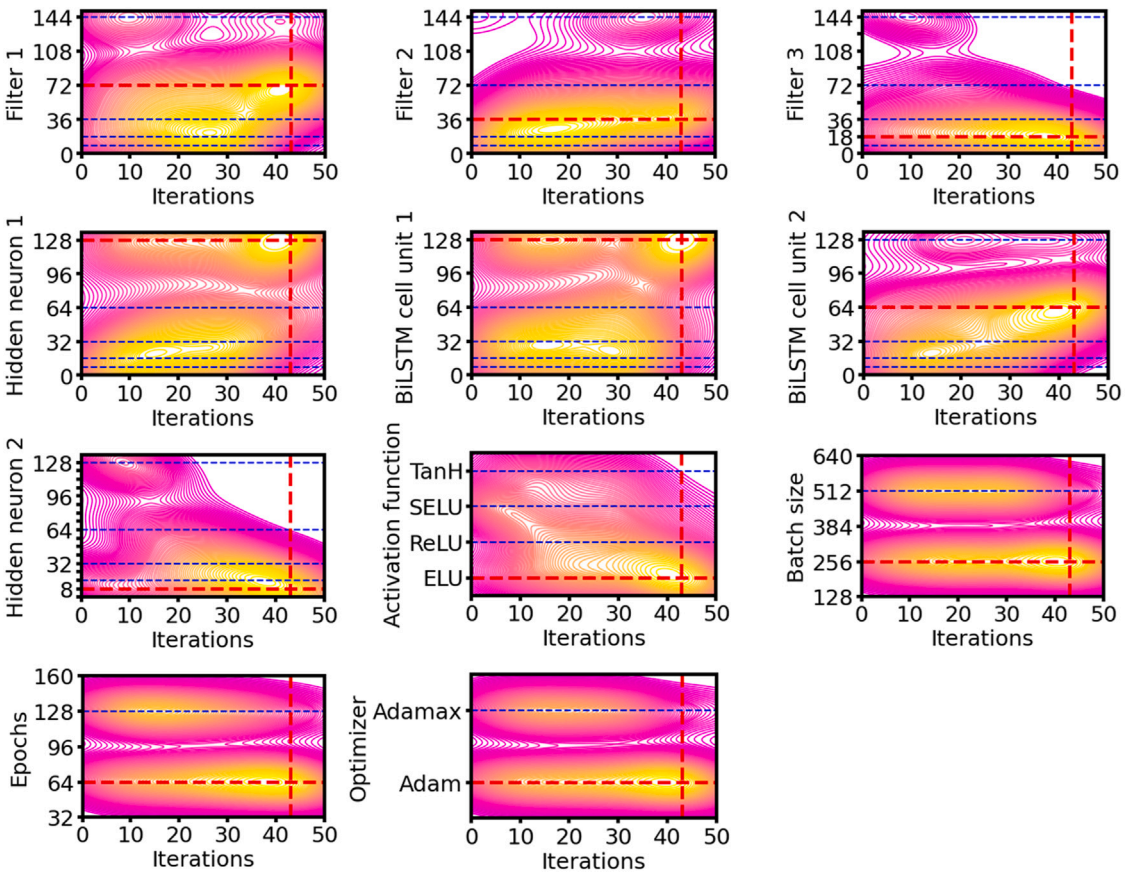


Fig. 5. Two-dimensional contour plots of Bayesian Optimization and HyperBand (BOHB) algorithm-based hyperparameter (HP) selection for the objective 3P-CBiLSTM model devised for Rakiraki (RK) site. The horizontal red dashed lines represent the selected HP(s) and the vertical red dashed lines correspond to the iteration at which the HP was selected based on the lowest validation mean square error (*MSE*) (ms^{-1}).

helps evaluate all deviations between WS^F and WS^O equally since it is not weighted towards higher or lower errors [81].

Also, the expression of *MAE* and *RMSE* in the units of *WS* (i.e., ms^{-1}) makes both measures useful for physical interpretation. Table 7 summarizes model performances based on these measures, where 3P-CBiLSTM obtained the highest *r* (0.963 – 0.990) and the lowest *MAE* (0.149 – 0.308) and *RMSE* (0.197 – 0.420) for all studied stations. Large *r* values (i.e., ≈ 1) show a strong linear association between WS^F and WS^O. The combined usage of *MAE* and *RMSE* is also handy for interpreting the variation in the errors [82]. For instance, there is a minute difference between *MAE* and *RMSE* values for the 3P-CBiLSTM model for all four sites. It indicates that all errors are of similar magnitude. However, the difference between *MAE* (0.438 – 0.915) and *RMSE* (0.576 – 1.240) is higher for the poor-performing 1P-DTR for all four sites. It means there is a higher variance in the individual errors of the predicted sample. The metrics presented in Table 7 favour the superiority of 3P-CBiLSTM over the benchmark models. However, model performance needs to be evaluated using a diverse set of metrics since no evaluation measure is solely perfect [83,84]. Even *r*, *MAE*, and *RMSE* have certain limitations. For instance, *r* standardizes the observed and forecasted means and variances. Also, it can assign higher correlation values to mediocre models [85,86]. *MAE* and *RMSE* are absolute error indicators, which are not suitable for model comparison across geographically disparate sites [87]. This is because a site with higher *WS* will give larger absolute error values compared to a site with lower *WS* regardless of model performance.

The relative error measures *MAPE* and *RRMSE* were used to assess model bias across different sites. *MAPE* and *RRMSE* uses percentage criteria to classify the models as: excellent (error < 10%), good (10% < error < 20%), fair (20% < error < 30%), and poor (error ≥ 30%) [88]. *MAPE* in Fig. 6 shows that the proposed model can be classified as excellent for RK and TK sites, and good for SG and RW. The best-performing 3P-CBiLSTM achieved 67.28%, 60.36%, 63.82%, and 41.07% improvements (i.e., reduction) in *MAPE* over the worst-performing models for RK, SG, TK, and RW sites, respectively. The standalone 1P-DTR was the worst-performing model for RK, SG, and TK sites; and 1P-RFR performed poorly for RW site. Additionally, based on *RRMSE* in Fig. 7, the proposed model is categorized as excellent for RK and good for SG, TK, and RW. Similar to *MAPE*, high percentage improvements in *RRMSE* were accomplished by the 3P-CBiLSTM over the tree-based 1P-DTR and 1P-RFR models. *MAPE* and *RRMSE* measures validate the superior performance of the DL-based CBiLSTM model over the ML-based DTR and RFR models. Further predictive enhancement to the CBiLSTM algorithm was facilitated by TMGWO FS and BOHB HPO.

Additional metrics E_{NS}, *WI*, and *LM* are furnished in Table 8 to study the predictive performance of 3P-CBiLSTM. E_{NS} is a normalized measure that determines the *MSE* in the model corresponding to the variance in the observed data [89]. For physical interpretation: E_{NS} = 1 shows a perfect match between WS^F and WS^O, E_{NS} = 0 reveals that the predictions are as accurate as the mean of WS^O (i.e., poor model), and E_{NS} < 0 indicates that the mean of WS^O is a better predictor of *WS* compared to the model (i.e., worst model). It is a good indicator to measure a model's ability to predict values different from the mean. The E_{NS} results in Table 8 favour 3P-CBiLSTM by registering values closer to unity for all sites (e.g., RK: 0.980, SG: 0.975, TK: 0.970, and RW: 0.926). However, like *MSE* and *RMSE*, E_{NS} tends to overestimate the higher *WS* outliers and neglect the lower values [90]. *WI* overcomes this by considering the ratio of *MSE* instead of the differences [91]. It is beneficial for detecting the additive and proportional differences in the forecasted and observed means and variances. *WI* ranges from 0 –

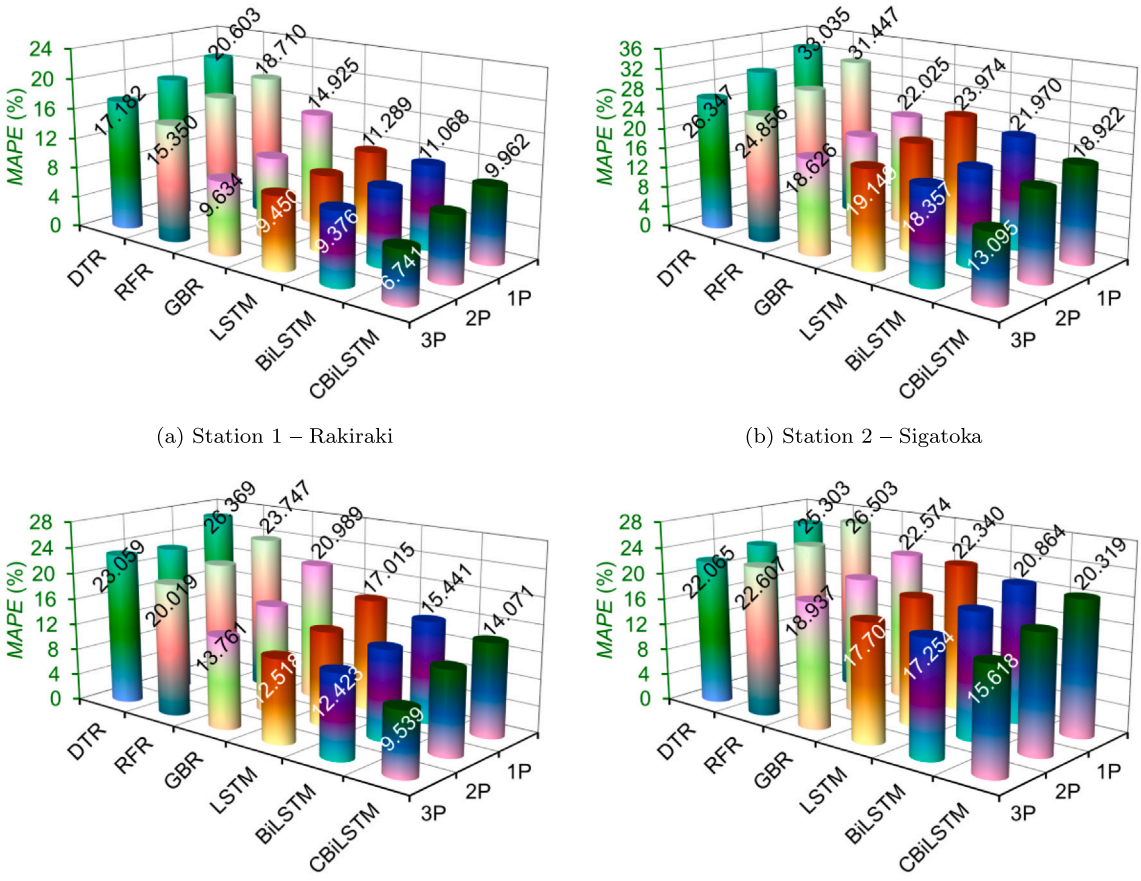
Table 6
Selected hyperparameters (HPs) for the hybrid three-phase (i.e., 3P) models obtained via Bayesian Optimization and HyperBand (BOHB) algorithm. (Note: HPs for hybrid two-phase (i.e., 2P) and standalone (i.e., 1P) models were retrieved using random search (RS) method.)

Predictive models	Model hyperparameters	Hyperparameter search space	RK	SG	TK	RW
3P-CBiLSTM	Filter 1	{9, 18, 36, 72, 144}	72	144	144	144
	Filter 2	{9, 18, 36, 72, 144}	36	72	144	144
	Filter 3	{9, 18, 36, 72, 144}	18	18	144	36
	Hidden neuron 1	{8, 16, 32, 64, 128}	128	16	128	32
	BiLSTM cell unit 1	{8, 16, 32, 64, 128}	128	128	128	128
	BiLSTM cell unit 2	{8, 16, 32, 64, 128}	64	32	8	128
	Hidden neuron 2	{8, 16, 32, 64, 128}	8	8	128	8
	Activation function	{'ELU', 'ReLU', 'SELU', 'TanH'}	'ELU'	'SELU'	'SELU'	'ELU'
	Batch size	{256, 512}	256	256	512	256
	Epochs	{64, 128}	64	64	64	64
	Optimizer	{'Adam', 'Adamax'}	'Adam'	'Adam'	'Adam'	'Adam'
3P-BiLSTM	BiLSTM cell unit 1	{8, 16, 32, 64, 128}	64	128	128	64
	BiLSTM cell unit 2	{8, 16, 32, 64, 128}	32	128	64	32
	Hidden neuron 1	{8, 16, 32, 64, 128}	32	128	128	128
	Hidden neuron 2	{8, 16, 32, 64, 128}	8	32	8	16
	Activation function	{'ELU', 'ReLU', 'SELU', 'TanH'}	'ReLU'	'ELU'	'SELU'	'SELU'
	Batch size	{256, 512}	256	256	512	256
	Epochs	{64, 128}	64	64	64	64
	Optimizer	{'Adam', 'Adamax'}	'Adam'	'Adam'	'Adam'	'Adam'
3P-LSTM	LSTM cell unit 1	{8, 16, 32, 64, 128}	128	128	128	128
	LSTM cell unit 2	{8, 16, 32, 64, 128}	128	128	128	16
	Hidden neuron 1	{8, 16, 32, 64, 128}	16	128	64	128
	Hidden neuron 2	{8, 16, 32, 64, 128}	16	32	16	32
	Activation function	{'ELU', 'ReLU', 'SELU', 'TanH'}	'SELU'	'SELU'	'SELU'	'SELU'
	Batch size	{256, 512}	256	256	256	256
	Epochs	{64, 128}	64	64	64	64
	Optimizer	{'Adam', 'Adamax'}	'Adam'	'Adam'	'Adam'	'Adam'
3P-GBR	Maximum depth of individual regression estimators	{1 – 24}	13	8	14	17
	Maximum features to consider for best split	{'auto', 'sqrt', 'log2', None}	'auto'	'auto'	'auto'	'auto'
	Minimum samples to split internal node	{2 – 30}	7	13	12	9
3P-RFR	Number of trees in the forest	{40, 60, 80, 100, 200}	60	80	80	60
	Maximum features to consider for best split	{'auto', 'sqrt', 'log2', None}	'auto'	'auto'	'auto'	'auto'
	Minimum number of samples required to be at leaf node	{1 – 30}	15	13	13	18
	Minimum samples to split internal node	{2 – 30}	8	7	10	14
3P-DTR	Maximum depth of the tree	{1 – 24}	7	7	7	7
	Minimum samples to split internal node	{2 – 30}	18	8	9	6
	Maximum features to consider for best split	{'auto', 'sqrt', 'log2', None}	'auto'	'auto'	'auto'	'auto'
	Strategy to choose the split at each node	{'best', 'random'}	'best'	'best'	'best'	'best'

Table 7
Statistical evaluation of the proposed 3P-CBiLSTM model against benchmark models in the test phase using correlation coefficient (r), mean absolute error (*MAE*), and root mean squared error (*RMSE*). The optimal results are italicized.

