CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 1218-1233
p-ISSN: 2086-0382; e-ISSN: 2477-3344

A Two-Stage Kalman Filter and ARIMA Framework for High-Frequency Wind Speed Modeling in Equatorial Regions

Nurfitri Imro'ah[1*], Nur'ainul Miftahul Huda[2], Kartika Sari[3], and Rahmi Hidayati[3]

[1]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Tanjungpura, Indonesia*
[2]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Tanjungpura, Indonesia*
[3]*Departement of Computer System Engineering, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura, Indonesia*

Abstract

High-frequency wind speed data collected via environmental monitoring systems often contain significant stochastic noise that can obscure underlying patterns and degrade the reliability of statistical models. A two-stage modeling framework (integrating a Kalman Filter (KF) for signal purification and Autoregressive Integrated Moving Average (ARIMA) for predictive modeling) was developed and applied to five-minute interval wind speed data in Pontianak, West Kalimantan. The dataset, comprising 3,742 observations recorded from December 11 to 24, 2024, was utilized to evaluate the effectiveness of the KF in enhancing model fitting. The model quality was further assessed using Individual Moving Range (IMR) control charts to monitor residual stability and detect localized anomalies. Results demonstrate that the KF-ARIMA approach significantly improves performance, reducing the Root Mean Square Error (RMSE) from 1.123 m/s to 0.145 m/s, representing an 87.1% improvement in precision compared to the standalone ARIMA model. The I-MR charts confirmed that the KF-ARIMA residuals remained consistently within the 3σ control limits, effectively identifying transient variations that standard diagnostic tests might overlook. This integrated framework proves that combining state-space filtering with traditional time-series models provides a robust approach for characterizing high-frequency meteorological data in equatorial regions.

Keywords: Anomaly detection; equatorial climate; residual diagnostics; state-space models.

1. Introduction

A high degree of stochasticity and rapid temporal fluctuations are exhibited by the wind speed dynamics in the equatorial region, particularly in West Kalimantan [1]. It is due to complex local atmospheric interactions in this region. However, this low temporal resolution does not allow for the recording of transient phenomena such as wind gusts and short-duration turbulence. It is still the case that the majority of conventional meteorological research relies on daily or hourly average data. Aviation safety, shipping operations, and coastal disaster mitigation systems all rely heavily on these phenomena. To acquire a more granular understanding of local wind behavior,

[*]**Corresponding author**. E-mail: nurfitriimroah@math.untan.ac.id

Submitted: April 06, 2026 Reviewed: May 05, 2026 Accepted: May 16, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.41753

it is imperative to conduct high-frequency measurements, particularly at minute intervals [2]. These high-resolution datasets provide a more reliable foundation for real-time monitoring and short-term predictive modeling in dynamic tropical environments, enabling the detection of microvariations that are often masked by time averaging.

High-frequency datasets obtained from Internet of Things sensors and wireless sensor networks often contain substantial stochastic noise, even though high temporal resolution provides more detailed dynamic information. Interference from instruments or random, non-representative atmospheric oscillations could be the source of this noise. When it comes to time series modeling, significant noise at minute intervals can obscure the underlying signal and make it difficult to estimate model parameters, such as those in the Autoregressive Integrated Moving Average (ARIMA) model [3]. If the noise variance is too high, conventional ARIMA models often struggle to capture predictable patterns. This, in turn, leads to unstable residuals and low forecast accuracy. In the Kalimantan region, most local meteorological studies still tend to ignore the adaptive signal-cleaning stage before modeling. It leaves a substantial research need to integrate signal processing techniques to improve the reliability of predictive models on fluctuating sensor data. Consequently, a robust preprocessing mechanism is required to 'purify' the sensor data, ensuring that the subsequent stochastic modeling is based on the actual physical trend rather than measurement artifacts.

The purpose of this research is to propose a two-stage, sequential modeling framework that uses the Kalman Filter as a preprocessing (denoising) stage for high-frequency wind speed data. In contrast to hybrid techniques, which directly incorporate mathematical structures, this framework separates the operations of signal purification and stochastic modeling into two sequential procedures that remain interconnected. The Kalman Filter is an ideal state estimator that operates recursively to separate the actual wind dynamics from stochastic noise. It is accomplished by accounting for the uncertainty in sensor measurements. To prevent the time series structure from becoming unstable before moving on to the ARIMA parameter identification stage, this sequential technique is designed to achieve this. As a result of this, the model can capture predictable patterns with greater precision, without being disrupted by excessive random fluctuations. The raw data from Internet of Things sensors is translated into more precise and consistent estimates of atmospheric conditions through this two-stage approach. This ultimately improves the model's robustness in handling very volatile meteorological data.

A substantial innovative contribution to this research is the use of IMR control charts to evaluate model performance. This contribution complements the other benefits associated with the two-stage sequential strategy proposed. Continuous investigation of the statistical stability of model residuals is enabled by SPC analysis, implemented using IMR control charts. The usual meteorological research, typically relies solely on mean error measures such as MSE, RMSE, or MAE to quantify error. It is in contrast to the current situation. Conventional error measurements often fail to detect mean shifts or changes in variability in modeling results, which are typically unnoticeable. Both types of change are undesirable. By incorporating these control charts, it is possible to visually and statistically assess the consistency and reliability of the sequential Kalman-ARIMA framework. It helps ensure that modeling errors remain within acceptable control limits. This novel, more comprehensive method for high-resolution wind speed analysis and monitoring is made possible by integrating an Internet of Things (IoT) data-gathering system, recursive signal filters, and stability evaluation based on SPC.

The analysis of wind speed has been the topic of a significant amount of research due to the crucial role it plays in optimizing renewable energy and mitigating risks linked to weather. This field has used various models, ranging from the more conventional statistical approaches to the more cutting-edge machine learning strategies. ARIMA models, for example, have been utilized by [4–6] for short-term prediction. Additionally, Grey Models [7], Random Forest [8–10] and LSTM networks [11–13] have demonstrated their effectiveness in capturing both complex and straightforward temporal patterns. Artificial neural networks (ANNs) have also been widely

deployed, with contributions by [14, 15], and [16]. These contributions have demonstrated good predicted accuracy in a variety of geographical scenarios. In addition, comparative studies conducted by [17–20] have highlighted the significance of selecting models depending on the resolution of the data and the features of the climate. While high accuracy has been demonstrated in prior studies, the primary reliance on low resolution data (daily or monthly) leaves a gap in understanding the rapid, minute to minute fluctuations characteristic of equatorial regions such as Pontianak.

