forecasting

MDPI

Article

A Hybrid Linear–Gaussian Process Framework with Adaptive Covariance Selection for Spatio-Temporal Wind Speed Forecasting

Thinawanga Hangwani Tshisikhawe [1], Caston Sigauke [1,*], Timotheous Brian Darikwa [2] and Saralees Nadarajah [3]

[1] Department of Mathematical and Computational Sciences, University of Venda, Thohoyandou 0950, South Africa; thina.tshisikhawe@univen.ac.za
[2] Department of Statistics and Operations Research, University of Limpopo, Polokwane 0727, South Africa; timotheus.darikwa@ul.ac.za
[3] Department of Mathematics, Manchester University, Manchester M13 9PL, UK; saralees.nadarajah@manchester.ac.uk
* Correspondence: caston.sigauke@univen.ac.za

Highlights

What are the main findings?

- The accuracy of the developed wind speed forecasting model based on a Gaussian process (GP) for residual modelling largely depends on the local wind regime.
- The GP model improved forecast accuracy by up to 16.3% at high-elevation inland stations. In contrast, the model was ineffective at the low-elevation coastal stations.

What are the implications of the main findings?

- The success of sophisticated residual modelling techniques, including the Gaussian process model, cannot be generalised and should be adapted to particular geographical and meteorological regimes.
- The forecasting process should account for the clustering of monitoring stations by geographical location and elevation, e.g., inland or coastal areas.

Abstract

Accurate wind speed forecasting is essential for the efficient integration of wind energy into power systems, as it directly influences generation scheduling, grid stability, and energy market operations. Forecast errors can lead to significant economic losses, including increased balancing costs, inefficient dispatch of backup generation, and penalties in electricity markets. However, wind behaviour is highly complex due to the influence of synoptic weather systems, terrain variability, and turbulence, which makes accurate prediction particularly challenging. This paper proposes a hybrid modelling framework that combines a linear regression mean model with Gaussian process (GP) residual modelling to improve forecast accuracy. Monitoring stations were grouped based on geographic coordinates and elevation, with cluster validation using the Hopkins statistic and silhouette analysis. The results show that for high-elevation inland stations (cluster 2), GP residual modelling improves forecast accuracy by up to 16.3%. In contrast, for low-elevation coastal stations (cluster 1), the GP approach does not yield improvements, indicating that its effectiveness depends strongly on the underlying wind regime.

Keywords: cluster validation; Gaussian processes; hybrid models; wind speed forecasting

Check for updates

Academic Editor: Sonia Leva

Received: 2 March 2026
Revised: 22 April 2026
Accepted: 25 April 2026
Published: 26 April 2026

Copyright: © 2026 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.

1. Introduction

1.1. Overview

Wind speed forecasting plays a critical role in the integration of wind energy into modern power systems, particularly as countries transition towards low-carbon energy sources. Accurate forecasts are essential for generation scheduling, grid stability, and participation in electricity markets. Forecast errors can have significant economic and operational consequences: overestimation of wind power may lead to unnecessary curtailment and financial penalties, while underestimation can result in reliance on costly backup generation and increased balancing costs. Furthermore, the inherent variability and intermittency of wind pose substantial challenges for grid integration, requiring system operators to manage uncertainty across multiple spatial and temporal scales. Despite its importance, wind behaviour remains difficult to model due to its complex, intermittent spatial and temporal dynamics, which are influenced by atmospheric conditions, topography, and local turbulence. Linear models are often effective in capturing large-scale trends but fail to account for localised dependencies, while purely data-driven approaches may achieve high predictive accuracy at the expense of interpretability and uncertainty quantification.

The use of GPs is advantageous for spatio-temporal forecasting because the method is probabilistic, allowing for uncertainty and covariance to be explicitly incorporated into the modelling process. But GPs have limitations: they are computationally intensive when applied to large datasets, and they may become redundant when applied to raw wind speed measurements. The proposed alternative here is to use GPs in the residual part of a mean model. There are also issues of variability across different wind regimes by geographic location; there is a significant difference in wind characteristics between low-elevation coastal areas and high-elevation inland areas. It would negatively affect forecast performance if the spatial heterogeneity of the wind regime was not taken into account. The current study, therefore, suggests using a combined approach combining linear regression and Gaussian process models for residual analysis.

1.2. Literature Review

Accurate forecasting of wind speeds is critical for integrating wind energy into the grid and for various meteorological needs. Unfortunately, the chaotic behaviour of wind, influenced by large-scale atmospheric conditions, local topography, and turbulence, makes wind speed forecasting a complex problem. This review traces the chronological and thematic developments of wind speed forecasting techniques, from earlier physical and statistical models to modern hybrid and probabilistic models. It focuses on the developments of Gaussian process regression (GPR) and the utilisation of clustering techniques, both of which are significant to the proposed research. A brief overview of the major models is presented in Table 1.

Table 1. Comparison of wind speed forecasting methods.

Method	Temporal Focus	Spatial Awareness	Uncertainty Quantification	Computational Cost	Key Limitations
NWP Models [1]	Medium-long range	Low	No	High	Systematic errors, limited resolution
SARIMA [2]	Short term	None	No	Low	Poor with nonlinear data
Kalman Filters [3]	Short term	Low	Limited	Low-medium	Linearity assumptions

Table 1. *Cont.*

Method	Temporal Focus	Spatial Awareness	Uncertainty Quantification	Computational Cost	Key Limitations
ML/DL [4,5]	Short-medium	None-medium	No	High	Black box, site-specific, data-hungry
Bayesian Models [6,7]	Short-medium	None-medium	Yes	High–very high	Computationally intensive, calibration needed
GPR [8,9]	Short term	Medium	Yes	Very high ($O(n^3)$)	Scalability issues
Hybrid Linear-GPR [9,10]	Short term	Medium	Yes	Medium-high	Assumes uniform spatial benefit
Spatio-temporal GPR/ Kriging [9,11]	Short term	High	Yes	Very high	Data density requirements
Cluster-based Hybrid (Proposed)	Short term	High	Yes	Medium	Regime identification, clustering sensitivity

1.2.1. Classical Methods: Physical and Statistical Models

Traditionally, wind speed forecasting was carried out using Numerical Weather Prediction (NWP) models, which simulate atmospheric physics to predict wind speeds. These are certainly essential in wind speed forecasting; however, these models are subject to systematic errors and lack spatial resolution, which are often overcome by post-processing these results to increase their accuracy locally [1]. To overcome these problems, statistical time-series models such as ARIMA and SARIMA were widely applied. These models are effective in capturing linear trends and seasonality in wind speeds; however, they are not effective in capturing rapid, nonlinear changes in wind speeds, which are characteristic of wind speeds in general [2]. More adaptive models, such as Kalman Filters, were then applied to correct errors in short-term forecasts; however, these filters rely on linear assumptions and are prone to noise, limiting their application to wind speed forecasting over longer time scales and in spatially heterogeneous domains [3].

1.2.2. The Ascent of Machine Learning and Deep Learning

The limitations of linear models triggered the development and application of machine learning (ML) and deep learning (DL) techniques. ML and DL models, including random forests, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM), are very effective at learning complex, nonlinear relationships directly from historical data, often resulting in significant accuracy gains in forecasts [4]. The hybrid model, combining NWP model results and LSTM, is also very effective for forecasting complex terrain, including nonlinear relationships and temporal dependencies [5]. However, ML/DL-based models are computationally intensive, require careful tuning, and often struggle to generalise across different locations. Additionally, ML/DL models are considered "black box" models, which limits their interpretability, making interpretability a driving need for finding models that are effective yet transparent and physically plausible [4,5].

1.2.3. Probabilistic Forecasting and Gaussian Process Regression

As the need to quantify uncertainty in grid operations increases, probabilistic forecasting has become a crucial area of study. Unlike point forecasts, probabilistic forecasting can offer predictive distributions for better risk-informed decision-making [6]. Bayesian methods, including Bayesian neural networks and Bayesian-Optimised models, inherently offer uncertainty quantification. However, their site-specific calibration can be a hindrance for real-time applications [7].

Within this paradigm, Gaussian process regression (GPR) has emerged as a prominent nonparametric tool for probabilistic forecasting. GPR is particularly prized for its capacity to yield a mean estimate and a well-founded interval estimate, which capture the essential randomness of wind phenomena [8]. The efficacy of GPR is critically dependent on its covariance function (kernel), which embodies assumptions regarding the underlying process, e.g., smoothness, periodicity, etc. GPR has been successfully utilised in hybrid frameworks, e.g., coupling GPR with support vector regression to post-process NWP forecasts for short-term forecasting (1 to 6 h ahead) [9]. Another hybrid model, which is a focus of this research, is a linear mean model coupled with a GPR model for forecasting residuals. This model balances interpretability with the power of GPR for modelling complex, spatially correlated error structures. However, standard GPR is known to scale poorly to large datasets, requiring a cubic time complexity ($\mathcal{O}(n^3)$) for its implementation [9,10].

1.2.4. Spatio-Temporal Modelling and the Role of Clustering

The spatial correlations in wind fields suggest that using data from adjacent areas can improve local predictions. A more advanced version of GPR, known as kriging, which is analogous to geostatistics, has been employed to interpolate wind fields while accounting for spatial correlations. It provides predictions along with their uncertainties [9]. More recent models seek to address complex spatio-temporal interactions. For example, dynamic versions of GPR that incorporate local clustering for improving accuracy have been presented [11]. Bayesian models for spatio-temporal forecasting for renewable energy forecasting have also been presented, but their success relies on the availability of dense data [12].

A notable result of the recent research is the recognition that wind speed statistical properties are not necessarily homogeneous over vast geographical areas. Various regimes, depending on the presence or absence of factors such as coastal areas and altitude, exhibit unique properties, including levels of non-stationarity, turbulence, and autocorrelation. This means that rather than seeking an optimal model applicable over the global domain, the regime-based approach of dividing the domain into homogeneous areas is more advisable. Grouping stations by their physical properties, such as geographic coordinates and altitude, is a statistically sound approach for dividing the domain.

Recent research emphasizes the potential of using the latest GP variants. For example, the application of hybrid GP approaches, which integrate spatial correlation with corrected NWP forecast information, has been found to offer better accuracy than the standard GP model [13]. Another GP model variant, the multi-task GP, is also found to be beneficial, particularly when applied to data-scarce locations, where information from data-rich locations can be exploited [14]. However, the current GP model variants, including the complex residual model, have the limitation that their added value is assumed to be applicable uniformly across the study area.

1.3. Summary of Literature Review

This research addresses the limitation of the complex residual model directly by proposing a hybrid linear-GP model, not over the entire study area but over geographically and elevation-based clusters. This is to evaluate the added value of the complex residual model across regimes, providing a physically informed, computationally efficient tool to improve the accuracy of wind speed forecasts.

A summary of the comparison of wind speed forecasting methods, including the key studies discussed, highlighting their temporal focus, uncertainty quantification, computational cost and limitations, is given in Table 1.

Despite significant advances in wind speed prediction techniques, several critical shortcomings remain. Traditional statistical models have drawbacks associated with their linear nature and their inability to incorporate more complex spatial relationships. Although machine learning and deep learning algorithms improve accuracy, they are not always interpretable and cannot provide proper uncertainty estimates. In turn, Gaussian processes solve both problems; however, they are computationally expensive. Finally, although hybrid and spatio-temporal models can offer more accurate predictions, they implicitly assume that improvements in model performance can be applied universally, regardless of local conditions such as terrain height and proximity to the coast. As a result, no algorithms can deliver high performance across interpretability, computation time, uncertainty estimation, and spatial heterogeneity. This gap motivates the proposed hybrid framework, which combines a linear mean model with Gaussian process residual modelling while explicitly incorporating regime-based clustering to account for spatial variability in wind behaviour.

1.4. Contribution and Research Highlights

The main contribution of the study under discussion is to demonstrate that hybrid models of linear and Gaussian processes for wind speed prediction must be used selectively, depending on the wind regime. Using clustering of elevations and coordinates, it is established that the Gaussian process residuals can be highly beneficial in high-variability areas (with up to 16.3% RMSE improvement) but potentially harmful in homogeneous clusters.

The main research highlights of this study are as follows:

1. The hybrid linear–Gaussian process model considers the wind speed as a sum of two components, namely, the linear mean and GP residual, ensuring both clarity and spatial interactions between locations.
2. Cluster analysis based on geographic coordinates and elevation allows for location-dependent analysis of the wind speed.
3. The probabilistic nature of the GP component gives information about predictive uncertainty, making risk-based decisions possible for wind energy management.
4. Systematic selection and optimisation of hyperparameters ensure capturing of spatial and temporal interaction in the wind regime.
5. The study shows strong dependence of GP residual modelling effectiveness on wind regime.

The rest of the paper is organised as follows. Section 2 briefly discusses materials and methods, including the Hopkins variable selection method, temporal GP regression, linear temporal regression, the hybrid, and the evaluation metrics. Section 4 presents the results, the discussion is presented in Section 5, and the conclusion is in Section 6.

2. Materials and Methods

Flowchart of the process

Figure 1 shows the flowchart of the structure of the paper. The process starts with collecting the required data from wind sensors, after which clustering is performed based on geo-coordinates and altitude. Cluster quality is assessed using the Hopkins and Silhouette measures. Clusters are created based on altitude and residual magnitude. Each type of residual model (low- and high-altitude) uses a Gaussian Process (GP) model with adaptive covariance selection.

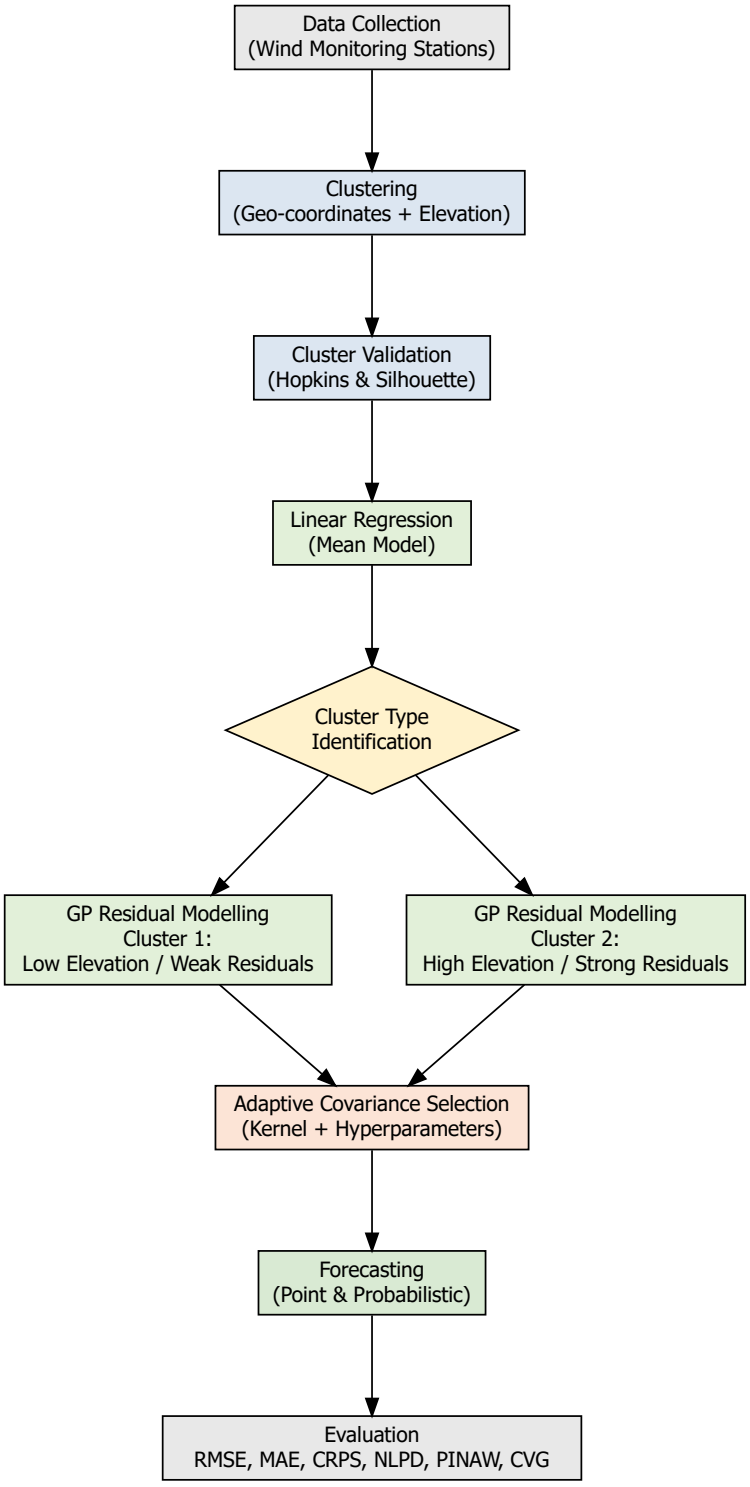


Figure 1. Flowchart of the structure of the paper.