Models	RK			SG			TK			RW		
	r	*MAE* (ms^{-1})	*RMSE* (ms^{-1})	r	*MAE* (ms^{-1})	*RMSE* (ms^{-1})	r	*MAE* (ms^{-1})	*RMSE* (ms^{-1})	r	*MAE* (ms^{-1})	*RMSE* (ms^{-1})
3P-CBiLSTM	*0.990*	*0.308*	*0.420*	*0.988*	*0.149*	*0.197*	*0.986*	*0.218*	*0.302*	*0.963*	*0.267*	*0.361*
3P-BiLSTM	0.982	0.417	0.565	0.975	0.214	0.280	0.975	0.279	0.391	0.951	0.306	0.414
3P-LSTM	0.982	0.427	0.576	0.967	0.240	0.320	0.974	0.280	0.392	0.950	0.310	0.418
3P-GBR	0.979	0.431	0.600	0.975	0.218	0.282	0.971	0.299	0.415	0.944	0.330	0.440
3P-RFR	0.947	0.704	0.957	0.933	0.342	0.450	0.942	0.421	0.585	0.918	0.398	0.530
3P-DTR	0.935	0.781	1.060	0.925	0.361	0.475	0.928	0.471	0.655	0.922	0.388	0.517
2P-CBiLSTM	0.983	0.401	0.543	0.979	0.195	0.257	0.976	0.272	0.379	0.947	0.322	0.431
2P-BiLSTM	0.978	0.461	0.623	0.972	0.230	0.299	0.968	0.313	0.438	0.939	0.344	0.461
2P-LSTM	0.977	0.469	0.631	0.960	0.264	0.352	0.968	0.316	0.441	0.938	0.345	0.462
2P-GBR	0.974	0.479	0.667	0.972	0.234	0.308	0.959	0.353	0.491	0.933	0.359	0.478
2P-RFR	0.934	0.785	1.068	0.918	0.378	0.497	0.921	0.491	0.679	0.906	0.427	0.569
2P-DTR	0.924	0.844	1.144	0.907	0.402	0.529	0.912	0.515	0.713	0.914	0.407	0.543
1P-CBiLSTM	0.978	0.472	0.630	0.972	0.228	0.297	0.969	0.315	0.437	0.933	0.362	0.481
1P-BiLSTM	0.972	0.526	0.707	0.960	0.277	0.356	0.958	0.361	0.501	0.923	0.385	0.515
1P-LSTM	0.971	0.540	0.720	0.951	0.292	0.389	0.956	0.371	0.514	0.922	0.393	0.524
1P-GBR	0.954	0.642	0.893	0.959	0.278	0.358	0.938	0.438	0.608	0.919	0.395	0.527
1P-RFR	0.924	0.847	1.148	0.898	0.421	0.554	0.904	0.540	0.747	0.888	0.465	0.621
1P-DTR	0.911	0.915	1.240	0.890	0.438	0.576	0.886	0.589	0.815	0.897	0.446	0.595

1, where values closer to unity indicate a better agreement of WS^F to WS^O. Table 8 shows that 3P-CBiLSTM recorded the highest *WI* (0.980 – 0.995) for all tested sites. Although *WI* is an improvement over r and E_{NS}, it is still sensitive to peak residuals due to the squaring of residuals in the numerator [92]. Hence, it can assign higher values to mediocre models. *LM* resolves this issue by removing the squaring effect of terms [90]. This way, the outliers are not exaggerated, making *LM* insensitive to extreme *WS* values. Thus, *LM* in Table 8 reasserts that 3P-CBiLSTM displayed the highest accuracy. When compared against the standalone 1P-CBiLSTM, the proposed hybrid model showed 8.15%,



(a) Station 1 – Rakiraki

(b) Station 2 – Sigatoka

(c) Station 3 – Tokotoko

(d) Station 4 – Rarawai

Fig. 6. Mean absolute percentage error (*MAPE* in %) bar plots of all models (i.e., proposed and benchmark) developed for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase.

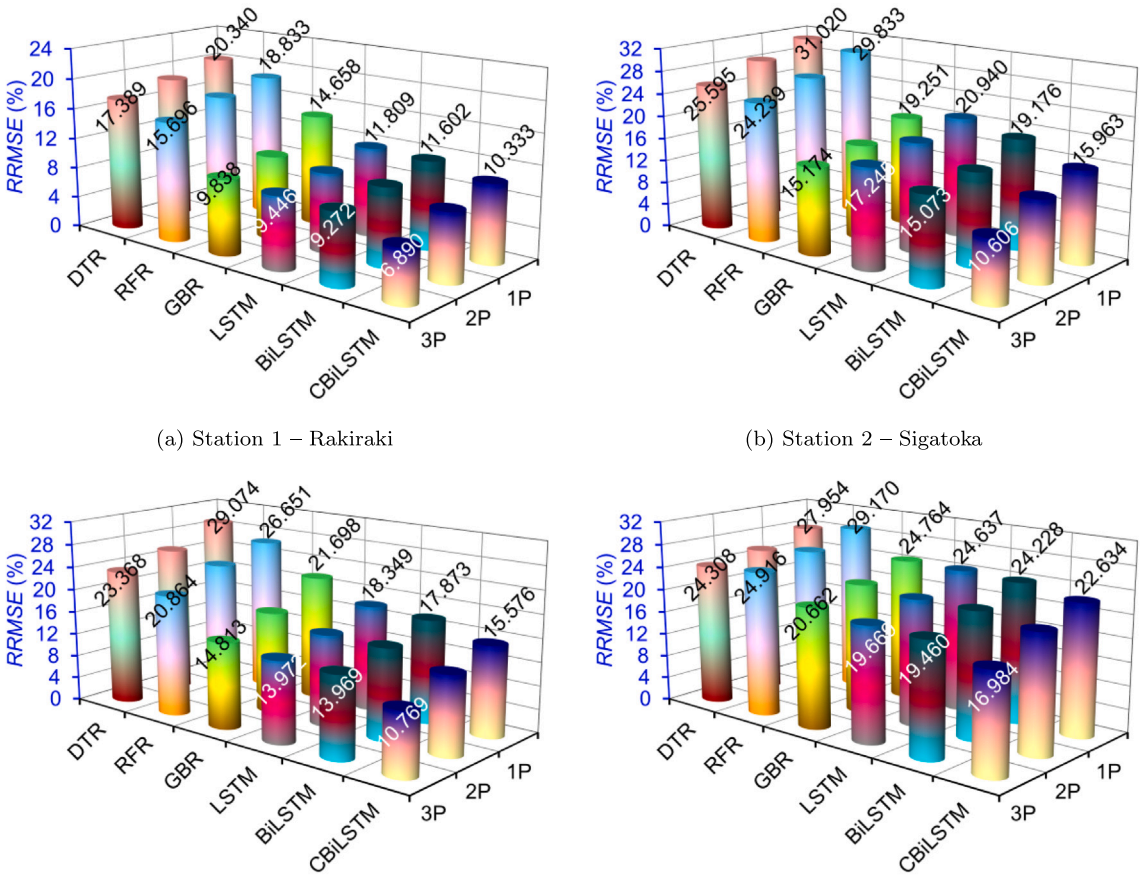
9.73%, 8.25%, and 13.17% increase in *LM* for RK, SG, TK, and RW, respectively. Similar to *MAE* and *RMSE*, the combined use of E_{NS}, *WI*, and *LM* is recommended as they present valuable information regarding the forecasted outliers in the sample [93]. For instance, upper extreme outliers result in low *LM* but high E_{NS} and *WI*, whereas lower extreme outliers result in low *LM*, E_{NS}, and *WI* [89–91]. Based on these criteria, the poor-performing 1P-DTR model mainly predicted incorrect large values, as evidenced by high E_{NS} (0.787 – 0.825) and *WI* (0.940 – 0.954) and low *LM* (0.580 – 0.632) for all sites. Furthermore, *GPI* was used to unify the results of all eight statistical metrics. Fig. 8 presents the *GPI* values and the respective model ranks, where 3P-CBiLSTM was ranked the best for all sites.

Moreover, diagnostic plots were used to further examine the aptness of the proposed model. To visually demonstrate the goodness-of-fit, Fig. 9 illustrates the best and worst three ranked density scatter plots of the WS^O and WS^F. The model performance was appraised using the coefficient of determination (R^2) value of a linear fit model, $WS^F = m(WS^O) + c$. The proposed 3P-CBiLSTM registered the highest R^2 of 0.9802, 0.9758, 0.9716, and 0.9267 for sites RK, SG, TK, and RW, respectively (Fig. 9). This showed that 3P-CBiLSTM had the least variance between WS^O and WS^F compared to Rank 2 and 3 models at the tested sites. For a thorough comparison, the R^2, m, and c values of all models (i.e., proposed and benchmark) are summarized in Table B.4. The competence of the proposed model was finally evaluated using the spread of forecasting errors (FE) to visually examine the biases. Fig. 10 displays the best and worst three ranked histogram plots. The probability distribution of $|FE|$ were yielded in error brackets of 0.25 step-sizes for all sites. The best three ranked models registered smaller spreads in prediction error for all sites over the worst ranked models. A detailed assessment of the probability of $|FE|$ for all models further confirmed the efficacy of 3P-CBiLSTM by acquiring the percentage of least errors in bigger error brackets and the most error of 53.4%, 81.8%, 68.3%, and 59.1% were in the first bin ($0 \leq FE \leq 0.25$) for RK, SG, TK, and RW, respectively. Aligned with other indicators used, the spread of errors revealed that the hybridization of CBiLSTM with TMGWO and BOHB outweigh the benchmark models for hourly *WS* forecasting.

4.2. Model training and testing run time

The training and testing run time (i.e., compute time) of all models are summarized in Table 9. The DL-based models recorded higher compute time than the ML-based models. Although more complex and time-consuming, the DL models offer better predictive accuracy, which is required in the wind energy sector for better decision-making. Amongst the DL models, CBiLSTM has a faster execution time. This is accredited to its CNN layer, which reduces the number of parameters, making the model concise. Furthermore, the comparison between hybrid and standalone counterparts indicates the following order of time efficiency: three-phase > two-phase > one-phase. The standalone models do not employ TMGWO; hence, have 24 (RK), 23 (SG), and 25 (TK and RW) inputs. The use of TMGWO for the two-phase models reduces the inputs to 13 (RK), 14 (SG), and 15 (TK and RW). The two-phase models trained with fewer inputs result in lower compute time.

For instance, the 2P-CBiLSTM model had a 27.83% (RK), 31.15% (SG), 25.18% (TK), and 27.04% (RW) reduction in the computation



(a) Station 1 – Rakiraki

(b) Station 2 – Sigatoka

(c) Station 3 – Tokotoko

(d) Station 4 – Rarawai

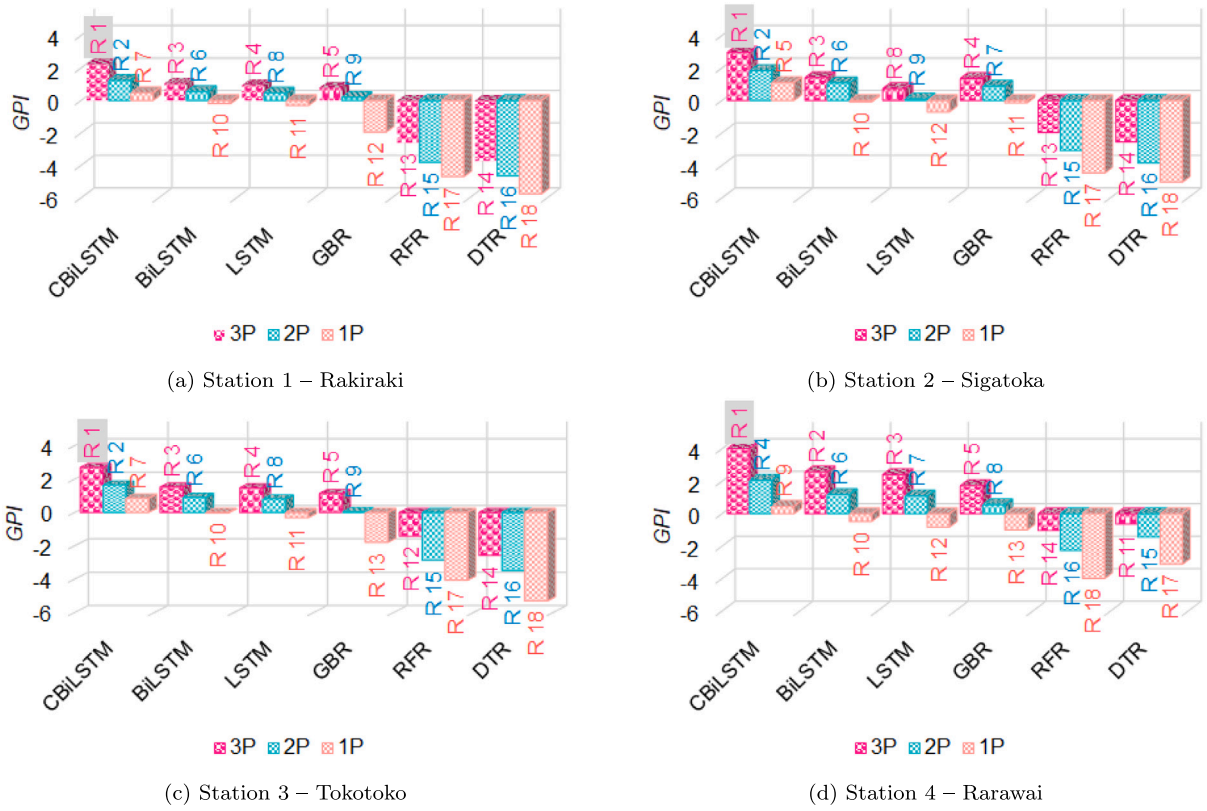
Fig. 7. Relative root mean square error (*RRMSE* in %) bar plots of all models (i.e., proposed and benchmark) developed for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase.

