

Article



Leveraging Advanced Data-Driven Approaches to Forecast Daily Floods Based on Rainfall for Proactive Prevention Strategies in Saudi Arabia

Anwar Ali Aldhafiri ¹, Mumtaz Ali ^{2,3,*} and Abdulhaleem H. Labban ⁴

- ¹ Department of Mathematics and Statistics, Faculty of Science, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia; aaaldhafeeri@kfu.edu.sa
- ² UniSQ College, University of Southern Queensland, Toowoomba, QLD 4305, Australia
- ³ Scientific Research Centre, Al-Ayen University, Nasiriyah 64001, Thi-Qar, Iraq
- ⁴ Department of Meteorology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; alabban@kau.edu.sa

* Correspondence: mumtaz.ali@unisq.edu.au

Abstract: Accurate flood forecasts are imperative to supervise and prepare for extreme events to assess the risks and develop proactive prevention strategies. The flood timeseries data exhibit both spatial and temporal structures and make it challenging for the models to fully capture the embedded features due to their complex stochastic nature. This paper proposed a new approach for the first time using variational mode decomposition (VMD) hybridized with Gaussian process regression (GPR) to design the VMD-GPR model for daily flood forecasting. First, the VMD model decomposed the (t - 1) lag into several signals called intrinsic mode functions (IMFs). The VMD has the ability to improve noise robustness, better mode separation, reduced mode aliasing, and end effects. Then, the partial auto-correlation function (PACF) was applied to determine the significant lag (t - 1). Finally, the PACF-based decomposed IMFs were sent into the GPR to forecast the daily flood index at (t - 1) for Jeddah and Jazan stations in Saudi Arabia. The long shortterm memory (LSTM) boosted regression tree (BRT) and cascaded forward neural network (CFNN) models were combined with VMD to compare along with the standalone versions. The proposed VMD-GPR outperformed the comparing model to forecast daily floods for both stations using a set of performance metrics. The VMD-GPR outperformed comparing models by achieving R = 0.9825, RMSE = 0.0745, MAE = 0.0088, E_{NS} = 0.9651, KGE = 0.9802, IA = 0.9911, U_{95%} = 0.2065 for Jeddah station, and R = 0.9891, RMSE = 0.0945, MAE = 0.0189, E_{NS} = 0.9781, KGE = 0.9849, IA = 0.9945, U_{95%} = 0.2621 for Jazan station. The proposed VMD-GPR method efficiently analyzes flood events to forecast in these two stations to facilitate flood forecasting for disaster mitigation and enable the efficient use of water resources. The VMD-GPR model can help policymakers in strategic planning flood management to undertake mandatory risk mitigation measures.

Received: 25 April 2025 Revised: 28 May 2025

Academic Editor: Paolo Mignosa

Revised: 28 May 2025 Accepted: 31 May 2025 Published: 3 June 2025

Citation: Aldhafiri, A.A.; Ali, M.; Labban, A.H. Leveraging Advanced Data-Driven Approaches to Forecast Daily Floods Based on Rainfall for Proactive Prevention Strategies in Saudi Arabia. *Water* **2025**, *17*, 1699. https://doi.org/10.3390/w17111699

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/license s/by/4.0/). Keywords: flood; rainfall; forecast; VMD; GPR; LSTM; BRT; CFNN

1. Introduction

Rainfall and floods are increasingly severe natural catastrophes aggravated by climate change, urbanization, and environmental degradation. The intensity of heavy rainfall, driven by global climate warming, significantly affects several areas worldwide, causing floods, droughts, water quality degradation, and landslides. Flooding and its frequencies are happening more often around the globe due to climate change and impact the greatest number of populations. Floods account for 39% of all natural disasters with increased economic losses and affect 94 million people annually [1,2]. Saudi Arabia faces serious challenges due to climate change and global warming, resulting in scarce water resources and reserves [3,4]. The Saudi Arabian Kingdom has experienced prolonged periods of drought and flooding due to climate change, which affects the frequency and intensity of extreme weather events such as rainfall [3,5]. Severe floods affected Saudi Arabia between 2008 and 2009, with estimated economic losses of approximately \$ 1300 million USD [3]. Flood forecasting is crucial and significant to manage and prepare for extreme events and scenarios to assess risks of rainfall and floods for proactive prevention strategies.

The flood time-series data exhibit both spatial and temporal structures, and datadriven prediction can be modeled as a sequence problem to exploit spatial and temporal features. Machine learning models can aptly approximate the embedded nonlinear characteristics of predictors within a system without any a priori knowledge or boundary conditions whilst forecasting the response variable. For example, an operational framework was designed using machine learning models for real-time flood forecasting [6], which consists of four subsystems: data validation, stage forecasting, inundation modeling, and alert distribution. Here, stage forecasting was carried out using the long short-term memory (LSTM) and the linear models, whereas flood inundation was calculated with the thresholding and the manifold models. The former determines the inundation extent, and the latter figures out both inundation extent and depth. The LSTM displayed better performance than the linear model, while the thresholding and manifold models achieved similar accuracy for modeling inundation extent. The data includes the stream gauge measurements of the water stage and satellite-derived precipitation utilized in this work.

Another study was conducted in Thailand to forecast flooding with machine learning methods such as linear regression, neural network regression, Bayesian linear regression, and boosted decision tree regression [7] with the MIKE11 (i.e., MIKE-NAM) model. The hybrid model based on MIKE11 and a machine learning technique produced a better forecast as compared to the single model MIKE11. The Bayesian linear regression showed improved performance as compared to other methods for runoff forecasting and flood water levels. Motta de Castro Neto [8] proposed a mixed approach for urban hourly flood prediction by integrating random forest and GIS for management and resilience planning, where a risk index was computed using scores and hot spot analysis. The proposed approach was helpful in using sensible factors and risk indices for the occurrence of floods at the city level, which could be influential in outlining a long-term strategy for smart cities. Random Forest with a Matthew's Correlation Coefficient of 0.77 and an Accuracy of 0.96 appeared to be the best. The GIS model was adopted to locate areas with a higher likelihood of being flooded under critical weather conditions. Finally, the predictions obtained from the random forest model and the hot spot analysis were then combined to create a flood risk index. The data used in this work comprise local weather measurements and fire department emergency records to classify flood occurrences.

Recently, Rajab, Farman [9] utilized historic climatic records of Bangladesh to forecast flood with polynomial regression, random forest regression, multiple linear regression, decision tree, k-nearest neighbor, support vector machines, AdaBoost Regressor, Stacking Regressor, artificial neural network, recurrent neural networks and long shortterm memory (LSTM). The polynomial regression, random forest, and LSTM provided the highest performance accuracy. The data contains maximum temperature, minimum temperature, rainfall, relative humidity, wind speed, cloud cover, and brilliant sunshine, the weather station numbers, latitude, longitude, and altitude to forecast rainfall, flood water levels, and velocities. Shafizadeh-Moghadam, Valavi [10] designed a novel forecasting approach using a combination of machine learning and statistical models for flood susceptibility mapping in Iran. Several models' artificial neural networks, classification and regression trees, flexible discriminant analysis, generalized linear model, generalized additive model, boosted regression trees (BRT), multivariate adaptive regression splines, and maximum entropy were implemented to construct the ensembles to reduce the uncertainty. BRT was the most accurate individual model, while Median ensemble forecasting was the most accurate in the group of ensemble models. Here, the data employed in this research include slope degree, curvature, elevation, topographic wetness index (TWI), stream power index (SPI), distance to river, river density, land use, Normalized Difference Vegetation Index (NDVI), rainfall, and lithology to predict flood susceptibility. Flood prediction based on weather parameters using a deep learning model compared with support vector machine (SVM), K-nearest neighbor (KNN), and Naïve Bayes were conducted [11] in Bihar and Orissa, India. The results showed that the deep neural network can be proficiently used for flood forecasting by achieving the highest accuracy based on monsoon parameters only before flood occurrence. The datasets include precipitation maximum and minimum temperatures for different regions to predict floods. The flood makes it challenging for the models to fully capture the embedded features and patterns due to its uncertain and complex stochastic nature. Moreover, the "black box" nature of the machine learning and deep learning models limits their ability to provide reasonable explanations and interpretations during the learning process.