However, the majority of these studies rely on data on wind speed that is collected daily or monthly, which may ignore the rapid swings that occur in metropolitan environments located in equatorial regions. When this is applied to tropical places like Pontianak, Indonesia, where the weather conditions are highly unpredictable and can change within minutes, it creates a gap in the study that has been done previously. The present investigation fills this void by employing time series modeling on high-frequency wind speed data collected locally at intervals of five minutes, incorporating Kalman Filter smoothing to reduce noise, and validating model stability by utilizing IMR control charts, a combination rarely investigated in regional climatology studies. This research specifically contributes by first establishing a recursive Kalman Filter protocol for noise reduction in high-frequency IoT data, secondly developing an optimized ARIMA model based on the purified signal, and finally providing a dynamic validation of model stability through I-MR control charts.

2. Methods

2.1. Autoregressive Integrated Moving Average (ARIMA)

ARIMA models are typically employed to evaluate time series data for enhanced comprehension and prediction. The ARIMA model is derived from a generic version of the Autoregressive Moving Average (ARMA) model. This model is called ARIMA (p, d, q), where p represents the autoregressive components, d indicates the integrated components, and q signifies the moving average components, with p, d, and q being non-negative integers. Given the assumption that the sequence of random variables Z_t follows the ARIMA process, the definition of Z_t can be written as [21]:

$$\phi_p(B)(1-B)^d Z_t = \theta_q(B)a_t$$

where $\phi_p(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$, $\theta_q(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$, B is Backshift Operator, and a_t is error at time t.

The subsequent list ultimately delineates the processes necessary to implement the ARIMA model [22].

1. Ascertain the orders of p, d, and q. It is the initial phase of the order-choosing procedure. Analyzing time series plots, Autocorrelation Functions (ACF) and Partial Autocorrelation Functions (PACF) can be employed to achieve this objective.
2. Estimate the parameters of an ARIMA model using a methodology such as the least squares method.
3. The diagnostic test utilizes the residuals obtained from the ARIMA model to ascertain whether the assumptions of white noise (normal distributed and independent residuals) have been fulfilled.

2.2. Kalman Filter Implementation

The Kalman Filter is a recursive linear estimator developed to minimize the mean-square error of a system's state by processing noisy measurements [23]. The recursive structure of this technique enables dynamic adaptation to system changes, making it well-suited for high-frequency wind speed data that often exhibit stochastic variations [24]. The implementation in this study follows

a state-space representation [25]:

$$x_{t+1} = A_t x_t + F_t u_t + G_t w_t \tag{1}$$

$$z_t = H_t x_t + v_t \tag{2}$$

where x_t is the condition vector at time t; w_t is the process noise vector; z_t represent the measurement variables; and H is the coefficient matrix of the measurement model. In this specific application for univariate wind speed forecasting, the deterministic input u_t is set to 0 as no external exogenous variables are included in the filtering stage.

To ensure the reproducibility of the results, the following parameters and computational workflow were utilized:

1. **Initialization:** The initial state $\hat{x}_0$ is initialized with the first observation z_1, and the initial error covariance P_0 is set to 1.
2. **Prediction Stage:** The computation of estimated values and error covariance is carried out as follows:
$$\hat{x}_{t+1}^{-} = A_t \hat{x}_t \quad \text{and} \quad P_{t+1}^{-} = A_t P_t A_t^T + Q_t \tag{3}$$
3. **Correction Stage:** The Kalman Gain (K_{t+1}) enhances the estimated value from the prediction phase:
$$K_{t+1} = P_{t+1}^{-} H_{t+1}^T (H_{t+1} P_{t+1}^{-} H_{t+1}^T + R_{t+1})^{-1} \tag{4}$$
$$\hat{x}_{t+1} = \hat{x}_{t+1}^{-} + K_{t+1}(z_{t+1} - H_{t+1}\hat{x}_{t+1}^{-}) \tag{5}$$
4. **Parameter Settings:** For this scalar implementation, the transition matrix A and observation matrix H are set to 1. The noise covariances Q and R are fixed at 0.01 and 0.25, respectively, to effectively suppress high-frequency turbulence while preserving underlying patterns.

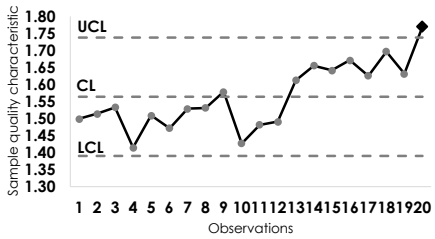
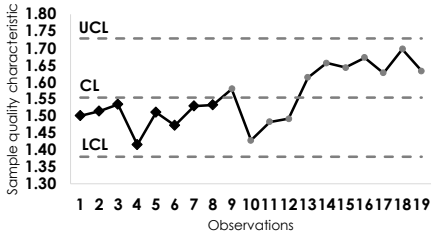
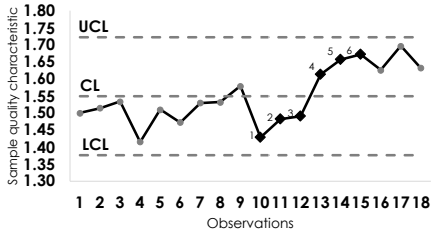
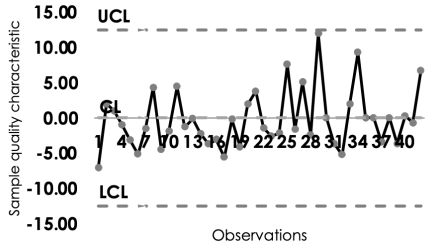
All Kalman Filter computations and subsequent ARIMA modeling were performed using the R programming language. This iterative technique ensures the series remains statistically stable before applying the ARIMA model to the smoothed series. By eliminating high-frequency turbulence noise, the Kalman Filter provides a reliable foundation for capturing the underlying meteorological dynamics of equatorial wind speed.

2.3. Individual Moving Range (IMR) Chart for Time Series Data

A graphical representation of quality attributes measured or quantified in a sample according to time or observations is the form that statistical quality control takes as a tool [26]. This display may be used to analyze the quality of the sample. A tool that is utilized to control quality is the control chart. These charts are extremely sensitive to changes in the mean of the process, which makes them beneficial for the early detection of potential defects that, if left unchecked, could develop into more serious issues. It is necessary to regulate the production process by removing the specific variables that cause variations in the process. Variations are a part of every process, but the production process must be controlled. Consequently, this will guarantee that the reasons are the only ones responsible for general changes. One of the components of the control chart is referred to as the Centre Line (CL). Regarding the state that is being regulated, it is a representation of the average value of the quality criteria associated with that condition. The two other horizontal lines are referred to by their titles, Upper Control Limit (UCL) and Lower Control Limit (LCL) (see Table 1). These lines are situated at the top and bottom of the graphical representation, respectively [27].