Table 8
Statistical evaluation of the proposed 3P-CBiLSTM model against benchmark models in the test phase using Nash–Sutcliffe coefficient (E_{NS}), Willmott's Index (*WI*), and Legates and McCabe's Index (*LM*). The optimal results are italicized.

Models	RK			SG			TK			RW		
	E_{NS}	*WI*	*LM*	E_{NS}	*WI*	*LM*	E_{NS}	*WI*	*LM*	E_{NS}	*WI*	*LM*
3P-CBiLSTM	*0.980*	*0.995*	*0.876*	*0.975*	*0.994*	*0.857*	*0.970*	*0.992*	*0.853*	*0.926*	*0.980*	*0.756*
3P-BiLSTM	0.964	0.991	0.832	0.950	0.987	0.794	0.949	0.987	0.811	0.903	0.974	0.720
3P-LSTM	0.962	0.990	0.829	0.934	0.983	0.769	0.948	0.986	0.811	0.901	0.973	0.716
3P-GBR	0.959	0.990	0.827	0.949	0.987	0.791	0.943	0.985	0.798	0.891	0.971	0.698
3P-RFR	0.896	0.973	0.717	0.870	0.965	0.672	0.886	0.970	0.716	0.842	0.957	0.636
3P-DTR	0.872	0.966	0.686	0.855	0.961	0.653	0.857	0.963	0.682	0.849	0.959	0.645
2P-CBiLSTM	0.966	0.991	0.839	0.958	0.989	0.813	0.952	0.988	0.816	0.895	0.971	0.705
2P-BiLSTM	0.956	0.989	0.815	0.943	0.985	0.779	0.936	0.984	0.789	0.880	0.967	0.685
2P-LSTM	0.955	0.988	0.812	0.921	0.979	0.747	0.935	0.983	0.786	0.880	0.966	0.684
2P-GBR	0.949	0.987	0.807	0.939	0.984	0.776	0.920	0.979	0.761	0.871	0.965	0.671
2P-RFR	0.870	0.966	0.684	0.841	0.957	0.637	0.847	0.959	0.669	0.818	0.951	0.609
2P-DTR	0.851	0.961	0.661	0.820	0.951	0.614	0.831	0.954	0.652	0.834	0.955	0.627
1P-CBiLSTM	0.955	0.988	0.810	0.944	0.985	0.781	0.937	0.983	0.788	0.869	0.963	0.668
1P-BiLSTM	0.943	0.985	0.789	0.919	0.978	0.734	0.916	0.978	0.756	0.850	0.958	0.647
1P-LSTM	0.941	0.985	0.783	0.903	0.975	0.720	0.912	0.976	0.749	0.845	0.957	0.640
1P-GBR	0.909	0.977	0.742	0.918	0.977	0.733	0.877	0.968	0.704	0.844	0.955	0.638
1P-RFR	0.850	0.960	0.660	0.803	0.946	0.596	0.814	0.950	0.635	0.783	0.941	0.574
1P-DTR	0.825	0.954	0.632	0.787	0.942	0.580	0.779	0.940	0.602	0.801	0.946	0.591

time over 1P-CBiLSTM. Further improvement in time efficiency is evident for the three-phase models (Table 9), which are optimized via BOHB. E.g., BOHB selected 64 epochs for training all DL models (Table 6), while RS selected 128 epochs for all one-phase DL models, which increased the training time. This is evident as the 3P-CBiLSTM model had 46.84% (RK), 47.34% (SG), 43.48% (TK), and 49.65% (RW) reduction in compute time over 1P-CBiLSTM.

Moreover, the time required to optimize the 3P-CBiLSTM model (2.41E+03 – 3.45E+03s) via BOHB was much lower compared to 2P-CBiLSTM (1.14E+04 – 1.54E+04s) and 1P-CBiLSTM (1.76E+04 –



(a) Station 1 – Rakiraki

(b) Station 2 – Sigatoka

(c) Station 3 – Tokotoko

(d) Station 4 – Rarawai

Fig. 8. Overall ranking of the objective 3P-CBiLSTM and benchmark models using global performance indicator (*GPI*) for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase. (Note: Acronym R indicates the model ranks, where R1 and R18 are the best and worst ranked models, respectively.)

Table 9
Training and testing run time (in seconds) of all models (i.e., proposed and benchmark) developed for the four studied stations.

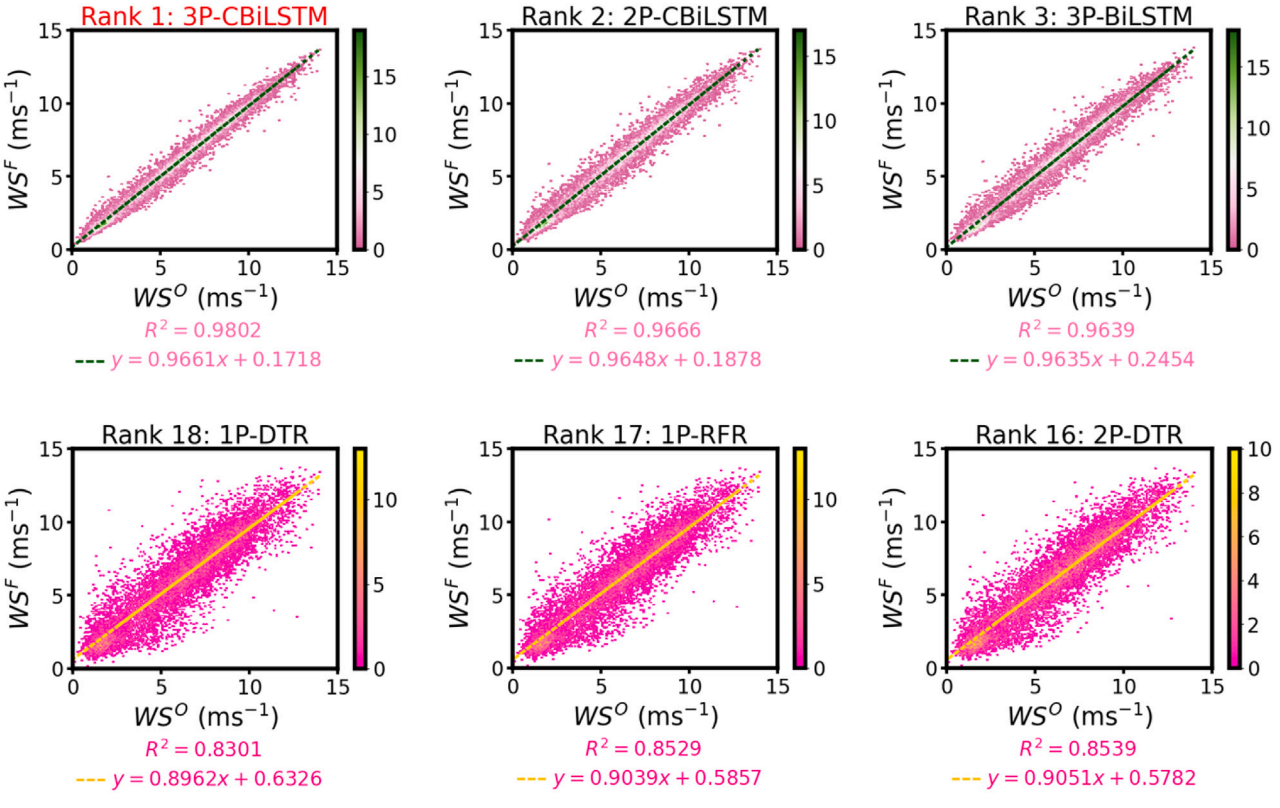
Models	RK		SG		TK		RW	
	Train time (s)	Test time (s)	Train time (s)	Test time (s)	Train time (s)	Test time (s)	Train time (s)	Test time (s)
3P-CBiLSTM	1.48E+02	7.28E−01	1.90E+02	9.22E−01	2.40E+02	1.36E+00	2.08E+02	9.80E−01
3P-BiLSTM	3.35E+02	1.67E+00	6.07E+02	2.54E+00	6.79E+02	2.72E+00	4.77E+02	2.05E+00
3P-LSTM	4.65E+02	2.20E+00	3.25E+02	1.71E+00	4.27E+02	2.13E+00	3.95E+02	1.76E+00
3P-GBR	1.27E+01	5.04E−02	1.83E+01	9.69E−02	3.06E+01	2.06E−01	2.57E+01	1.48E−01
3P-RFR	2.40E+00	2.44E−02	4.06E+00	2.71E−02	4.80E+00	4.14E−02	4.59E+00	3.99E−02
3P-DTR	1.04E−01	1.35E−03	1.14E−01	1.79E−03	1.21E−01	5.88E−03	1.17E−01	3.26E−03
2P-CBiLSTM	2.01E+02	9.99E−01	2.48E+02	1.20E+00	3.18E+02	1.79E+00	3.01E+02	1.40E+00
2P-BiLSTM	4.53E+02	2.28E+00	8.01E+02	3.28E+00	9.17E+02	3.57E+00	6.83E+02	2.91E+00
2P-LSTM	6.19E+02	2.94E+00	4.38E+02	2.19E+00	5.86E+02	2.77E+00	5.62E+02	2.45E+00
2P-GBR	2.61E+01	7.14E−02	4.90E+01	2.17E−01	7.21E+01	2.72E−01	5.02E+01	2.23E−01
2P-RFR	4.84E+00	4.60E−02	6.86E+00	5.78E−02	8.75E+00	7.40E−02	8.07E+00	7.33E−02
2P-DTR	2.92E−01	2.62E−03	3.38E−01	3.43E−03	4.32E−01	7.51E−03	4.18E−01	4.99E−03
1P-CBiLSTM	2.78E+02	1.37E+00	3.61E+02	1.73E+00	4.24E+02	2.39E+00	4.13E+02	1.92E+00
1P-BiLSTM	5.94E+02	3.08E+00	1.12E+03	4.73E+00	1.20E+03	4.77E+00	9.31E+02	3.99E+00
1P-LSTM	8.14E+02	4.00E+00	6.04E+02	3.17E+00	7.63E+02	3.71E+00	7.63E+02	3.35E+00
1P-GBR	6.71E+01	1.25E−01	1.13E+02	2.63E−01	1.55E+02	4.80E−01	1.39E+02	3.32E−01
1P-RFR	9.44E+00	6.51E−02	1.16E+01	9.14E−02	1.30E+01	1.85E−01	1.25E+01	1.15E−01
1P-DTR	4.17E−01	3.15E−03	6.05E−01	4.98E−03	7.62E−01	8.53E−03	6.07E−01	5.50E−03

2.31E+04s) models tuned via RS for all sites. Therefore, the use of TMGWO and BOHB helped improve the computational efficiency of the proposed 3P-CBiLSTM model.

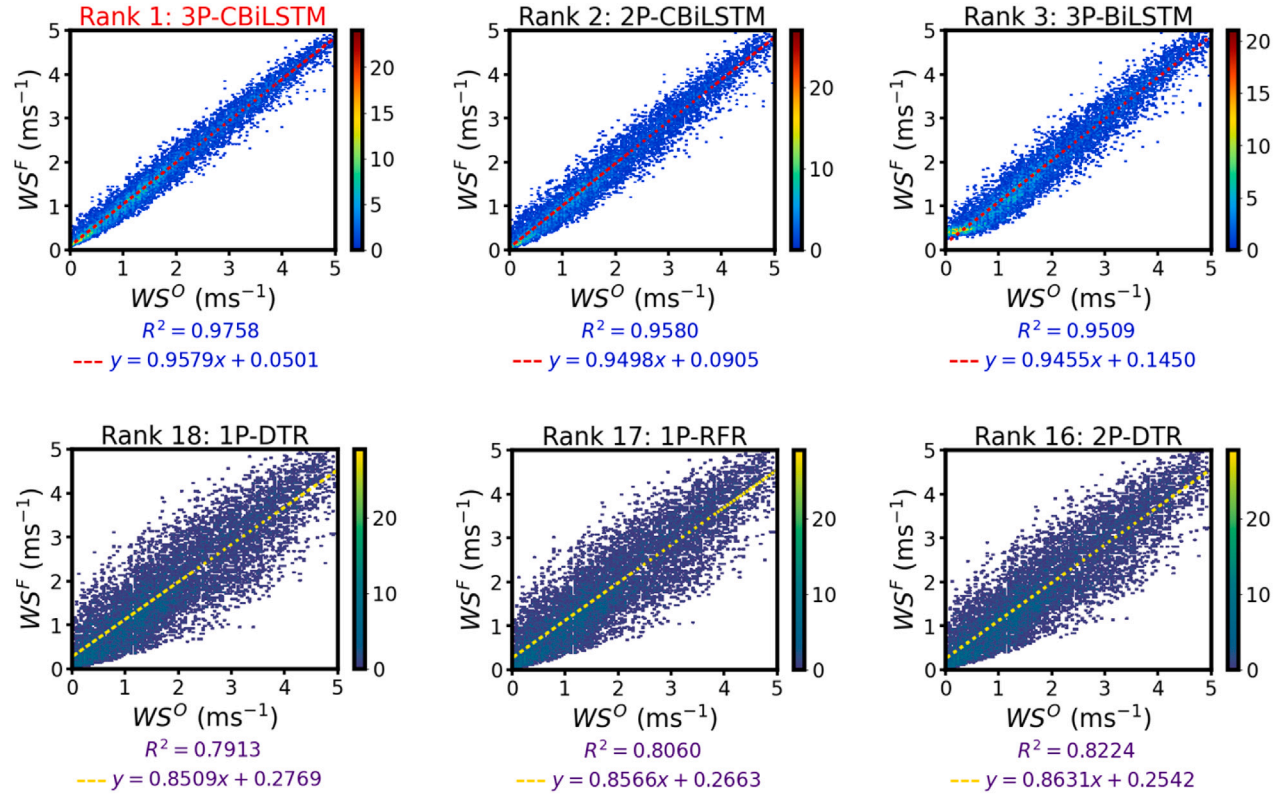
4.3. Advantages of the proposed 3P-CBiLSTM model

This section summarizes how the integration of CBiLSTM, TMGWO, and BOHB is advantageous for *WS* forecasting.