The efficiency of the model in terms of precision is significantly impacted by failing to capture all the features and information properly. To handle this problem, a multiresolution analysis is useful to uncover and present the embedded features within the timeseries data. To abstract the core and fundamental sub-frequencies, sequential decomposition is more suitable and advantageous [12]. Literature shows that Fourier spectra analysis [13], discrete wavelet transformation [14–20], Empirical Mode Decomposition (EMD) [21], Ensemble EMD (EEMD) [22], complete ensemble EMD with adaptive noise (CEEMDAN) [23] and improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) [24] are the frequently adopted methods. However, these methods suffer from the major issues of decomposing an individual input at a time [25,26] and require a sequential decomposition [12] to uncover and present the embedded features within the time-series data in terms of core and fundamental sub-frequencies. To overcome this problem, the variational mode decomposition (VMD) method is applied as an adaptive and non-recursive decomposition tool [13]. VMD has the capacity to simultaneously decompose the time-series signal and instantaneously resolve the embedded subfrequency components (i.e., IMFs) in input predictors without loss of any information. It is a signal processing technique applied to decompose a complex signal (i.e., time series data) into a set of band-limited modes, namely intrinsic mode functions (IMFs). It iteratively solves a constrained variational problem and aims to find the optimal solution that minimizes the bandwidth of each mode while ensuring they collectively reproduce the original signal.

Up until now, a VMD-based hybrid decomposition model to forecast flood scenarios has never been proposed in Saudi Arabia. The novelty and contribution of this work centered around the implementation of the signal processing variational mode decomposition (VMD) method for flood forecasting. By utilizing a proper multiresolution analyzing scheme, the VMD can simultaneously extract the deep "important" forecast information from several non-stationary and nonlinear historical data. Another innovation of the current work lies in the development of Gaussian process regression (GPR) for flood forecasting. This study extends the development and application of the GPR model, which is a non-parametric probabilistic method for regression problems and offers a more flexible framework to model flood scenarios and provides uncertainty estimates. This study aims to assess the VMD method unified into forecasting models that include GPR, long shortterm memory (LSTM), boosted regression tree (BRT), and cascaded forward neural network (CFNN). Incorporating VMD with GPR led to the development of the VMD-GPR model. The newly established VMD-GPR model is compared against the hybrid versions VMD-LSTM, VMD-BRT, VMD-CFNN, and standalone GPR, LSTM, BRT, and CFNN models. The proposed VMD-GPR, along with the benchmarking models, were developed using the data from Jeddah and Jazan stations in Saudi Arabia. The next section of the paper outlines the material and methods comprising the study area and data description, followed by model performance evaluation and model development. Then, the next sections present results, further discussions, and conclusions.

Background

Moreover, the VMD is entirely data-dependent and involves minimal external participation during the MRA process [27]. Additionally, the VMD has several other advantages, including improved noise robustness, better mode separation, and reduced mode aliasing and end effects, making it a more reliable and versatile signal decomposition technique. Fu, Li [28] integrated the VMD method with support vector machine (SVM) and grey wolf optimizer (GWO) to create VMD-SVM-GWO for predicting monthly evapotranspiration in the Tengger Desert, China. The approach consisted of three sections, including data pre-processing, parameter optimization, and estimation. The results demonstrated that the VMD-SVM-GWO method achieved superior computational performance compared to the SVM and hybrid versions of discrete wavelet transform (DWT) and ensemble empirical mode decomposition (EEMD) using only historical evapotranspiration (ET) data to predict ET. Wang, Liu [29] applied wavelet transformation (WT), VMD, and back propagation neural network (BP) optimized by differential evolution (DE) to construct a WT-VMD-DE-BP model for day ahead PM2.5 concentration forecasting. This approach first uses WT to decompose the PM2.5 data into frequency subsets, then applies VMD to further decompose these subsets into variational modes. The DE-BP model is then used to forecast each VMD-based mode, and the aggregated results are used to predict the original PM2.5 concentrations in Wuhan and Tianjin, China. The proposed WT-VMD-DE-BP model shows superiority in forecasting PM2.5 concentrations against the benchmarking BP, DE-BP, WT-DE-BP, VMD-DE-BP, and WT-VMD-DE-BP models.

A new decomposition-optimization-based VMD, Backtracking Search Algorithm (BSA), and Regularized Extreme Learning Machine (RELM) were designed to forecast short-term wind speed [30]. Here, the wind speed time-series data is decomposed using VMD into several modes. The BSA optimization method is employed to search the optimal parameters of the RELM to forecast multi-step wind speed in Sotavento Galicia, Spain. The results reveal that the proposed model achieves notably better performance than its rivals (i.e., ARIMA, RBF, GRNN, RELM, VMD-RELM, and BSA-RELM) both on single- and multi-step forecasting with at least 50% average improvement. Variational mode decomposition-based low-rank robust kernel extreme learning machine (RMWK) was created by [31] for solar irradiation forecasting. In this research, the VMD and EMD methods are used to decompose the solar irradiation time series data. Moreover, to reduce the computational overhead, the size of the kernel matrix is lowered, and a variable weighting factor is used for each error residual to achieve robustness. A comparative study between VMD-RMWK and EMD-RMWK for 15 min ahead prediction in different weather conditions shows clearly that the hybrid VMD-based robust Morlet Wavelet Kernel outperforms the EMD-based Morlet Wavelet Kernel.

2. Materials and Methods

2.1. Study Area and Data Description

This research study was conducted in Saudi Arabia, where the data for the selected stations, Jeddah and Jazan, was acquired from the Department of Meteorology. The required data includes the rainfall on a daily basis from 1 January 1978 to 30 December 2013. Jeddah city is stationed on the western shoreline of Saudi Arabia with extremely hot desert and scorched summers, frequently suffering over 40 °C of temperature, and mild winters with 15 °C due to its coastal position. The city of Jeddah is experiencing extreme humidity levels. Jeddah is one of the flash flood-prone cities influenced by heavy rainfall incidents, mainly driven by rivers and small basins. This city hosts a major port and industrial zone and links the country to the international markets.

On the other hand, the city of Jazan is in the southwest region of Saudi Arabia and features a tropical climate. Temperatures in Jazan remain consistently high throughout the year, accompanied by moderate humidity levels. Summers in Jazan City are hot and humid, where the temperatures repeatedly rise to 40 °C during the summer seasons. The monsoon continues from June through September and brings substantial rainfall, causing sudden heavy flooding in recent times. The city is diverse, comprising diverse coastal plains, mountains, and valleys with sandy beaches and the Red Sea. Due to its geographic and diverse position, Jazan is a strategic trading center for the Arabian Peninsula. Figure 1 demonstrates the map of the Jeddah and Jazan stations.



Figure 1. Geographic position of the stations in Saudi Arabia.

