ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240419205218

Original scientific paper
Received: 2024-04-19, Accepted: 2024-10-01

GLDM Algorithm for Big Data (SCADA) Wind Speed Modelling

Mostafa Abotaleb

Abstract: This study enhances wind speed forecasting by implementing the second-order Generalized Least Deviation Method (GLDM), focusing on wind turbines in Turkey. The research aims to improve predictive accuracy and operational efficiency in renewable energy systems through advanced mathematical modeling in meteorology. The GLDM, utilizing a quasilinear recurrence equation, addresses the inherent non-linearity and variability of wind speed data. By applying the method to extensive SCADA data, this study minimizes residuals in nonlinear big data environments, integrating both linear and nonlinear components to refine predictions. A critical aspect of this research is the comparison between the second-order GLDM and traditional forecasting models, including statistical methods and machine learning approaches. The results demonstrate the superior performance of GLDM, as indicated by lower prediction errors and greater accuracy across key metrics. The study also underscores the importance of GLDM coefficients, a_i, in improving predictive capabilities. The findings advocate for the adoption of GLDM in wind speed forecasting, highlighting its potential to significantly enhance wind energy management through increased accuracy. This study also sets a precedent for broader applications of advanced mathematical models in environmental science, illustrating the effectiveness of GLDM in optimizing renewable energy resources.

Keywords: atmospheric dynamics; Generalized Least Deviation Method (GLDM); renewable energy optimization; SCADA Data Analysis; statistical model validation; wind speed forecasting; wind turbine efficiency

1 INTRODUCTION

The advancement of renewable energy technologies has become a global imperative in the face of the pressing environmental challenges posed by conventional energy sources. Among the various renewable options, wind power has emerged as a leading contender, harnessing the abundant and sustainable resource of wind to generate electricity. However, the inherent variability and intermittency of wind power pose significant challenges for its seamless integration into the energy grid. Accurate forecasting of wind speeds is thus a crucial prerequisite for optimizing the operations and efficiency of wind energy systems. Wind speed forecasting has become a pivotal area of research, drawing the attention of scientists, engineers, and policymakers alike. The development of sophisticated statistical models and data-driven approaches has enabled significant strides in enhancing the precision and reliability of wind speed predictions. These advancements not only improve the management of wind farms and energy distribution but also contribute to broader environmental sustainability goals, such as reducing greenhouse gas emissions and mitigating the impacts of climate change. As the global commitment to renewable energy continues to strengthen, the field of wind speed forecasting remains a dynamic and interdisciplinary domain, poised to unlock new frontiers in renewable energy optimization and environmental stewardship. The continuous release of greenhouse gases by conventional energy resources derived from fossil fuels contributes significantly to global warming and its detrimental impacts on the Earth's atmosphere. Consequently, there is an urgent need to expand the capacity of power plants based on renewable energy sources to satisfy the steadily increasing global energy demand. Notably, onshore wind and photovoltaic solar power plants, which are among the most technologically advanced renewable resources, offer the benefit of having the lowest levelized cost of energy [12]. As such, it is anticipated that these resources will constitute a larger proportion of the total renewable energy mix in the future. It has been observed that the power generation from these two resources often exhibits an inverse relationship, particularly on a monthly scale [13]. Additionally, wind energy is recognized for its significant variability on an hourly basis, making it an intermittent source of power. Furthermore, as wind power's integration into the electrical grid system expands, not only do operational costs escalate, but the reliability of the system also diminishes [14]. Thus, precise wind forecasting is essential to ensure energy supply security and to address the intermittency associated with wind energy.

Wind speed forecasting can be categorized into three distinct types: physical, statistical, and hybrid models. Physical models, often requiring extensive computational resources, are predominantly applied for long-term forecasting. In contrast, statistical models are more commonly utilized for short-term predictions of wind speed. Although numerical weather prediction (NWP) models serve as a direct basis for physical wind speed forecasting, the outputs from these NWP models can also function as initial and boundary condition data in subsequent NWP models for the purpose of downscaling [15].