CBiLSTM outperformed the standard BiLSTM in terms of both prediction accuracy and time. The added benefits came from the CNN layer, which helped mitigate the effects of noise by filtering out irrelevant information and focussing on the salient features derived from ground and satellite-based sources. The salient feature map generated by CNN was then passed onto the BiLSTM component of CBiLSTM, which effectively captured both past and future long-term dependencies from the historical sequential data. Both CBiLSTM and BiLSTM gave better predictive results than LSTM since it cannot capture context from both directions. However, BiLSTM has twice the number of parameters than LSTM, which resulted in higher compute time in majority cases. The ML tree-based GBR, RFR, and DTR models had the lowest compute

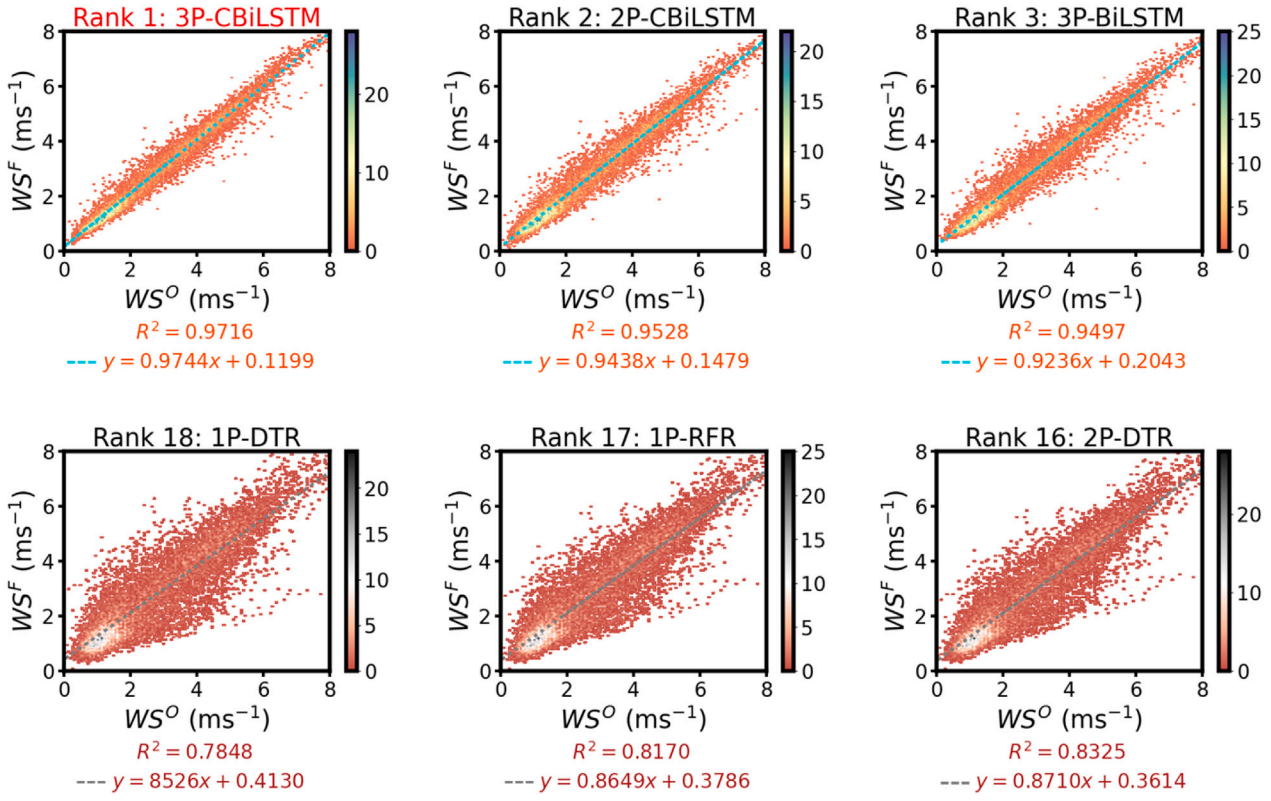


(a) Station 1 – Rakiraki

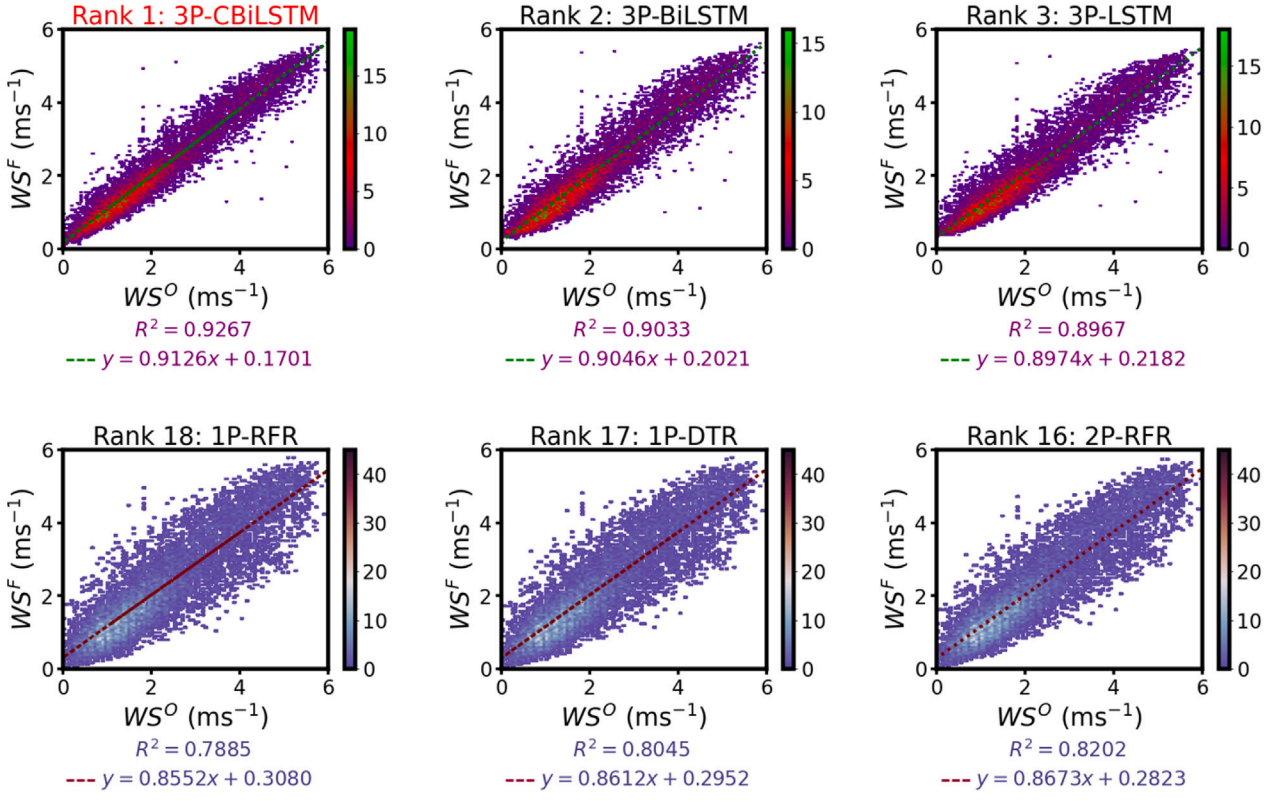


(b) Station 2 – Sigatoka

Fig. 9. Selected best and worst three ranked kernel density estimate (KDE) scatter plots of forecast models comparing the forecasted WS^F (ms^{-1}) and observed WS^O (ms^{-1}) for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase. The frequency of samples within the binning area of plots are depicted using the colour bars.

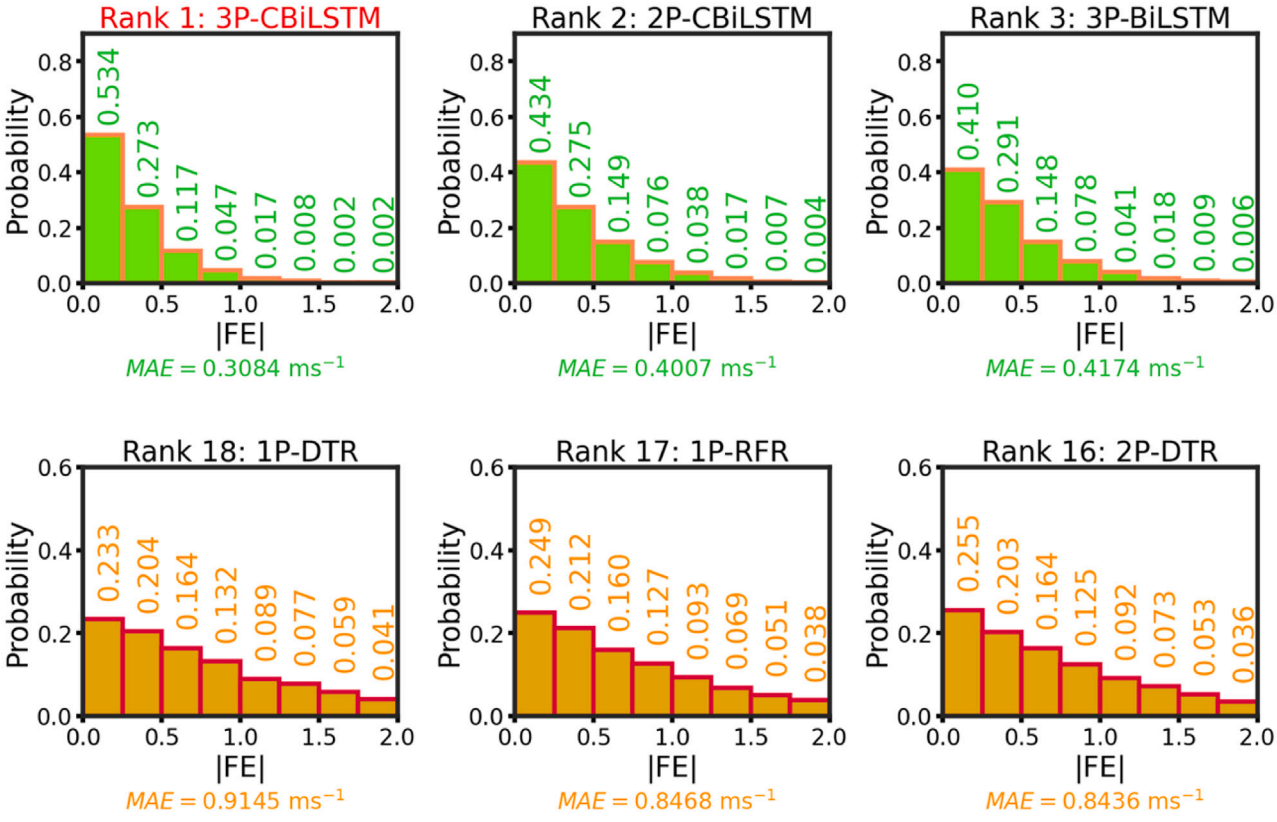


(c) Station 3 – Tokotoko

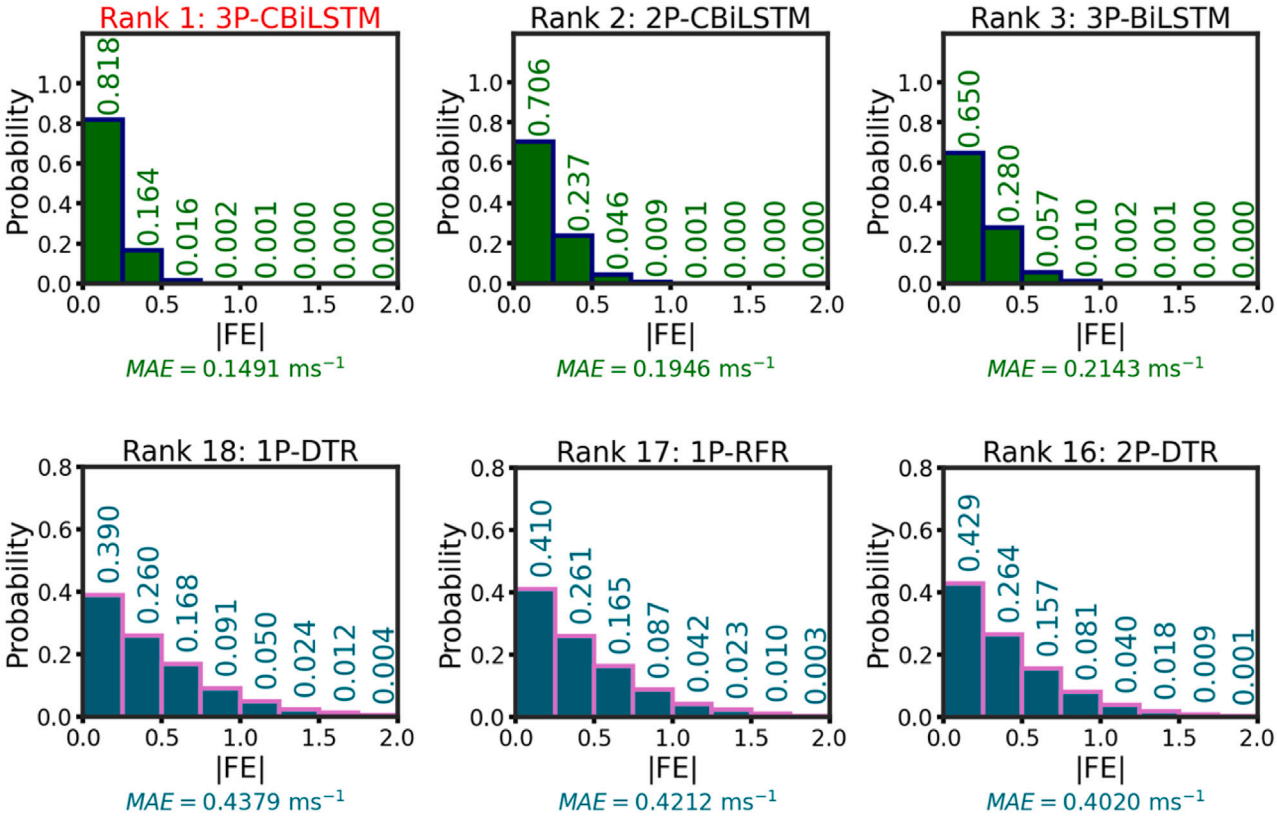


(d) Station 4 – Rarawai

Fig. 9. (*continued*).