The acquired dataset includes rainfall at daily time intervals from 1 January 1978 to 30 December 2013. The specific descriptive summary including longitude ($^{\circ}E$) = 42.551 $^{\circ}E$, 39.237° E, latitude (°S) = 16.889° N, 21.285° N, elevation = 12 and 40 m, minimum rainfall = 0, maximum rainfall = 777.7, mean = 0.27, 0.58, standard (Std.) deviation = 9.82, 13.90, skewness = 75.8, 53.3, and kurtosis = 5934.0, 2962.8 of the data for both Jeddah and Jazan stations, respectively. The maximum rainfall in both stations is approximately 777.7, whereas the skewness is positive, indicating that the right tail of the distribution is longer or heavier than the left tail, suggesting more extremely high values. The kurtosis of the rainfall data represents Leptokurtic distribution with a high value, implying heavier tails and the shape of a probability distribution by quantifying its tailedness in terms of thin or broad. The obtained rainfall datasets have a few missing values. To overcome this issue, these missing values were replaced with the average record of the previous and the next data point. The flood index [32] was computed using the daily rainfall data and based on the onset and severity of the current and antecedent day's rainfall. Figure 2 shows the time series graph of the daily flood trends of Jeddah and Jazan stations. There were extreme flood events during 2013 in both Jeddah and Jazan stations because the FI values were greater than 2.0.



Figure 2. Time series of daily flood trends in Jeddah and Jazan stations.

2.2. Variational Mode Decomposition (VMD)

The VMD is a non-recursive signal processing method designed by [27] to decompose non-stationary signals adaptively in the form of discrete bandwidth-limited modes called IMFs. Using the Wiener filtering groups, the VMD employs several filters. Each mode u_k (k = 1, 2, ..., k) is compacted to a central pulsation ω_k determined in the decomposition process, and the bandwidth of a modal function is computed by the following steps. Firstly, to obtain a unilateral frequency spectrum, the Hilbert transform computes the related analytic signal of each mode u_k . Then, by joining the model's frequency spectrum with an exponential, modified to its own evaluated center frequency, the spectrum was received into the "baseband". Then, the bandwidth modal function is projected, and

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k} k \left\| \delta_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}$$

$$s.t. \sum_{k} u_k(t) = f(t)$$
(1)

where the function f(t) indicates the *t*-th data of the predictor variable, $\{u_k 2 \coloneqq \{u_1, \dots, u_k\}$ and $\{\omega_k\} \coloneqq \{\omega_1, \dots, \omega_k\}$ denote the set of modes and corresponding center frequencies. The term $\delta(t)$ is the unit pulse function; *j* is an imaginary unit; δ_t is the partial derivative of a function with respect to time and * represents convolution operation.

The quadratic penalty function α and the Lagrange multiplier λ are merged to resolve the constraint with the augmented Lagrangian structure as follows:

$$L = (\{u_k\}, \{\omega_k\}, \{\lambda\}) = \alpha \sum_k \left\| \delta_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 +$$

$$\left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle$$
(2)

The modal factors and the central frequencies are drawn by the alternating direction method of multipliers (ADMM) iterative suboptimization sequence:

$$\hat{\lambda}_{(\omega)}^{n+1} = \hat{\lambda}^n(\omega) + \tau(\hat{f}(\omega) - \sum_k \hat{u}_k^{n+1}(\omega))$$
(3)

$$\hat{u}_{k}^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i=k} \hat{u}_{i}(\omega) + \frac{\lambda(\omega)}{2}}{1 + 2\alpha(\omega - \omega_{k})^{2}}$$
(4)

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \, |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega} \tag{5}$$

where τ is the updated parameters of Lagrange multipliers; $\hat{u}_{k}^{n+1}, \hat{f}_{(\omega)}, \hat{u}_{(\omega)}, \hat{\lambda}_{(\omega)}$ indicate the Fourier transforms of $u_{k}^{n+1}(t), f(t), u_{k}^{n}(t), and \lambda^{n}(t)$.

2.3. Gaussian Process Regression (GPR)

GPR is a non-parametric Bayesian method founded on a compelling theoretical background, handling the nonlinear regression issues even for small data [33–35], and has been successfully tested over the past years for real-world applications of supervised learning. The GPR method precisely describes prior probability distributions over latent functions and is completely explained by mean and covariance [36]. The model output in terms of posterior GP is determined with a mixed Gaussian likelihood. Suppose D(X, Y) indicates the *n* observations of the training data.

$$X = [x_1^T, x_2^T, \cdots, x_n^T]^T$$
(6)

$$Y = [y_1, y_2, \cdots, y_n]^T \tag{7}$$

The observation model is expressed as

$$y = f(\mathbf{x}) + \varepsilon_n, \varepsilon_n \sim N(0, \sigma_n^2)$$
(8)

where f(x) denotes the latent variable and σ_n^2 is the noise variance. The distribution of f[f = f(X)] can obtained by:

$$f = N(0, C(X, X)) \tag{9}$$

Consider f_* to be the predicted value of an unobserved point x_* . GP prior specifies that the combination of f and f_* follows a joint Gaussian as

$$p(f, f_*) = N\left(\begin{bmatrix} 0\\0 \end{bmatrix}, \begin{bmatrix} C[34,35 & C^*]\\C_*^T & \tilde{C} \end{bmatrix}\right)$$
(10)

where

$$_{*} = \mathcal{C}(x_{*}, X) \tag{11}$$

$$= \mathcal{C}(x_*, x_*) \tag{12}$$

Bayesian inference expresses the posterior distribution of the goal prediction f_* conditioned on the training set as

С

Ĉ

$$p(f_* \mid Y) = N(\mu_{f_*}, \sigma_{f_*}^2)$$
(13)

The mean and standard deviation provided by

$$\mu_{f_*} = C_* K^{-1} Y \tag{14}$$

$$\sigma_{f_*}^2 = \tilde{C} - C_*^T K^{-1} C_*$$
(15)

where I_n is the identity matrix of $n \times n$ and $K = C + \sigma_n^2 I_n$.

The covariance function is vital for the GPR's modeling process. Based on interpretation and adaptability, the squared exponential (SE) function is chosen to define the GPR:

$$C(x, x') = \eta^2 exp\left[-\frac{1}{2}\sum_{k=1}^d \left(\frac{x_k - x_k'}{l_k}\right)^2\right]$$
(16)

where *l* is the length scale and η^2 stands for the signal variance. By maximizing the likelihood function the hyperparameters $\theta = \{\eta, l\}$ are calculated:

$$\mathcal{L}(\Theta) = \frac{1}{2} Y^T K^{-1} Y + \frac{1}{2} \log |K| + \frac{n}{2} \log (2\pi)$$
(17)

2.4. Long Short-Term Memory (LSTM)

LSTM is a modified version of the Recurrent Neural Network (RNN) [37], equipped with a separate storage unit and a mechanism to control the data stream within the network [37]. LSTM reaches the optimal solution by optimizing the error function using gate cells and granting neurons to interact with each other [38]. The LSTM model is designed to capture nonlinear trends in time series data and learn from previous information over a long time. The standard structure of the LSTM contains three gates: forget, input, and output [38]. The LSTM learns from the information stored over a long-term period to resolve the vanishing gradient issue of RNN. The first layer of the memory gate is in control of removing redundant information from the cell state, and it is defined as follows:

$$f_t = \sigma \Big(w_f \times X_t + Y_f \times h_{t-1} + p_f \Big)$$
(18)

where f_t is the forgetting threshold at time t, w_f and Y_f are weights, σ is the sigmoid activation function, h_{t-1} is the output value at time t, X_t is the input value, and p_f is the bias term. The second input gate defines the information that should be placed in the cell state from the current input set [39]. This comprises decision i_t which altered *tanh* layer and the value to create a new state value C_t . It is expressed as

$$i_t = \sigma(w_i \times X_i + Y_i \times h_{t-1} + p_i) \tag{19}$$

$$C_t = \sigma(w_c \times X_c + Y_c \times h_{t-1} + p_c) \tag{20}$$

where i_t is the input threshold at time t, w_i , w_c , Y_i , Y denote to the weights, p_i , p_c are biased terms. The following expression is used to modify the state of the cell at time t:

$$C_t = f_t \times C_t + i_t \times p_o \tag{21}$$

The third layer is defined as output data in the current time step, and it is expressed as

$$O_t = \sigma(w_0 \times X_c + Y_o \times h_{t-1} + p_o) \tag{22}$$

where O_t refers to the output threshold at time *t*, then the output value of the cell is defined as

$$h_t = O_t \times \tanh(C_t) \tag{23}$$

where tanh represents the activation function, and h_t refers to the output value of the cell at time *t*. The data is passed through all three gates, so the important information is output, and the invalid information is eliminated.