Data Description and Clustering

The data used in this study are from the Wind Atlas South Africa (WASA) at https://wasadata.csir.co.za/wasa1/WASAData?pscreen=nwa&pemail=nimo@risoe.dtu.dk (accessed on 16 January 2025). The WASA data was obtained from the website [15]. Wind speed data from 10 stations were collected, each with multiple meteorological predictors (e.g., temperature and pressure). The variables included are as follows: Wind speed (m/s) at hub height (62 m). The data also feature the temporal aspect, which is the time of the day; the spatial aspect, which is the coordinates; and the meteorological aspect, which includes temperature, barometric pressure, relative humidity, and wind direction. These measurements were recorded every 10 min for all variables.

Clustering

Wind conditions in several neighbouring wind farms are likely similar to each other because they are subject to similar weather conditions. To reduce computation, several wind farms can be combined into clusters, enabling forecasts based on clusters rather than individual wind farms. There is an advantage to combining information from several wind farms in the cluster, as it may help iron out local weather extremes.

To capture regime-dependent wind dynamics, stations were clustered using geographical coordinates (longitude and latitude) and elevation. We first used the Hopkins test to assess the dataset's tendency to form clusters and determine whether the data are meaningfully clustered or randomly distributed (i.e., not clusterable). According to the method proposed by Hopkins and Skellam [16], the clustering tendency can be quantified using the Hopkins statistic H defined as

$$H = \frac{\sum_{i=1}^{n} u_i^d}{\sum_{i=1}^{n} u_i^d + \sum_{i=1}^{n} w_i^d}, \tag{1}$$

where $H > 0.75$ suggests strong evidence of the cluster structure, while $H < 0.5$ suggests little or no cluster tendency. Thereafter, we used hierarchical clustering to group the data. This is a method of cluster analysis that builds a hierarchy of clusters to uncover the underlying structure in a dataset. It does not rely on probability distributions, making it a nonparametric method, suitable even when the underlying data distribution is unknown. It also produces a dendrogram, a tree diagram, that provides a visual, flexible representation of how clusters form at different levels. This allows for users to choose the number of clusters by cutting the tree at the desired level.

Variable selection—Relaxed Lasso

The Relaxed Lasso is a two-stage modification of the standard Lasso regression that aims to reduce the bias introduced by Lasso's shrinkage, particularly in high-dimensional settings [17]. Lasso performs both variable selection and shrinkage by minimising Equation (2): Lasso estimates regression coefficients by solving Equation (2). As shown in Equation (2), the Lasso introduces an ℓ_1-penalty to enforce sparsity.

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{2}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the predictor matrix, and $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. The notation $\|\cdot\|_2$ denotes the ℓ_2-norm (Euclidean norm), and $\|\cdot\|_1$ denotes the ℓ_1-norm, defined as the sum of absolute values of the vector components. The regularisation parameter $\lambda \geq 0$ controls the trade-off between model fit and sparsity.

The factor $\frac{1}{2n}$ expresses the loss as average empirical risk; alternative scalings (e.g., $\frac{1}{2}$) are equivalent up to a rescaling of the tuning parameter λ. While effective for variable selection, the Lasso tends to over-shrink coefficient estimates, leading to biased results.

$$\hat{\beta}^{\text{relaxed}}(\lambda, \phi) = \phi \cdot \hat{\beta}^{\text{lasso}}(\lambda) + (1 - \phi) \cdot \hat{\beta}^{\text{LS}}(\lambda) \tag{3}$$

where $\hat{\beta}^{\text{lasso}}(\lambda)$ is the Lasso solution at regularisation level λ, and $\hat{\beta}^{\text{LS}}(\lambda)$ is the ordinary least squares estimator computed on the subset of variables selected by the Lasso. The parameter $\phi \in [0, 1]$ is a relaxation parameter controlling the degree of shrinkage. When $\phi = 1$, the estimator reduces to the standard Lasso; when $\phi = 0$, it corresponds to unregularised least squares on the Lasso-selected model. Intermediate values of ϕ provide a bias–variance trade-off.

The Relaxed Lasso addresses the tendency of the Lasso to over-shrink coefficients by decoupling variable selection and coefficient estimation. Specifically, it first performs variable selection using the Lasso and then refits the model with reduced shrinkage on the selected variables, improving estimation accuracy.

3. Methodology

3.1. GPR Models for Wind Forecasting

Gaussian processes are a powerful nonparametric Bayesian approach to regression and classification introduced by [18]. These processes are conceptualised as components of spatial–temporal modelling, specifically within a defined region where stations are located at different sites. Spatial analysis seeks to construct an optimal model for generating outputs by considering inputs from diverse locations over various time frames. Throughout this section, s_i denotes the spatial location of station i, and t denotes time. The observed response is denoted by $Y(s_i, t)$, while $X(s_i, t)$ represents the vector of covariates at location s_i and time t. The regression coefficients are denoted by β. The latent spatial–temporal Gaussian process is denoted by $w(s_i, t)$, and $\varepsilon(s_i, t)$ represents independent Gaussian measurement noise.

Equation (4) describes a spatio-temporal GP model:

1. Model equation:

$$Y(s_i, t) = \mathbf{x}^\top(s_i, t)\boldsymbol{\beta} + w(s_i, t) + \varepsilon(s_i, t), \tag{4}$$

where $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T$, $Y(s_i, t)$ is the observed data at location s_i and time t. and

$$w(s_i, t) \sim GP\big(0,\ C\big((s_i, t), (s_j, t')\big)\big), \tag{5}$$

$$\varepsilon(s_i, t) \sim \mathcal{N}(0, \sigma^2). \tag{6}$$

- $\mathbf{x}(s_i, t) = x_1(s_i, t), \ldots, x_p(s_i, t)$ is the p-dimensional covariate vector of s at (s_i, t).
- $\boldsymbol{\beta}$ is the vector of regression coefficients.
- $w(s_i, t)$ is the latent Gaussian process capturing spatio-temporal correlation.
- $\varepsilon(s_i, t)$ is independent Gaussian noise with variance σ_ε^2.

2. Prior to the latent process:

The independent Gaussian process (GP) model is specified hierarchically by

$$Y_{it} = O_{it} + \varepsilon_{it}, \tag{7}$$

$$O_{it} = X_{it}\beta + w_{it}, \tag{8}$$

for each $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where we assume that ε_{it} and w_{it} are independent, and each is normally distributed with its respective parameters. The notation

$\varepsilon_{it} = (\varepsilon(s_1,t),\ldots,\varepsilon(s_n,t))^\top$ will be used to denote the so-called nugget effector the pure error term, which is assumed to be independently normally distributed as

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I),$$

where σ_ε^2 is the unknown pure error variance and I is the identity matrix.

The spatio-temporal random effects will be denoted by $w_{it} = (w(s_1,t),\ldots,w(s_n,t))^\top$, and these will be assumed to follow

$$w_{it} \sim \mathcal{N}(0, \Sigma)$$

independently in time

Let $\mathcal{O}$ denote all the random effects O_{it}, for $i = 1,\ldots,n$ and $t = 1,\ldots,T$. Let

$$\Theta = (\beta, \sigma_\varepsilon^2, \sigma_w^2, \rho, \nu)$$

denote all the parameters of this model, and let $\pi(\Theta)$ denote the prior distribution.

The logarithm of the joint posterior distribution of the parameters and the missing data for this GP model is given by

$$\begin{aligned}\log \pi(\Theta, \mathcal{O}, y^* \mid y) \propto & -\frac{N}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-O_{it})^\top(Y_{it}-O_{it}) \\ & -\sum_{i=1}^{n}\frac{T}{2}\log\left|\sigma_w^2\right| - \frac{1}{2\sigma_w^2}\sum_{i=1}^{n}\sum_{t=1}^{T}(O_{it}-X_{it}\beta)^\top S_w^{-1}(O_{it}-X_{it}\beta) \\ & +\log\pi(\Theta). \end{aligned} \tag{9}$$

where $N = nT$ denotes the total number of observations. The matrix S_w denotes the spatial covariance matrix induced by the Gaussian process kernel. The term y^* represents latent or unobserved quantities associated with the posterior formulation.

3. Linear models:

The very first attempt at modelling a spatio-temporal response variable is to consider the linear regression models. The multiple linear regression model is written in Equation (10) as

$$Y(s_i,t) = \beta_1 x_1(s_i,t) + \cdots + \beta_p x_p(s_i,t) + \varepsilon(s_i,t), \quad \text{for } i = 1,\ldots,n;\ t = 1,\ldots,T, \tag{10}$$

where $\varepsilon(s_i,t)$ is the spatio-temporal error term. The error term $\varepsilon(s_i,t)$ is assumed to be a zero-mean spatio-temporal Gaussian process with a separable covariance structure, given by

$$\mathrm{Cov}\big(\varepsilon(s_i,t_k),\varepsilon(s_j,t_l)\big) = \sigma^2\,\rho_s(\|s_i-s_j\|;\theta_s)\,\rho_t(|t_k-t_l|;\theta_t),$$

where $\rho_s(\|s_i-s_j\|;\theta_s)$ is an anisotropic correlation function of the spatial distance $\|s_i-s_j\|$ between two locations s_i and s_j, which may depend on the parameter θ_s, which may be more than one in number.

3.2. Hybrid Model

The proposed hybrid GP regression combines a linear regression model with a Gaussian process (GP) applied to residuals, enabling both interpretable trend estimation and flexible modelling of spatio-temporal dependencies.

Hybrid Model Formulation

Let $Y(\mathbf{s_i}, t)$ denote the wind speed observed at spatial location $\mathbf{s}$ and time t. The proposed model is defined as

$$Y(\mathbf{s_i}, t) = \mathbf{x}(\mathbf{s_i}, t)^\top \boldsymbol{\beta} + f(\mathbf{s_i}, t) + \varepsilon(s_i, t), \tag{11}$$

with

$$f(s_i, t) \sim GP\big(0,\ k\big((s_i, t), (s_j, t')\big)\big), \tag{12}$$

and

$$\varepsilon(s_i, t) \sim \mathcal{N}(0, \sigma^2). \tag{13}$$

In the hybrid formulation, the Gaussian process $f(s_i, t)$ is applied to the residual component of the linear model. This allows for the linear term to capture large-scale trends, while the GP accounts for nonlinear spatio-temporal dependencies in the residual structure.

As shown in Equation (11), the model combines a linear regression component with a Gaussian process for residuals. This two-stage formulation enables a clear separation between deterministic structure and stochastic variability. The linear component captures large-scale spatio-temporal patterns using meteorological and temporal covariates. Model parameters are estimated using ordinary least squares. The fitted linear model serves as a baseline and provides residuals for subsequent GP modelling.

Gaussian Process Residual Model

The residual process ($f(\mathbf{s}, t)$) is modelled using a zero-mean GP with covariance function (k). Multiple covariance structures are considered, including Matérn covariance with smoothness parameters ($\nu = 0.5, 1.5, 2.5$), exponential covariance, and Gaussian (squared exponential) covariance. Each covariance structure encodes different assumptions about the smoothness of the residual wind speed field.

Covariance Selection Procedure

For every validation split of residuals, GP models with different covariance functions are employed. Estimation of the hyperparameters is conducted through maximum likelihood optimisation. When selecting the best covariance function, a choice is made based on the validation RMSE and probabilistic performance measures.

Validation Strategy and Evaluation Metrics

A leave-one-station-out cross-validation method was used to evaluate the model's performance. In each iteration, validation was performed on one station, and training was conducted on the remaining stations. This repetition continued until all stations were used for validation once. Hence, for four stations, 25% of the data was utilised for validation; for six stations, 16.7% was utilised. This checks the model's ability to make predictions at unobserved locations of wind measurement.

3.3. Validation Metrics

Bayesian evaluation metrics are used to evaluate the performance of Bayesian models, such as Bayesian neural networks. These metrics often go beyond traditional accuracy measures to incorporate uncertainty estimates, which are a key feature of Bayesian methods. The following techniques are used to evaluate the models and are classified as deterministic and probabilistic. The deterministic metrics are the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), and the probabilistic metrics are the continuous ranked probability score (CRPS), the coverage probability (CVG), the negative log predictive density (NLPD), PINAD, and PINAW.

RMSE (Root Mean Squared Error):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{14}$$

MAE (Mean Absolute Error):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{15}$$

where y_i is the wind speed, $\hat{y}_i$ is the forecast of the wind speed, and N is the number of forecasts. RMSE measures the average magnitude of forecast errors, giving a higher weight to larger errors, whereas MAE measures the average absolute difference between predicted and actual values.

The probabilistic metrics are given below:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbf{1}\{y \le z\}\right)^2 dz \tag{16}$$

The continuous ranked probability score (CRPS) can be written in expectation form [19] as

$$\text{CRPS}(F, y) = \mathbb{E}_F\left[|Y - y|\right] - \frac{1}{2}\mathbb{E}_F\left[|Y - Y'|\right], \quad Y, Y' \sim F \text{ i.i.d.} \tag{17}$$

where F is the predictive cumulative distribution function (CDF), y is the observed value $Y \sim F$, $Y' \sim F$ are independent random variables drawn from the forecast distribution F, and $\mathbb{E}_F$ denotes the expectation under the distribution F. CRPS evaluates the quality of a probabilistic forecast by comparing the forecast's cumulative distribution function (CDF) with the observed value. CVG measures the proportion of true observations that fall within a predicted confidence interval of a given nominal level (e.g., 95%). The predictive coverage of the $1 - \alpha$ confidence interval is measured using the coverage (CVG) metric. For n observations, let $\hat{y}_i^L$ and $\hat{y}_i^U$ denote the lower and upper bounds of the predicted $(1 - \alpha) \times 100\%$ confidence interval for the ith observation. The empirical coverage is then

$$\text{CVG}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{y_i \in [\hat{y}_i^L, \hat{y}_i^U]\}, \tag{18}$$

where y_i is the observed value at test point i, and $\mathbf{1}[\cdot]$ is the indicator function, returning 1 if the condition is true and 0 otherwise. CVG represents the proportion of observations falling within their corresponding predictive intervals. A well-calibrated model should achieve $\text{CVG}_{1-\alpha} \approx 1 - \alpha$, reflecting accurate uncertainty quantification.