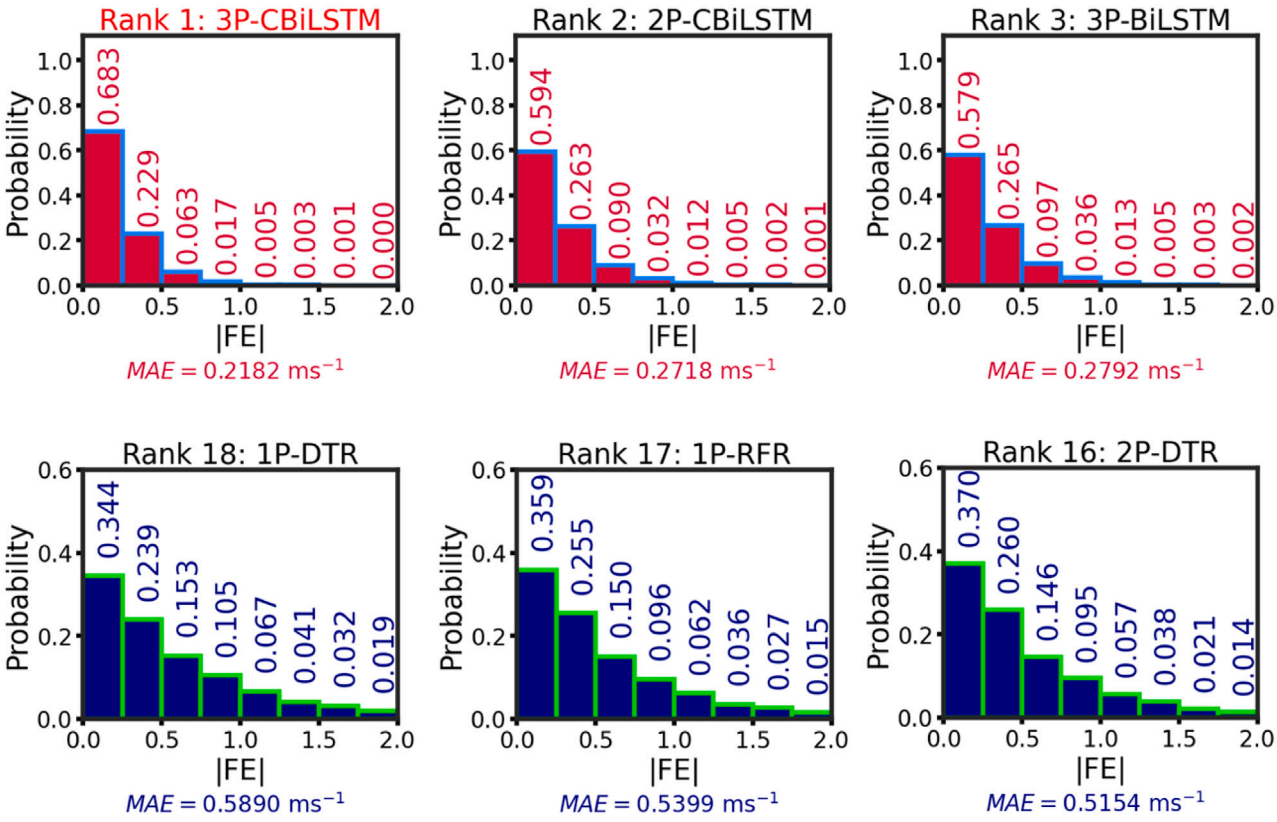


(a) Station 1 – Rakiraki

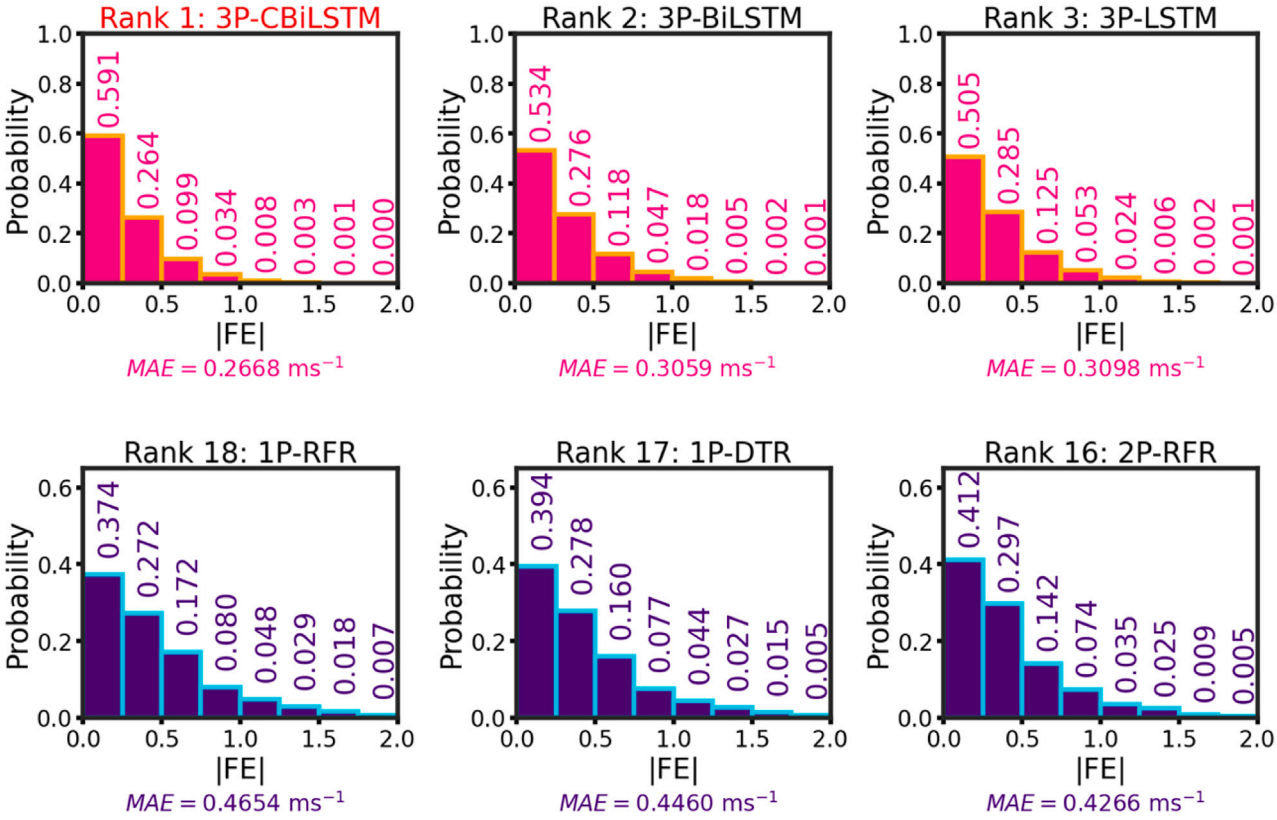


(b) Station 2 – Sigatoka

Fig. 10. Selected best and worst three ranked histograms of forecast models displaying the probability of the absolute 1-hour forecasting error $|FE|$ for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase.



(c) Station 3 – Tokotoko



(d) Station 4 – Rarawai

Fig. 10. (*continued*).

time. However, CBiLSTM outperformed these models given their limited ability in capturing complex patterns in the time series data. Also, tree-based models perform poorly when extrapolating outside the range of the training data and are prone to overfitting.

TMGWO was used for FS to enhance the performance of 1P-CBiLSTM. TMGWO helped in identifying and selecting relevant features whilst reducing overfitting and increasing model efficiency that improved the model interpretability and generalization. TMGWO is beneficial for the following reasons: (1) It efficiently explores the feature space by using a two-phase mutation process that allows the algorithm to explore both local and global optima. (2) It uses an adaptive mutation rate that allows the algorithm to converge faster to the optimal subset of features. (3) It is robust to noisy data since it uses a probabilistic approach to mutation. To further emphasize the importance of TMGWO, a detailed comparison is done with GWO and four state-of-the-art optimizers (i.e., WOA, SSA, SCA, and PSO) in Table B.2. The strengths of TMGWO outweigh these optimizers for FS. Experimental evaluation of these six optimizers is shown in Table B.5. Considering all four sites, the TMGWO registered an average increase in *LM* by 1.41%, 2.19%, 3.42%, 3.59%, and 3.83% when compared against GWO, WOA, SSA, SCA, and PSO, respectively.

BOHB was used for HPO to further improve the performance of 2P-CBiLSTM. BOHB combined the benefits of BO and HB algorithms. In its original form, HB uses a RS to screen its HP search space, which results in low efficiency. BOHB replaces the RS with BO, which helped achieve both high performance and low execution time in this study. BOHB is essential for the following reasons: (1) It efficiently locates optimal HPs, which improves the models' generalization ability. (2) It helps the model converge faster towards the best set of HPs, which reduces the time and computational resources required for training and validation. (3) It eliminates the need for manual HP selection, which is time-consuming and require domain expertise. To further highlight the importance of BOHB, an elaborate comparison is done against five popular HPO algorithms (i.e., BO, HB, RS, GS, and GA) in Table B.1. The benefits of BOHB helps overcome the drawbacks of these HPO methods. Experimental evaluation of these six HPO methods is furnished in Table B.6. For all four stations, the BOHB achieved an average increase in *LM* by 3.5%, 4.54%, 5.93%, 4.7%, and 6.94% when compared against BO, HB, RS, GS, and GA, respectively.

4.4. Model explainability

xAI was used to increase the authenticity of the proposed "black-box" model. Recently, *xAI* has gained popularity in many fields and different methods have been compared. LIME and SHAP are the two popular *xAI* methods used to provide insight into the features of a model that are most important in making a prediction. LIME is a versatile and efficient method that provides easily interpretable explanations, while SHAP is a more accurate and consistent method that provides both global and local explanations [94]. Based on the suitability and characteristics of these methods, LIME was used for local explainability and SHAP was used for global model interpretability.

LIME was used to locally interpret the respective (i^{th}) predicted instances of the test datasets. The results of five selected instances: i = 1, 2,187, 4,375, 6,562, and 8,749 are presented. The bar graphs in Fig. 11 represent the contribution of the best five predictors to the prediction of i^{th} instances for RK. The red bars represent negative LIME values, which favour lower *WS* prediction, whereas the green bars depict positive LIME values, which favour higher *WS* prediction. For all four sites, three out of five predicted instances have higher *WS* compared to the WS^O (Table 10). These predictions were strongly driven by the top features. For example, considering the first instance of RK, the best five predictors influencing the prediction were *WSa*, *WD*, *QV2M*, *Tmax*, and *CSWDiR* (Fig. 11(a)). *WSa* and *WD* had a higher combined positive LIME score, which favoured a higher WS^F. For all sites, the antecedent *WS* (*WSa*) registered the largest LIME scores; thus, had the greatest impact on the predictions. The interpretations regarding *WSa* are as follows: $WSa \leq 5.87$ resulted in lower WS^F and $WSa > 5.87$ led to higher WS^F (RK), $WSa \leq 1.79$ caused lower WS^F and $WSa > 1.79$ favoured higher WS^F (SG), $WSa \leq 2.89$ led to lower WS^F and $WSa > 2.89$ influenced higher WS^F (TK), and $WSa \leq 1.85$ favoured lower WS^F and $WSa > 1.85$ led to higher WS^F (RW). The LIME results were valuable in providing simplified local explanations.

Furthermore, SHAP was used to globally interpret the proposed model through feature importance and summary plots (Fig. 12). The feature importance bar plots used the mean absolute Shapley values to rank attributes from high to low relevance. These plots only represent the feature ranks and give no other information. Conversely, the summary plots are more practical in giving better interpretations. It integrates feature importance with feature effects, where each point on the beeswarm plot is a Shapley value for the respective feature at that instance [56]. The colour of each instance depicts the feature value from low (in blue) to high (in pink). Fig. 12 reveals that *WSa* has the highest importance for all sites. Other common best features at these sites were *CSWDiR*, *WD*, *RH*, *Tmax*, *T2M*, and *CPARtot*. For RK (Fig. 12(a)), higher values of *WSa* and *CSWDiR* had positive Shapley values indicating that higher values of these features favour higher WS^F, while lower values led to lower WS^F. For SG and TK (Fig. 12(b) and 12(c)), higher values of *WSa*, *CPARtot*, and *CSWDiR* had positive Shapley values meaning that higher values of these features favour higher WS^F. For RW (Fig. 12(d)), higher values of *WSa*, *Radn*, and *CSWDiR* had positive Shapley scores, whereas higher values of *RH* and *Tmax* had negative Shapley values. This indicates that at higher *RH* and *Tmax*, the WS^F is lower. These global interpretations are significant for practical applications, which makes the proposed model highly reliable.