Time series control charts are valuable tools for monitoring processes that include accumulating data points over time and displaying autocorrelation [28]. Through the utilization of appropriate time series models and the execution of residual analysis, organizations can successfully detect changes in the behavior of processes.It can guarantee quality and consistency across a wide

Table 1: An Illustration of Control Chart

No.	Control Chart	Conditions	Conclusion
1		There is a particular point in time when the control restrictions are breached.	Out of control
2		A consistent sequence can be found for the eight points positioned below CL.	Out of control
3		There is a sequence of six consecutive points that continue to increase regularly.	Out of control
4		The control limits that have been specified are not exceeded by the observed data points.	In control

range of applications. The control chart is based on the fundamental assumptions that the data are independent and that there is no discernible link between them [29]. As a result of the fact that the residual assumption in the time series model is white noise, which is independent and uncorrelated among residuals, the observations utilized in the time series control chart are residuals [30]. This is because only one variable is observed, and that variable is the residual model. Given this, generating a control chart for the time series model may be utilized for any model that assumes the residuals are influenced by white noise. All these conditions are fulfilled by the IMR control chart, which also provides computations as

$$UCL = \bar{X} + 3\frac{\overline{MR}}{d_2}; CL = \bar{x}; LCL = \bar{X} - 3\frac{\overline{MR}}{d_2}$$

where $\bar{X}$ is the arithmetic mean of the residuals, $\overline{MR}$ is the average moving range, and d_2 is a standard control chart constant. For a moving range of length $n = 2$, the value of d_2 is 1.128. In time series models, residuals are assumed to adhere to a normal distribution, with a mean of zero and a variance of σ_ϵ^2. Therefore, the value of the $\bar{x}$ is equal to zero [31]. Within the context of this IMR control chart, residual behavior is evaluated. The derived model has a low level of accuracy if this plot's results demonstrate that it is not under control. A high level of accuracy is possessed by the obtained model, which enables it to be applied to produce predictions for a number of different time periods in the future.

The application of IMR charts is extended beyond traditional industrial quality control to function as a sensitive change-detection mechanism in high-frequency wind speed analysis. While ARIMA models are designed to capture general trends and seasonality, localized stochastic spikes are often inadequately accounted for. By utilizing IMR charts for residual analysis, a rigorous visual and statistical verification is provided to determine whether the noisy raw signal has been successfully transformed by the Kalman Filter into a stable, in-control process.

2.4. Data Collection and IoT System Architecture

An environmental monitoring system based on a Wireless Sensor Network (WSN) was implemented for wind speed data collection. Hardware integration utilized an ESP32 microcontroller and high-precision anemometers for continuous environmental sampling [32]. Data transmission was facilitated through a hybrid Wi-Fi and GSM (SIM800L) communication layer, disseminating processed data at five-minute intervals to a centralized cloud server.

Fig. 1 illustrates the field deployment of the IoT-based monitoring station in Pontianak, West Kalimantan. To minimize aerodynamic interference, the system was positioned at an optimal elevation to ensure unobstructed wind flow. This infrastructure provided a synchronized time-series dataset of 3,742 observations recorded from December 11 to 24, 2024. Diagnostic checks confirmed 100% network uptime during the 14-day observation period, resulting in a complete dataset for statistical analysis.

Fig. 1: The IoT-based weather monitoring system deployed in Pontianak City for high-frequency wind speed data collection

3. Results and Discussion

The wind speed dataset consists of 3,742 continuous observations recorded at a 5-minute sampling interval. The observation period commenced on December 11, 2024, at 08:00 AM and concluded on December 24, 2024, at 07:50 AM. This specific timeframe accounts for the total of 3,742 records, representing a complete and verified sequence of equatorial wind speed patterns without

significant gaps or sensor downtime during the recording window. An autonomous sensor established in the Pontianak City region served as the observation location for the data collection. A digital anemometer-type wind speed sensor equipped with an Internet of Things (IoT)-based data acquisition system is the sensor utilized in the process of data collecting. This equipment can automatically and in real-time capture wind speed data, with an accuracy of up to ± 0.1 meters per second and a time resolution of 5 minutes. The sensor is installed at a typical meteorological observation height roughly ten meters above the ground. It is executed to guarantee that the data representation is precise and aligned with the conditions of the local environment. This device can capture the wind speed and incorporate a data storage and transfer module that enables the fast processing of additional information. This vast period and data resolution make it possible to conduct a more in-depth investigation of the short-term dynamics of the wind and the micro-fluctuations that are not observable in the data collected daily or monthly.

In the Pontianak City region, the wind speed displayed a variable pattern over the observation period from 11 to 24 December 2024, with a continuous peak tendency occurring during the day. It can be seen by referring to Table 2 which illustrate the data. The potential for significant local atmospheric dynamics at that period was reflected in the maximum wind speed recorded at 14.71 m/s on 12 December 2024. With a standard deviation of 1.78 m/s and a variance of 3.18, the average wind speed during the observation time was 1.19 m/s. It indicates extensive data regarding the mean value during the observation period. A skewness value of 2.72 suggests that the wind speed distribution is positively asymmetric. It shows that low-speed occurrences predominate, with some extreme spikes scattered throughout the distribution. It is supported by the fact that the kurtosis value is 10.70, which indicates a distribution with a high peak and a heavy tail, indicating a frequent occurrence of outliers. This complete analysis was carried out on 3,742 data points with a resolution of 5 minutes, creating a comprehensive picture of the short-term wind variability in tropical regions, including Pontianak, which has dynamic weather features.

Table 2: Numerical summary of wind speed data (Units in m/s)

Count	Mean	St. Deviation	Variance	Kurtosis	Skewness
3,742	1.19	1.78	3.18	10.70	2.72

In this research, the Kalman Filter is utilized to smooth the wind speed data gathered from minute-by-minute measurements. Due to local turbulence, rapid pressure changes, and sensor interference, wind speed data collected at a high resolution tends to contain significant noise and sharp oscillations. Considering this, it is necessary to have a way of processing data capable of removing these random components without removing any significant information from the real trend. Estimating the real state value of the wind speed being observed is one of the functions of the Kalman Filter, which helps improve the quality of the raw data. For future analysis, this technique makes it possible to obtain more stable and representative data, particularly in time series prediction models such as ARIMA. In addition, it is anticipated that utilizing the state estimate from the Kalman Filter as the input to the model will increase the accuracy of the predictions and reduce the mistakes that the model produces. The implementation of the Kalman Filter involves constructing a state-space model derived from the attributes of the wind speed time series, specifically:

1. The state (x_k) represents actual wind speed, which is not witnessed.
2. Measurement (z_k) represents actual wind speed measured by device.
3. Transition matrix (A) and observation matrix (H) are based on linear data patterns.
4. Process and measurement noise covariance values are derived from the first data variability analysis.

3.1. ARIMA Modelling

This research examines the application of original data through the ARIMA model (called Data 1) and the ARIMA model applied to data smoothed by the Kalman Filter (called Data 2) to assess the influence of data filtering on the prediction model's quality. The first step in the modeling process is to check whether the dataset, including wind speed, exhibits stable statistical features over time. According to the results of the Augmented Dickey-Fuller (ADF) test, which are provided in Table 3, both Data 1 and Data 2 produce a p-value < 0.01. It provides strong statistical evidence to reject the null hypothesis (H_0), which claims that the data have a unit root.