Negative log predictive density (NLPD) evaluates uncertainty calibration. For a single forecast-observation pair,

$$\text{NLPD} = -\log p(y \mid \mathcal{D}), \tag{19}$$

where y is the observed value, and $p(y \mid \mathcal{D})$ is the predictive probability density at y given training data $\mathcal{D}$.

For N test points, the average NLPD is

$$\text{NLPD}_{\text{avg}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i \mid \mathcal{D}) \tag{20}$$

The negative log predictive density (NLPD) measures how likely the model thinks the observed data are, accounting for both the mean and the uncertainty. The prediction interval normalised average width (PINAW) measures the average width of the prediction interval, normalised by the range of observed values. Narrower intervals indicate higher precision. It is defined as

$$\text{PINAW} = \frac{1}{k \cdot R} \sum_{i=1}^{k} (U_i - L_i), \tag{21}$$

where k is the number of validation points, U_i and L_i are the upper and lower bounds of the prediction interval for the i-th observation, and $R = \max(y_{\text{true}}) - \min(y_{\text{true}})$ is the range of the observed values.

The prediction interval normalised average deviation (PINAD) measures the average deviation of the midpoint of the prediction interval from the true values, normalised by the range. It indicates the calibration of the prediction intervals. It is defined as

$$\text{PINAD} = \frac{1}{k \cdot R} \sum_{i=1}^{k} \left| \frac{U_i + L_i}{2} - y_{\text{true},i} \right|, \tag{22}$$

where the symbols have the same meaning as above, and lower values of PINAD indicate that the prediction interval is well centred around the observed values.

3.4. Summary of Experimental Setup

The objective of the experiments is to evaluate the effectiveness of the proposed hybrid linear–Gaussian process (GP) model for short-term wind forecasting across two distinct wind regimes. Wind speed data from South African monitoring stations were used, clustered based on geographic coordinates and elevation. This regime-aware clustering aims to capture location-specific wind patterns and topographic effects that influence forecast accuracy.

Clusters: Two clusters were identified:

- Cluster 1: Low-to-mid elevation, coastal stations.
- Cluster 2: High elevation, interior stations.

Models Compared:

- Linear regression baseline (mean model).
- Standard GP models on raw data.
- Hybrid linear–GP model (GP applied to residuals of the linear model).

GP Specifications: Candidate kernels included Gaussian and Matérn variants; hyperparameters were optimised via grid search. Covariance selection was performed adaptively for each cluster to match residual structures.

Validation and Metrics: Models were evaluated using three-fold temporal validation splits. Point forecasts were evaluated using RMSE to quantify prediction accuracy, while probabilistic forecasts were assessed through their predictive variance and reliability, capturing the uncertainty associated with each prediction.

4. Results

In this section, the exploratory data analysis will be outlined and evaluated. R statistical software version 4.5.2 was used for the data analysis.

4.1. Exploratory Data Analysis (EDA)

As part of analysing wind speed data, exploratory data analysis (EDA) uses statistical measures and various forms of data visualisation. EDA, in particular, becomes im-

portant when forming spatio-temporal data models, as it reveals inherent data trends, structures, and patterns, providing the foundation for selecting an ideal model and specifying parameters.

Clustering

In spatio-temporal predictions, clustering the locations balances efficiency, accuracy, and interpretability. This makes clustering a powerful strategy for large-scale forecasting tasks such as weather, traffic, and energy production.

Hopkins Test

The Hopkins statistic is a method for evaluating whether a dataset has a meaningful clustering structure, i.e., its tendency to form clusters. It is often used before applying clustering algorithms to spatial data. The Hopkins test is used to assess the tendency for clustering in a dataset.

Null Hypothesis (H_0). *The data is uniformly randomly distributed in the feature space (i.e., there is no meaningful cluster structure).*

Alternative Hypothesis (H_1). *The data is not uniformly random and exhibits a clustering tendency.*

Decision Rule:

- If the Hopkins statistic $H \approx 0.5$, we fail to reject H_0—the data is likely random.
- If $H \gg 0.5$ (typically $H > 0.75$), we reject H_0—the data shows a significant cluster tendency.

Since the Hopkins statistic is $H = 0.6680092 \gg 0.5$, we reject the null hypothesis and conclude that the data is clusterable.

Clustering

Since we established that our data is clusterable, the next step is to estimate the optimal or maximum number of clusters that can be meaningfully formed. The silhouette measures how well each point fits within its own cluster relative to other points. It provides a more quantitative evaluation and suggests that two clusters can be used for the data. Table 2 shows the stations grouped into two clusters according to their geographic location and elevation.

Table 2. Station data with clusters.

Stn No	Station Code	Station Name	Lon	Lat	Elev	Cluster
1	WM1	Alexander Bay	16.66441	28.60188	152	1
2	WM2	Calvinia	19.36075	31.52494	824	1
3	WM3	Vredendal	18.41992	31.73051	242	1
4	WM5	Napier	19.69245	34.61192	288	1
5	WM6	Sutherland	20.69124	32.55680	1581	2
6	WM7	Beaufort West	22.55667	32.96672	1047	2
7	WM8	Humansdorp	24.51436	34.10997	110	1
8	WM9	Noupoort	25.02838	31.25254	1806	2
9	WM12	Eston	30.52871	29.85026	770	2
10	WM19	Upington	20.56833	27.72670	848	1

Spatial distribution of clusters

The scatterplot Figure 2 illustrates the spatial distribution of clusters based on station coordinates in terms of latitude and longitude. The two clusters are visually dis-

tinct—cluster 1 (shown in red) is primarily located in the western region with lower longitude values. In contrast, cluster 2 (shown in blue) is positioned toward the eastern part of the study area. Lastly, the coverage of the spread of cluster 1 stations reveals a rather wide latitude range, demonstrating greater dispersion, unlike the cluster 2 stations, which are few and geographically concentrated. This spatial differentiation is reflected in the clusters obtained, underscoring the importance of the geographical factor in their formation and highlighting climatic differences among the WASA stations.

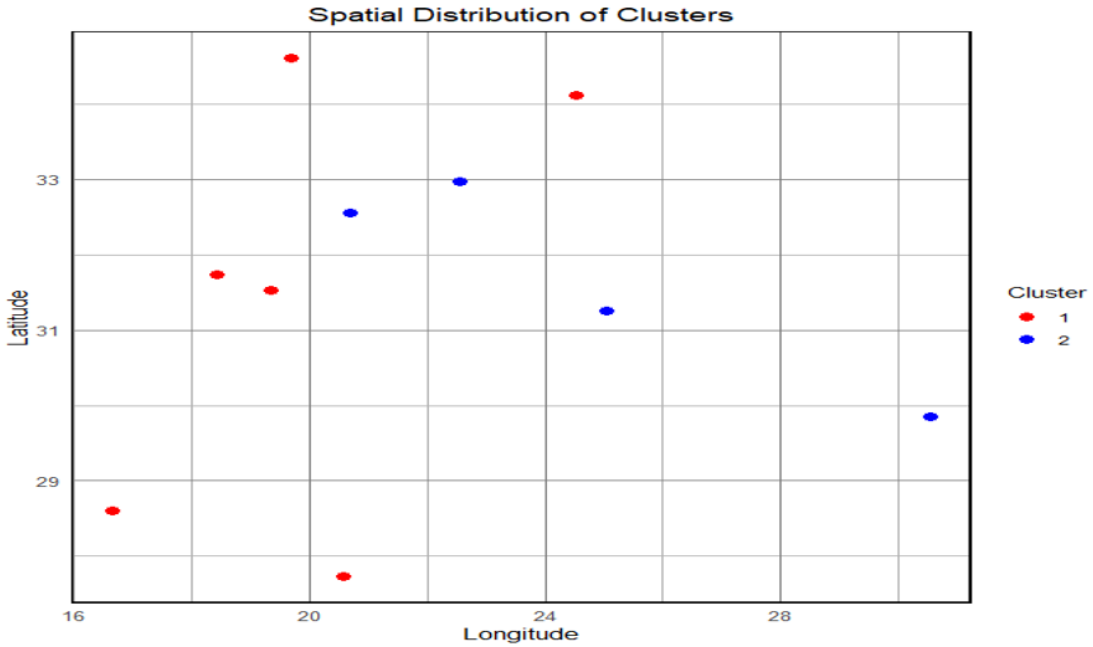


Figure 2. Spatial distribution of station clusters.

Figure 3 displays the spatial and topographical characteristics of two clusters based on their longitude, latitude, and elevation status. The longitude plot shows that stations in cluster 1 are primarily in a lower-latitude zone, whereas those in cluster 2 are located farther east, with higher longitudes. The plot showing the distribution based on latitude shows that while cluster 1 covers a broader zone, extending further into the northern latitudes, cluster 2 stations are concentrated within a narrower zone of lower latitudes. Finally, based on elevation, a notable difference between the two clusters is observed in Figure 3. In this regard, stations within cluster 1 are concentrated in lower-elevation zones, whereas those within cluster 2 are concentrated in higher-elevation zones. Overall, Figure 3 shows notable differences between clusters, thus supporting a spatial coherence between clusters based on longitude, latitude, and elevation status.

Figure 4 includes two clusters, 1 and 2, with each cluster included in a polygon that denotes the cluster boundary. Here, Dim1 accounts for 43.7% of the variance, while Dim2 accounts for 34%. These values add up to over 77%. Thus, it is evident that this plot includes most of the clusters. In Figure 4, it is evident that the spatial distribution of the clusters is demonstrated through hierarchical methods. All the clusters generated by the dendrogram are clearly separated, demonstrating that hierarchical clustering effectively distinguishes between them. In this plot, it is evident that the clusters are well separated in Dim1, indicating that it is an important dimension for distinguishing between them.

Figure 5 shows a dendrogram for clustering ten different stations labelled from 1 to 10. The vertical axis is labelled "Height," indicating the degree of dissimilarity or the distance at which clustering occurs. The degree of dissimilarity is higher for points clustered at higher positions than for those clustered at lower positions. From the given dendrogram,

two clusters are marked within dashed boxes. The first cluster consists of stations 9, 8, 5, and 6, whereas the second cluster consists of stations 1, 10, 2, 3, 4, and 7.

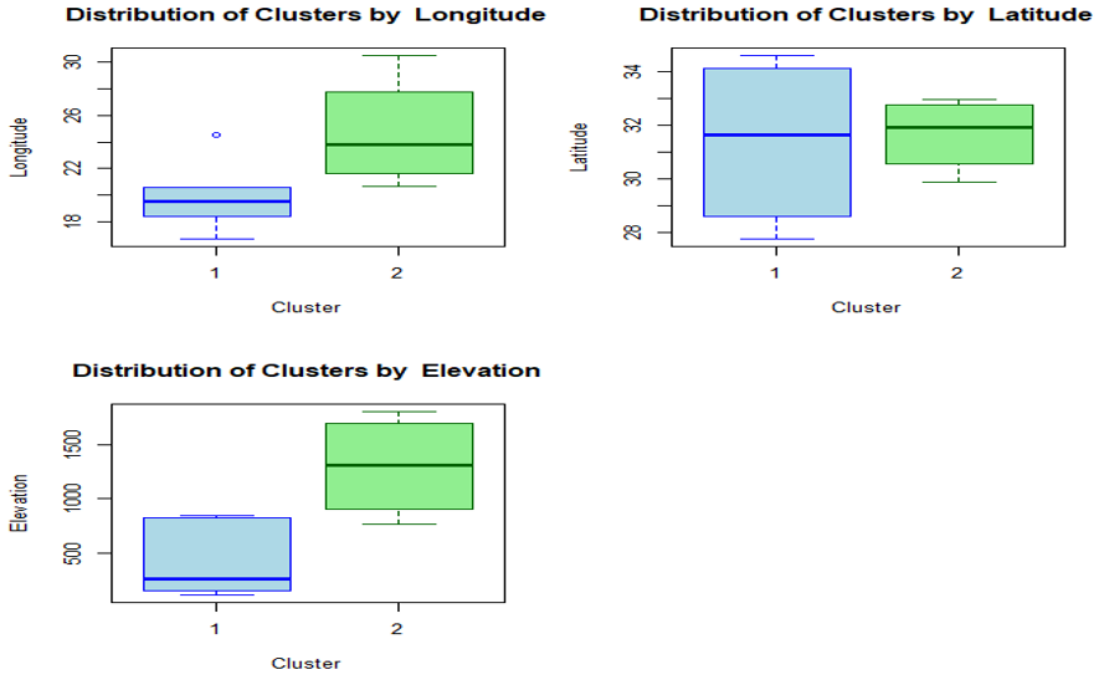


Figure 3. Distribution of clusters by coordinates and elevation.

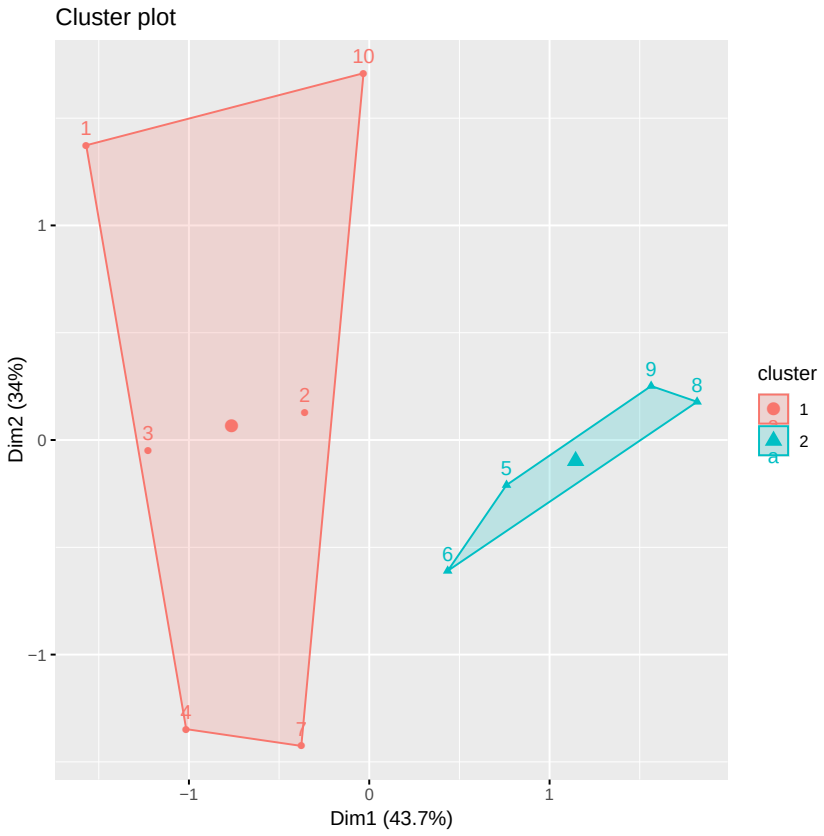


Figure 4. Hierarchical cluster plot.