Moreover, the feature dependence plots for RK (Fig. 13) reveal how the model predictions vary with the feature values and their corresponding interactions. Fig. 13(a) illustrates the feature interaction between *WSa* and *WD*, which reveals the following: the range $100 < WD \leq 190$ deg and lower *WSa* values favour a higher WS^F (i.e., positive SHAP values), the same *WD* range with higher *WSa* favour a lower WS^F (i.e., negative SHAP values), and $WD > \approx 280$ deg with lower *WSa* values favour a higher WS^F (i.e., positive SHAP values). Fig. 13(b) shows the interaction between *WSa* and *Tmax*. This relationship reveals that $Tmax \leq \approx 25.9$ °C and higher *WSa* values lead to higher WS^F, whereas $Tmax > \approx 25.9$ °C result in lower WS^F. Fig. 13(c) explains the interaction between *WSa* and *CSWDiR*. Here, $CSWDiR > \approx 14.46$ Whm^{-2} and higher *WSa* forecasts higher *WS*, whereas minimal to no *CSWDiR* (e.g., during night time) favour lower WS^F. The interaction between *WSa* and *Tmin* is displayed in Fig. 13(d), where $Tmin \leq \approx 25.4$ °C and $Tmin > \approx 25.48$ °C favour higher and lower WS^F, respectively. Also, $Tmin \leq \approx 22$ °C and higher *WSa* values lead to higher WS^F. Therefore, feature dependence plots offer detailed model interpretation by revealing how different predictors interact to generate predictions.

4.5. Application of the proposed explainable approach

A real-life application of the proposed explainable approach is shown in Fig. 14. The pre-trained explainable 3P-CBiLSTM model would continuously be fine-tuned, trained, and updated by the labelled training dataset stored in the database. This is to ensure that the model does not get outdated. New historical data (i.e., $t_{L-1}, \ldots, t_{L-n}$) would be fed as unlabelled testing data to the proposed pre-trained model. The model would then predict the 1-hour ahead *WS* (i.e., t_{L+1}) and reveal how accurate the prediction is. Additionally, the local and global explanations would be shown on the user interface to increase model interpretability, trust, and reliability. All these information can be used by wind farm operators to help with optimal decision-making to enhance the stability, reliability, and the security of the wind power systems and avoid unwarranted power brownouts. By doing so, wind

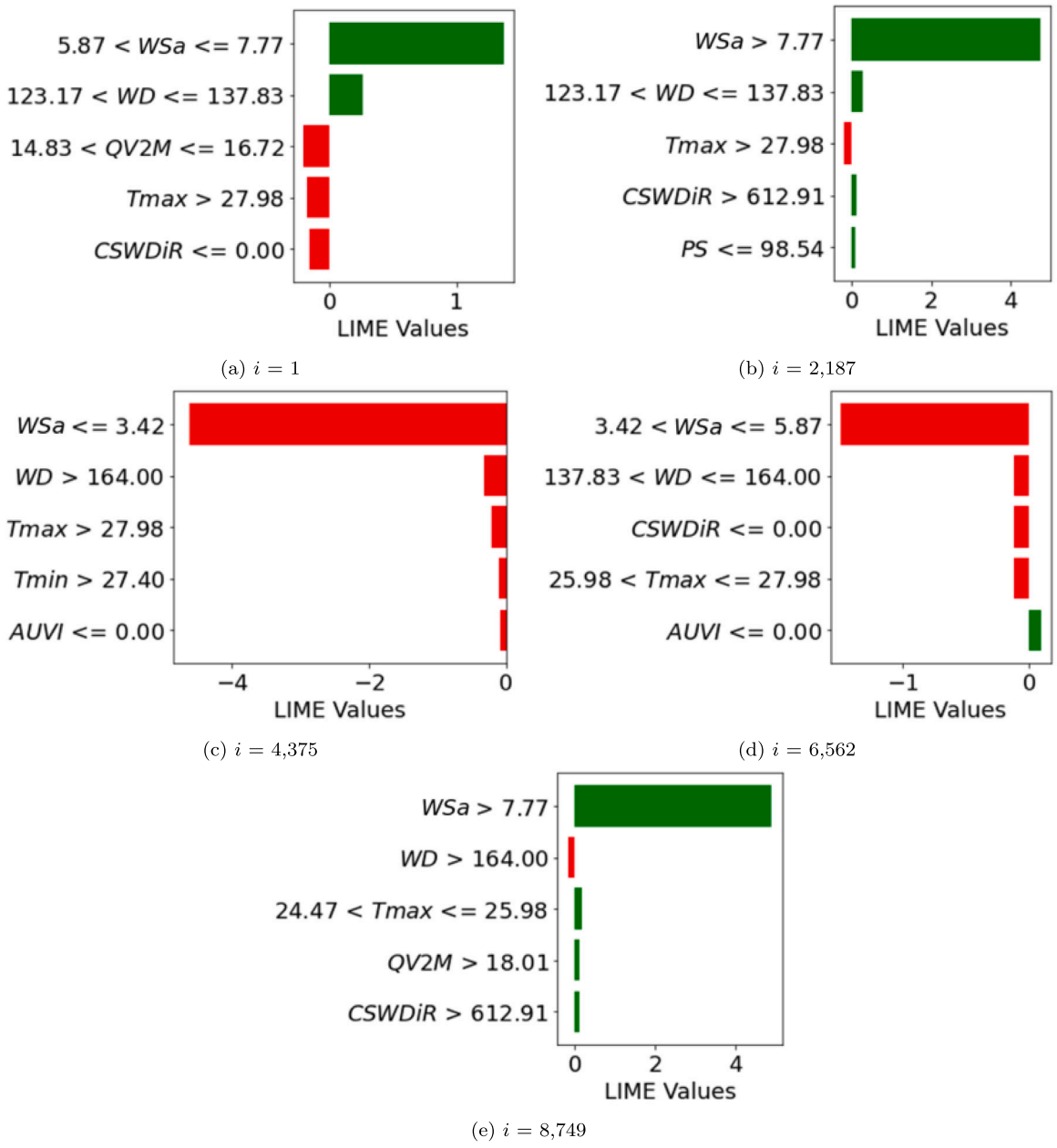
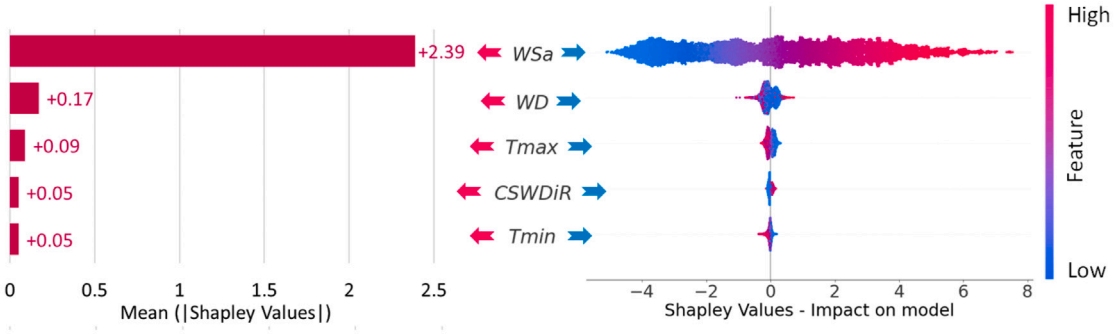


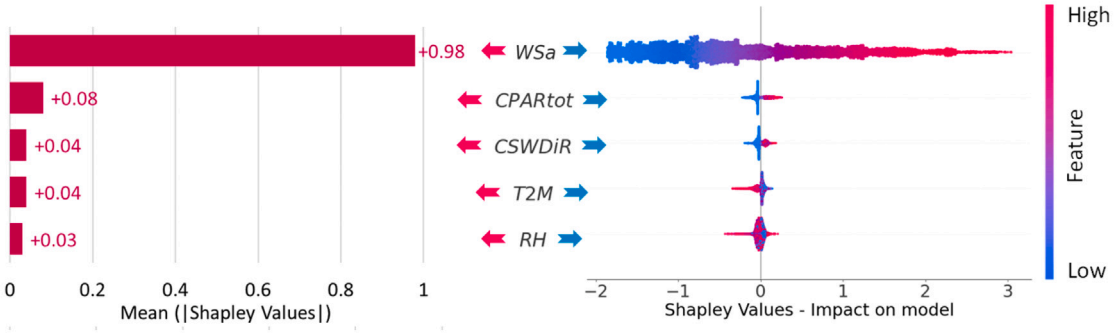
Fig. 11. Bar plots representing the LIME scores for local instances i = (a) 1 (i.e., first data point in the test phase), (b) 2,187, (c) 4,375, (d) 6,562, and (e) 8,749 (i.e., last or 100^{th} percentile data point) for studied station 1 – Rakiraki (RK). Note: The green bar (i.e., positive LIME score) indicates that the predictor favours a higher forecasted *WS* and the red bar is when the predictor favours a lower forecasted *WS*.

Table 10
Local interpretability of the five selected instances using Local Interpretable Model-Agnostic Explanations (LIME) in the test phase. (Note: The explanations are based on the top predictors with highest LIME scores.)

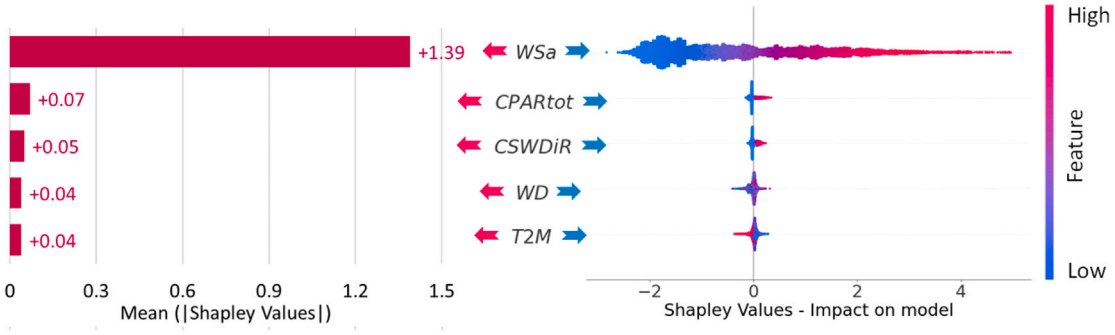
Percentile (%)	Instance (i)	RK			SG		
		WS^O	WS^F	LIME explainer indicators	WS^O	WS^F	LIME explainer indicators
0	1	7.73	7.83	*WSa* and *WD* lead to $WS^F > WS^O$	4.00	4.10	*WSa*, *CPARtot*, and *CSWDiR* lead to $WS^F > WS^O$
25	2,187	9.75	9.86	*WSa*, *WD*, *CSWDiR*, and *PS* lead to $WS^F > WS^O$	4.72	4.85	*WSa*, *CPARtot*, and *CSWDiR* lead to $WS^F > WS^O$
50	4,375	1.35	1.23	*WSa*, *WD*, *Tmax*, *Tmin*, and *AUVI* lead to $WS^F < WS^O$	0.87	0.75	*WSa*, *CPARtot*, *CSWDiR*, and *RH* lead to $WS^F < WS^O$
75	6,562	4.08	3.98	*WSa*, *WD*, *CSWDiR*, and *Tmax* lead to $WS^F < WS^O$	1.53	1.42	*WSa* and *Tmin* lead to $WS^F < WS^O$
100	8,749	9.48	9.59	*WSa*, *Tmax*, *QV2M*, and *CSWDiR* lead to $WS^F > WS^O$	4.60	4.73	*WSa*, *CPARtot*, *CSWDiR*, and *AUVI* lead to $WS^F > WS^O$
		TK			RW		
0	1	6.27	6.39	*WSa*, *CPARtot*, and *CSWDiR* lead to $WS^F > WS^O$	5.09	5.21	*WSa*, *Radn*, *RH*, and *CSWDiR* lead to $WS^F > WS^O$
25	2,187	4.17	4.31	*WSa*, *CPARtot*, *CSWDiR*, *WD*, and *ACI* lead to $WS^F > WS^O$	5.05	4.89	*WSa*, *Tmax*, and *CSWDiR* lead to $WS^F < WS^O$
50	4,375	2.60	2.49	*WSa* and *T2M* lead to $WS^F < WS^O$	0.78	0.59	*WSa*, *Radn*, *RH*, and *CSWDiR* lead to $WS^F < WS^O$
75	6,562	4.03	4.13	*WSa*, *CPARtot*, and *CSWDiR* lead to $WS^F > WS^O$	3.23	3.37	*WSa*, *Radn*, *RH*, and *CSWDiR* lead to $WS^F > WS^O$
100	8,749	0.95	0.83	*WSa*, *CPARtot*, *CSWDiR*, and *ACI* lead to $WS^F < WS^O$	4.23	4.35	*WSa*, *Radn*, *RH*, and *CSWDiR* lead to $WS^F > WS^O$



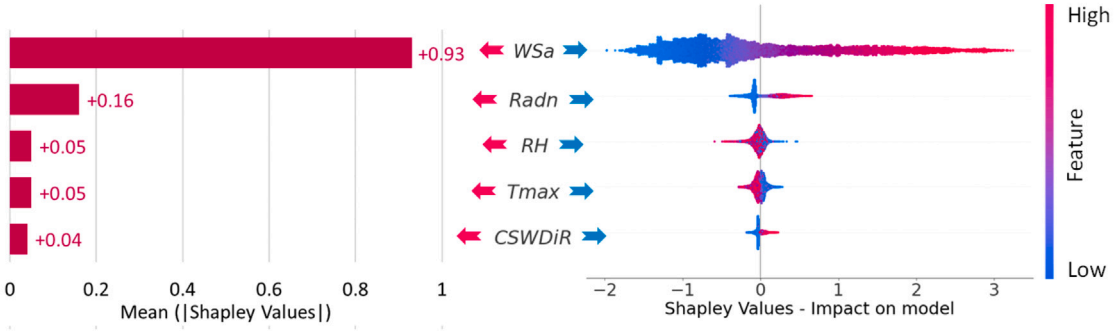
(a) Station 1 – Rakiraki



(b) Station 2 – Sigatoka



(c) Station 3 – Tokotoko



(d) Station 4 – Rarawai

Fig. 12. Bar SHAP feature importance plots and beeswarm SHAP summary plots for the four studied stations: (a) Rakiraki (RK), (b) Sigatoka (SG), (c) Tokotoko (TK), and (d) Rarawai (RW) in their test phase.