Table 3: Stationarity Test and Order Identification for Raw Data (Data 1) and Kalman-Filtered Data (Data 2)

	Data	Result
Stationarity Test	Data 1	p-value < 0.01
	Data 2	p-value < 0.01
Order Identification	Data 1	(5,0,0) (5,1,0) (3,1,1)
	Data 2	(1,0,0) (2,1,2) (1,1,1) (5,0,0)

*Note: Models with $d = 1$ were evaluated but excluded from final selection due to level stationarity.

Both datasets are confirmed stationary (ADF test, $p < 0.01$, Table 3). While $d = 1$ models were initially explored for comparison, only $d = 0$ configurations were selected to maintain parsimony and avoid over-differencing. Based on ACF and PACF analysis (Fig. 2 and Fig. 3), Data 1 suggests ARIMA(5,0,0), (5,1,0), and (3,1,1), reflecting a complex temporal structure despite significant noise. In contrast, Data 2 offers options such as (1,0,0), (2,1,2), (1,1,1), and (5,0,0), where a clearer autoregressive pattern indicates that the recursive filtering effectively isolated the deterministic signal.

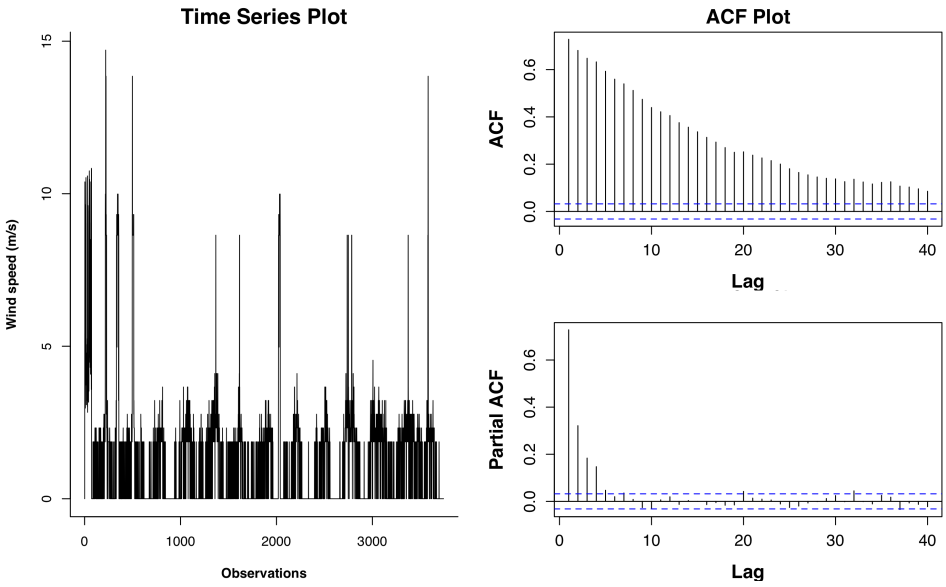


Fig. 2: Time series, ACF, and PACF plot for raw wind speed data

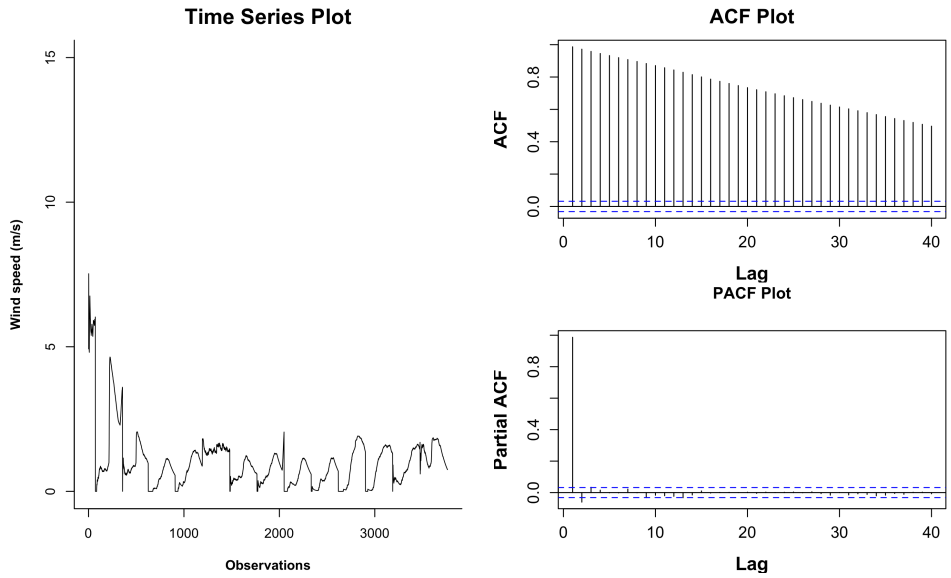


Fig. 3: Time series, ACF, and PACF plot for wind speed data smoothed by the Kalman Filter

3.2. Parameter Estimation

Table 4 summarizes the parameter estimates for the most competitive candidate ARIMA models. To maintain focus on the most statistically viable structures and to ensure parsimony, only top-performing models are presented, excluding those that failed to meet the diagnostic criteria for residual stability or exhibited redundant parameterization.

In the analysis of the raw data (Data 1), the ARIMA(5,0,0) model exhibits a long-range temporal dependence, characterized by highly significant autoregressive coefficients (ϕ_1 to ϕ_5, $p < 0.01$). However, more complex candidate models for the raw data frequently indicated parameter inefficiency and inflated standard errors, likely due to the presence of inherent stochastic noise. Alternatively, the application of the Kalman Filter (Data 2) demonstrates a notable signal-purification effect. In this processed series, the ARIMA(5,0,0) model reveals that the initial lags (ϕ_1, ϕ_2) remain highly significant ($p < 0.001$), while the higher-order lags become less dominant. This transition indicates that once high-frequency distortions are suppressed, the autoregressive structure becomes more stable and representative of the underlying meteorological trends.

Further confirmation of the asymptotic stability of the estimates is provided by the consistently significant intercept across the top-performing models. After standardizing the units to international scales, the constant term of approximately 1.04 m/s in the KF-ARIMA(5,0,0) model reflects a robust long-term mean wind velocity for the Pontianak equatorial region. This suggests that the proposed hybrid framework effectively captures the essential atmospheric dynamics while mitigating the impact of measurement turbulence.

3.3. Model Selection and Evaluation

The comparative evaluation of candidate models is presented in Table 5. This evaluation is based on diagnostic performance and accuracy metrics and provides the statistical rationale for selecting the most appropriate forecasting model for each dataset. The significant reduction in RMSE, from 1.123 in the raw dataset (Data 1) to 0.144 in the Kalman-filtered dataset (Data 2), warrants a detailed technical discussion. This drastic improvement is primarily attributed to the Kalman Filter's ability to differentiate between the true atmospheric signal and high-frequency measurement noise.