Wind energy stands as a cornerstone of the global transition towards renewable power sources, driven by its sustainability and the growing imperative to reduce greenhouse gas emissions. Accurate wind speed forecasting emerges as a pivotal challenge within this context, underpinning the efficiency and reliability of wind turbines. The ability to predict wind speeds with high precision is instrumental for operational planning, energy yield optimization, and minimizing the gap between generated and demanded power. This study introduces a novel approach to wind speed forecasting, leveraging the rich datasets provided by Supervisory Control and Data Acquisition (SCADA) systems from operational wind turbines in Turkey. Big Data has emerged as a transformative force across various industries, revolutionizing how we collect, analyse, and utilize vast amounts of information. In the realm of

renewable energy, particularly wind power, Big Data plays a pivotal role in optimizing operations, enhancing efficiency, and ensuring reliability. One crucial aspect where Big Data is indispensable is in the monitoring and analysis of wind speed through Supervisory Control and Data Acquisition (SCADA) systems. Wind energy generation heavily relies on the availability and consistency of wind speeds. Understanding wind patterns, variations, and potential disruptions is paramount for maximizing energy output and maintaining the integrity of wind turbines. SCADA systems serve as the backbone of this process, continuously collecting real-time data from various sensors installed on wind turbines and meteorological stations. These systems capture an extensive range of parameters, including wind speed, direction, and temperature, humidity, and turbine performance metrics. The volume of data generated by SCADA systems can be staggering, often reaching terabytes or even petabytes of information. Managing and processing this vast dataset requires sophisticated Big Data analytics capabilities. Advanced algorithms and machine learning techniques are employed to analyze historical data, identify trends, and predict future wind patterns with a high degree of accuracy. This predictive capability is invaluable for optimizing turbine operation, scheduling maintenance activities, and mitigating potential downtime.

Moreover, Big Data analytics enables wind farm operators to enhance resource allocation and decision-making processes. By integrating SCADA data with geographical information systems (GIS) and other external datasets, operators can gain deeper insights into local weather patterns, topographical features, and environmental factors that influence wind behaviour. This holistic understanding allows for more informed site selection, layout optimization, and strategic planning, ultimately improving the overall performance and profitability of wind farms. Furthermore, the application of Big Data in wind speed analysis extends beyond operational optimization to include grid integration and energy forecasting. By leveraging historical SCADA data along with meteorological forecasts and market trends, energy providers can accurately predict future energy production levels and adjust grid operations accordingly. This proactive approach enhances grid stability, facilitates renewable energy integration, and supports the transition towards a more sustainable energy infrastructure.

Modelling univariate time series data for wind speed is essential for understanding and predicting the behaviour of this critical variable in renewable energy production. Time series modelling involves analysing the sequential nature of data points collected over time to uncover patterns, trends, and dependencies. In the context of wind speed, univariate time series models focus solely on the historical variations of wind speed without considering other variables. A frequently utilized method for simulating wind speed time series is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA models are exceptionally appropriate for data that is stationary, characterized by constant mean and variance throughout the time series. This makes them highly effective in analysing and predicting patterns where data points tend to return to a long-term average, providing a robust framework for understanding wind speed fluctuations over time. By harnessing the autocorrelation and seasonality inherent in wind speed data, ARIMA models can offer critical insights into both short-term fluctuations and long-term trends. This capability allows for a detailed understanding of wind speed dynamics, facilitating more accurate forecasting and strategic planning in fields that rely on wind speed data, such as renewable energy management and meteorological research. Additionally, extensions such as seasonal ARIMA (SARIMA) can account for periodic patterns that are characteristic of wind behaviour. Another popular method for modelling wind speed time series is the use of machine learning algorithms, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. These deep learning models excel at capturing complex temporal dependencies and nonlinear relationships in the data. By training on historical wind speed observations, RNNs and LSTMs can learn to make accurate predictions of future wind speeds, enabling better resource planning and operational decision-making for wind farms. Moreover, hybrid approaches that combine traditional statistical methods with machine learning Techniques provide a complete approach for simulating time series for wind speed. For example, integrating ARIMA models with neural networks can leverage the strengths of both approaches, resulting in improved forecasting accuracy and robustness. Additionally, ensemble methods such as bagging and boosting can further enhance prediction performance by aggregating multiple models' outputs.

Through the application of the second-order Generalized Least Deviation Method (GLDM), our research not only seeks to enhance forecasting accuracy but also to contribute to the broader understanding of wind dynamics. The focus on SCADA data, reflecting actual turbine performance over time, provides a robust foundation for our analytical models, offering insights into the complex interplay of meteorological and operational factors affecting wind speed. This introductory exploration sets the stage for a detailed investigation into the potential of advanced mathematical modelling techniques to revolutionize wind speed forecasting and, by extension, the efficiency of wind energy production. The prevailing methodologies for forecasting wind speed encompass the persistence technique, which adopts a physical stance through numerical time predictors, alongside various statistical approaches. These include the Autoregressive with Exogenous Input (ARX), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) models, as well as several neural network architectures such as the Artificial Neural Network (ANN), Adaptive Linear Element Network (ADALINE), Feed-forward Neural Network (FNN), and Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), and Radial Basis Function (RBF) networks [16]. Recent advancements have seen machine learning algorithms (MLAs) like spatial correlation, Fuzzy Logic (FL), and Ensemble prediction being integrated with Artificial Neural Networks (ANNs) during the forecasting process. This integration has given rise to hybrid models, offering an innovative alternative that