2.5. Boosted Regression Tree (BRT)

The BRT is another non-parametric regression model that integrates boosting and regression trees [40] and avoids prior information between the input and target [41] to improve the efficiency of several individual models [42]. The BRT model principally works on (*a*) CART regression tree, (*b*) the formation and integration of a series of models via the boosting technique. The BRT resolves the obstacle of the single decision trees to produce the primary tree using training, and the remaining data is employed to build the subsequent trees [43]. Boosting methods are devoted to improving the regression tree's prediction capability. It looks like the model averaging process, except for the boosting operation in a step-by-step manner to fit the models to a subset of the training set [44]. The effectiveness of the BRT is greatly dependent on two regularization parameters: (i) the number of additive terms or tress (n_t) and (ii) the learning rate (LR). The LR parameter is employed to decrease the impact of every single tree in the model, which ranges from 0.1–0.0001. The smaller LR value indicates a reduced loss function, but this needs the presence of additional tress (n_t) to the model [45]. The BRT involves the ability to quickly evaluate the larger dataset and is less susceptible to overfitting [46].

2.6. Cascaded Forward Neural Network (CFNN)

The CFNN model is introduced by [47], which is a variant of a traditional artificial neural network (ANN) and engages a parallel information processing system containing input, hidden, and output neuron layers. The structure of CFNN resembles FFNN, except the input data is attached to each concealed layer behind it through a weight matrix. The difference rests in the neurons of their hidden layer. A new hidden neuron is included in these networks at each successive stage. Each new neuron captures information from the input neurons and all previously hidden neurons exiting beforehand to the input of each output neuron where the input and output neurons are connected. Apart from the first hidden layer, all hidden layers in CFNN surround at least two weight matrices to regulate the output signal of the top layer and the input signal of the network. This scheme can supply more degrees of freedom to the training for increasing the network's nonlinear mapping facility. The weight and bias matrix optimization are carried out through the BP algorithm during training. Its intention is to produce the actual output of the network as near to the predicted output as possible, as measured by the mean squared error. First, the no. of hidden layers and neurons must be specified for typical neural networks, and therefore, reliable detection of optimal design is normally difficult and normally involves trial and error [48]. In the opening step, cascade networks are trained using input and output neurons. Training will finish if the error is acceptable after a predetermined number of iterations. If not, the model will be re-run at each stage by adding a new neuron and appropriately training the network to reduce residual error [47]. This process will last until the error rate falls below the target threshold [49]. The following is the central equation for CFNN:

$$Out_{CFNN}(k) = f_{act}\left(\sum_{k=0}^{N} \left[HN(j) \times W_j(j,k) + I(i) \times W_i(i,k)\right] + b(k)\right)$$
(24)

where $Out_{CFNN}(k)$ is the output neuron, $W_j(j, k)$ and $W_i(i, k)$ are the vectors of weights, b(k) the bias weight, f_{act} is the activation function, I(i) is the input value, and HN(j) is the hidden neuron.

2.7. Model Performance Evaluation Criteria

Evaluating model performance is fundamental during the construction stage, which implies comparing the forecasts of the models with their actual values employing statistical metrics to examine how well the proposed model simulates the actual output. This research utilizes the following metrics to assess the model efficiency in terms of accuracy along with the comparing models: r (Correlation Coefficient), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), IA (Willmott's Index of agreement) [50], ENS (Nash-Scuttle estimator) [51], KGE (Kling-Gupta efficiency) [52], and uncertainty coefficient with 95% confidence level (U95%). The mathematical formulation of these metrics can be expressed in the following equations:

$$r = \frac{\sum_{i=1}^{N} (FI_ob_{o,i} - \overline{FI_ob_o}) (FI_fc_{o,i} - \overline{FI_fc_o})}{\sqrt{\sum_{i=1}^{N} (FI_ob_{o,i} - \overline{FI_ob_{o,i}})^2 \sum_{i=1}^{N} (FI_fc_{o,i} - \overline{FI_o})^2}}$$
(25)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (FI_{ob_{o,i}} - FI_{fc_{o,i}})^{2}}$$
(26)

$$IA = 1 - \frac{\sum_{i=1}^{N} (FI_{-}ob_{o,i} - FI_{-}fc_{o,i})^{2}}{\sum_{i=1}^{N} (|FI_{-}ob_{o,i} - FI_{-}ob_{o,i}| + |FI_{-}ob_{o,i} - FI_{-}ob_{o,i}|)^{2}}$$
(27)

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |FI_{ob_{o,i}} - FI_{fc_{o,i}}|$$
(28)

$$E_{NS} = 1 - \frac{\sum_{i=1}^{N} (FI_{o}b_{o,i} - FI_{-}fc_{o,i})^{2}}{\sum_{i=1}^{N} (FI_{-}ob_{o,i} - FI_{-}ob_{o,i})^{2}}$$
(29)

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$$
(30)

$$U_{95\%} = 1.96\sqrt{Standard\ deviation^2 - RMSE^2} \tag{31}$$

where $FI_fc_{o,i}$ is the forecasted value of the flood index and $FI_ob_{o,i}$ is the actual value. $\overline{FI_fc_{o,i}}$ is the average of the forecasted results while $\overline{FI_ob_{o,i}}$ is the average of actual values. The term *N* indicates the total number of collected data points, and α displays the relative variability of the forecasted and actual values, whereas β is the ratio between the forecasted and observed mean values. The correlation coefficient r measure of the strength and direction of a linear relationship between predicted and observed data, ranging between -1 and 1, with 1 indicating a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation. A lower RMSE indicates a better model, with values closer to 0 signifying more accurate predictions, whereas a lower MAE shows better model performance, with a value of zero representing no error. The IA demonstrates the differences between forecasted and computed means and variances, which reflect sensitivity to outliers in the observation data and insensitivity to additional and proportional variances between expected and calculated values. The value of IA ranges from 0 to + 1, with + 1 being the ideal value. The E_{NS} is used to compare model performance (range from $-\infty$ to +1), and its best value is 1. Regarding this metric, the performance of the model is scored as follows: great (E_{NS} > 0.75), good (0.65 < E_{NS} < 0.75), satisfactory (0.50 < E_{NS} < 0.65), acceptable (0.40 < E_{NS} < 0.50), and inadequate (E_{NS} < 0.4). KGE assesses how well a model's simulations align with observed data, considering correlation, variability, and bias. It ranges from — infinity to 1, with 1 representing perfect agreement. KGE values greater than —0.41 indicate the model performs better than using the mean as a benchmark, while the U_{95%} uncertainty coefficient tells that if an experiment or study is repeated many times, U_{95%} of the resulting confidence intervals would contain the true predictions.