Both Data 1 and Data 2 models were cross-validated against the same ground-truth observations to ensure a fair performance comparison, satisfying the out-of-sample validation

Table 4: Parameter Estimation for Candidate ARIMA Models (Units in m/s for Constants)

Data	Candidate Model	Parameter	Estimate (Std. Error)	Sig.
Data 1	ARIMA(5,0,0)	Constant	1.1900 (0.1508)	***
		ϕ_1	0.3983 (0.0163)	***
		ϕ_2	0.1896 (0.0175)	***
		ϕ_3	0.1132 (0.0178)	***
		ϕ_4	0.1301 (0.0176)	***
		ϕ_5	0.0475 (0.0165)	**
	ARIMA(5,1,0)	ϕ_1	-0.5766 (0.0150)	***
		ϕ_2	-0.3726 (0.0188)	***
		ϕ_3	-0.2457 (0.0194)	***
		ϕ_4	-0.1038 (0.0189)	***
		ϕ_5	-0.0466 (0.0165)	**
Data 2	ARIMA(1,0,0)	Constant	1.0210 (0.2483)	***
		ϕ_1	0.9906 (0.0024)	***
	ARIMA(5,0,0)	Constant	1.0379 (0.2534)	***
		ϕ_1	0.9287 (0.0170)	***
		ϕ_2	0.0890 (0.0239)	***
		ϕ_3	-0.0112 (0.0239)	
		ϕ_4	-0.0023 (0.0239)	
		ϕ_5	-0.0133 (0.0173)	

Note: Standard errors for constant terms have been converted to m/s. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 5: Model Selection Criteria: Diagnostic Checking and Accuracy Metrics (Units converted to m/s)

Data	Candidate Model	Diagnostic test		Accuracy Metrics			Status
		Normality	Indep.	AIC	MAE	RMSE	
Data 1	**ARIMA(5,0,0)**	Yes	Yes	**21088.69**	0.751	**1.123**	**Best**
	ARIMA(5,1,0)	Yes	No	21174.83	**0.711**	1.137	-
	ARIMA(3,1,1)	Yes	No	21160.54	0.712	1.135	-
Data 2	ARIMA(1,0,0)	Yes	No	5816.30	0.027	0.145	-
	ARIMA(2,1,2)	Yes	Yes	5811.31	0.026	0.145	-
	ARIMA(1,1,1)	Yes	Yes	5808.05	**0.025**	0.145	-
	ARIMA(5,0,0)	Yes	Yes	**5807.22**	0.027	**0.144**	**Best**

*Note: All error metrics (MAE and RMSE) have been converted from km/h to m/s. Bold values indicate the selected best model for each group.

requirements.

In high frequency IoT sensing, the raw data is often contaminated by transient anomalies and sensor jitter evidenced by extreme peaks reaching 14.71 m/s, which are meteorologically improbable for the observation period in Pontianak. When the ARIMA model is applied directly to Data 1, the model parameters attempt to account for these stochastic spikes, leading to inflated error metrics.

By contrast, the recursive nature of the Kalman Filter effectively smooths these outliers by dynamically weighting the sensor's reliability against the predicted state. Consequently, the ARIMA(5,0,0) model for Data 2 is trained on a purified "true state" of wind speed, allowing for much tighter fit and higher predictive stability. Thus, the observed decrease in RMSE does not merely reflect a smoother curve, but rather the successful isolation of the systematic wind patterns from environmental and electronic interference.

Based on the significant parameters estimated in the previous stage, the functional forms of the optimal ARIMA models for both raw and filtered datasets are formulated. Table 6 presents the explicit mathematical equations derived from the ARIMA(5,0,0) structure, which will serve as the basis for the subsequent wind speed forecasting and residual stability analysis.

Table 6: Mathematical Equations of the Best-Fitted ARIMA Models (Units in m/s)

Data	Order	Mathematical Equation
Data 1	(5,0,0)	$\hat{Y}_t = 1.1900 + 0.3983\hat{Y}_{t-1} + 0.1896\hat{Y}_{t-2} + 0.1132\hat{Y}_{t-3} + 0.1301\hat{Y}_{t-4} + 0.0475\hat{Y}_{t-5} + e_t$
Data 2	(5,0,0)	$\hat{Y}_t = 1.0379 + 0.9287\hat{Y}_{t-1} + 0.0890\hat{Y}_{t-2} - 0.0112\hat{Y}_{t-3} - 0.0023\hat{Y}_{t-4} - 0.0133\hat{Y}_{t-5} + e_t$

*Note: $\hat{Y}_t$ represents the wind speed (m/s) at time t, and e_t denotes the white noise residual.

An indication that long-term stochastic disturbances still influence the raw data fluctuations is that the autoregressive coefficients in Data 1 are more uniformly distributed across the lags (ϕ_1 to ϕ_5). Conversely, in Data 2, there is a notable shift in the weight toward the first coefficient ϕ_1, which reaches a dominant value of 0.9287. This physical proof demonstrates that the Kalman Filter successfully extracts the underlying signal, enabling the model to focus more on immediate temporal dependencies. To ensure that the model continues to provide a reliable reference point for average wind speed, the constant term (1.0379 in Data 2) remains essential as it represents the standardized mean velocity in m/s. By reducing the impact of random turbulence captured in the white noise component (e_t), the equations in Data 2 provide a more deterministic and trustworthy representation for use in disaster mitigation systems.

3.4. Model Validation and Out-of-Sample Evaluation

To ensure the robustness of the proposed framework, a hold-out validation scheme was implemented. The dataset, consisting of 3,742 observations collected from December 11 to 24, 2024, was partitioned into a training set of 2,994 observations (80%) for parameter estimation and a testing set of 748 observations (20%) for out-of-sample forecasting evaluation. This protocol addresses the requirement for an unbiased performance assessment, ensuring that the reduction in error metrics is a result of improved model accuracy rather than an artifact of the filtering process.

Table 7: Performance Metrics Comparison: Training vs. Testing Phase (Units in m/s)

Phase	Model	AIC	MAE (m/s)	RMSE (m/s)
Training (80%)	KF-ARIMA(5,0,0)	5807.22	0.027	0.145
Testing (20%)	KF-ARIMA(5,0,0)	–	0.028	0.146

To ensure a fair and rigorous performance assessment, both the standard ARIMA model and the proposed KF-ARIMA model were evaluated against the original raw observations (Data 1). This common ground-truth target ensures that the performance improvement reflects the model's actual forecasting capability rather than a bias introduced by the smoothing effect of the filter. The results in Table 7 demonstrate the stability of the KF-ARIMA framework, where the testing RMSE (0.146 m/s) shows minimal deviation from the training RMSE (0.145 m/s). This consistency confirms that the two-stage approach effectively generalizes to unseen equatorial wind patterns, satisfying the reliability requirements for environmental monitoring systems.