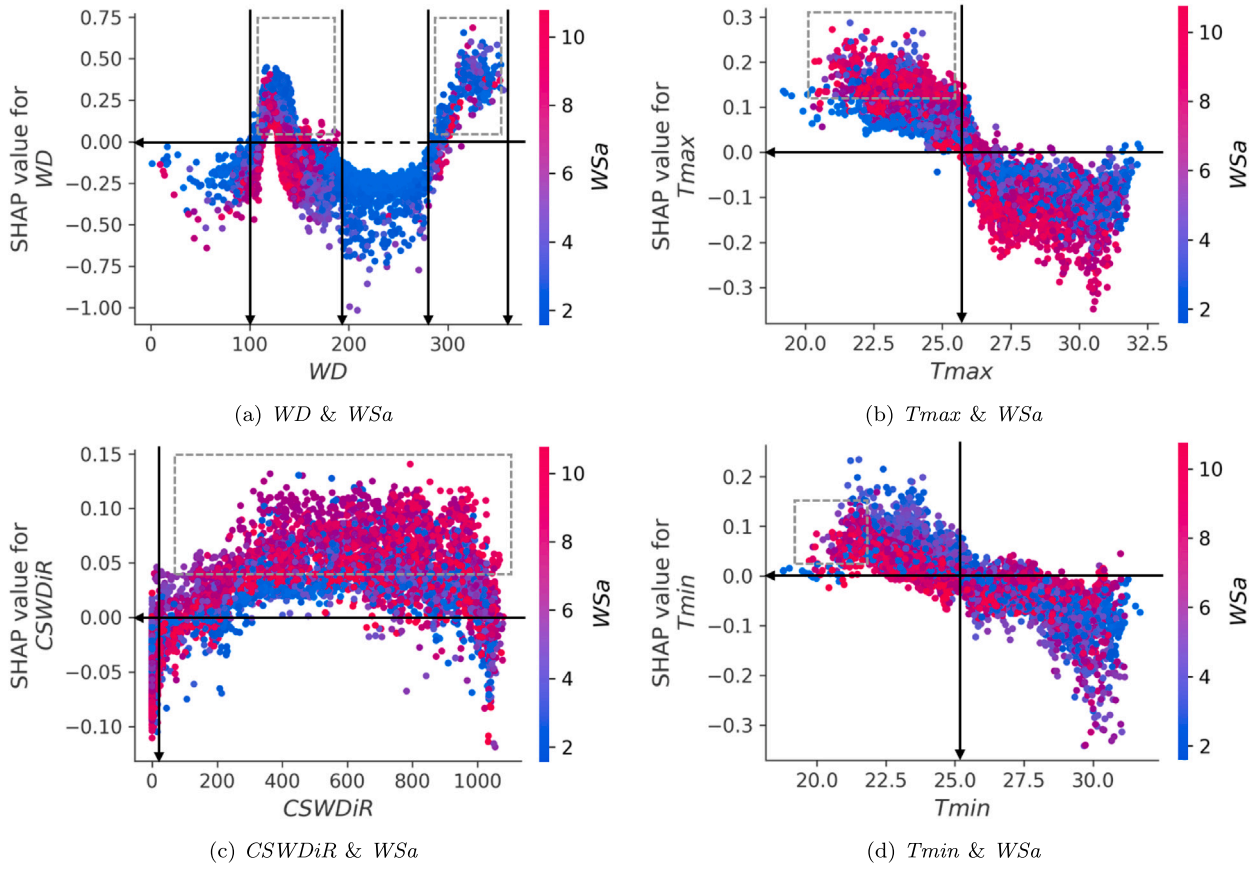


Fig. 13. SHAP dependence plots for interaction between attributes (a) wind direction (*WD*) & antecedent wind speed (*WSa*), (b) maximum temperature (*Tmax*) & *WSa*, (c) Clear Sky Surface Shortwave Downward Irradiance (*CSWDiR*) & *WSa*, and (d) minimum temperature (*Tmin*) & *WSa* for studied station 1 – Rakiraki (RK) in the test phase.

farm operators can optimize energy production, reduce costs, improve operational safety during extreme events, and contribute to a more sustainable energy future.

5. Conclusions

In this paper, an explainable three-phase hybrid CBiLSTM (3P-CBiLSTM) prediction framework is developed for hourly *WS*. To improve the predictive accuracy and lower the computational complexity of CBiLSTM, a two-phase mutation grey wolf optimizer (TMGWO) is used for dimensionality reduction and Bayesian Optimization and HyperBand (BOHB) algorithm is used for efficient hyperparameter selection. The performance of 3P-CBiLSTM is evaluated against powerful ML and DL-based benchmark models. Diverse statistical metrics and diagnostic plots confirm the excellent predictive capability of the proposed 3P-CBiLSTM model over other counterpart models. The objective model recorded the highest *r* (0.963 – 0.990), and the lowest *MAE* (0.149 – 0.308) and *RMSE* (0.197 – 0.420) for all four sites. It also registered the largest proportion of forecast errors (≈ 53.4 – 81.8%) in the smallest bin ≤ $|0.25|$ ms^{-1} amongst all evaluated sites. Furthermore, *xAI* is used to interpret the underlying architecture of the proposed "black-box" model to showcase its authenticity. LIME *xAI* achieved local interpretability and SHAP *xAI* enhanced the global model interpretability. These tools effectively explain how different meteorological variables affect *WS* at different locations. For instance, LIME and SHAP point out the highest contributing features, where the top five features are studied in this work. Both local and global *xAI* analysis reveal antecedent *WS* (*WSa*) to be the best predictor. The proposed prediction method can help the wind farm operators with quality decision-making to help: maximize wind energy generation, reduce the turbine failure rate; hence, minimize the maintenance costs, prevent sudden fluctuations in the capacity factor, and enhance the security of the electric power systems and avoid unwarranted power brownouts. These benefits can help make wind energy a more feasible and sustainable option for meeting the rising energy demand.

5.1. Limitations and future research directions

The proposed explainable framework has shown remarkable results. However, there remain a few limitations that are to be addressed in future studies. These are summarized as follows:

(i) This study adopted a single-step prediction strategy that does not predict *WS* at longer forecast horizon than 1-hour. In future, a multiple-input multiple-output (MIMO) strategy should be tested to predict *WS* at longer forecast horizon.

(ii) The single-step prediction outputs were expressed as point forecasts. Future research should explore interval and probabilistic forecasting strategies to provide a range of possible outcomes to facilitate better decision-making. These methods also take into account the uncertainty inherent in future events.

(iii) This study investigated the effect of various predictor variables on the output. Therefore, use of univariate decomposition methods like ICEEMDAN would have been time-consuming. However, multivariate decomposition methods like MEMD and stationary wavelet transform (SWT) can be explored to decompose numerous variables simultaneously.

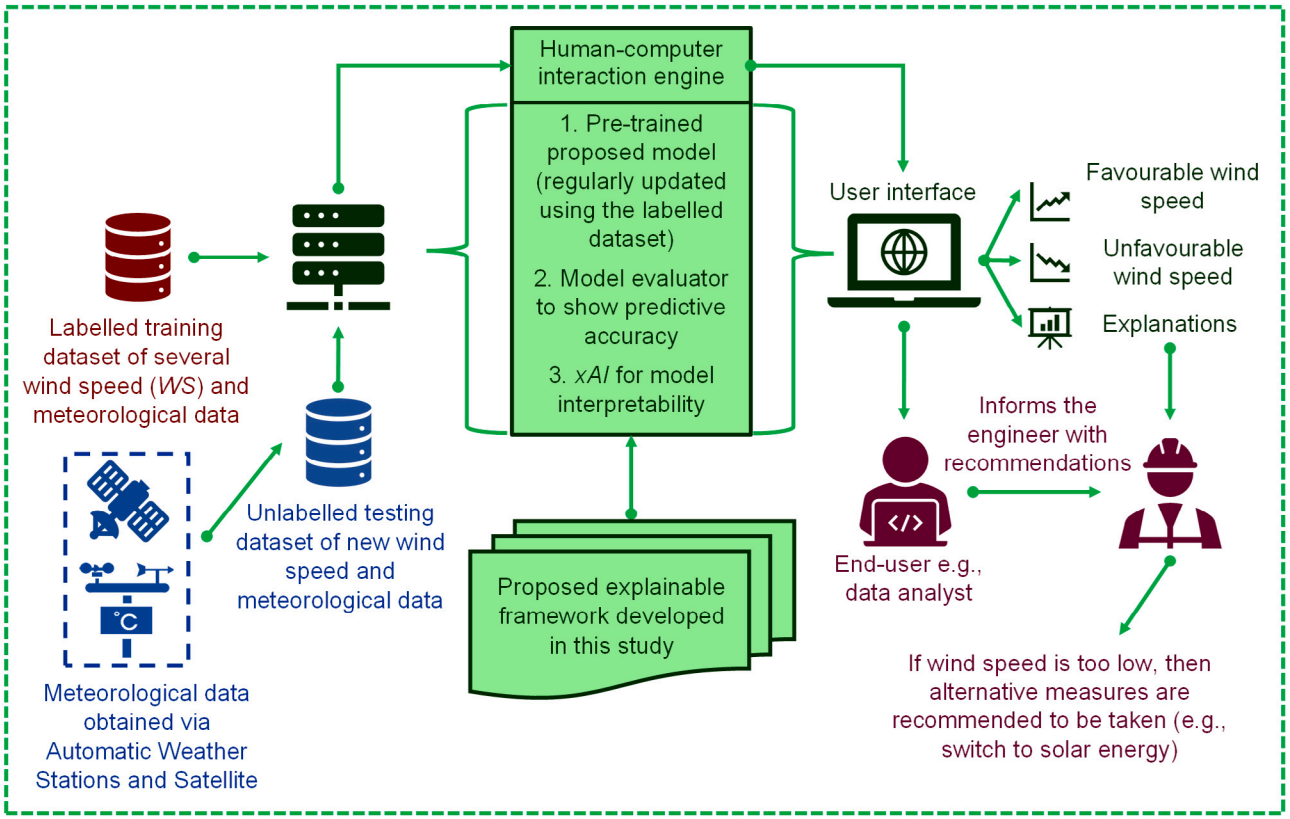


Fig. 14. Schematic representation of the proposed explainable 3P-CBiLSTM predictive model for practical application in the wind energy sector.

(iv) CBiLSTM model is complex and highly parametric. In future, model pruning techniques can be used to reduce the model size, while maintaining its accuracy.
(v) The proposed CBiLSTM is a non-interpretable "black-box" model. Hence, model-agnostic *xAI* methods: LIME and SHAP were used. In future, model-specific interpretable models like TabNet and N-BEATS should be tested.
(vi) Data availability is a pressing issue for a SIDS like Fiji. The wind power data of Butoni wind farm is not accessible; hence, cannot be studied for forecasting purpose. In future, wind power data for a similar SIDS can be used.
(vii) Also, offshore Fijian sites have better wind regime than onshore sites. However, there are no offshore wind monitoring towers in Fiji. Therefore, the offshore data can be obtained via satellite-based sources for future studies.

CRediT authorship contribution statement

Lionel P. Joseph: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ravinesh C. Deo:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **David Casillas-Pérez:** Writing – review & editing. **Ramendra Prasad:** Writing – review & editing. **Nawin Raj:** Writing – review & editing. **Sancho Salcedo-Sanz:** Writing – review & editing.

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.

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgements

The authors would like to acknowledge the Fiji Meteorological Services (FMS) for providing data that enabled this study. This research was supported by USQ International PhD Stipend and International PhD Tuition Fee Scholarships awarded to the first author by the University of Southern Queensland (UniSQ), managed by Graduate Research School (GRS). This research was partially supported by project PID2020-115454GB-C21 of Spanish Ministry of Science and Innovation (MICINN), to continually build research synergies between Professors Ravinesh Deo (UniSQ, Australia) and Sancho Salcedo-Sanz (UAH, Spain).

Appendix A. Acronyms list

See Table A.1.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.apenergy.2024.122624.

Table A.1
List of acronyms.

Acronym	Full name
ADF	Augmented Dickey–Fuller
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
BiLSTM	Bidirectional LSTM
BOHB	Bayesian Optimization (BO) and HyperBand (HB)
CBiLSTM	CNN-BiLSTM
3P-CBiLSTM	Three-Phase Hybrid CBiLSTM (i.e., Proposed Model)
CCF	Cross-Correlation Function
CNN	Convolutional Neural Network
CSA	Crow Search Algorithm
DL	Deep Learning
DTR	Decision Tree Regressor
ED	Evolutionary Decomposition
E_{NS}	Nash–Sutcliffe Efficiency
EWT	Empirical WT
FC	Fully Connected
FE	Forecasting Errors
FMS	Fiji Meteorological Services
FS	Feature Selection
FWA	Fireworks Algorithm
GA	Genetic Algorithm
GBR	Gradient Boosting Regressor
GPI	Global Performance Indicator
GS	Grid Search
GWO	Grey Wolf Optimizer
HGNDO	Hybrid Generalized Normal Distribution Optimizer
HPO	Hyperparameter (HP) Optimization
ICEEMDAN	Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
ISCA	Improved Sine and Cosine Algorithm (i.e., Improved SCA)
KNN	K-Nearest Neighbours
LIME	Local Interpretable Model-Agnostic Explanations
LM	Legates and McCabe Index
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MEMD	Multivariate EMD
MH	*Meta*-Heuristic
MI	Mutual Information
ML	Machine Learning
PACF	Partial Auto-Correlation Function
POA	Population-based Optimization Algorithm
PSIDS	Pacific Small Island Developing States
PSO	Particle Swarm Optimization
r	Pearson's Correlation Coefficient
RE	Renewable Energy
RFR	Random Forest Regressor
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RRMSE	Relative Root Mean Square Error
RS	Random Search
SHAP	SHapley Additive exPlanations
SSA	Salp Swarm Algorithm
SSD	Singular Spectrum Decomposition
TMGWO	Grey Wolf Optimizer integrated with a Two-Phase Mutation
TPE	Tree-structured Parzen Estimator
TVFEMD	Time-Varying Filter based EMD
WI	Willmott's Index of Agreement
WOA	Whale Optimization Algorithm
WS	Wind Speed
WS^F	Forecasted Wind Speed
WS^O	Observed Wind Speed
WTD	Wavelet Transform Decomposition
xAI	eXplainable Artificial Intelligence

References

[1] Lv S-X, Wang L. Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model. Energy 2023;263:126100.