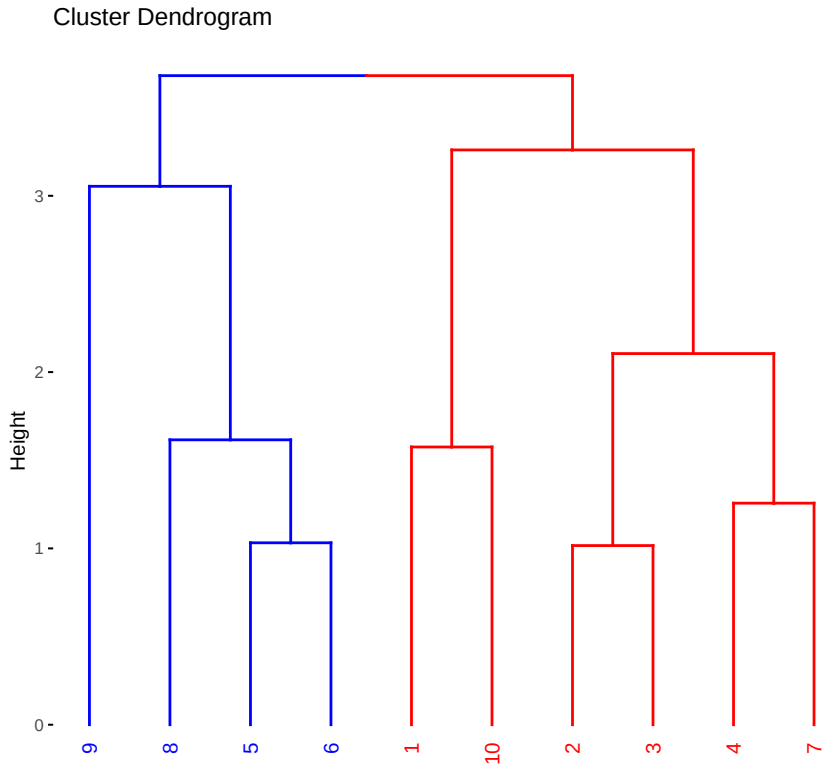


Figure 5. The dendrogram.

Multivariate exploratory data analysis

Table 3 lists the summary statistics of wind data across stations. Many stations exhibit moderate to high variability, with median wind speeds typically 6–7.5 m/s. The highest wind activity comes from stations 3 and 4, while Station 9 is the calmest. The maximum wind speed is close to 24 m/s, suggesting occasional high-wind events.

Also, the average values across the stations are quite different as indicated in Table 3. Alternatively, for Station 1 (stn1), the mean is higher than the median, suggesting the data may be right-skewed. Also, the difference between Q3 and the maximum is very high, suggesting the presence of outliers or a long tail.

Table 3. Summary statistics for wind speed data (stn1 to stn10).

Statistic	stn1	stn2	stn3	stn4	stn5	stn6	stn7	stn8	stn9	stn10
Min	0.3818	0.2533	0.3869	0.2075	0.2406	0.6305	0.2344	0.4235	0.2112	0.6037
1st Qu.	4.4349	3.3906	4.7480	4.9616	4.4436	4.6944	4.8155	5.3201	3.1822	4.4880
Median	7.1031	5.4118	7.5978	7.5278	6.1062	7.1340	6.7533	7.4886	4.8019	6.0382
Mean	7.5087	5.8444	7.8269	7.9377	6.6211	6.9727	7.0867	7.8101	5.0163	6.1670
3rd Qu.	9.9906	7.5868	10.7213	10.9338	8.4525	8.9410	9.2498	10.0568	6.4848	7.6730
Max	21.6708	20.5868	23.9921	20.3895	18.5664	17.7714	20.0661	19.3474	18.4819	16.5804

Figure 6 is a boxplot showing the distribution of wind speed (in m/s) across 10 stations (stn1 to stn10). Each box shows the central 50% of the data (interquartile range), the line within the box shows the median wind speed, and the dots represent outliers. Stations such as stn4, stn6, and stn7 exhibit higher wind speeds and greater variability, while others, such as stn1, stn2, and stn8, have lower, more consistent wind speeds.

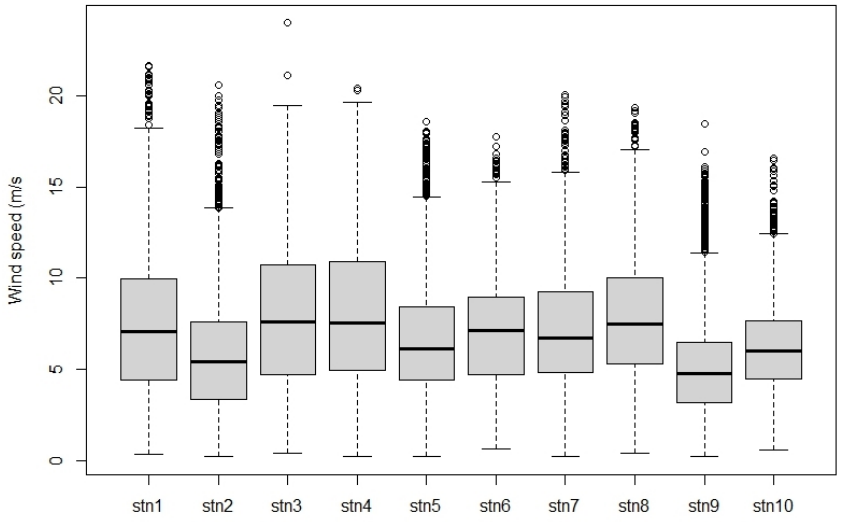


Figure 6. Boxplot of the distribution of wind speed (in m/s) across all stations.

The plot in Figure 7 is a time-series line plot showing the wind speed (in m/s) recorded at 10 different stations over a series of observations. A sample of 500 was taken to improve visualisation of wind speed. Each line represents the variation in wind speed at a specific station, showing temporal fluctuations and variability. The overlapping lines illustrate how wind speed trends differ and coincide between stations during the observed period.

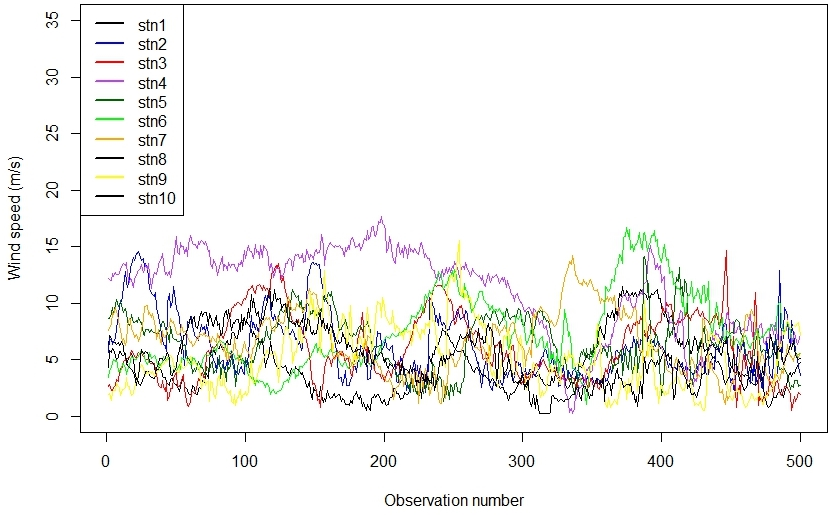


Figure 7. Time-series plot for all the stations.

Figure 8 shows a scatterplot matrix combined with histograms and correlation coefficients for wind speed data. This plot is useful for understanding the pairwise relationships and dependencies among wind speed measurements across different stations. The histograms show the wind speed distribution for each station. The scatterplots show pairwise relationships between wind speeds at different stations. Each plot has a fitted smooth curve (blue line) to show trends. Correlation coefficients between wind speeds at pairs of stations are also presented. These values quantify the strength and direction of linear relationships (positive, negative, or near zero). The correlation coefficients are mostly positive but low to moderate (generally below 0.4), indicating a weak linear association. Stations such as stn1 and stn2, or stn2 and stn3, have somewhat higher correlations.

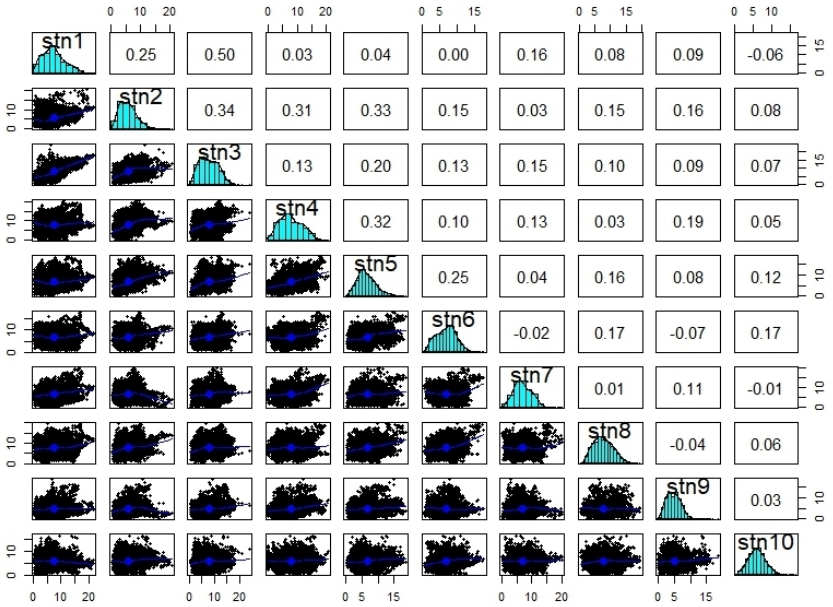


Figure 8. Correlation plot for all the stations.

Figure 9 gives a clear visualisation of how wind speed is distributed at each station. Most stations experience wind speeds typically between 0 and 15 m/s, with varying frequencies. Differences in histogram shape suggest variations in local wind speed patterns across stations.

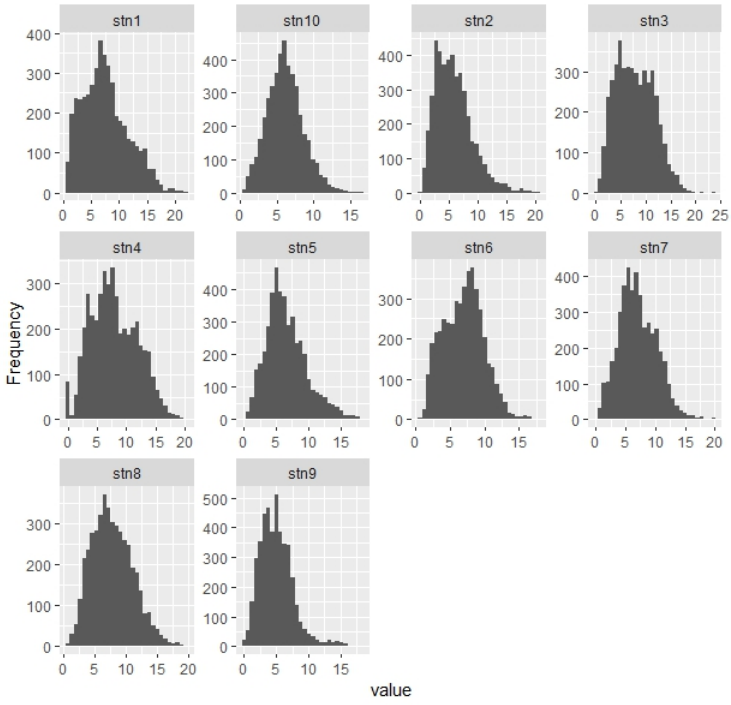


Figure 9. Histogram of all the stations.

4.2. Results for the Models

The results for the models are presented in the following sections in two stages. Firstly, Section 3.1 presents the results for the linear and Gaussian process (GP) models applied to all possible combinations of training and testing stations, providing a comprehensive evaluation of predictive performance and metrics such as RMSE, NLPD, CVG, and acceptance

rate ϕ. These results establish the baseline behaviour of the GP models across clusters and highlight the effects of station selection and spatial coverage.

In contrast, Section 3.2 evaluates the hybrid linear–Gaussian process framework. Here, a linear regression model is first estimated to capture large-scale spatio-temporal trends, and a GP is subsequently applied to the residuals to model local deviations. Within this second stage, a variety of covariance structures are rigorously assessed, including Matérn functions with smoothness parameters $\nu = 0.5, 1.5, 2.5$, as well as exponential and Gaussian covariances. The final covariance choice is determined based on validation performance. Because the hybrid model explicitly separates linear and residual components, the train–test splits, validation folds, and reported metrics differ from Section 3.1. Metrics such as ϕ are not always relevant for residual modelling, and certain summary metrics (e.g., CRPS) are computed only where meaningful for the hybrid framework. This structure ensures that the evaluation in Section 3.2 reflects the performance of the hybrid approach in capturing both large-scale trends and local residuals rather than replicating the exhaustive GP-only analysis.

4.2.1. Results for the Linear and GP Models

Variable selection

Relaxed Lasso is a variation of the Lasso (Least Absolute Shrinkage and Selection Operator) regression method, designed to improve variable selection by reducing the bias introduced by Lasso's strong penalty. In the first step, Lasso is applied to select a subset of relevant variables by shrinking some coefficients exactly to zero. Then, in the second step, a less-penalised or unpenalised regression (like OLS) is fit only on the selected variables, allowing the model to estimate their coefficients more accurately. This two-stage process helps maintain Lasso's variable selection strength while improving the accuracy and interpretability of the prediction by relaxing the amount of shrinkage applied.

Forecast evaluation

We perform modelling using the spTimer package in R, and we use an MCMC approach to estimate the posterior distributions of the model parameters. The following MCMC settings were used: number of iterations (N): 2000; burn-in period: 300 iterations; thinning interval: 1; number of chains: 3; and reporting interval (report): 3. Such settings mean that the program was run for 2000 iterations, of which the initial 300 were considered as burn-in to minimise the impact of initial values. Thinning of 1 means that the entire set of post-burn samples is utilised for posterior analysis. Running the chains enables convergence diagnosis, ensuring proper parameter evaluation. The report = 3 settings ensure that the algorithm reports every third iteration, thus enabling tracking of the execution of the sampling function. Such settings ensure an appropriate trade-off between computational efficiency and the sample size required to approximate the posterior distribution.