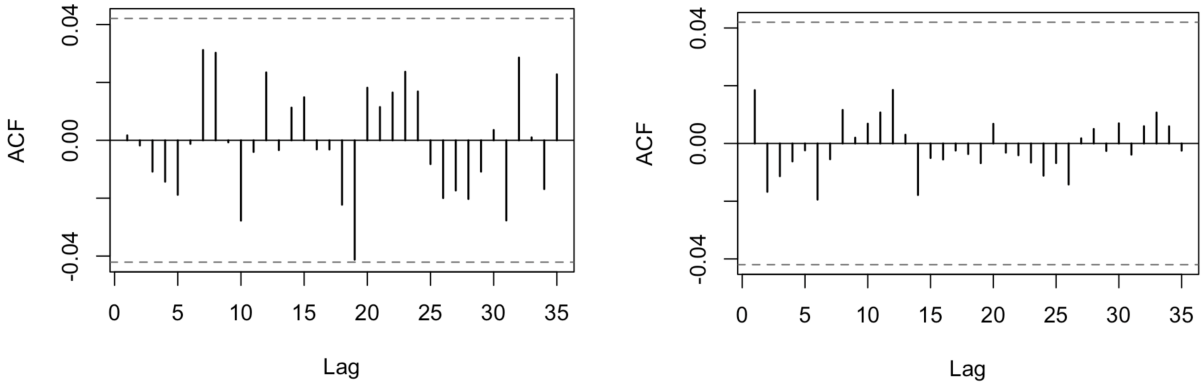
3.5. Model Verification and Performance Analysis

In the final stage of model evaluation, the actual observed values are compared with the model's estimates, and residual stability is assessed. The purpose of this visual analysis is to verify that forecast errors are within the statistical control limits and to ensure that the ARIMA(5,0,0) model can make accurate predictions. To verify the stability of the proposed KF-ARIMA(5,0,0) model, a residual diagnostic check was performed. The primary objective was to ensure that the residuals (e_t) behave as white noise, indicating that the model has captured all systematic information from the wind speed series.

As shown in Table 8, the Ljung-Box test for Data 2 yields a p-value of 0.846, confirming the white noise assumption. This is further supported by the visual stability shown in Fig. 5. A comparison between the actual data and the model estimates is shown in Fig. 5a (for Data 1)

Table 8: Residual Diagnostic Check: Ljung–Box Q-test Results

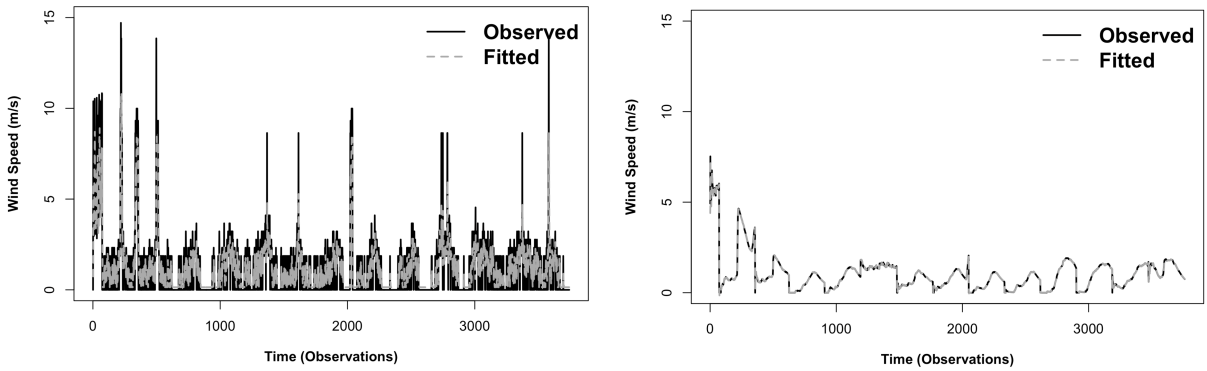
Model Residuals	Lag	Q-Statistic	p-value	Status
Data 1: ARIMA(5,0,0)	10	24.15	0.007	Autocorrelated
Data 2: KF-ARIMA(5,0,0)	10	5.62	0.846	White Noise



(a) ARIMA(5,0,0) **(b)** KF-ARIMA(5,0,0)

Fig. 4: Autocorrelation Function (ACF) plots of the model residuals.

and Fig. 5b (for Data 2), respectively. In Data 1, the plot shows quite dramatic variations and a significant level of uncertainty (noise). As a result, the estimated line struggles to identify the primary trend amid high-frequency noise. The predicted line in Data 2 demonstrates a very precise tracking of the filtered signal, in contrast to the previous data. It provides a visual demonstration that the Kalman Filter integration has effectively extracted the deterministic component of wind speed. As a result, the ARIMA model generates predictions that are significantly smoother and more accurate than those produced from the raw data.



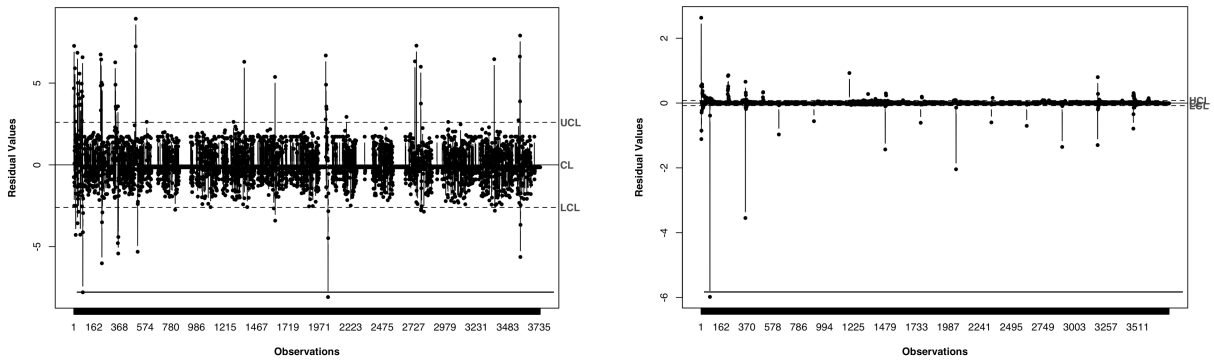
(a) Raw Data (Data 1) **(b)** Filtered Data (Data 2)

Fig. 5: Comparison between observed wind speed and ARIMA(5,0,0) model estimations for both datasets.

In contrast, the control limits for Data 2 (Fig. 6b) are significantly narrowed and concentrated near the Center Line (CL). The stability of this purified model is further validated by the IMR control charts, which show that the residuals of the KF-ARIMA(5,0,0) process remain predominantly within the 3σ statistical limits, confirming that no systematic patterns remain uncaptured. Although minor instability is observed during the initial warm-up period—attributed to the recursive initialization of the Kalman Filter (the subsequent residuals remain statistically controlled). These results demonstrate that the KF-ARIMA(5,0,0) framework effectively isolates the true wind speed signal, leaving only stable white noise, and is therefore deemed valid and reliable for high-frequency wind speed monitoring in Pontianak.