2.8. Model Development

In this study, several new hybrid VMD-GPR, VMD-LSTM, VMD-BRT, and VMD-CFNN models were designed to forecast the daily flood index at (t – 1) for Jeddah and Jazan stations in Saudi Arabia. Additionally, the standalone versions of the hybrid models GPR, LSTM, BRT, and CFNN were also constructed to compare the performance of each model. All the models were developed and executed in the MATLAB R2023a environment using an Intel Core i5-8400, 2.80 GHz CPU series, and 8 GB RAM. Each step and corresponding details during the model development are explained below:

Step 1: Decomposition via VMD method

In this primary pre-processing step, the VMD method starts operating to concurrently decompose the daily flood index data of both stations to obtain IMFs and residual factors (i.e., signals/modes). The optimal number of modes (i.e., K) was selected to be 6 for both Jeddah and Jazan stations, and this was obtained using the trial-and-error process. The setting of parameters of the VMD method are $\alpha = 2000$, $\tau = 0$, DC = 3, init = 1, tol = 1e - 7, k = 6 for both stations.

Step 2: Determination of lagged-time components

The statistically significant lags were determined using the PACF to the decomposed IMFs (in step 1) to examine the relation between the IMFs at t and (t - 1) for both Jeddah and Jazan stations, as shown in Figure 3. The time-lagged input variables at (t - 1) were used for forecasting the daily flood because these lags at (t - 1) show a high relationship with the target variable (i.e., flood). PACF determines the correlation between observations at two time points while accounting for the influence of all shorter lags. It helps isolate the direct relationship between observations at different time lags, removing the indirect influence of intervening observations.

Step 3: Preparations of models feeding

The statistically significant decomposed IMFs were then supplied directly into the GPR model to create the novel VMD-GPR model. Here, the rate of partitioning of data is performed into 70% for training and the remaining 30% for testing purposes independently [53]. As this is time series data that has a time dependency, it is important to preserve the order to avoid the temporal disorder and cause data leakage, e.g., unintentionally inferring the trend of future samples. Therefore, any randomized strategy cannot be applied here because observations from the future must not be seen by the model. Here, the data sequence was kept in order and used the most recent records as our validation/test set and the earlier observations as the training set.



Figure 3. Partial auto-correlation function (PACF) of the corresponding IMFs for (**a**) Jeddah station and (**b**) Jazan station.

The normalization and denormalization processes of the training and testing data were also carried out within the unit interval to speed up the convergence of the models. As the core objective of this research work is to discover how the pre-processing technique influences precision during daily flood forecasting, it is critical to draw a comparison of the novel VMD-GPR model with other state-of-the-art models. Therefore, the VMD method was integrated with LSTM, BRT, and CFNN models to construct VMD-LSTM, VMD-BRT, and VMD-CFNN models. The results of the VMD-GPR model were compared with the hybrid versions as well as with the counterpart standalone models, GPR, LSTM, BRT, and CFNN. Figure 4 exhibits the topological structure of the proposed VMD-GPR model along with the comparing models.

Step 4: Tuning and hyperparameter setting

One of the most fundamental stages in constructing data-driven models is the modification and adjustment of the hyperparameters called tuning, which significantly impacts the accuracy. Numerous approaches can be adopted to acquire the optimal hyperparameters, and this research uses the traditional trial and error procedure. The optimum hyperparameters here were determined using the RMSE as the convergence criterion in the MATLAB environment. Details on the hyperparameters are summarized in Table 1, where the GPR model has log-likelihood, basis function, Kernel function, Beta, iteration, etc. are important. For the LSTM model, the main hyperparameters are hidden units, optimizer, verbose, gradient threshold, batch size, and epochs. The BRT model possesses the learn rate value and ensemble strategy (LSBoost) as the significant ones, whereas CFNN model has the neuron number in the hidden layer and the best training algorithm are the important hyperparameters.

Stations	Models	Tuning Parameters				
Jeddah Station	GPR	 Hybrid and Standalone Structure Log Likelihood = 4.509; Basis Function = Linear Kernel Function = Squared Exponential; Beta = 0 Active Set Size = 2000; Max. Iteration = 10,000 Ontimizer = Quasi power of Varbase = 0 				
	LSTM	 Optimizer - Quasi newton, verbose - 0 Hidden units = 10; Optimizer = Adam, Verbose = 0 Gradient Threshold = 1, Initial Learn Rate = 0.005 Learn Rate Drop period = 200; Batch Size = 32 Learn Rate Drop Factor = 0.1; Epochs = 250 				
	BRT	 Learn rate = 0.194, Method = LSBoost, N Learn= 100, Combine Weights = Weighted Sum Learner name = Tree 				
	CFNN	 Hybrid Structure: 6-9-1; Standalone Structure: 1-9-1 Epoch = 18 iterations, Validation checks = 6 Mu = 0.001, Training = Levenberg-Marquadt 				
Jazan Station	GPR	 Hybrid and Standalone Structure Log Likelihood = 4.509; Basis Function = Linear Kernel Function = Squared Exponential; Beta = 0 Active Set Size = 2000; Max. Iteration = 10,000 Optimizer = Quasi newton; Verbose = 0 				
	LSTM	 Hidden units = 10; Optimizer = Adam, Verbose = 0 Gradient Threshold = 1, Initial Learn Rate = 0.005 Learn Rate Drop period = 200; Batch Size = 32 Learn Rate Drop Factor = 0.1; Epochs = 250 				
	BRT	 Learn rate = 0.194, Method = LSBoost, N Learn= 100, Combine Weights = Weighted Sum Learner name = Tree 				
	CFNN	 Hybrid Structure: 6-9-1; Standalone Structure: 1-9-1 Epoch = 45 iterations, Validation checks = 6 Mu = 0.001, Training = Levenberg-Marquadt 				

Table 1. Parameter setting of the models forecasting flood index.

The accuracy of the VMD-GPR models in training and testing periods based on MSE, RMSE, MARE, and time-series trend plots is presented in Figure 5. The VMD-GPR model attained consistent and stable performance accuracy over both training and testing data for the Jeddah station to forecast daily floods. Similarly, the VMD-GPR was also better in terms of these metrics for the Jazan station against the counterpart comparing models in Figure 5 for the Jeddah station.



Figure 4. Schematic view of the proposed modelling framework.



Figure 5. Training and testing performance of the VMD-GPR model in terms of MSE, RMSE, MARE, and time-series trend plot for Jeddah station.

3. Results and Discussion

The proposed hybrid VMD-GPR approach is assessed and benchmarked with VMD-LSTM, VMD-BRT, VMD-CFNN, GRP, LSTM, BRT, and CFNN models using the evaluation metrics r, RMSE, MAE, ENS, KGE, IA, and U95% and analytical diagnostic plots for daily flood index forecasting.

Table 2 shows the performance of the models using r, RMSE, and MAE to predict the daily flood index for Jeddah and Jazan stations. The newly designed VMD-GPR model is superior in terms of efficiency as compared to VMD-LSTM, VMD-BRT, VMD-CFNN, GRP, LSTM, BRT, and CFNN models by achieving the highest value of r and lowest RMSE and MAE errors as shown in Table 2. For the Jeddah station, the recorded metrics of VMD-GPR model are (r = 0.9825, RMSE = 0.0745, MAE = 0.0088), followed by VMD-CFNN mode with (r = 0.9788, RMSE = 0.0870, MAE = 0.0109). From the group of standalone models, the GPR appeared to be the best in terms of (r = 0.9678, RMSE = 0.1006, MAE = 0.0051), followed by CFNN (r = 0.9674, 0.1012, 0.0062), LSTM and BRT models. Overall, it can be seen in Table 2 that the VMD-GPR is the best model in both hybrid and standalone groups of the models to forecast the daily flood index in Jeddah station.