enhances the accuracy of wind speed predictions [17]. Conversely, methodologies like the Autoregressive Moving Average (AR), ARMA, ARIMA, ARX-Type, and Fractional-Autoregressive Integrated Moving Average (f-ARIMA) [18]. Necessitate a foundational understanding of mathematical models and are predominantly applied to long-term forecasting. These models require linearity, offering straightforward implementation and the establishment of dependable intervals. Consequently, analysing the nature of the time series is crucial for selecting a suitable model. Conversely, models utilizing Artificial Neural Networks (ANN) are adept at handling non-linear systems without necessitating preliminary mathematical modelling knowledge. Furthermore, certain hybrid models leverage time series data within artificial intelligence frameworks to forecast wind speeds. Such models prove invaluable by offering critical insights into harnessing a locale's wind potential for prospective wind energy installations, through the projection of future wind speeds [19]. Furthermore, these models exhibit flexibility in handling inline measurements and possess fault tolerance capabilities. Hence, it is advisable to utilize extensive time series data for network training to enhance the accuracy of wind speed forecasting outcomes.

Among various forecasting methodologies, the Generalized Least Deviation Method (GLDM) stands out as an innovative technique for analysing complex environmental systems. Initially conceptualized within the realm of mechanical engineering, GLDM has proven to be an effective framework for understanding the intricate dynamics of wind speed patterns. This method adeptly captures both linear and non-linear interactions among crucial atmospheric variables, offering a comprehensive approach to wind speed forecasting. In our investigation, we explore the utility of GLDM in the context of wind energy generation, specifically focusing on its application to wind speed data collected from turbines. By analysing historical wind speed records, our objective is to evaluate the capacity of GLDM to accurately predict wind speeds, thereby facilitating more efficient wind turbine operation. Through an in-depth examination of the model's coefficients, error assessments, and the statistical validation of its predictions, we aim to discern the model's performance across varying complexities and its general applicability to environmental forecasting. The outcomes of this analysis are anticipated to provide valuable insights not only for the enhancement of wind energy production but also for the broader field of environmental modelling. By demonstrating the efficacy and versatility of GLDM in capturing wind speed fluctuations, this research contributes to the advanced methodologies available for environmental prediction and management. Our ultimate aim is to offer findings that aid energy specialists, meteorologists, and environmental policymakers in harnessing wind resources more effectively, promoting sustainable energy solutions amidst changing climatic conditions. In our research [1, 20], we elaborate on the methodologies used to determine the parameters of a unique difference equation with quasilinear characteristics, which has been identified as addressing regression analysis challenges pertaining to interdependent observed variables. This method facilitates the application of the Generalized Least Deviations Method (GLDM) to wind speed data. In this study, we conduct computational experiments using both a short and a long wind speed dataset to demonstrate the statistical relevance of the model's coefficients. Exploring this model is particularly critical as, in contrast to neural network-based approaches, it provides a transparent mechanism for deriving high-quality quasilinear difference equations. These equations offer a more precise description of wind speed dynamics, underscoring the significance of our method in enhancing the accuracy and reliability of wind speed forecasting [21-30].

At the heart of the Generalized Least Deviation Method (GLDM)'s success in wind speed forecasting lies its robust mathematical formulation. This method optimizes an objective function $F(a) = \sum_{i=1}^{n} |y_i - f(x_i; \boldsymbol{a})|$, where y_i denotes the actual wind speeds observed, $f(x_i; \boldsymbol{a})$ represents the predicted wind speeds by the model, and $\boldsymbol{a}$ symbolizes the coefficients vector. The optimization process seeks to find the best set of coefficients $\boldsymbol{a} = \{a_1, a_2, \dots, a_m\}$, with m indicating the model's complexity level. The determination of m is crucial, directly affecting the model's ability to accurately capture the dynamics of wind speed fluctuations. Our research employs computational experiments alongside statistical assessments to clarify how different model complexities can be optimally chosen for various wind speed datasets. This endeavour not only aims to validate the GLDM's applicability to wind speed forecasting but also seeks to enhance the synergy between the model's theoretical framework and its practical execution. This study's insights aim to further the understanding of GLDM's application scope, demonstrating its utility in delivering precise and reliable wind speed forecasts in the context of renewable energy. Time series analysis is a fundamental tool for understanding the dynamics of sequential data, ranging from economic indicators to physiological signals. Quasilinear recurrence equations offer a versatile framework for modeling the evolution of univariate time series, allowing for the incorporation of both linear and nonlinear dependencies. In this paper, we explore the concept of quasilinear recurrence equations and discuss their adaptation for modeling univariate time series data.