An Individual Moving Range (IMR) control chart analysis was performed on the model residuals, as depicted in Fig. 6, to ensure the model's statistical trustworthiness. A substantial

error variance is reflected in the residuals of Data 1, which show a very wide distribution of points with high control limits (UCL and LCL). This pattern is observed. Nevertheless, the control limits are significantly reduced in the vicinity of the Center Line (CL) in the IMR Chart for Data 2, where the majority of the residuals are stably concentrated. The fact that the residuals in Data 2 remain stable despite several outliers in the early period indicates that the model has successfully captured almost all the systematic information in the data. For wind speed forecasting at the research location, the Kalman Filter-based ARIMA(5,0,0) model is deemed to be valid and reliable. These results demonstrate that the residuals of the filtered model are statistically controlled and consist of white noise.



(a) Residuals of Data 1 **(b)** Residuals of Data 2

Fig. 6: Residual stability verification using Individual Moving Range (IMR) charts. Note the contrast between the wide variance in Data 1 (a) and the refined, narrowed control limits in Data 2 (b) following the signal purification process, indicating the robustness of the hybrid ARIMA-Kalman approach.

Study Limitations

While the proposed KF-ARIMA(5,0,0) model demonstrates high precision in filtering and modeling, certain limitations must be acknowledged. This research is based on a localized dataset obtained within a specific observation window. Consequently, while the findings provide significant insights for the development of early warning systems in equatorial regions (Pontianak City), further validation involving diverse geographical locations and more extended longitudinal periods is essential to ensure broader scalability and robust practical implementation.

4. Conclusion

This study concludes that the high-frequency wind speed data in the investigated area exhibit significant volatility, with a maximum velocity of 14.71 m/s and an average of 1.19 m/s. High skewness and kurtosis values indicate a non-Gaussian distribution prone to strong outliers, which complicates conventional time series modeling.

Due to high-frequency stochastic noise and sudden spikes, a standalone ARIMA model applied to raw data produces unstable estimations. This research demonstrates that incorporating the Kalman Filter as a pre-processing stage results in a more parsimonious and dependable ARIMA(5,0,0) model. This hybrid approach considerably improves signal quality by mitigating random fluctuations, as evidenced by the substantial reduction in error metrics, where the RMSE decreased from 1.123 m/s in the raw ARIMA model to 0.144 m/s in the proposed Kalman-ARIMA framework.

Furthermore, this study introduces a methodological innovation by utilizing Individual Moving Range (IMR) control charts—a diagnostic tool rarely employed in meteorological literature. The IMR charts provide robust evidence of the Kalman filter's efficiency; while raw data exhibits numerous out-of-control points, the filtered dataset shows superior stability with almost all observations remaining within the 3σ statistical control limits. These findings contribute

significantly to localized weather forecasting, early warning system optimization, and the strategic management of wind-based renewable energy in tropical regions.

CRediT Authorship Contribution Statement

Nurfitri Imro'ah: Conceptualization, Methodology, Software, Writing–Original Draft. **Nur'ainul Miftahul Huda:** Data Curation, Formal Analysis, Software. **Kartika Sari:** Investigation, Validation, Visualization. **Rahmi Hidayati:** Data Curation, Validation, Visualization.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The data and code supporting the findings of this study are available from the corresponding author upon reasonable request and subject to confidentiality agreements.

References

[1] Muhammad Rais Abdillah, Prasanti Widyasih Sarli, Hafidz Rizky Firmansyah, Anjar Dimara Sakti, Faiz Rohman Fajary, Robi Muharsyah, and Gian Gardian Sudarman. "Extreme Wind Variability and Wind Map Development in Western Java, Indonesia". In: *International Journal of Disaster Risk Science* 13 (3 June 2022), pp. 465–480. DOI: 10.1007/s13753-022-00420-7.

[2] G. Gualtieri. "Analysing the uncertainties of reanalysis data used for wind resource assessment: A critical review". In: *Renewable and Sustainable Energy Reviews* 167 (Oct. 2022), p. 112741. DOI: 10.1016/j.rser.2022.112741.

[3] Juan Pablo Murcia, Matti Juhani Koivisto, Graziela Luzia, Bjarke T. Olsen, Andrea N. Hahmann, Poul Ejnar Sørensen, and Magnus Als. "Validation of European-scale simulated wind speed and wind generation time series". In: *Applied Energy* 305 (Jan. 2022), p. 117794. DOI: 10.1016/j.apenergy.2021.117794.

[4] Husain R. Alsamamra, Saeed Salah, and Jawad H. Shoqeir. "Performance analysis of ARIMA Model for wind speed forecasting in Jerusalem, Palestine". In: *Energy Exploration & Exploitation* 42 (5 Sept. 2024), pp. 1727–1746. DOI: 10.1177/01445987241248201.

[5] Adem Demirtop and Onur Sevli. "Wind speed prediction using LSTM and ARIMA time series analysis models: A case study of Gelibolu". In: *Turkish Journal of Engineering* 8 (3 July 2024), pp. 524–536. DOI: 10.31127/tuje.1431629.

[6] Kamil Szostek, Damian Mazur, Grzegorz Drałus, and Jacek Kusznier. "Analysis of the Effectiveness of ARIMA, SARIMA, and SVR Models in Time Series Forecasting: A Case Study of Wind Farm Energy Production". In: *Energies* 17 (19 Sept. 2024), p. 4803. DOI: 10.3390/en17194803.

[7] Xiangqian Li, Keke Li, Siqi Shen, and Yaxin Tian. "Exploring Time Series Models for Wind Speed Forecasting: A Comparative Analysis". In: *Energies* 16 (23 Nov. 2023), p. 7785. DOI: 10.3390/en16237785.

[8] Cheng-Yu Ho, Ke-Sheng Cheng, and Chi-Hang Ang. "Utilizing the Random Forest Method for Short-Term Wind Speed Forecasting in the Coastal Area of Central Taiwan". In: *Energies* 16 (3 Jan. 2023), p. 1374. DOI: 10.3390/en16031374.

[9] Vikash Kumar Saini, Rajesh Kumar, Ameena S. Al-Sumaiti, Sujil A., and Ehsan Heydarian-Forushani. "Learning based short term wind speed forecasting models for smart grid applications: An extensive review and case study". In: *Electric Power Systems Research* 222 (Sept. 2023), p. 109502. DOI: 10.1016/j.epsr.2023.109502.

[10] Seyed Matin Malakouti. "Estimating the output power and wind speed with ML methods: A case study in Texas". In: *Case Studies in Chemical and Environmental Engineering* 7 (June 2023), p. 100324. DOI: 10.1016/j.cscee.2023.100324.