[2] GWEC. Global wind report 2022. Brussels, Belgium: Global Wind Energy Council; 2022.

[3] Baïle R, Muzy J-F. Leveraging data from nearby stations to improve short-term wind speed forecasts. Energy 2023;263:125644.

[4] Li M, Yang Y, He Z, Guo X, Zhang R, Huang B. A wind speed forecasting model based on multi-objective algorithm and interpretability learning. Energy 2023;269:126778.

[5] Lv S-X, Wang L. Deep learning combined wind speed forecasting with hybrid time series decomposition and multi-objective parameter optimization. Appl Energy 2022;311:118674.

[6] Di Z, Ao J, Duan Q, Wang J, Gong W, Shen C, et al. Improving WRF model turbine-height wind-speed forecasting using a surrogate-based automatic optimization method. Atmos Res 2019;226:1–16.

[7] Hao Y, Yang W, Yin K. Novel wind speed forecasting model based on a deep learning combined strategy in urban energy systems. Expert Syst Appl 2023;219:119636.

[8] Jiang Z, Che J, He M, Yuan F. A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition. Renew Energy 2023;203:802–27.

[9] Ahmadi A, Nabipour M, Mohammadi-Ivatloo B, Amani AM, Rho S, Piran MJ. Long-term wind power forecasting using tree-based learning algorithms. IEEE Access 2020;8:151511–22.

[10] Joseph LP, Deo RC, Prasad R, Salcedo-Sanz S, Raj N, Soar J. Near real-time wind speed forecast model with bidirectional LSTM networks. Renew Energy 2023.

[11] Ortiz-García EG, Salcedo-Sanz S, Pérez-Bellido ÁM, Gascón-Moreno J, Portilla-Figueras JA, Prieto L. Short-term wind speed prediction in wind farms based on banks of support vector machines. Wind Energy 2011;14(2):193–207.

[12] Salcedo-Sanz S, Ortiz-Garcı EG, Pérez-Bellido ÁM, Portilla-Figueras A, Prieto L, et al. Short term wind speed prediction based on evolutionary support vector regression algorithms. Expert Syst Appl 2011;38(4):4052–7.

[13] Santamaría-Bonfil G, Reyes-Ballesteros A, Gershenson C. Wind speed forecasting for wind farms: A method based on support vector regression. Renew Energy 2016;85:790–809.

[14] Salcedo-Sanz S, Pérez-Bellido ÁM, Ortiz-García EG, Portilla-Figueras A, Prieto L, Paredes D. Hybridizing the fifth generation mesoscale model with artificial neural networks for short-term wind speed prediction. Renew Energy 2009;34(6):1451–7.

[15] Salcedo-Sanz S, Perez-Bellido AM, Ortiz-García EG, Portilla-Figueras A, Prieto L, Correoso F. Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks. Neurocomputing 2009;72(4–6):1336–41.

[16] Liu H, Chen C, Tian H-q, Li Y-f. A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. Renew Energy 2012;48:545–56.

[17] Salcedo-Sanz S, Ortiz-García E, Pérez-Bellido ÁM, Portilla-Figueras A, Prieto L, Paredes D, et al. Performance comparison of multilayer perceptrons and support vector machines in a short-term wind speed prediction problem. Neural Network World 2009;19(1):37.

[18] Neshat M, Nezhad MM, Abbasnejad E, Mirjalili S, Tjernberg LB, Garcia DA, et al. A deep learning-based evolutionary model for short-term wind speed forecasting: A case study of the lillgrund offshore wind farm. Energy Convers Manage 2021;236:114002.

[19] Memarzadeh G, Keynia F. A new short-term wind speed forecasting method based on fine-tuned LSTM neural network and optimal input sets. Energy Convers Manage 2020;213:112824.

[20] Altan A, Karasu S, Zio E. A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. Appl Soft Comput 2021;100:106996.

[21] Shao B, Song D, Bian G, Zhao Y. Wind speed forecast based on the LSTM neural network optimized by the firework algorithm. Adv Mater Sci Eng 2021;2021:1–13.

[22] Jaseena K, Kovoor BC. Decomposition-based hybrid wind speed forecasting model using deep bidirectional LSTM networks. Energy Convers Manage 2021;234:113944.

[23] Zhang S, Chen Y, Xiao J, Zhang W, Feng R. Hybrid wind speed forecasting model based on multivariate data secondary decomposition approach and deep learning algorithm with attention mechanism. Renew Energy 2021;174:688–704.

[24] Zhang C, Ma H, Hua L, Sun W, Nazir MS, Peng T. An evolutionary deep learning model based on TVFEMD, improved sine cosine algorithm, CNN and BiLSTM for wind speed prediction. Energy 2022;254:124250.

[25] Nguyen THT, Phan QB. Hourly day ahead wind speed forecasting based on a hybrid model of EEMD, CNN-Bi-LSTM embedded with GA optimization. Energy Rep 2022;8:53–60.

[26] Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. Appl Energy 2021;304:117766.

[27] Chen J, Liu H, Chen C, Duan Z. Wind speed forecasting using multi-scale feature adaptive extraction ensemble model with error regression correction. Expert Syst Appl 2022;207:117358.

[28] Wang Z, Zhang J, Zhang Y, Huang C, Wang L. Short-term wind speed forecasting based on information of neighboring wind farms. IEEE Access 2020;8:16760–70.

[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.

[30] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 2005;18(5–6):602–10.

[31] Qian Z, Pei Y, Zareipour H, Chen N. A review and discussion of decomposition-based hybrid models for wind energy forecasting applications. Appl Energy 2019;235:939–53.

[32] Huang C-L, Wang C-J. A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl 2006;31(2):231–40.

[33] Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: A big comparison for NAS. 2019, arXiv preprint arXiv:1912.06059.

[34] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(2).

[35] Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 2017;18(1):6765–816.

[36] Močkus J. On Bayesian methods for seeking the extremum. In: Optimization techniques IFIP technical conference: novosibirsk, July 1–7, 1974. Springer; 1975, p. 400–4.

[37] Eggensperger K, Feurer M, Hutter F, Bergstra J, Snoek J, Hoos H, et al. Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In: NIPS workshop on bayesian optimization in theory and practice, vol. 10. 2013, p. 1–5.

[38] Wang J, Xu J, Wang X. Combination of hyperband and Bayesian optimization for hyperparameter optimization in deep learning. 2018, arXiv preprint arXiv:1801.01596.

[39] Falkner S, Klein A, Hutter F. BOHB: Robust and efficient hyperparameter optimization at scale. In: International conference on machine learning. PMLR; 2018, p. 1437–46.

[40] Haris M, Hasan MN, Qin S. Early and robust remaining useful life prediction of supercapacitors using BOHB optimized deep belief network. Appl Energy 2021;286:116541.

[41] Rong M, Gong D, Gao X. Feature selection and its use in big data: Challenges, methods, and trends. IEEE Access 2019;7:19709–25.

[42] Li Y, Peng T, Zhang C, Sun W, Hua L, Ji C, et al. Multi-step ahead wind speed forecasting approach coupling maximal overlap discrete wavelet transform, improved grey wolf optimization algorithm and long short-term memory. Renew Energy 2022;196:1115–26.

[43] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Appl Soft Comput 2018;62:441–53.

[44] Kennedy J, Eberhart RC, Shi Y. The particle swarm. Swarm Intell 2001;287–325.

[45] Mirjalili S. SCA: A Sine cosine algorithm for solving optimization problems. Knowl-Based Syst 2016;96:120–33.

[46] Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. Salp swarm algorithm: A bio-inspired optimizer for engineering design problems. Adv Eng Softw 2017;114:163–91.

[47] Mirjalili S, Lewis A. The whale optimization algorithm. Adv Eng Softw 2016;95:51–67.

[48] Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. Adv Eng Softw 2014;69:46–61.

[49] Abdel-Basset M, El-Shahat D, El-Henawy I, De Albuquerque VHC, Mirjalili S. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Syst Appl 2020;139:112824.

[50] Guo M, Wang J-S, Zhu L, Guo S-S, Xie W. An improved grey wolf optimizer based on tracking and seeking modes to solve function optimization problems. IEEE Access 2020;8:69861–93.

[51] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy 2020;23(1):18.

[52] Alves MA, Castro GZ, Oliveira BAS, Ferreira LA, Ramírez JA, Silva R, et al. Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Comput Biol Med 2021;132:104335.

[53] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 1135–44.

[54] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Mach Intell 2020;2(1):56–67.

[55] Shapley LS. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the theory of games II. Princeton University Press; 1953, p. 307–17.

[56] Joseph LP, Joseph EA, Prasad R. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. Comput Biol Med 2022;151:106178.

[57] Kumari P, Toshniwal D. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. Appl Energy 2021;295:117061.

[58] Zhang C, Peng T, Nazir MS. A novel integrated photovoltaic power forecasting model based on variational mode decomposition and CNN-BiGRU considering meteorological variables. Electr Power Syst Res 2022;213:108796.

[59] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;1189–232.

[60] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43(6):1947–58.

[61] Xu M, Watanachaturaporn P, Varshney PK, Arora MK. Decision tree regression for soft classification of remote sensing data. Remote Sens Environ 2005;97(3):322–36.

[62] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst 2011;24.

[63] Jamieson K, Talwalkar A. Non-stochastic best arm identification and hyperparameter optimization. In: Artificial intelligence and statistics. PMLR; 2016, p. 240–8.

[64] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Amer Statist 1992;46(3):175–85.

[65] Mirjalili S, Lewis A. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. Swarm Evol Comput 2013;9:1–14.

[66] Mafarja M, Eleyan D, Abdullah S, Mirjalili S. S-shaped vs. V-shaped transfer functions for ant lion optimization algorithm in feature selection problem. In: Proceedings of the international conference on future networks and distributed systems. 2017, p. 1–7.
[67] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
[68] Stackhouse PW, Zhang T, Westberg D, Barnett AJ, Bristow T, Macpherson B, et al. POWER release 8.0. 1 (with GIS applications) methodology (data parameters, sources, & validation). Data Version 2018;8(1).
[69] Quansah AD, Dogbey F, Asilevi PJ, Boakye P, Darkwah L, Oduro-Kwarteng S, et al. Assessment of solar radiation resource from the NASA-POWER reanalysis products for tropical climates in ghana towards clean energy application. Sci Rep 2022;12(1):10684.
[70] Rocha PAC, de Sousa RC, de Andrade CF, da Silva MEV. Comparison of seven numerical methods for determining Weibull parameters for wind energy generation in the northeast region of Brazil. Appl Energy 2012;89(1):395–400.
[71] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[72] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. In: Osdi, vol. 16. Savannah, GA, USA; 2016, p. 265–83.
[73] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2623–31.
[74] 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.
[75] Dickey DA, Fuller WA. Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 1979;74(366a):427–31.
[76] Ghimire S, Deo RC, Casillas-Perez D, Salcedo-Sanz S. Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms. Appl Energy 2022;316:119063.
[77] Ghimire S, Nguyen-Huy T, Deo RC, Casillas-Perez D, Salcedo-Sanz S. Efficient daily solar radiation prediction with deep learning 4-phase convolutional neural network, dual stage stacked regression and support vector machine CNN-REGST hybrid model. Sustain Mater Technol 2022;32:e00429.
[78] Nadimi-Shahraki MH, Taghian S, Mirjalili S. An improved grey wolf optimizer for solving engineering problems. Expert Syst Appl 2021;166:113917.
[79] Li J, Stones RJ, Wang G, Li Z, Liu X, Xiao K. Being accurate is not enough: New metrics for disk failure prediction. In: 2016 IEEE 35th symposium on reliable distributed systems. IEEE; 2016, p. 71–80.
[80] Behar O, Khellaf A, Mohammedi K. Comparison of solar radiation models and their validation under Algerian climate–The case of direct irradiance. Energy Convers Manage 2015;98:236–51.
[81] Deo RC, Downs N, Parisi AV, Adamowski JF, Quilty JM. Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. Environ Res 2017;155:141–66.
[82] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geosci Model Dev 2014;7(3):1247–50.
[83] Ritter A, Muñoz-Carpena R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 2013;480:33–45.
[84] Althoff D, Rodrigues LN. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. J Hydrol 2021;600:126674.
[85] Deo RC, Şahin 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.
[86] Prasad R, Deo RC, Li Y, Maraseni T. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. Soil Tillage Res 2018;181:63–81.
[87] Hora J, Campos P. A review of performance criteria to validate simulation models. Expert Syst 2015;32(5):578–95.
[88] 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 Manage 2013;70:139–48.
[89] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol 1970;10(3):282–90.
[90] Legates DR, McCabe Jr GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35(1):233–41.
[91] Willmott CJ. On the validation of models. Phys Geogr 1981;2(2):184–94.
[92] Krause P, Boyle D, Bäse F. Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 2005;5:89–97.
[93] Deo RC, Şahin M, Adamowski JF, Mi J. Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over Australia to forecast global solar radiation: A new approach. Renew Sustain Energy Rev 2019;104:235–61.
[94] Moscato V, Picariello A, Sperlí G. A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 2021;165:113986.