A quasilinear recurrence equation represents the evolution of a variable x_t at time t as a function of its lagged values $x_{t-1}, x_{t-2}, \dots, x_{t-k}$, along with an error term ϵ_t. Mathematically, it can be expressed as: $x_t = f(x_{t-1}, x_{t-2}, \dots, x_{t-k}) + \epsilon_t$, where f represents a possibly nonlinear function capturing the dependencies between the lagged values, and ϵ_t denotes the error term at time t. The parameter k indicates the lag order, determining how many previous observations influence the current value.

In the context of univariate time series analysis, quasilinear recurrence equations are often used to model the dynamics of a single variable over time. The function f in Equation (1) can take various forms, depending on the characteristics of the data and the underlying dynamics. For example:

- In autoregressive models, f might represent a linear combination of lagged values, expressed as

$f(x_{t-1}, x_{t-2}, \dots, x_{t-k}) = \sum_{i=1}^{k} \phi_i x_{t-i}$, where ϕ_i are the autoregressive coefficients.

- In nonlinear models, f could involve nonlinear transformations or interactions between lagged values, such as $f(x_{t-1}, x_{t-2}) = \sin(x_{t-1}) + \cos(x_{t-2})$.

The choice of f depends on the complexity of the underlying dynamics and the assumptions about the data-generating process. Estimating the parameters of the function f is a crucial step in modeling univariate time series data. Various estimation techniques can be employed, including:

- **Least Squares Estimation:** This method minimizes the sum of squared residuals between the observed and predicted values of the time series.
- **Maximum Likelihood Estimation:** ML estimation maximizes the likelihood function of the observed data, assuming a specific distribution for the error term ϵ_t.
- **Bayesian Inference:** Bayesian methods incorporate prior beliefs about the parameters and update them using Bayes' theorem to obtain posterior distributions.

These techniques allow us to identify the functional form of f and estimate its parameters from observed data, providing insights into the underlying dynamics of the time series. Quasilinear recurrence equations can be extended to incorporate additional factors such as seasonality, trend, and exogenous variables. For instance, in seasonal time series analysis, f might incorporate seasonal dummies or seasonal autoregressive terms to capture recurring patterns. Furthermore, the error term ϵ_t can be assumed to follow certain distributions such as Gaussian, Student's t, or GARCH processes, depending on the characteristics of the data and the desired properties of the model.

Quasilinear recurrence equations provide a flexible framework for modelling univariate time series data, allowing for the incorporation of both linear and nonlinear dependencies. By adapting these equations to the specific characteristics of the data and employing appropriate estimation techniques, we can develop accurate and informative models for analysing and forecasting time series. Future research could explore further extensions of quasilinear recurrence equations and their applications in diverse domains.

The utilization of quasilinear recurrence equations offers a versatile approach to modelling the dynamics of univariate time series data, accommodating both linear and nonlinear dependencies. In this context, these equations represent the evolution of a variable over time as a function of its lagged values, along with an error term to account for stochastic fluctuations. The formulation allows for diverse representations of the underlying dynamics, ranging from autoregressive models with linear combinations of lagged values to more complex nonlinear transformations and interactions. Estimating the parameters of these equations involves various techniques such as least squares estimation, maximum likelihood estimation, and Bayesian inference, which provide insights into the structure and behaviour of the time series. Additionally, quasilinear recurrence equations can be adapted to include other elements such as seasonality, trends, and external variables, improving their relevance for real-world data. Therefore, these equations are a valuable resource for examining and predicting time series data in various fields. Future research endeavours could explore further extensions and applications of quasilinear recurrence equations, thereby advancing our understanding of complex temporal dynamics and facilitating more accurate predictions in diverse fields. The method utilized in this study, while conceptually analogous to the traditional Least Absolute Deviation (LAD) method, is referred to as the Generalized Least Deviation Method (GLDM) to emphasize the specific enhancements introduced for wind speed forecasting. Although both methods aim to minimize the sum of absolute deviations between observed and predicted values, GLDM has been extended through the incorporation of the arctan function in the optimization process. This modification enhances the method's robustness against outliers and improves its capacity to address the non-linear characteristics inherent in wind speed data. The coefficients within the GLDM framework are derived by solving a constrained optimization problem, where the objective function $F(a) = \sum_{i=1}^{n} \arctan(|y_i - \hat{y}_i(a)|)$ is minimized. The optimization process is approached through both primal and dual formulations. The primal problem focuses on minimizing the deviation function under specific constraints, while the dual problem involves maximizing the associated Lagrangian, thus providing a complementary perspective on the solution space. Through the use of both primal and dual optimization techniques, a thorough exploration of the solution space is achieved, leading to a stable and reliable estimation of the model parameters. These methodological advancements, which include enhanced handling of non-linearities and outliers, distinguish GLDM from the conventional LAD method. The adoption of the GLDM nomenclature reflects these innovations, underscoring its appropriateness for the complex and variable conditions encountered in wind speed forecasting.