Evaluation for cluster 1

Training and testing for cluster 1

In cluster 1, we use the ratio $\frac{6}{2} = 3$ to generate all possible combinations of data for the test set, with test set sizes ranging from one to three samples. We used the configurations $\binom{6}{1} = 6$, $\binom{6}{2} = 15$ and $\binom{6}{3} = 20$, which resulted in 41 unique train–test partitions. In each configuration, the selected subset of stations served as the test set and the remaining stations were used for training. The objective of this exhaustive enumeration is to evaluate the robustness of spatial generalisation with respect to the specific choice of held-out stations. By considering every possible hold-out combination, we eliminate randomness due to fold assignment and provide a complete assessment of how sensitive model performance is

to station selection. The combinations resulting from the testing and training sets are in Table A1 in the Appendix.

Results for cluster 1

Table 4 summarises predictive performance across all 41 spatial hold-out configurations. When a single station is held out (leave-one-station-out validation), the model is trained on five stations and evaluated on one. As expected, this setting yields the lowest RMSE, MAE, NLPD, and CRPS values because the training set is largest and the extrapolation domain is minimal. When two stations are held out, predictive performance degrades moderately across all proper scoring rules. This reflects both the reduced amount of spatial information available for training and the increased geographic extent of the test domain. The most challenging setting occurs when three stations are held out. In this case, the model is trained on only half of the available stations and must extrapolate to a substantially larger unseen spatial region. Consequently, RMSE, MAE, CRPS, and NLPD increase consistently in relation to cases of one and two stations.

This monotonic degradation pattern indicates that performance is sensitive to the degree of spatial data removal rather than to any specific station combination. Importantly, coverage (CVG) remains stable across configurations, suggesting that predictive uncertainty is well-calibrated even as extrapolation difficulty increases. The variability observed within each group (for example, across different three-station subsets) reflects heterogeneity in spatial information content across stations, highlighting that some station combinations are inherently more informative for spatial interpolation than others. Therefore, Table 4 quantifies how predictive accuracy deteriorates as the spatial extrapolation task becomes more challenging. In this sense, the table illustrates the sensitivity of model performance to the degree of spatial data removal, providing a structured robustness assessment of spatial extrapolation performance. The metric ϕ quantifies the relative contribution of each train–test combination within the fusion scheme. Specifically, for combination j, ϕ_j is defined as

$$\phi_j = 100 \times \frac{w_j}{\sum_k w_k}, \tag{23}$$

where w_j represents the weight or influence assigned to combination j based on its predictive performance. CVG and other metrics are used to determine w_j in a manner that emphasizes well-performing combinations. The resulting ϕ_j values are expressed as percentages, highlighting which combinations contribute the most to the overall fused prediction. It represents the acceptance rate in the GPR fusion scheme, highlighting the train–test combinations that contribute most reliably to the final predictive ensemble. The metric ϕ further highlights the combinations that contribute the most within the fusion scheme, with certain station sets, such as 2, 3, 3–5, 3–6, and 2–5–6, receiving higher relevance scores, suggesting that these combinations provide more informative predictors for the model.

Prediction plots

Pairwise combinations tend to report higher NLPD, wider prediction intervals, and greater RMSEs as model complexity increases due to greater spatial variability. This phenomenon is reinforced when visualised within Figure 10, where combined stations report wider bands of conformal prediction and less smooth trends for the mean trajectories, while the prediction intervals report perfectly well-calibrated performance across the different combinations, reporting the desired true coverage of 95% as affirmed by the CVG values reported within Table 4.

Table 4. Performance metrics for different station combinations (compact view).

Combination	NLPD	PINAW	PINAD	RMSE	MAE	CRPS	CVG	ϕ (%)
1	1.9772	0.2690	−0.0156	1.4581	1.1639	0.8857	0.9500	22.0000
2	1.8799	0.2134	0.0367	1.2051	0.9906	0.7701	0.9500	92.0500
3	1.8407	0.1984	−0.0194	1.2138	0.9260	0.7491	0.9500	67.5500
4	2.0453	0.3163	−0.0301	1.6529	1.3192	0.9794	0.9500	22.3500
5	1.9173	0.2382	0.0035	1.2164	0.9596	0.7881	0.9500	26.3500
6	1.9280	0.3274	0.0339	1.3307	1.0546	0.8254	0.9500	24.3500
1–2	2.1802	0.2640	0.0161	1.4966	1.2269	1.0074	0.9500	41.4400
1–3	2.3404	0.4258	−0.0366	2.4553	1.9306	1.3918	0.9500	23.1500
1–4	2.2288	0.3677	−0.0321	1.9697	1.5561	1.1681	0.9500	23.0500
1–5	2.2467	0.3432	−0.0107	1.8504	1.4508	1.1399	0.9500	20.2500
1–6	2.1970	0.3022	0.0059	1.6391	1.3080	1.0537	0.9500	18.5000
2–3	2.1310	0.2640	0.0088	1.6682	1.3634	1.0251	0.9500	18.4500
2–4	2.3002	0.4221	0.0162	2.2910	1.8898	1.3245	0.9500	20.4000
2–5	2.1372	0.2512	0.0273	1.3396	1.0761	0.9412	0.9500	32.8000
2–6	2.2779	0.3443	0.0768	1.9823	1.7042	1.2087	0.9500	32.3500
3–4	2.1903	0.3035	−0.0388	1.8522	1.4662	1.1113	0.9500	53.7500
3–5	2.1291	0.2397	−0.0206	1.3993	1.0840	0.9480	0.9500	36.7500
3–6	2.1276	0.2310	0.0016	1.4057	1.0887	0.9475	0.9500	40.0000
4–5	2.4069	0.5244	−0.0276	2.6420	2.1088	1.5004	0.9500	23.2500
4–6	2.3480	0.4647	−0.0063	2.4049	1.9453	1.3858	0.9500	19.2000
5–6	2.3853	0.5165	0.0261	2.5984	2.0662	1.4700	0.9500	26.7000
1–2–3	2.6265	0.5341	0.0010	3.2953	2.6461	1.8635	0.9500	25.9500
1–2–4	2.4450	0.4465	−0.0037	2.4452	1.9569	1.4516	0.9500	20.5500
1–2–5	2.4384	0.3841	0.0179	2.1756	1.7505	1.3661	0.9500	18.9500
1–2–6	2.4819	0.4118	0.0482	2.3775	1.9584	1.4631	0.9500	19.4500
1–3–4	2.5240	0.5029	−0.0570	2.9707	2.3361	1.6811	0.9500	23.3000
1–3–5	2.5704	0.5219	−0.0346	3.0764	2.4197	1.7442	0.9500	23.6000
1–3–6	2.4954	0.4413	−0.0099	2.5856	2.0518	1.5300	0.9500	22.0000
1–4–5	2.4914	0.4787	−0.0374	2.5171	1.9611	1.5019	0.9500	21.1500
1–4–6	2.4247	0.3955	−0.0178	2.1086	1.6635	1.3356	0.9500	20.0000
1–5–6	2.4990	0.4158	0.0050	2.2558	1.7919	1.4351	0.9500	26.4000
2–3–4	2.4937	0.4780	−0.0137	2.9242	2.3500	1.6535	0.9500	20.9500
2–3–5	2.3653	0.3322	0.0066	2.0923	1.6908	1.2898	0.9500	18.4500
2–3–6	2.4715	0.4477	0.0390	2.8127	2.3023	1.6068	0.9500	24.3000
2–4–5	2.5664	0.5905	0.0107	3.0813	2.5013	1.7589	0.9500	26.6500
2–4–6	2.7089	0.6601	0.0401	3.5722	2.9373	2.0517	0.9500	25.0000
2–5–6	2.5753	0.5879	0.0690	3.1232	2.5483	1.7855	0.9500	27.8000
3–4–5	2.4724	0.4564	−0.0501	2.6810	2.0913	1.5440	0.9500	21.0000
3–4–6	2.4396	0.3782	−0.0241	2.2505	1.7764	1.3871	0.9500	19.4000
3–5–6	2.4585	0.4047	−0.0005	2.4462	1.9551	1.4616	0.9500	24.6500
4–5–6	2.6000	0.5781	−0.0017	2.9612	2.3805	1.7331	0.9500	27.7500

Importantly, the adaptive extension of the prediction interval in the regime of high wind variability, evident from Figure 10, shows the capability of the split conformal method in handling the dynamic uncertainty without sacrificing the guarantees. These examples reveal an interesting balance between the sharpness of predictions and the level of aggregation, indicating that although aggregation can be valuable, it might also entail a penalty in the loss of local wind dynamics, which is of principal importance.

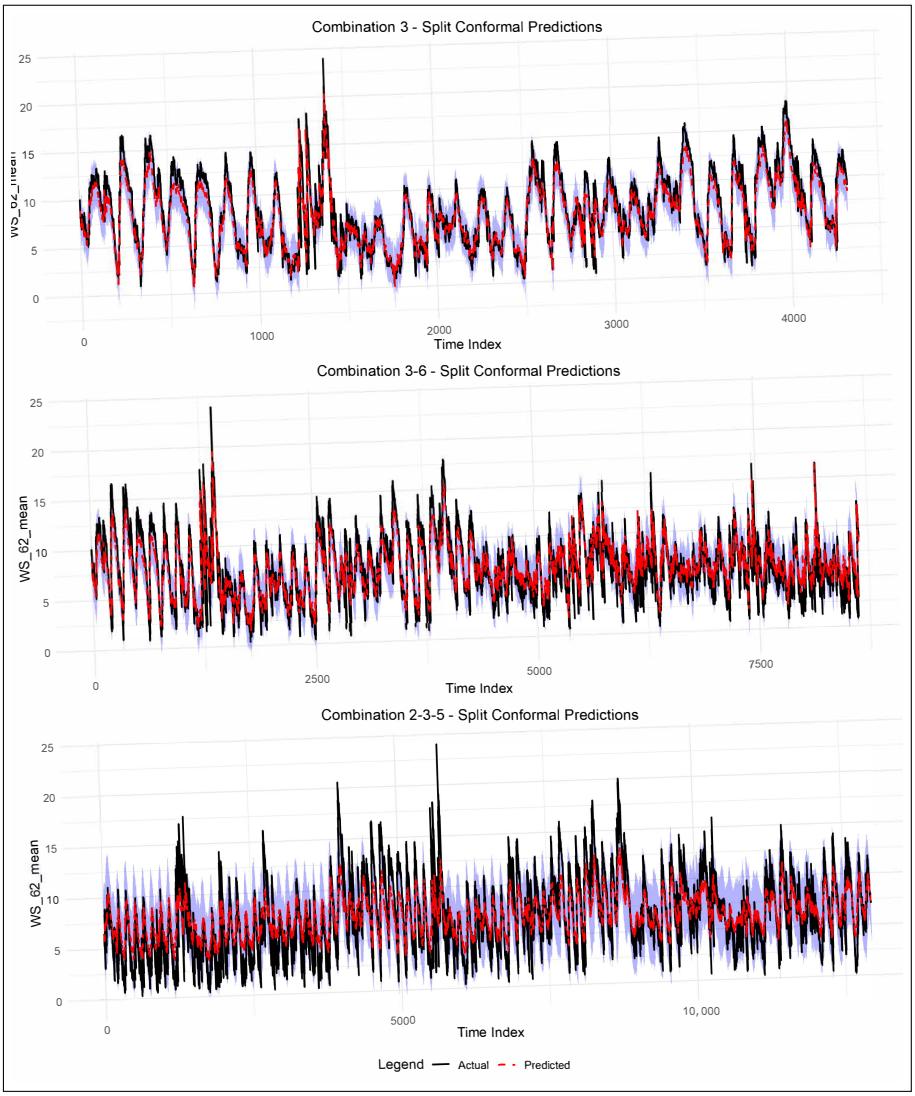


Figure 10. Prediction plots for cluster 1 high-performing stations.

Evaluation for cluster 2

Testing and Training sets

The training set is the data used to train the model, allowing it to learn patterns and relationships. In contrast, the testing set is a separate portion of data used to assess how well the model performs on new, unseen inputs. By splitting data into these two sets, we can ensure the model learns effectively without simply memorising the data, helping to avoid overfitting and giving a more accurate measure of how the model will perform in real-world scenarios. Table A2 in the Appendix shows the test–train sets for cluster 2.

To evaluate the model's performance and generalisation, two data-splitting strategies were employed, each using combinations of four spatial locations. We used the two configurations $\binom{4}{1} = 4, \binom{4}{2} = 6$, which denote combinations in which subsets of four components are split into training and testing sets. These combinations provide a systematic way to assess the model's robustness under varying data availability scenarios and to evaluate predictive performance across different spatial partitions.

Results for Cluster 2

Table A3 presents a comparative evaluation of the performance of the model in various training and testing configurations using key statistical metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), continuous ranked probability score (CRPS),

probability of prediction interval coverage (PICP), negative log predictive density (NLPD), normalised average width of prediction interval (PINAW), and normalised average deviation of prediction interval (PINAD).

From Table 5, among all configurations, Combination 1 achieves the best overall performance, yielding the lowest values of NLPD, RMSE, MAE, and CRPS, together with the narrowest prediction intervals (PINAW). Combination 2 shows comparable but slightly inferior performance, while Combination 3 and Combination 4 exhibit noticeably higher error metrics and wider prediction intervals. Pairwise station combinations generally lead to higher prediction errors and greater uncertainty than single-station cases, with Combination 2–3 showing the weakest performance across most metrics.

Table 5. Performance metrics for different combinations (best in bold).

Combination	NLPD	PINAW	PINAD	RMSE	MAE	CRPS	CVG	Phi
1	**1.8956**	**0.2527**	**0.0093**	**1.1498**	**0.8755**	**0.7592**	**0.9500**	**29.8000**
2	1.9205	0.2880	−0.0267	1.2615	1.0077	0.8033	**0.9500**	29.1000
3	2.1498	0.4514	−0.0635	2.0573	1.5824	1.1532	**0.9500**	21.7000
4	*2.4542*	*0.5583*	*0.1041*	*2.7843*	*2.3337*	*1.6078*	**0.9500**	27.3500
1–2	2.3199	0.5117	−0.0071	2.4423	1.9678	1.3852	**0.9500**	26.0500
1–3	2.3980	0.5917	−0.0579	2.6567	1.9977	1.4638	**0.9500**	21.4000
1–4	2.4421	0.5595	0.0794	2.7456	2.2495	1.5682	**0.9500**	28.1500
2–3	*2.7928*	*0.6799*	−0.0715	*3.1687*	*2.5026*	*1.8312*	**0.9500**	*23.5500*
2–4	2.4647	0.6033	0.0632	2.8752	2.3328	1.6309	**0.9500**	25.9500
3–4	2.4293	0.5452	0.0209	2.7039	2.1907	1.5378	**0.9500**	26.5000

Across all configurations, the empirical coverage (CVG) remains close to the nominal level of 0.95, indicating that the probabilistic forecasts are well-calibrated regardless of the station combination. The relative improvement metric Φ indicates consistent performance gains over the linear baseline, with the largest improvements observed for single-station configurations.