	Jeddah Station			Jazan Station			
	r	RMSE	MAE	r	RMSE	MAE	
GPR	0.9678	0.1006	0.0051	0.9834	0.1227	0.0179	
VMD-GPR	0.9825	0.0745	0.0088	0.9891	0.0945	0.0189	
LSTM	0.9603	0.1902	0.0588	0.9809	0.3126	0.1140	
VMD-LSTM	0.9348	0.2213	0.0655	0.9802	0.2895	0.1098	
BRT	0.8661	0.2316	0.0330	0.8164	0.5311	0.1416	
VMD-BRT	0.8485	0.2787	0.0700	0.7943	0.5625	0.1636	
CFNN	0.9674	0.1012	0.0062	0.9827	0.1205	0.0156	
VMD-CFNN	0.9788	0.0870	0.0109	0.9726	0.1349	0.0187	

Table 2. Testing performance of the VMD-GPR vs. VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models using *r*, *RMSE*, and *MAE*.

For Jazan station, the VMD-GPR model again received the highest accuracy by gaining (r = 0.9891, RMSE = 0.0945, MAE = 0.0189) as compared to VMD-LSTM, VMD-BRT, and VMD-CFNN models to forecast daily flood index. Similarly, the VMD-GPR model is also outperformed against the standalone GPR, LSTM, BRT, and CFNN models in Jazan station. Thus, Table 3 suggests that the VMD-GPR model is better for forecasting the daily flood index for both stations as compared to the VMD-LSTM, VMD-BRT, VMD-CFNN, GRP, LSTM, BRT, and CFNN models. Table 3 also confirms that the VMD notably enhanced the accuracy of the models, particularly when combined with the GPR model. Overall, the VMD-GPR model is showing better precision in forecasting daily floods in Jeddah and Jazan stations.

Table 3. The performance of VMD-GPR, VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models based on assessment metrics E_{NS}, KGE, IA, and U_{95%}. Note that the best model is boldfaced (blue).

Jeddah Station				Jazan Station				
	Ens	KGE	IA	U95%	Ens	KGE	IA	U95%
GPR	0.9364	0.9646	0.9836	0.2790	0.9631	0.9003	0.9899	0.3390
VMD-GPR	0.9651	0.9802	0.9911	0.2065	0.9781	0.9849	0.9945	0.2621
LSTM	0.7731	0.5690	0.9111	0.5146	0.7608	0.3483	0.9034	0.8372
VMD-LSTM	0.6927	0.5034	0.8665	0.6043	0.7948	0.3995	0.9208	0.7753
BRT	0.6636	0.5529	0.8573	0.6393	0.3098	0.0339	0.5521	1.4471
VMD-BRT	0.5129	0.3757	0.7420	0.7677	0.2257	-0.0480	0.4559	1.5308
CFNN	0.9357	0.9598	0.9833	0.2806	0.9644	0.9285	0.9905	0.3332
VMD-CFNN	0.9525	0.9019	0.9867	0.2409	0.9455	0.9690	0.9861	0.3740

The VMD-GPR and comparing models' performance was further assessed and investigated in Table 3 based on ENS, KGE, IA, and U95% for both stations. The highest accuracy generated by the VMD-GPR model (ENS = 0.9651, KGE = 0.9802, IA = 0.9911, U95% = 0.2065) for Jeddah station, and (ENS = 0.9781, KGE = 0.9849, IA = 0.9945, U95% = 0.2621) against the comparing VMD-LSTM, VMD-BRT, VMD-CFNN, GRP, LSTM, BRT, and CFNN models. Table 3 again endorsed that the VMD-GPR is an excellent model for forecasting daily floods as compared to the counterpart models. Based on ENS, KGE, IA, and U95% metrics, it is evident that the VMD-GPR model has better analytical capabilities to achieve precise forecasts for both Jeddah and Jazan stations. This is also correct by the declaration that the models recorded an ENS \leq 0.800 are determined to be 'unsatisfactory', while ENS between 0.800–0.900 portrays the models are 'fairly good' and ENS \geq 0.900 are viewed to be 'very satisfactory' [54]. Therefore, the newly established VMD-GPR model can be ranked as 'very satisfactory' for forecasting daily floods for both Jeddah and Jazan stations.

Figure 6 demonstrates the scatter diagram between the daily forecasted and observed flood index of the proposed VMD-GPR model vs. VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models for both stations. In Jeddah station, the forecasted and observed flood index generated by VMD-GPR appeared to be the highest precise model with ($r^2 = 0.977$) as compared to VMD-CFNN ($r^2 = 0.972$), VMD-LSTM ($r^2 = 0.942$), and VMD-BRT ($r^2 = 0.787$) models. On the other hand, the standalone BRT model acquired good accuracy for Jeddah station in terms of $r^2 = 0.970$, followed by GPR, CFNN, and LSTM models. Similarly, for Jazan station, the VMD-GPR model again recorded better accuracy ($r^2 = 0.991$), with VMD-CFNN in 2nd place, VMD-LSTM in 3rd position, and VMD-BRT in 4th position for forecasting daily flood index at (t – 1) as compared to standalone models. The outcomes in Figure 6 are also supported by Tables 2 and 3, which prove that the VMD-GPR model is better for forecasting the daily flood index for both stations at (t – 1).

The Swarm plots in Figure 7 display the distribution of the absolute forecasted errors |FE| produced by the VMD-GPR vs. VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models during the daily flood forecast at both stations. Figure 7 yielded that the VMD-GPR model expresses lower |FE| errors between the observed and forecasted flood index for both Jeddah and Jazan stations as compared to VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models. Therefore, the Swarm plots approved the forecasting accuracy was higher for the VMD-GPR model against comparing models.

Figure 8 characterizes the empirical cumulative distribution function (ECDF) of forecasted and observed daily flood index to draw a clear illustration of the model's efficiency. For both Jeddah and Jazan stations, the ECDF of VMD-GPR demonstrated a very close profile with the observed ECDF against hybrid version VMD-LSTM, VMD-BRT, VMD-CFNN, and standalone GPR, LSTM, BRT, and CFNN models. Overall, the ECDF in Figure 8 further confirms the accuracy of the VMD-GPR model in forecasting the daily flood index of Jeddah and Jazan stations.

A Taylor diagram specifies an organized, efficient, and thorough exam of the models' forecasting capability on a wider scale [55]. Figure 9 depicts a more finite and compelling connection between the forecasted and observed flood index grounded on correlation coefficient and standard deviation. It is amazingly established that the VMD-GPR model was sitting close to the observed daily flood index, indicating the forecasting efficiency was remarkably better for both Jeddah and Jazan stations. On the other hand, for the Jeddah station, the VMD-BRT, VMD-LSTM, and BRT models were separated far away while VMD-BRT, VMD-CFNN, and BRT in the Jazan station were positioned in a distant corner proving poor performance. Thus, Figure 9 settles that the VMD-GPR model is realistically accurate in forecasting the daily flood index for both stations as compared to the VMD-LSTM, VMD-LSTM, BRT, and CFNN models.



Figure 6. Scatter chart between the forecasted and observed ET_0 using MVMD-BRT and benchmark comparing models.





Figure 7. Swarm plot of the absolute forecasted errors |FE| generated by the VMD-GPR vs. VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models.