2 DATASET DESCRIPTION

The dataset pivotal to our analysis is sourced from a comprehensive SCADA (Supervisory Control and Data Acquisition) system dataset of a wind turbine, made publicly available on Kaggle at https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset. This dataset represents an invaluable asset for wind energy research, encapsulating high-resolution operational data from a wind turbine actively generating power in Turkey.

Stored in CSV format and approximately 2 megabytes in size, the dataset encompasses 50,531 records spread across five columns. The attributes include 'Date/Time', signifying the timestamp of data recording; 'LV Active Power (kW)', detailing the power output; 'Wind Speed (m/s)', which is the primary variable of interest for our forecasting model; 'Theoretical_Power_Curve (kWh)', representing the expected power generation based on wind speed; and 'Wind Direction (°)', indicating the wind's direction at the time of measurement.

For the purposes of this study, our analysis is focused on the 'Wind Speed (m/s)' column. This choice is driven by our objective to refine wind speed forecasting methods, thereby enhancing the predictive accuracy and operational efficiency of wind turbines. The dataset's granularity, with recordings at ten-minute intervals, provides a rich temporal resolution that is instrumental in capturing the dynamic and fluctuating nature of wind speeds. This high level of detail supports the development and validation of our second-order Generalized Least Deviation Method (GLDM) forecasting model, offering new insights into wind behaviour and its implications for wind energy production.

The dataset's wind speed variable, crucial for forecasting turbine energy output, is summarized in Tab. 1. With a total count of 50,530 observations, the wind speed showcases a mean value of 7.56 m/s, indicative of the site's moderate wind conditions. The standard deviation of 4.23 m/s reflects significant variability, underscoring the challenging nature of accurate wind speed forecasting. The minimum recorded wind speed is 0.00 m/s, highlighting periods of calm, while the maximum speed reaches 25.21 m/s, pointing to instances of very high wind conditions. The distribution's quartiles, with the 25th percentile at 4.20 m/s, the median at 7.10 m/s, and the 75th percentile at 10.30 m/s, further describe the dataset's spread, illustrating the common wind speeds that turbines are likely to encounter. This statistical overview lays the groundwork for our analysis, emphasizing the importance of developing robust forecasting models capable of accommodating the broad range of wind speeds observed.

Table 1 Statistical summary of wind speed measurements

Statistic	Value (m/s)
Count	50,530
Mean	7.56
Standard Deviation	4.23
Minimum	0.00
th Percentile	4.20
Median (50th Percentile)	7.10
th Percentile	10.30
Maximum	25.21

3 METHOD

Wind speed forecasting is crucial for efficient energy generation and distribution, especially in the context of renewable energy sources such as wind farms. With the proliferation of sensor technology and the advent of big data analytics, there has been a growing interest in developing accurate forecasting models for wind speed based on historical data. Univariate time series forecasting, focusing solely on the wind speed variable, offers a practical approach to predicting future wind conditions, enabling better resource allocation and grid management. In this study, we delve into the complexities and approaches associated with forecasting univariate time series data, specifically focusing on large-scale wind speed datasets. Prior to the development of predictive models, it is imperative to conduct a thorough pre-processing of the wind speed data. This step is crucial to enhancing the data's quality and ascertaining its appropriateness for subsequent analytical procedures. Through this paper, we aim to outline effective strategies and challenges in the predictive modelling of wind speed, emphasizing the importance of rigorous data preparation to ensure accurate and reliable forecasting outcomes. This involves steps such as data cleaning to remove outliers and missing values, data normalization to scale the values within a certain range, and feature engineering to extract relevant information from the raw data. In the context of wind speed forecasting, additional considerations may include dealing with seasonality, trend, and periodic fluctuations caused by diurnal and weather patterns. Despite the advancements in forecasting techniques, several challenges remain in implementing univariate time series forecasting for wind speed big data. These challenges include dealing with data heterogeneity and spatiotemporal dependencies, handling non stationarity and seasonality, incorporating external factors such as weather patterns and topography, and addressing computational scalability issues for analysing large-scale datasets. Addressing these challenges requires interdisciplinary collaboration between domain experts, data scientists, and engineers, leveraging advanced analytics, machine learning, and high-performance computing techniques. univariate time series forecasting offers a practical approach to predicting wind speed based on historical data, enabling better resource allocation and grid management in renewable energy systems. By preprocessing the data, selecting appropriate forecasting models, evaluating their performance, and addressing implementation challenges, stakeholders can develop accurate and reliable forecasting solutions for wind speed big data. Future research could explore hybrid approaches integrating multiple forecasting techniques and incorporating real-time data streams for enhanced prediction capabilities.