Figure 11 presents the split conformal prediction results for Combination 1 and Combination 1–2, while the corresponding quantitative performance metrics are reported in Table 5. For Combination 1, the predicted mean closely tracks the observed wind speed, with relatively narrow prediction intervals, consistent with the low RMSE (1.15), MAE (0.88), CRPS (0.76), and PINAW (0.25) values reported in the table. In contrast, Combination 1–2 exhibits wider prediction intervals and increased variability in the predicted mean, which aligns with the higher RMSE (2.44), MAE (1.97), CRPS (1.39), and PINAW (0.51) observed in Table 5. In both cases, the empirical coverage shown in Figure 11 is consistent with the nominal level, in agreement with the CVG value of approximately 0.95 reported in the table. The split conformal prediction method provides prediction intervals (shown as shaded regions) that effectively capture the variability in the observed wind speed data (black line). The predicted values (red dashed line) closely follow the actual measurements, and the uncertainty bands adapt over time, widening during more volatile periods and narrowing during more stable periods. This shows that split conformal prediction generates well-calibrated, dynamic-width intervals that reflect the true uncertainty in predictions across disparate time indices.

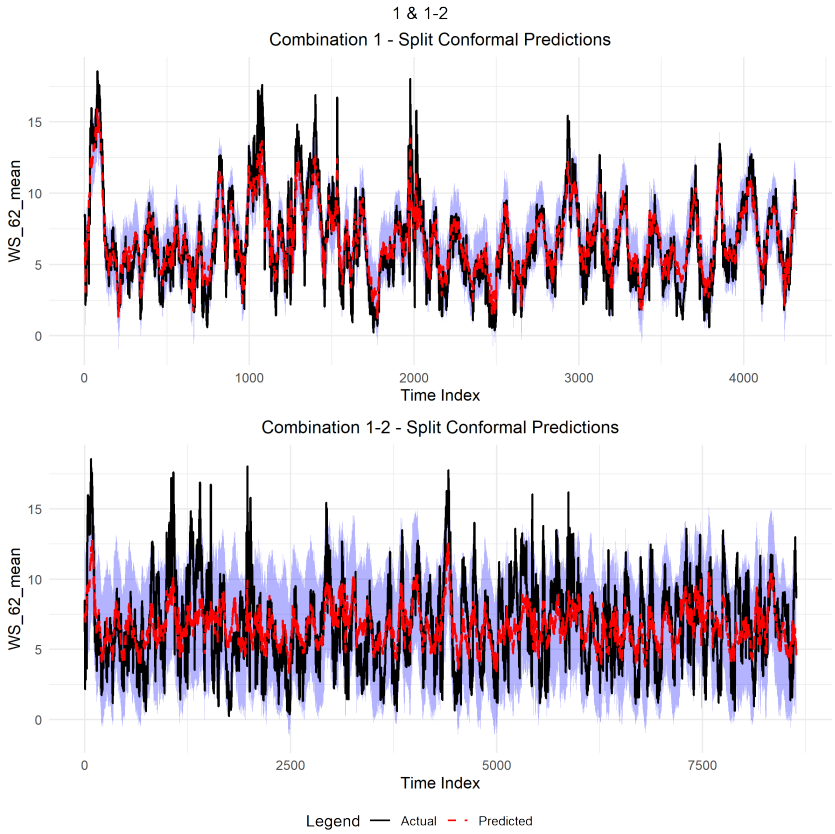


Figure 11. Prediction plots for cluster 2 high-performing stations.

4.2.2. Results for the Hybrid Linear–Gaussian Process

Cluster setup

The dataset used for the hybrid model was divided into two clusters based on elevation, longitude, and latitude to account for spatial heterogeneity in wind behaviour. Cluster 1 consists of lower to mid-elevation stations (152–848 m), while cluster 2 contains higher-elevation stations (770–1806 m). This is important for accounting for heterogeneity in wind speed dynamics. Model performance was evaluated separately for each cluster using multiple validation splits. This strategy allows for assessment of whether the proposed hybrid framework performs consistently across regimes with different statistical and physical characteristics.

Validation Setup

The proposed hybrid modelling framework was evaluated using a two-fold validation strategy. A leave-one-station-out cross-validation approach was used to evaluate model performance. In each iteration, data from one wind station were withheld for validation, while the remaining stations were used for model training. This procedure was repeated until each station had served as the validation site once. For cluster 1 (total sample size 25,914), each validation set comprised 16.7% of the available observations (4319 samples per split), and for cluster 2 (total sample size 17,276), each validation set comprised 25% of the available observations (4319 samples per split). For each split, a two-stage process was used. First, a linear regression model was estimated to account for large-scale spatio-temporal effects, followed by a GP analysis of the residuals. A variety of covariance structures were rigorously evaluated within the second stage, including several Matérn covariance functions with smoothness $u = 0.5, 1.5$, and 2.5; an exponential covariance; and a Gaussian covariance. The best-performing covariance structure is based on validation performance.

Results for Cluster 1: Weak or Unstable Residual Structure

Predictive Performance

In respect of cluster 1, the improved linear–Gaussian process (GP) hybrid model performance was reported as mixed, with varying levels of improvement depending on the residual errors it could model relative to the purely linear model. From the results presented in Table 6, the three distinct validation set RMSEs were 3.127, 21.779, and 2.828, while the linear model reported values were 3.125, 21.431, and 2.844. Here, the enhancements reported ranged from −1.622 to 0.576, indicating that this approach is indeed useful for modelling the nonlinear residuals. However, the overall improvement is limited by the varying spatial residuals in cluster 1. In one of the folds, it reports an improvement of 0.576, despite the overall spatial correlation in the residuals.

Table 6. Model performance metrics by clusters and validations.

Metric	C1 Val1	C1 Val2	C1 Val3	C2 Val1	C2 Val2
Covariance Type	Gaussian	Gaussian	Gaussian	Matern_0.5	Gaussian
NLPD	2.949	185.534	11.375	2.402	2.474
PINAW	0.510	2.511	0.465	0.539	0.490
PINAD	0.048	−1.053	−0.017	0.020	−0.019
RMSE	3.127	21.779	2.828	2.506	2.309
MAE	2.649	21.403	2.293	2.015	1.866
CRPS	1.931	20.763	2.000	1.429	1.345
CVG	0.950	0.950	0.950	0.950	0.950
Linear RMSE	3.125	21.431	2.844	2.994	2.386
GP Improvement (%)	−0.065	−1.622	0.576	16.279	3.242

Covariance Selection and Residual Structure

In all three model validation cases, the Gaussian (squared exponential) covariance was selected as optimal. This suggests that the residuals of cluster 1 could be best modelled as having a stationary covariance structure. Additionally, the negative improvement in the GP fold indicates that the residuals' spatial structure is neither strong nor stationary, resulting in little improvement from the GP method. Covariance-based feature selection cannot always adequately compensate for residual heterogeneity.

Probabilistic Forecast Quality

The hybrid method provided well-calibrated uncertainty estimates. The coverage (CVG) was kept at 0.95 for all the folds, and the PINAW values ranged from 0.465 to 2.511. Low PINAD values across most folds indicate little bias in interval centring. This shows that even when point prediction was not significantly improved, the GP component still provided good predictions; hence, decisions could be made accordingly.

Implications for Wind Speed Forecasting

The results for cluster 1 should be kept in mind for the practical application of the wind forecast method. In particular, the usefulness of the hybrid GP modelling method depends on the spatial coherence of the residuals; a lack of heterogeneity in the residuals could prevent efficient RMSE improvement within the clusters. The selection of covariance should be based on validation performance, but this may require non-stationary and/or multi-scale kernels for more accurate terrain-effect modelling. Probabilistic forecasts remain robust, and a small improvement in point predictions underscores the scope for risk-informed decision-making with hybrid models, especially in wind energy applications. The findings from cluster 1 highlight that modelling strategies should vary across clusters,

accounting for trends and residual spatial structures, especially in complex terrain and heterogeneous winds.

Results for Cluster 2: Strong Spatio-Temporal Dependence

Predictive performance

In both sets of validation data, the proposed hybrid linear–GP model significantly performed better than the linear-only model, validating the presence of underlying spatio-temporal correlation in the residual wind speed field (Table 6). For the first validation set, the proposed model showed improved performance, achieving an RMSE of 2.51, compared with the linear model, which achieved an RMSE of 2.99. The performance improvements in the proposed model were hence 16.3%. On the second set, the improvements were slightly lower, at 3.2%, as indicated by the reduced RMSE from 2.39 to 2.31.

Covariance Selection and Residual Structure

Different optimal covariance structures were selected based on the validation splits reported by the machine learning algorithm, using the automatic covariance structure selection method implemented in the code. In the first validation set, the Matérn covariance structure with $u = 0.5$ was consistently selected, while in the second set, the Gaussian covariance structure was preferred. The Matérn $u = 0.5$ covariance structure essentially corresponds to a rough, even non-differentiable, stochastic process, which is typically preferred for modelling abrupt changes in the underlying dynamics, for instance, those characterising localised variability in the wind field. The optimal structure found for the first set suggests that the dynamics of the residual wind field were characterised by particularly rough variability, potentially linked to factors such as gustiness and rapid regime changes. On the contrary, the optimal choice of the Gaussian kernel was observed in the second set, where the proportion of the original variability explained by the linear model was also particularly large.

Probabilistic Forecast Quality

Aside from point prediction accuracy, the hybrid models also demonstrated excellent probabilistic performance. Also, the Matérn $u = 0.5$ performed best in terms of the smallest negative log predictive density, well-centred prediction intervals, and almost nominal empirical coverage of about 95%. Even though the prediction intervals generated by the Gaussian kernel were narrower and the CRPS was slightly reduced, the model also showed some underdispersion, indicating overconfidence in the predictions. These findings reveal a balance between point prediction accuracy and probabilistic calibration, with rougher covariance structures yielding a more realistic quantification of uncertainty in wind speed predictions.

Implications for Wind Speed Forecasting

The results show the capability of hybrid linear-GP approaches with adaptive covariance selection in spatio-temporal prediction of wind speed while being physically consistent. The linear model demonstrates its capability to capture larger trends, while the GP model captures the localised stochastic component. The tendency toward rough covariance structures supports the non-smooth nature of the underlying wind speed processes, justifying the use of Matérn kernels in this context. The presented model offers greater prediction accuracy, more precise quantification of uncertainty, and greater interpretability of predictions than traditional linear models.

As shown in Figure 12, the figure compares the predictive performance of the linear and hybrid GP regression models across two data clusters and various validation sets. The left-hand side of the figure illustrates a representation of the Root Mean Square Error (RMSE) for the linear model and the hybrid GP model. In cluster 1, the RMSE improvements are inconsistent: one validation fold shows a small gain, while the others show

minor degradation relative to the linear model. In cluster 2, all validation folds show meaningful RMSE reductions with the GP component, emphasizing the strong spatio-temporal correlations in high-elevation stations. As is apparent, the RMSE is a vital measure that effectively quantifies error. As such, lower error values are desirable compared to higher ones. This figure effectively compares the relative performance of including the GP model in improving overall wind speed prediction across different data clusters. On the other hand, the right-hand side of Figure 12 demonstrates the percentage improvement in the hybrid GP model over the basic linear model.

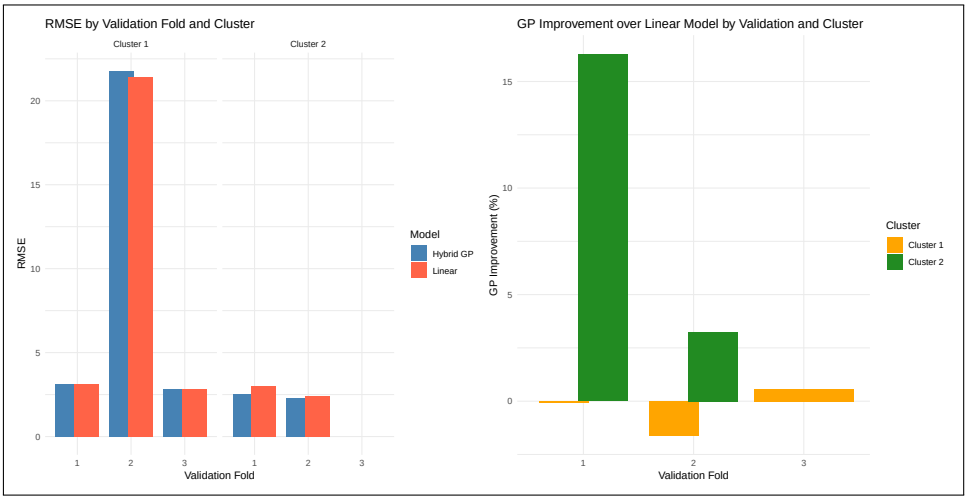


Figure 12. Comparison of predictive performance between linear and hybrid GP models across clusters and validation sets.

From the Appendix, Figure A1 presents the time-series plot for the independent variable wind speed. Figures A2–A4 show the results of the hybrid model from both clusters with the covariances selected as well as the improvement from the linear model.

5. Discussion

5.1. Discussion of Results

The suggested hybrid approach was evaluated on two different clusters based on wind speed data. The hybrid linear–Gaussian process approach showed different levels of performance across the two clusters distinguished by clustering.

Cluster 1 (low elevation/weak residuals): In the three validation groups from cluster 1, GP residual modelling yielded contrasting results. For validation 1 and 2, the use of the Gaussian process slightly undermined forecast accuracy, with RMSE changes of -0.065% and -1.622%, respectively. This finding agrees with the presence of high levels of noise in the residual terms in validation 2, as evidenced by the high NLPD value of 185.534. Only for validation 3, the GP showed slight gains (+0.576%) due to lower RMSE and MAE values. This suggests that, in a loosely structured wind regime, there is little to learn from residuals, and adding a Gaussian process will not necessarily improve the results. Predictive performance was still satisfactory overall, with well-calibrated probabilistic predictions, as indicated by a stable calibration of 0.95.

Cluster 2 (high altitude/residuals are strong): On the other hand, for cluster 2, the hybrid model always performed well. In the two validation rounds, the GP model reduced the RMSE by 16.279% and 3.242% compared to the linear model. Small PINAW and PINAD values were observed, indicating that the hybrid model's predictive intervals were well-calibrated. It is important to note that the covariance function mattered because it determined the model's predictive power. Specifically, the use of the $Matern_{\nu} = 0.5$ kernel

outperformed in validation 1, while validation 2 only needed a Gaussian kernel. The probabilistic criteria for the CRPS and NLPD indicated that the hybrid model provided accurate uncertainty estimates.

On the whole, these findings highlight the importance of regime sensitivity when evaluating the predictive power of GP residual modelling. While additional complexity will be meaningless for regions with low spatial correlation (cluster 1), it becomes highly valuable for regimes with high variability (cluster 2). In practical terms, these findings underscore the importance of regime-aware modelling methodologies. Overusing sophisticated stochastic models can lead to unnecessary computational costs and reduced forecasting accuracy.

5.2. Limitations and Future Work

Firstly, the assumption of stationarity within the clusters generated by the Gaussian process residual model does not account for regime shifts, which may be inherent to atmospheric processes. This problem might be especially relevant for cluster 1, based on the validation outcomes. The second potential drawback of the Gaussian process approach is its sensitivity to sparsity/outliers, which can be observed occasionally during model validation. It must be acknowledged that the problem can be alleviated by implementing adaptive covariance estimation. Finally, although the two-stage technique offers greater interpretability and computational efficiency, no physical constraints were incorporated into the model.