Figure 8. Empirical cumulative distribution function (ECDF) of the forecasted and observed Flood index generated by the VMD-GPR vs. other benchmarking models for each station.

Figure 10 draws a comparison between the observed and forecasted Flood index (FI) generated by the VMD-GPR model of the rainy seasons from 2008 to 2013 for Jeddah station. It is evident that the forecasted FI closely follows the observed FI for each year from 2008 to 2012. For the year 2012–2013, the forecasted FI of the VMD-GPR model is slightly lower than the observed FI, and this shows a sudden increase in rainfall during the season. Overall, the VMD-GPR model is performing well during the rainy season to forecast daily floods in Jedda station.



Figure 9. Taylor diagram of the daily forecasted and observed Flood index generated by the VMD-GPR vs. the benchmarking models.



Figure 10. Comparison of observed and forecasted Flood index (FI) generated by the VMD-GPR model of the rainy seasons from 2008 to 2013 for Jeddah station.

This research study proposed a hybrid VMD-GPR model that has been designed and assessed to forecast daily flood at (t-1) for Jeddah and Jazan stations in Saudia Arabia. The VMD-GPR model is evaluated and benchmarked against VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models and registered its dominance for daily flood forecasting, thus establishing that the VMD-GPR has the better forecasting ability based on several assessment metrics. The outcomes of this work proved that the VMD-GPR model was efficient and successful in demarcating the inputs via the VMD technique, which can later enhance the accuracy of the GPR model for daily flood forecasting.

The innovative VMD-GPR model was an effective method for daily flood forecasts in the regions of Saudi Arabia against the comparing models, but further suggestions and recommendations needed to be studied in the future. This work only utilized the significant lags of the flood index based on precipitation in the VMD-GPR model for forecasting purposes; however, several other climates, meteorological, and hydrological data as input predictors can be employed to augment the accuracy. Additionally, the satellite-derived data can be further used as an alternative, which can notably improve the forecasting facility of the VMD-GPR model, and thus, another method could be a prospective strategy to bring in more physical data aspects for the daily flood forecasts. The computational time for the VMD-GPR is slightly higher as compared to other models.

Advanced machine learning (ML) models prevail in the area of forecasting, but their black-box properties restrict their capability and power, which makes it challenging to comprehend and assess the complex associations of the inputs during the learning process. Subsequently, the fusion of ML with numerical weather prediction models can be an interesting area of research. Moreover, the VMD-GPR model can also be improved in terms of optimization with the help of Bayesian Model Averaging [56] and bootstrapping techniques [57] to capture the underlying model's uncertainties.

Additionally, the VMD benefits the GPR model by improving the precision due to concurrently attaining the non-stationary and non-linearity features within the flood data and handling the mode mixing problems [58]. Thus, it is proven that VMD-GPR can potentially be a feasible advance data-driven model for hydrological sciences to deliver

supportive insights on water resource management to design better proactive prevention strategies in Saudi Arabia.

4. Conclusions

Floods are becoming severely destructive, influenced by recent climate change, significantly affecting several areas around the world and impacting the greatest number of populations. This research work proposed a hybrid VMD-GPR model, and the novelty centered around the fusion of variational mode decomposition and Gaussian process regression integrated to forecast daily floods. The VMD decomposed the input data into IMFs, and then the PACF determined the significant lags, which were later used in the GPR to forecast the daily flood index for Jeddah and Jazan stations in Saudi Arabia. The outcomes prove that the VMD-GPR model substantially improves the daily flood forecast for both Jeddah and Jazan stations in Saudia Arabia as compared to the VMD-LSTM, VMD-BRT, VMD-CFNN, GPR, LSTM, BRT, and CFNN models using valuation scores R, RMSE, MAE, ENS, KGE, IA, and U95%.

The VMD-GPR model designed in this research work was state-of-the-art in terms of the VMD and GPR models, where the VMD substantially advances the forecasting accuracy by tackling the non-stationarity and non-linearity produced by the complicated and complex nature of the daily flood. The results authenticate that the VMD-GPR model is exceptionally good in daily flood forecasting against the comparing models. The proposed VMD-GPR model can play a crucial role in municipal flood mitigation and risk reduction by allowing for proactive measures such as evacuation and resource mobilization. It relies on accurate flood risk assessments and forecasts and helps to identify areas at high risk and inform decisions about infrastructure development and land use planning. Moreover, precise forecasts by VMD-GPR can provide a lead time for preparation, allowing communities to take required safety measures. To widen the scope and potentiality, the proposed VMD-GPR model can be implemented in areas such as atmospheric and environment research such as climate change, renewable and sustainable energy, water resource management, and agriculture sectors for better decision making.

Author Contributions: Conceptualization, A.A.A. and M.A.; methodology, A.A.A. and M.A.; validation, M.A.; formal analysis, A.A.A. and M.A.; investigation, A.A.A., M.A. and A.H.L.; resources, A.H.L.; data curation, A.H.L.; writing—original draft preparation, A.A.A. and M.A.; writing—review and editing, A.A.A., M.A. and A.H.L.; visualization, M.A.; supervision, A.H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. KFU250577].

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors are thankful to the Department of Meteorology, Saudi Arabia, for providing the relevant datasets. This work was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. KFU250577].

Conflicts of Interest: The authors declare no conflict of interest.

References

- 1. Wu, W.; Emerton, R.; Duan, Q.; Wood, W.A.; Wetterhall, F.; Robertson, E.D. Ensemble flood forecasting: Current status and future opportunities. *Wiley Interdiscip. Rev. Water* **2020**, *7*, e1432.
- 2. Hou, S.; Wei, J.; Hou, M.; Xu, J.; Han, L. A hydrological knowledge-informed LSTM model for monthly streamflow reconstruction using distributed data: Application to typical rivers across the Tibetan plateau. *J. Hydrol.* **2025**, *649*, 132409.
- 3. Baig, M.B.; Alotibi, Y.; Straquadine, S.G.; Alataway, A. Water resources in the Kingdom of Saudi Arabia: Challenges and strategies for improvement. In *Water Policies in MENA Countries*; Springer: Cham, Switzerland, 2020; pp. 135–160.
- Erica DeNicola, E.; Aburizaiza, S.O.; Siddique, A.; Khwaja, H.; Carpenter, O.D. Climate change and water scarcity: The case of Saudi Arabia. Ann. Glob. Health 2015, 81, 342–353.
- 5. Almazroui, M. Rainfall trends and extremes in Saudi Arabia in recent decades. Atmosphere 2020, 11, 964.
- Nevo, S.; Morin, E.; Rosenthal, G.A.; Metzger, A.; Barshai, C.; Weitzner, D.; Voloshin, D.; Kratzert, F.; Elidan, G.; Gideon Dror, G.; et al. Flood forecasting with machine learning models in an operational framework. *Hydrol. Earth Syst. Sci.* 2022, 26, 4013–4032.
- Noymanee, J.; Theeramunkong, T. Flood forecasting with machine learning technique on hydrological modeling. *Procedia Comput. Sci.* 2019, 156, 377–386.
- 8. Motta, M.; de Castro Neto, M.; Sarmento, P. A mixed approach for urban flood prediction using Machine Learning and GIS. *Int. J. Disaster Risk Reduct.* **2021**, *56*, 102154.
- 9. Rajab, A.; Farman, A.; Islam, N.; Syed, D.; Elmagzoub, A..; Shaikh, A.; Akram, M.; Alrizq, A. Flood forecasting by using machine learning: A study leveraging historic climatic records of Bangladesh. *Water* **2023**, *15*, 3970.
- Shafizadeh-Moghadam, H.; Valavi, R.; Shahabi, H.; Chapi, K.; Shirzadi, A. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. J. Environ. Manag. 2018, 217, 1–11.
- 11. Sankaranarayanan, S.; Sahoo, A.; Satapathy, D.P. Flood prediction based on weather parameters using deep learning. *J. Water Clim. Chang.* **2020**, *11*, 1766–1783.
- 12. Prasad, R.; Deo, R.C. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena* **2019**, *177*, 149–166.
- Soman, K.P.; Athira, S.; Harikumar, K. Recursive Variational Mode Decomposition Algorithm for Real Time Power Signal Decomposition. *Procedia Technol.* 2015, 21, 540–546.
- 14. Mallat, S.G. A Wavelet Tour of Signal Processing; Academic: New York, NY, USA, 1998.
- 15. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693.
- Nourani, V.; Aida Hosseini Baghanam, H.A.; Adamowski, J.; Kisi, O. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. J. Hydrol. 2014, 514, 358–377.
- 17. Nourani, V.; Komasi, M.; Mano, A. A multivariate ANN-wavelet approach for rainfall–runoff modeling. *Water Resour. Manag.* **2009**, *23*, 2877–2894.
- Deo, R.C.; Tiwari, M.K.; Adamowski, J.F.; Quilty, J.M. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. *Stoch. Environ. Res. Risk Assess.* 2016, *31*, 1211–1240.
- 19. Deo, R.C.; 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*, 568–593.
- Krishna, B.; Rao, Y.R.S.; Nayak, P.C. Time Series Modeling of River Flow Using Wavelet Neural Networks. J. Water Resour. Prot. 2011, 03, 50–59.
- 21. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A* **1998**, *454*, 903–995.
- Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* 2009, 1, 1–41.
- Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, F. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011.
- Colominas, M.A.; Schlotthauer, G.; Torres, M.E. Improved complete ensemble EMD: A suitable tool for biomedical signal processing. *Biomed. Signal Process. Control* 2014, 14, 19–29.