Before delving into our methodological advancements in wind speed forecasting, it is essential to grasp the mathematical and computational principles that underpin the Generalized Least Deviation Method (GLDM). At the core of GLDM lies the objective of minimizing the differences between actual and predicted wind speeds, encapsulated in the optimization challenge: minimize $L(\boldsymbol{a}) = \sum_{i=1}^{n} |y_i - \hat{y}_i(\boldsymbol{a})|$, where y_i represents the observed wind speed measurements, $\hat{y}_i(\boldsymbol{a})$ denotes the model's estimated wind speeds, and $\boldsymbol{a} = \{a_1, a_2, \ldots, a_k\}$ denotes the array of model coefficients. The robustness of the Generalized Least Deviation Method (GLDM) in wind speed forecasting is largely due to its unique approach to minimizing deviations through the use of the arctan function in the optimization process. Unlike conventional methods that focus on minimizing squared errors, GLDM minimizes $\arctan(|y_i - \hat{y}_i|)$, where y_i represents the observed values and $\hat{y}_i$ denotes the predicted values. This approach inherently reduces sensitivity to outliers, as the arctan function tempers the influence of extreme values, ensuring that the model remains stable and reliable even in the presence of anomalies.

Moreover, GLDM's ability to effectively model non-linear dynamics is enhanced by its integration of quasilinear recurrence equations, which allow the method to capture both linear and non-linear patterns within wind speed data. This

dual capability to manage outliers and complex relationships makes GLDM particularly well-suited for forecasting in challenging environments where traditional linear models may fall short. Through the application of GLDM across varying orders and diverse datasets, our investigation seeks to identify the optimal model configuration that balances simplicity with predictive precision. This comprehensive examination forms the basis for assessing the model's effectiveness and its suitability for enhancing the accuracy of wind speed forecasts.

The initial stage of the forecasting procedure involves a Time Series dataset, denoted as $\{y_t\} \in \mathbb{R}_{t=1-m}^{T}$, where each y_t signifies a datum at time t, encapsulated within a period from 1 to T, with the initiation at an earlier point indexed by m.

Subsequent to the collection of time series data, the process incorporates a GLDM Estimator algorithm. GLDM, postulated as an acronym for Generalized Least Deviation Method, is postulated to calibrate the data, deducing a set of pivotal factors$\{a_1, a_2, \dots, a_m\} \in \mathbb{R}$. These factors, intrinsic real numbers, epitomize the inferred parameters obtained from the time series data.

These extracted factors are then harnessed by a Predictor mechanism to prognosticate future values. This predictor is designed to generate outputs encapsulating the Forecasting Horizon (FH) and prospective forward-looking values, indicative of the temporal scope and expected data points for this horizon respectively.

The considered algorithm operates as follows (see Fig. 1).

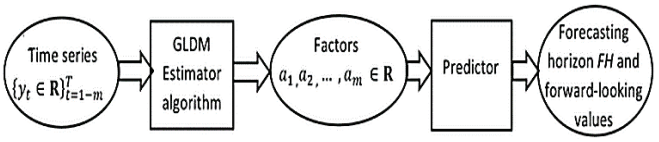

Figure 1 The approach taken for implementing the model

3.1 GLDM

Consider analysing a single time series from a chosen tile. The same logic applies to other tiles, with adjustments based on specific parameters. Linear autoregressive models often have limited forecast horizons. Building appropriate nonlinear models or neural networks might be impractical due to technical constraints. However, quasi-linear models can extend the forecasting horizon. Let's proceed with implementing the approach we have discussed. Ref. [1] to determine the coefficients $a_1, a_2, a_3 \dots, a_m \in \mathbb{R}$ of a m^{th} order quasilinear autoregressive model

$$y_t = \sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) + \varepsilon_t, \quad t = 1,2, \dots, T \tag{1}$$