6. Conclusions

This study used a hybrid linear–Gaussian process model for short-term prediction of wind speed data. The proposed model was tested for clustering wind speed datasets based on elevation and geographic coordinates. The main results obtained show that the applicability of the hybrid model strongly relies on the properties of the cluster. In particular, residual modelling via Gaussian processes did not yield substantial benefits in cluster 1; on the contrary, in several validation exercises, it caused even minor performance losses due to the low-residual structure of the cluster.

On the other hand, in cluster 2, consistent advantages of using Gaussian processes were observed across both point and probabilistic predictions, with up to a 16.3% improvement in RMSE. Another important finding was the role of kernel selection and the need to adapt this process to each cluster, as evidenced by the superior performance of the adaptive Gaussian process over the adaptive Matérn kernel with $\nu = 0.5$. As for probabilistic prediction, good calibration was observed across all experiments (coverage probability of 0.95). This demonstrates that the hybrid model provides reliable uncertainty estimates suitable for operational decision-making.

Operationally speaking, the findings suggest that regime-specific modelling is valid. In other words, there is no need to employ Gaussian process residuals indiscriminately in regions with consistent wind regimes, where this might not only prove superfluous but even detrimental. At the same time, there is immense potential to utilise them in regions with prominent spatial variation. Overall, the hybrid approach is a balanced solution that maintains transparency, considers spatial dependence, and ensures accurate uncertainty estimation.

Author Contributions: Conceptualisation, T.H.T. and C.S.; methodology, T.H.T.; software, T.H.T.; validation, T.H.T., C.S., T.B.D. and S.N.; formal analysis, T.H.T.; investigation, T.H.T., C.S., T.B.D. and S.N.; data curation, T.H.T.; writing—original draft preparation, T.H.T.; writing—review and editing, T.H.T., C.S., T.B.D. and S.N.; visualisation, T.H.T.; supervision, C.S., T.B.D. and S.N.; project

administration, C.S., T.B.D. and S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data in brief can be accessed at https://github.com/csigauke/A-Hybrid-Linear--Gaussian-Process-Framework-for-Spatio-Temporal-Wind-Speed-Forecasting (accessed on 16 January 2025).

Acknowledgments: The authors thank the anonymous reviewers for their helpful comments on this paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

MAE	Mean Absolute Error
MASE	Mean Absolute Scaled Error
RMSE	Root Mean Squared Error
MCMC	Markov Chain Monte Carlo
CRPS	Continuous Ranked Probability Score
PICP	Probability of Prediction Interval Coverage
NLPD	Negative Log Predictive Density
PINAW	Normalised Average Width of Prediction Interval
PINAD	Normalised Average Deviation of Prediction Interval

Appendix A. Supplementary Results

Table A1. Testing and training set combinations (grouped into 6C1–6C2 and 6C3).

6C1 and 6C2		
ID	Testing Set	Training Set
1	1	2, 3, 4, 5, 6
2	2	1, 3, 4, 5, 6
3	3	1, 2, 4, 5, 6
4	4	1, 2, 3, 5, 6
5	5	1, 2, 3, 4, 6
6	6	1, 2, 3, 4, 5
7	1, 2	3, 4, 5, 6
8	1, 3	2, 4, 5, 6
9	1, 4	2, 3, 5, 6
10	1, 5	2, 3, 4, 6
11	1, 6	2, 3, 4, 5
12	2, 3	1, 4, 5, 6
13	2, 4	1, 3, 5, 6
14	2, 5	1, 3, 4, 6
15	2, 6	1, 3, 4, 5
16	3, 4	1, 2, 5, 6
17	3, 5	1, 2, 4, 6
18	3, 6	1, 2, 4, 5
19	4, 5	1, 2, 3, 6
20	4, 6	1, 2, 3, 5
21	5, 6	1, 2, 3, 4

Table A1. *Cont.*

6C3		
ID	Testing Set	Training Set
22	1, 2, 3	4, 5, 6
23	1, 2, 4	3, 5, 6
24	1, 2, 5	3, 4, 6
25	1, 2, 6	3, 4, 5
26	1, 3, 4	2, 5, 6
27	1, 3, 5	2, 4, 6
28	1, 3, 6	2, 4, 5
29	1, 4, 5	2, 3, 6
30	1, 4, 6	2, 3, 5
31	1, 5, 6	2, 3, 4
32	2, 3, 4	1, 5, 6
33	2, 3, 5	1, 4, 6
34	2, 3, 6	1, 4, 5
35	2, 4, 5	1, 3, 6
36	2, 4, 6	1, 3, 5
37	2, 5, 6	1, 3, 4
38	3, 4, 5	1, 2, 6
39	3, 4, 6	1, 2, 5
40	3, 5, 6	1, 2, 4
41	4, 5, 6	1, 2, 3

Table A2. Testing and training set combinations for cluster 2.

No.	Testing Set	Training Set
1	1	2, 3, 4
2	2	1, 3, 4
3	3	1, 2, 4
4	4	1, 2, 3
5	1, 2	3, 4
6	1, 3	2, 4
7	1, 4	2, 3
8	2, 3	1, 4
9	2, 4	1, 3
10	3, 4	1, 2

Table A3. Model performance for LM and GPR across station combinations.

s.index	Model	RMSE	MAE	CRPS	PICP	NLPD	PINAW
1	LM	0.8198	0.6321	0.4530	0.9454	1.2241	0.1728
	GP	1.1313	0.8616	1.3826	0.9990	1.8957	0.5070
2	LM	0.8283	0.6421	0.4585	0.9474	1.2346	0.1852
	GP	1.2536	1.003	1.3806	0.9990	1.9227	0.5429
3	LM	0.8199	0.6212	0.4495	0.9379	1.2241	0.1658
	GP	2.1206	1.6249	1.2953	0.9506	2.1743	0.4679
4	LM	1.2005	0.7890	0.5963	0.8946	1.8832	0.1708
	GP	2.8746	2.4061	1.3586	0.9215	2.4990	0.5225
1,2	LM	0.8202	0.6300	0.4521	0.9463	1.2245	0.1732
	GP	2.4598	1.9836	1.5065	0.9024	2.6392	0.5685
1,3	LM	1.5695	1.3962	1.0527	0.5676	2.8405	0.1596
	GP	2.1250	2.6561	1.9941	0.8330	2.9306	0.5220

Table A3. *Cont.*

s.index	Model	RMSE	MAE	CRPS	PICP	NLPD	PINAW
1,4	LM	1.1129	0.7302	0.5433	0.9166	1.7300	1.1689
	GP	2.7990	2.2896	1.7487	0.9099	2.7441	0.6694
2,3	LM	0.8231	0.6303	0.4536	0.9420	1.2285	1.6688
	GP	3.1901	2.5197	1.1572	0.7214	3.3072	0.4266
2,4	LM	1.0158	0.9910	0.5221	0.9263	1.5252	0.1717
	GP	2.9163	2.3631	1.4808	0.8449	2.8834	0.5617
3,4	LM	0.7123	0.7123	0.5266	0.9124	1.5011	0.1633
	GP	2.7110	2.1946	1.7376	0.9026	2.7573	0.6300

Hybrid model plots

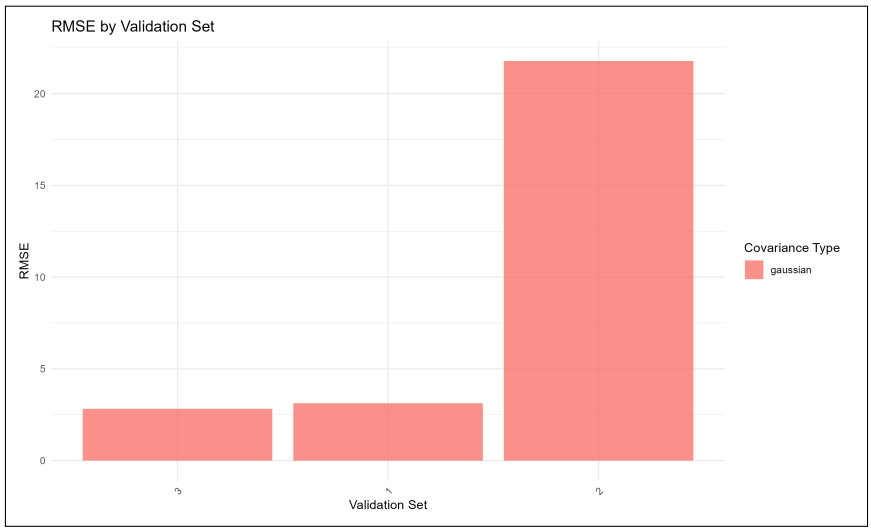


Figure A1. RMSE for cluster 1.

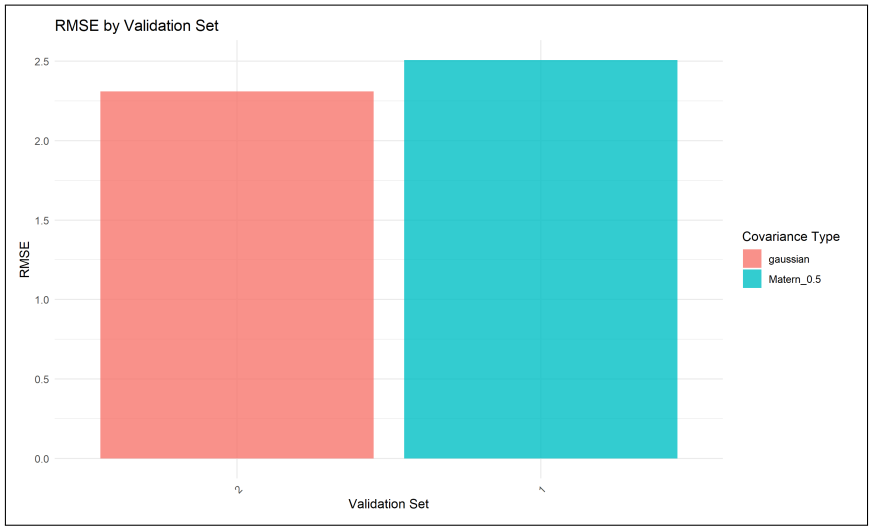


Figure A2. RMSE for cluster 2.

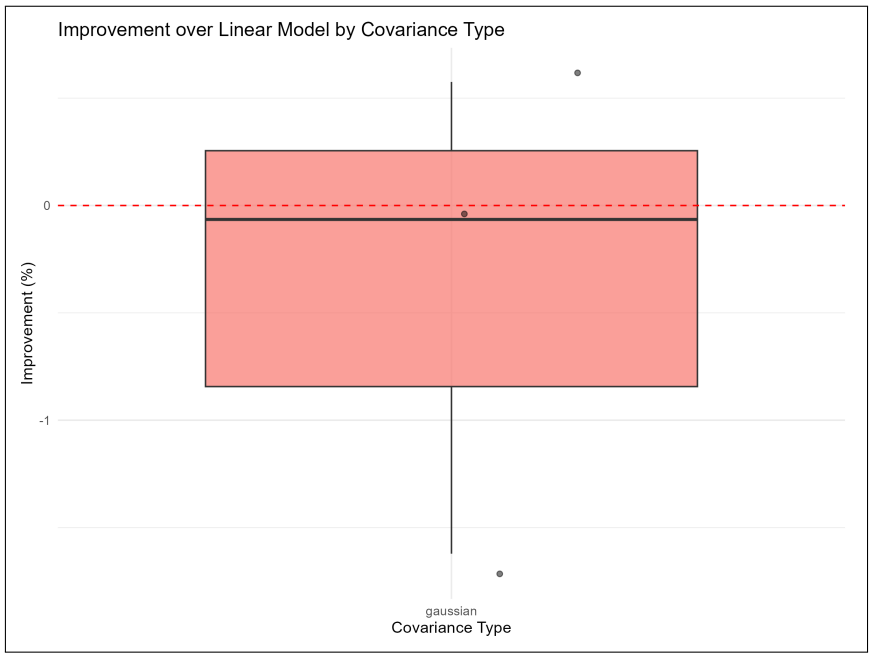


Figure A3. Improvement for cluster 1.

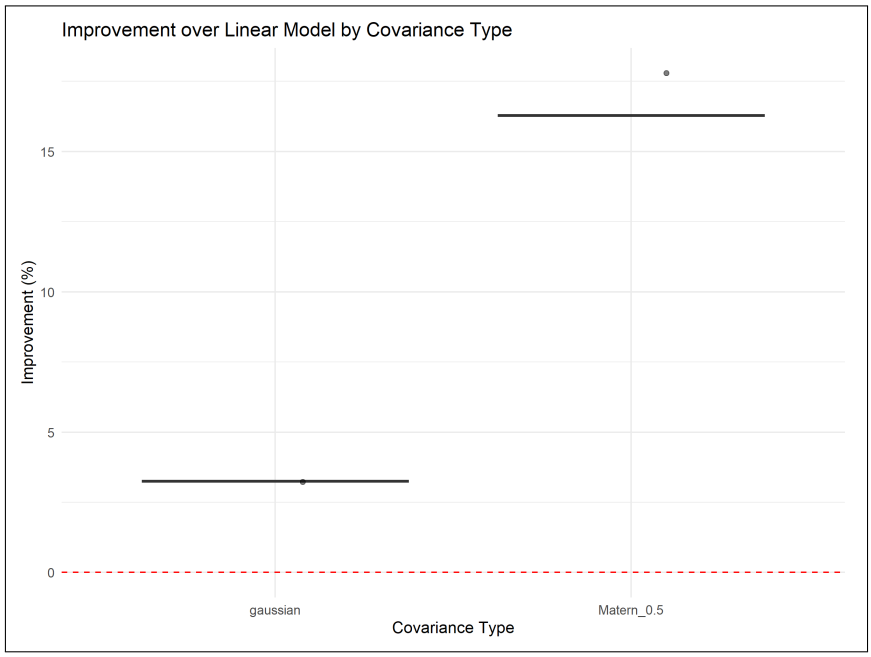


Figure A4. Improvement for cluster 2.

Appendix B. Modelling Specification: An Extensive Hierarchical Bayesian Formulation with MCMC Sampling

Appendix B.1. Hybrid Model Specification

Let $y(\mathbf{s}, t)$ denote the response variable wind speed at location $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^2$ and time $t \in \mathcal{T} \subset \mathbb{R}$. The complete model is

$$y(\mathbf{s}, t) = \underbrace{\mathbf{x}(\mathbf{s}, t)^\top \boldsymbol{\beta}}_{\text{Linear trend}} + \underbrace{f(\mathbf{s}, t)}_{\text{GP component}} + \epsilon(\mathbf{s}, t) \tag{A1}$$

where

- $\mathbf{x}(\mathbf{s}, t) \in \mathbb{R}^p$ is a vector of covariates;

- $\boldsymbol{\beta} \in \mathbb{R}^p$ are regression coefficients;
- $f(\mathbf{s}, t) \sim GP(0, K(\cdot, \cdot; \boldsymbol{\theta}))$ is a zero-mean Gaussian process;
- $\epsilon(\mathbf{s}, t) \sim N(0, \tau^2)$ is the measurement error (nugget effect);
- $K(\cdot, \cdot; \boldsymbol{\theta})$ is the covariance function with parameters $\boldsymbol{\theta}$.