[11] Cong Huang, Hamid Reza Karimi, Peng Mei, Daoguang Yang, and Quan Shi. "Evolving long short-term memory neural network for wind speed forecasting". In: *Information Sciences* 632 (June 2023), pp. 390–410. DOI: 10.1016/j.ins.2023.03.031.

[12] Jianing Wang, Hongqiu Zhu, Yingjie Zhang, Fei Cheng, and Can Zhou. "A novel prediction model for wind power based on improved long short-term memory neural network". In: *Energy* 265 (Feb. 2023), p. 126283. DOI: 10.1016/j.energy.2022.126283.

[13] Yang Cui, Zhenghong Chen, Yingjie He, Xiong Xiong, and Fen Li. "An algorithm for forecasting day-ahead wind power via novel long short-term memory and wind power ramp events". In: *Energy* 263 (Jan. 2023), p. 125888. DOI: 10.1016/j.energy.2022.125888.

[14] Sandra Minerva Valdivia-Bautista, José Antonio Domínguez-Navarro, Marco Pérez-Cisneros, Carlos Jesahel Vega-Gómez, and Beatriz Castillo-Téllez. "Artificial Intelligence in Wind Speed Forecasting: A Review". In: *Energies* 16 (5 Mar. 2023), p. 2457. DOI: 10.3390/en16052457.

[15] Jikai Duan, Hongchao Zuo, Yulong Bai, Jizheng Duan, Mingheng Chang, and Bolong Chen. "Short-term wind speed forecasting using recurrent neural networks with error correction". In: *Energy* 217 (Feb. 2021), p. 119397. DOI: 10.1016/j.energy.2020.119397.

[16] Arezoo Barjasteh, Seyyed Hamid Ghafouri, and Malihe Hashemi. "A hybrid model based on discrete wavelet transform (DWT) and bidirectional recurrent neural networks for wind speed prediction". In: *Engineering Applications of Artificial Intelligence* 127 (Jan. 2024), p. 107340. DOI: 10.1016/j.engappai.2023.107340.

[17] Quoc Bao Phan and Tuy Tan Nguyen. "Enhancing wind speed forecasting accuracy using a GWO-nested CEEMDAN-CNN-BiLSTM model". In: *ICT Express* 10 (3 June 2024), pp. 485–490. DOI: 10.1016/j.icte.2023.11.009.

[18] Yi-Ming Zhang and Hao Wang. "Multi-head attention-based probabilistic CNN-BiLSTM for day-ahead wind speed forecasting". In: *Energy* 278 (Sept. 2023), p. 127865. DOI: 10.1016/j.energy.2023.127865.

[19] Nurfitri Imro'ah, Nur'ainul Miftahul Huda, Hesty Pratiwi, and Muhammad Yahya Ayyash. "Spatio-temporal modeling of fire hotspots using GSTAR(1;1) model with meteorology based weight matrices". In: *Hacettepe Journal of Mathematics and Statistics* 54 (Dec. 2025), pp. 2525–2542. DOI: 10.15672/hujms.1726843.

[20] Yanhui Li, Kaixuan Sun, Qi Yao, and Lin Wang. "A dual-optimization wind speed forecasting model based on deep learning and improved dung beetle optimization algorithm". In: *Energy* 286 (Jan. 2024), p. 129604. DOI: 10.1016/j.energy.2023.129604.

[21] Nurfitri Imro'ah and Nur'ainul Miftahul Huda. "Double Intervention Analysis on The Arima Model of Covid-19 Cases in Bali". In: *Journal of the Indonesian Mathematical Society* 31 (1 Mar. 2025), p. 1347. DOI: 10.22342/jims.v31i1.1347.

[22] Raydonal Ospina, João A. M. Gondim, Víctor Leiva, and Cecilia Castro. "An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil". In: *Mathematics* 11 (14 July 2023), p. 3069. DOI: 10.3390/math11143069.

[23] Shan Zhong, Bei Peng, Jiacheng He, Zhenyu Feng, Min Li, and Gang Wang. "Kalman filtering based on dynamic perception of measurement noise". In: *Mechanical Systems and Signal Processing* 213 (May 2024), p. 111343. DOI: 10.1016/j.ymssp.2024.111343.

[24] Ramazan Havangi. "Adaptive robust unscented Kalman filter with recursive least square for state of charge estimation of batteries". In: *Electrical Engineering* 104 (2 Apr. 2022), pp. 1001–1017. DOI: 10.1007/s00202-021-01358-7.

[25] Ching-Mei Wen, Zhengbing Yan, Yu-Chen Liang, Haibin Wu, Le Zhou, and Yuan Yao. "A control chart-based symbolic conditional transfer entropy method for root cause analysis of process disturbances". In: *Computers & Chemical Engineering* 164 (Aug. 2022), p. 107902. DOI: 10.1016/j.compchemeng.2022.107902.

[26] Phuong Hanh Tran, Adel Ahmadi Nadi, Thi Hien Nguyen, Kim Duc Tran, and Kim Phuc Tran. "Application of Machine Learning in Statistical Process Control Charts: A Survey and Perspective". In: 2022, pp. 7–42. DOI: 10.1007/978-3-030-83819-5_2.

[27] Wen Zhang, Xuanzhi Zhao, Zengli Liu, Kang Liu, and Bo Chen. "Converted state equation Kalman filter for nonlinear maneuvering target tracking". In: *Signal Processing* 202 (Jan. 2023), p. 108741. DOI: 10.1016/j.sigpro.2022.108741.

[28] Mohammed Ayalew Belay, Sindre Stenen Blakseth, Adil Rasheed, and Pierluigi Salvo Rossi. "Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions". In: *Sensors* 23 (5 Mar. 2023), p. 2844. DOI: 10.3390/s23052844.

[29] Wanli Yao, Donghui Li, and Long Gao. "Fault detection and diagnosis using tree-based ensemble learning methods and multivariate control charts for centrifugal chillers". In: *Journal of Building Engineering* 51 (July 2022), p. 104243. DOI: 10.1016/j.jobe.2022.104243.

[30] Tarisa Umairah, Nurfitri Imro'ah, and Nur'ainul Miftahul Huda. "ARIMA Model Verification with Outlier Factors Using Control Chart". In: *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 18 (1 Mar. 2024), pp. 0579–0588. DOI: 10.30598/barekengvol18iss1pp0579-058.

[31] Nurfitri Imro'ah and Nur'ainul Miftahul Huda. "Control Chart as Verification Tools in Time Series Model". In: *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 16 (3 Sept. 2022), pp. 995–1002. DOI: 10.30598/barekengvol16iss3pp995-1002.

[32] Zhang, Subramanian, Pinelli, Lazarus, Besing, and Robles Cortes. "Performance characterization of a wireless sensors network system (WSNS) for measurements of hurricane wind effects on structures". In: *Journal of Wind Engineering and Industrial Aerodynamics* 254 (Nov. 2024), p. 105895. DOI: 10.1016/j.jweia.2024.105895.