- Ali, M.; Deo, R.C.; Maraseni, T.; Downs, N.J. Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms. J. Hydrol. 2019, 576, 164–184.
- Ali, M.; Prasad, R. Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition. *Renew. Sustain. Energy Rev.* 2019, 104, 281–295.
- 27. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. IEEE Trans. Signal Process. 2014, 62, 531-544.
- 28. Fu, T.; Li, X.; Jia, R.; Feng, L. A novel integrated method based on a machine learning model for estimating evapotranspiration in dryland. *J. Hydrol.* **2021**, *603*, 126881.
- Wang, D.; Liu, Y.; Luo, H.; Yue, C.; Cheng, S. Day-ahead PM2. 5 concentration forecasting using WT-VMD based decomposition method and back propagation neural network improved by differential evolution. *Int. J. Environ. Res. Public Health* 2017, 14, 764.
- Zhou, J.; Sun, N.; Jia, B.; Peng, T. A novel decomposition-optimization model for short-term wind speed forecasting. *Energies* 2018, 11, 1752.
- Majumder, I.; Dash, P.; Bisoi, R. Variational mode decomposition based low rank robust kernel extreme learning machine for solar irradiation forecasting. *Energy Convers. Manag.* 2018, 171, 787–806.
- Moishin, M.; Deo, R.C.; Prasad, R.; Raj, N.; Abdulla, S. Development of Flood Monitoring Index for daily flood risk evaluation: Case studies in Fiji. *Stoch. Environ. Res. Risk Assess.* 2021, 35, 1387–1402.
- Williams, C.; Rasmussen, C. Gaussian processes for regression. In Advances in Neural Information Processing Systems; MIT Press: Cambridge, MA, USA, 1995; Volume 8.
- Gao, J.; Ling, H.; Hu, W.; Xing, J. Transfer learning based visual tracking with gaussian processes regression. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014.
- 35. Williams, C.K.; Rasmussen, C.E. Gaussian Processes for Machine Learning; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
- Ghasemi, P.; Karbasi, M.; Nouri, A.Z.; Tabrizi, M.S.; Azamathulla, H.M. Application of Gaussian process regression to forecast multi-step ahead SPEI drought index. *Alex. Eng. J.* 2021, 60, 5375–5392.
- Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* 2019, *31*, 1235–1270.
- Staudemeyer, R.C.; Morris, E.R. Understanding LSTM—A tutorial into long short-term memory recurrent neural networks. *arXiv* 2019, arXiv:1909.09586.
- 39. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, 404, 132306.
- 40. Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232.
- Saha, S.; Arabameri, A.; Saha, A.; Blaschke, T.; Ngo, P.T.T.; Nhu, V.H.; Band, S.S. Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Sci. Total Environ.* 2021, 764, 142928.
- Faskari, S.A.; Falope, T.; Ojim, G.; Abdullahi, U.B.; Abba, S.I. A Novel Machine Learning based Computing Algorithm in Modeling of Soiled Photovoltaic Module. *Knowl.-Based Eng. Sci.* 2022, 3, 28–36.
- 43. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. J. Anim. Ecol. 2008, 77, 802–813.
- Naghibi, S.A.; Pourghasemi, H.R. A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water Resour. Manag.* 2015, 29, 5217–5236.
- Carty, D.M.; Young, T.M.; Zaretzki, R.L.; Guess, F.M.; Petutschnigg, A. Predicting and correlating the strength properties of wood composite process parameters by use of boosted regression tree models. *For. Prod. J.* 2015, 65, 365–371.
- Westreich, D.; Lessler, J.; Funk, M.J. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol. 2010, 63, 826–833.
- 47. Fahlman, S.; Lebiere, C. The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1989; Volume 2.
- Dharma, S.; Hassan, M.H.; Ong, H.C.; Sebayang, A.H.; Silitonga, A.S.; Kusumo, F.; Milano, J. Experimental study and prediction of the performance and exhaust emissions of mixed Jatropha curcas-Ceiba pentandra biodiesel blends in diesel engine using artificial neural networks. J. Clean. Prod. 2017, 164, 618–633.
- Mohammadi, M.-R.; Hemmati-Sarapardeh, A.; Schaffie, M.; Husein, M.M.; Ranjbar, M. Application of cascade forward neural network and group method of data handling to modeling crude oil pyrolysis during thermal enhanced oil recovery. *J. Pet. Sci. Eng.* 2021, 205, 108836.

- 50. Willmott, C.J. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 1982, 63, 1309–1313.
- 51. McCuen, R.H.; Knight, Z.; Cutter, A.G. Evaluation of the Nash–Sutcliffe efficiency index. J. Hydrol. Eng. 2006, 11, 597–602.
- Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F.; Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 2009, 377, 80–91.
- Fijani, E.; Barzegar, R.; Deo, R.C.; Tziritis, E.; Skordas, K. Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters. *Sci. Total Environ.* 2019, 648, 839–853.
- 54. Shamseldin, A.Y. Application of a neural network technique to rainfall runoff. J. Hydrol. 1997, 199, 272–294.
- 55. Xu, Z.; Hou, Z.; Han, Y.; Guo, W. A diagram for evaluating multiple aspects of model performance in simulating vector fields. *Geosci. Model Dev.* **2016**, *9*, 4365–4380.
- 56. Sloughter, J.M.; Gneiting, T.; Raftery, A.E. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Am. Stat. Assoc.* **2010**, *105*, 25–35.
- 57. Tiwari, M.K.; Chatterjee, C. A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. *J. Hydroinform.* **2011**, *13*, 500–519.
- 58. Gao, F.; Shao, X. A novel interval decomposition ensemble model for interval carbon price forecasting. Energy 2022, 243, 123006.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.