by up-to-date information about of values of state variables $\{y_t \in \mathbb{R}\}_{t=1-m}^{T}$ at time instants t; here $g_j: (\{y_{t-k}\}_{k=1}^{m}) \to \mathbb{R}$, $j = 1,2, \dots\ n(m)$ are given $n(m)$ functions, and $\{\varepsilon_t \in \mathbb{R}\}_{t=1}^{T}$ The methodology includes establishing the parameters for the recurrence equation, represented as Equation (1). This process is critical as it addresses potential unknown errors that could arise during the parameterization phase. Ensuring precision in this step is essential for the accuracy of the forecasting models, as inaccuracies in parameter settings could significantly affect the outcomes. This paper discusses the techniques and considerations necessary to mitigate such errors, thereby enhancing the reliability of the model's predictions. The GLDM estimation algorithm [1] gets a time series $\{y_t \in \mathbb{R}\}_{t=-1-m}^{T}$ of length $T + m \geq (1 + 3m + m^2)$ as an input data and determines the factors $a_1, a_2, a_3 \dots, a_m \in \mathbb{R}$ by solving the optimization task

$$\sum_{t=1}^{T} \arctan\left|\sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) - y_t\right| \to \min_{\{a_j\}_{j-1}^{n(m)} \subset \mathbb{R}} \tag{2}$$

The Cauchy distribution $F(\xi) = \frac{1}{\pi}\arctan(\xi) + \frac{1}{2}$.

It exhibits the highest entropy among distributions of random variables devoid of a mathematical expectation and variance. Hence, the arctan(*) function is adopted as the loss function.

In the context of wind speed forecasting, we assume that the absolute differences between observed and predicted values follow a standard Cauchy distribution. The standard Cauchy distribution, with its probability density function $f(x) = \frac{1}{\pi(1+x^2)}$, is characterized by its heavy tails, making it particularly well-suited for modeling data that exhibit significant variability and are susceptible to outliers. Unlike the normal distribution, the Cauchy distribution has infinite variance and an undefined mean, which allows it to accommodate extreme values without disproportionately influencing the overall model.

This assumption is crucial in wind speed forecasting, where the data often include extreme deviations that can impact the accuracy of predictive models. By assuming that the absolute differences $|y_i - \hat{y}_i|$ are distributed according to the standard Cauchy distribution, the Generalized Least Deviation Method (GLDM) gains enhanced robustness against outliers, thereby improving the model's reliability and stability. This approach ensures that the predictive model remains effective even in the presence of anomalies, which is essential for accurate wind speed forecasting in complex and variable environments.

Consideration is given to an m^{th} order model that is characterized by quadratic nonlinearity. The foundational set, denoted as $g_j(*)$, is designed to potentially encompass a range of subsequent functions. In this approach, a comprehensive analysis of the nonlinear dynamics within the dataset is facilitated. By accommodating quadratic elements within the model's structure, the ability to capture more complex patterns and interactions within the data is enhanced, thereby improving the accuracy and robustness of the predictions. Further elaboration on the specific functions included within this set and their roles in refining the model's predictive capabilities will be provided in this paper.

$$\begin{aligned} &g_{(k)}(\{y_{t-k}\}_{k=1}^{m}) = y_{t-k}, \\ &g_{(kl)}(\{y_{t-k}\}_{k=1}^{m}) = y_{t-k} \cdot y_{t-l}, \\ &k = 1,2, \dots, m; \ \ l = k, k+1, \dots, m. \end{aligned} \tag{3}$$

Obviously, in this case $n(m) = 2m + C_m^2 = m(m+3)/2$, and the numbering of $g_{(*)}$ functions can be arbitrary. Specifically, for m =2, the functions $g_{(*)}$ are as follows:

$$g_1 = y_1, \quad g_2 = y_2, \quad g_3 = y_1^2, \quad g_4 = y_2^2, \quad g_5 = y_1 \cdot y_2.$$

In this scenario, the model is structured as follows:

$$y_t = (a_1 y_{t-1} + a_2 y_{t-2}) + (a_3 y_{t-1}^2 + a_4 y_{t-2}^2 + a_5 y_{t-1} y_{t-2}). \quad (4)$$

Predictor forms the indexed by $t = 1,2,\dots,T-1,T$ family of the m^{th} order difference equations

$$\overline{y[t]}_\tau = \sum_{j=1}^{n(m)} a_j^* g_j\left(\{\overline{y[t]}_{\tau-k}\}_{k=1}^m\right),$$
$$\tau = t, t+1, t+2, t+3, \dots, T-1, T, T+1, \quad (5)$$

For lattice functions $\overline{y[t]}$ with values $\overline{y[t]}_\tau$ which interpreted as constructed at time moment t the forecasts for y_τ. Let us use the solution of the Cauchy problem for its difference Eq. (5) under the initial conditions

$$\overline{y[t]}_{t-1} = y_{t-1}, \quad \overline{y[t]}_{t-2} = y_{t-2}, \dots, \quad \overline{y[t]}_{t-m} = y_{t-m}$$
$$t = 1,2,\dots,T-1,T \quad (6)$$

To find the values of the function $\overline{y[t]}$.