Appendix B.2. Two-Stage Estimation Procedure

Appendix B.2.1. Stage 1: Linear Regression

The very first attempt at modelling a spatio-temporal response variable is to consider the linear regression models. First, estimate the linear component via ordinary least squares:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 \tag{A2}$$

The linear predictions are

$$\hat{y}_L(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^\top \hat{\boldsymbol{\beta}} \tag{A3}$$

Residuals for GP modelling:

$$r(\mathbf{s}, t) = y(\mathbf{s}, t) - \hat{y}_L(\mathbf{s}, t) \tag{A4}$$

Appendix B.2.2. Stage 2: Gaussian Process on Residuals

Model the residuals as follows:

$$r(\mathbf{s}, t) = f(\mathbf{s}, t) + \epsilon(\mathbf{s}, t) \tag{A5}$$

where $f(\mathbf{s}, t) \sim GP(0, K(\cdot, \cdot; \boldsymbol{\theta}))$.

Appendix B.3. Covariance Structures

Appendix B.3.1. General Form of Covariance Functions

For spatio-temporal modelling, we consider separable covariance functions:

$$Cov\big(f(\mathbf{s}, t), f(\mathbf{s}', t')\big) = \sigma^2 \cdot C_s(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}_s) \cdot C_t(t, t'; \boldsymbol{\theta}_t) \tag{A6}$$

where σ^2 is the marginal variance.

Appendix B.3.2. Matérn Family of Covariance Functions

The Matérn covariance function is defined as

$$K_{\text{Matern}}(h; \nu, \phi) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} h}{\phi} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} h}{\phi} \right) \tag{A7}$$

where

- $h = \|\mathbf{s} - \mathbf{s}'\|$ is the Euclidean distance;
- $\nu > 0$ is the smoothness parameter;
- $\phi > 0$ is the range parameter;
- $\sigma^2 > 0$ is the variance parameter;
- $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

Appendix B.3.3. Special Cases

1. Exponential covariance ($\nu = 0.5$):

$$K_{\text{Exp}}(h;\phi) = \sigma^2 \exp\left(-\frac{h}{\phi}\right) \tag{A8}$$

2. Matérn with $\nu = 1.5$:

$$K_{\nu=1.5}(h;\phi) = \sigma^2\left(1 + \frac{\sqrt{3}h}{\phi}\right)\exp\left(-\frac{\sqrt{3}h}{\phi}\right) \tag{A9}$$

3. Matérn with $\nu = 2.5$:

$$K_{\nu=2.5}(h;\phi) = \sigma^2\left(1 + \frac{\sqrt{5}h}{\phi} + \frac{5h^2}{3\phi^2}\right)\exp\left(-\frac{\sqrt{5}h}{\phi}\right) \tag{A10}$$

4. Gaussian/squared exponential ($\nu \to \infty$):

$$K_{\text{Gaussian}}(h;\phi) = \sigma^2 \exp\left(-\frac{h^2}{2\phi^2}\right) \tag{A11}$$

- Exponential covariance ($\nu = 0.5$) models rough processes with less smoothness.
- Gaussian covariance ($\nu \to \infty$) assumes very smooth processes.
- Matérn covariance with intermediate ν balances smoothness and flexibility.

Appendix B.3.4. Rational Quadratic Covariance

$$K_{\text{RQ}}(h;\alpha,\phi) = \sigma^2\left(1 + \frac{h^2}{2\alpha\phi^2}\right)^{-\alpha} \tag{A12}$$

where $\alpha > 0$ controls the smoothness.

Appendix B.4. Likelihood and Estimation

Appendix B.4.1. Gaussian Process Likelihood

For n observations $\mathbf{r} = (r_1, \ldots, r_n)^\top$, the log-likelihood is

$$\log p(\mathbf{r} \mid \boldsymbol{\theta}) = -\frac{1}{2}\left[n\log(2\pi) + \log|\mathbf{K}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}| + \mathbf{r}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \tau^2\mathbf{I})^{-1}\mathbf{r}\right] \tag{A13}$$

where $\mathbf{K}_{\boldsymbol{\theta}}$ is the $n \times n$ covariance matrix with entries $K_{\boldsymbol{\theta}}(\mathbf{s}_i, \mathbf{s}_j)$.

Appendix B.4.2. Maximum Likelihood Estimation

The MLE estimates are obtained by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{r} \mid \boldsymbol{\theta}) \tag{A14}$$

Appendix B.4.3. Predictive Distribution

For a new location $\mathbf{s}_0$, the predictive distribution is Gaussian:

$$p(r(\mathbf{s}_0) \mid \mathbf{r}) = N(\mu_0, \sigma_0^2) \tag{A15}$$

with

$$\mu_0 = \mathbf{k}_0^\top (\mathbf{K} + \tau^2 \mathbf{I})^{-1} \mathbf{r} \tag{A16}$$

$$\sigma_0^2 = K(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{k}_0^\top (\mathbf{K} + \tau^2 \mathbf{I})^{-1} \mathbf{k}_0 + \tau^2 \tag{A17}$$

where $\mathbf{k}_0 = [K(\mathbf{s}_0, \mathbf{s}_1), \ldots, K(\mathbf{s}_0, \mathbf{s}_n)]^\top$.

Appendix B.5. Hierarchical Bayesian Formulation

Appendix B.5.1. Complete Bayesian Model

$$\text{Likelihood:} \quad \mathbf{y} \mid \boldsymbol{\beta}, f, \tau^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}, \tau^2 \mathbf{I}) \tag{A18}$$

$$\text{GP Prior:} \quad \mathbf{f} \mid \sigma^2, \phi, \nu \sim N(\mathbf{0}, \mathbf{K}(\sigma^2, \phi, \nu)) \tag{A19}$$

$$\text{Priors:} \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \tag{A20}$$

$$\tau^2 \sim \text{IG}(a_\tau, b_\tau) \tag{A21}$$

$$\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma) \tag{A22}$$

$$\phi \sim \text{LN}(\mu_\phi, \sigma_\phi^2) \tag{A23}$$

$$\nu \sim \text{Gamma}(a_\nu, b_\nu) \tag{A24}$$

Appendix B.5.2. Posterior Distribution

$$p(\boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{f}, \tau^2) \cdot p(\mathbf{f} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\beta}) \cdot p(\boldsymbol{\theta}) \tag{A25}$$

Appendix B.5.3. Gibbs Sampling

1. Sample $\boldsymbol{\beta}$ (conjugate):

$$\boldsymbol{\beta} \mid \cdot \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \tag{A26}$$

$$\boldsymbol{\Sigma}_\beta = \left(\frac{1}{\sigma_\beta^2} \mathbf{I} + \frac{1}{\tau^2} \mathbf{X}^\top \mathbf{X} \right)^{-1} \tag{A27}$$

$$\boldsymbol{\mu}_\beta = \frac{1}{\tau^2} \boldsymbol{\Sigma}_\beta \mathbf{X}^\top (\mathbf{y} - \mathbf{f}) \tag{A28}$$

2. Sample $\mathbf{f}$ (conjugate):

$$\mathbf{f} \mid \cdot \sim N(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) \tag{A29}$$

$$\boldsymbol{\Sigma}_f = \left(\mathbf{K}^{-1} + \frac{1}{\tau^2} \mathbf{I} \right)^{-1} \tag{A30}$$

$$\boldsymbol{\mu}_f = \frac{1}{\tau^2} \boldsymbol{\Sigma}_f (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{A31}$$

3. Sample covariance parameters (Metropolis–Hastings):

$$\theta^{(t+1)} = \theta^{(t)} + \epsilon, \quad \epsilon \sim N(0, \sigma_{\text{mh}}^2) \tag{A32}$$

Accept with probability:

$$\alpha = \min\left(1, \frac{p(\mathbf{y} \mid \theta^*) p(\theta^*)}{p(\mathbf{y} \mid \theta^{(t)}) p(\theta^{(t)})}\right) \tag{A33}$$

Appendix B.6. Model Comparison and Selection

Appendix B.6.1. Information Criteria

$$\text{AIC} = 2k - 2\log \mathcal{L} \tag{A34}$$

$$\text{BIC} = k\log n - 2\log \mathcal{L} \tag{A35}$$

$$\text{WAIC} = -2\left(\sum_{i=1}^{n} \log E[p(y_i \mid \boldsymbol{\theta})] - \sum_{i=1}^{n} Var[\log p(y_i \mid \boldsymbol{\theta})]\right) \tag{A36}$$

Appendix B.6.2. Cross-Validation

For K-fold CV:

$$\text{CV} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}(y_i - \hat{y}_{-k,i})^2 \tag{A37}$$

Appendix B.7. Uncertainty Quantification

Prediction Intervals

For a new location $\mathbf{s}_0$, the predictive distribution of the response is Gaussian. The $(1-\alpha)$ prediction interval:

$$\hat{y}(\mathbf{s}_0) \pm z_{1-\alpha/2}\sqrt{Var[\hat{y}(\mathbf{s}_0)]} \tag{A38}$$

where $Var[\hat{y}(\mathbf{s}_0)] = \sigma_0^2 + \tau^2$. This provides both a point prediction (μ_0) and an uncertainty measure (σ_0^2) for the new location.

Appendix B.8. Conformal Prediction

1. Compute residuals: $e_i = |y_i - \hat{y}_i|, \quad i \in \mathcal{D}_{\text{cal}}$.
2. Find α-quantile: $q_{1-\alpha} = \text{Quantile}(1-\alpha; \{e_i\})$.
3. Prediction interval: $\text{PI}(\mathbf{s}_0) = [\hat{y}(\mathbf{s}_0) - q_{1-\alpha}, \hat{y}(\mathbf{s}_0) + q_{1-\alpha}]$.

Appendix B.9. Computational Aspects

Appendix B.9.1. Scalability Challenges

Exact GP: $\mathcal{O}(n^3)$. Solutions:

- Low-rank approximations: $\mathbf{K} \approx \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{nm}^{\top}$, $\mathcal{O}(nm^2)$.
- Sparse covariance methods.
- Stochastic Variational Inference.

Appendix B.9.2. Numerical Stability

- Add jitter: $\mathbf{K} + \delta\mathbf{I}$.
- Use Cholesky decomposition with pivoting.
- Compute log-determinant via Cholesky factors.

Appendix B.10. Theoretical Properties

Appendix B.10.1. Consistency

Under regularity conditions, as $n \to \infty$:

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{A39}$$

$$\hat{y}(\mathbf{s}) \xrightarrow{p} E[y(\mathbf{s}) \mid \mathcal{D}] \tag{A40}$$

$$\text{PI}_{1-\alpha}(\mathbf{s}) \xrightarrow{p} \text{True PI}_{1-\alpha}(\mathbf{s}) \tag{A41}$$

Appendix B.10.2. Asymptotic Normality

For fixed-domain asymptotics:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}) \tag{A42}$$

where $\mathcal{I}(\boldsymbol{\theta}_0)$ is the Fisher information matrix.

References

1. Yang, M.; Jiang, Y.; Che, J.; Han, Z.; Lv, Q. Short-Term Forecasting of Wind Power Based on Error Traceability and Numerical Weather Prediction Wind Speed Correction. *Electronics* **2024**, *13*, 1559. [CrossRef]
2. Szostek, K.; Mazur, D.; Drałus, G.; Kusznier, J. Analysis of the Effectiveness of ARIMA, SARIMA, and SVR Models in Time Series Forecasting: A Case Study of Wind Farm Energy Production. *Energies* **2024**, *17*, 4803. [CrossRef]
3. Xu, J.; Xiao, Z.; Lin, Z.; Li, M. System bias correction of short-term hub-height wind forecasts using the Kalman filter. *Prot. Control Mod. Power Syst.* **2021**, *6*, 37. [CrossRef]
4. Liu, Z.; Guo, H.; Zhang, Y.; Zuo, Z. A Comprehensive Review of Wind Power Prediction Based on Machine Learning: Models, Applications, and Challenges. *Energies* **2025**, *18*, 350. [CrossRef]
5. Che, G.; Zhou, D.; Wang, R.; Zhou, L.; Zhang, H.; Yu, S. Wind Energy Assessment in Forested Regions Based on the Combination of WRF and LSTM-Attention Models. *Sustainability* **2024**, *16*, 898. [CrossRef]
6. Bessa, R.J.; Miranda, V.; Gama, J. The Role of Predictive Distributions in Modern Wind Power Forecasting Frameworks. *Int. J. Forecast.* **2022**, *38*, 601–615. [CrossRef]
7. Elshaboury, N.; Elmousalami, H. Wind speed and power forecasting using Bayesian optimised machine learning models in Gabal Al-Zayt, Egypt. *Sci. Rep.* **2025**, *15*, 28500. [CrossRef] [PubMed]
8. Yang, Q.; Huang, G.; Li, T.; Xu, Y.; Pan, J. A Novel Short-Term Wind Speed Prediction Method Based on Hybrid Statistical-Artificial Intelligence Model with Empirical Wavelet Transform and Hyperparameter Optimization. *J. Wind Eng. Ind. Aerodyn.* **2023**, *239*, 105499. [CrossRef]
9. Cai, H.; Jia, X.; Feng, J.; Li, W.; Hsu, Y.-M.; Lee, J. Gaussian Process Regression for Numerical Wind Speed Prediction Enhancement. *Renew. Energy* **2020**, *146*, 2112–2123. [CrossRef]
10. Heaton, M.J.; Gelfand, A.E.; Vecchia, D.J.; Guinness, J.R. A Case Study Competition Among Methods for Analysing Large Spatial Data. *J. Agric. Biol. Environ. Stat.* **2019**, *24*, 398–425. [CrossRef] [PubMed]
11. Wang, B.; Yan, L.; Rong, Q.; Chen, J.; Shen, P.; Duan, X. Dynamic Gaussian Process Regression for Spatio-Temporal Data Based on Local Clustering. *Chin. J. Aeronaut.* **2024**, *37*, 245–257. [CrossRef]
12. Sigauke, C.; Chandiwana, E.; Bere, A. Spatio-Temporal forecasting of global horizontal Irradiance Using Bayesian Inference. *Appl. Sci.* **2023**, *13*, 201. [CrossRef] [CrossRef]
13. Hu, S.; Xiang, Y.; Zhang, H.; Xie, S.; Li, J.; Gu, C.; Sun, W.; Liu, J. Hybrid Forecasting Method for Wind Power Integrating Spatial Correlation and Corrected Numerical Weather Prediction. *Appl. Energy* **2021**, *293*, 116951. [CrossRef]
14. Liao, Q.; Cao, D.; Chen, Z.; Blaabjerg, F.; Hu, W. Probabilistic Wind Power Forecasting for Newly-Built Wind Farms Based on Multi-Task Gaussian Process Method. *Renew. Energy* **2023**, *217*, 119054. [CrossRef]
15. CSIR. WASA Data Portal. Available online: https://wasadata.csir.co.za/wasa1/WASAData?pscreen=nwa&pemail=nimo@risoe.dtu.dk (accessed on 16 January 2025).
16. Hopkins, B.; Skellam, J.G. A new method for determining the type of distribution of plant individuals. *Ann. Bot.* **1954**, *18*, 213–227. [CrossRef]
17. Meinshausen, N. Relaxed Lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393. [CrossRef]
18. Williams, C.K.I.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006. Available online: http://www.gaussianprocess.org/gpml/ (accessed on 23 July 2025).
19. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]

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.