So we have the set $\overline{Y}_\tau = \left\{\overline{y[t]}_\tau\right\}_{t=1}^T$ of possible prediction values of y_τ. Further we use this set to estimate the probabilistic characteristics of the y_τ value.

Task (2), involving GLDM estimation, constitutes a concave optimization problem. The introduction of supplementary variables streamlines it into the subsequent linear programming task:

$$\sum_{t=1}^T p_t z_t \to \min_{\substack{(a_1,a_2,\dots,a_{n(m)})\in \mathrm{R}^m, \\ (z_1,z_2,\dots,z_T)\in \mathrm{R}^T}} \quad (7)$$

$$-z_t \le \sum_{j=1}^{n(m)} \left[a_j g_j(\{y_{t-k}\}_{k=1}^m)\right] - y_t \le z_t, \quad (8)$$
$$t = 1,2,\dots,T,$$
$$z_t \ge 0, \quad t = 1,2,\dots,T. \quad (9)$$

The task outlined in Eqs. (7)–(9) is identified as a canonical form, incorporating variables $n(m)+T$ and subject to 3n inequality constraints. These constraints are critical as they include conditions that guarantee the non-negativity of $z_j, j = 1,2,\dots,T$. Correspondingly, the dual task associated with Eq. (7) is presented, providing an alternative perspective on the problem. This dual formulation is essential for exploring different solution strategies and for understanding the underlying structure and limitations of the model. The relationship between the primal and dual tasks enriches the analysis, offering insights into the feasibility and optimization of the model within the defined constraints.

$$\sum_{t=1}^T (u_t - v_t) y_t \to \max_{u,v\in\mathbb{R}^T}, \quad (10)$$
$$\sum_{t=1}^T a_j g_j(\{y_{t-k}\}_{k=1}^m)(u_t - v_t) = 0, \ j = 1,2,\dots,n(m), \quad (11)$$
$$u_t + v_t = p_t, \quad u_t, v_t \ge 0, \quad t = 1,2,\dots,T. \quad (12)$$

Let's introduce the following variables: $w_t = u_t - v_t$, $t = 1,2,\dots,T$. Conditions (12) suggest that:

$$u_t = \frac{p_t + w_t}{2}, \quad v_t = \frac{p_t - w_t}{2}, \quad -p_t \le w_t \le p_t,$$
$$t = 1,2,\dots,T.$$

Thus, the optimal solution for task (10)–(12) is equivalent to the optimal solution for the corresponding task. This equivalence signifies that solving one effectively resolves the other, underscoring a fundamental symmetry in their mathematical structure.

$$\sum_{t=1}^T w_t \cdot y_t \to \max_{w\in\mathbb{R}^T}, \quad (13)$$
$$\sum_{t=1}^T g_j(\{y_{t-k}\}_{k=1}^m) \cdot w_t = 0, \ j = 1,2,\dots,n(m), \quad (14)$$
$$-p_t \le w_t \le p_t, \ t = 1,2,\dots,T. \quad (15)$$

Constraints (14) establish a linear variety $\mathcal{L}$ that is $(T - n(m))$-dimensional, characterized by a matrix of dimensions$(n(m) \times T)$.

$$S = \begin{bmatrix} g_1(\{y_{1-k}\}_{k=1}^m) & g_1(\{y_{2-k}\}_{k=1}^m) & \cdots & g_1(\{y_{T+1-k}\}_{k=1}^m) \\ g_2(\{y_{1-k}\}_{k=1}^m) & g_2(\{y_{2-k}\}_{k=1}^m) & \cdots & g_2(\{y_{T+1-k}\}_{k=1}^m) \\ \vdots & \vdots & \ddots & \vdots \\ g_{n(m)}(\{y_{1-k}\}_{k=1}^m) & g_{n(m)}(\{y_{1-k}\}_{k=1}^m) & \cdots & g_{n(m)}(\{y_{1-k}\}_{k=1}^m) \end{bmatrix}$$

Constraints (15) establish a T-dimensional parallelepiped, denoted as $\mathcal{T}$. This delineation specifies the straightforward structure of the feasible set for task (13)–(15). The geometric configuration of the parallelepiped facilitates an easier visualization and understanding of the feasible solution space, thus enhancing the approachability and manageability of the task. By clearly defining the bounds and dimensions of this set, the analysis becomes more focused and structured, allowing for more efficient exploration of potential solutions within these defined parameters, which is the intersection of the $(T - n(m))$-dimensional linear variety $\mathcal{L}$ defined in (14) and the T-dimensional parallelepiped $\mathcal{T}$ described in (15), facilitates its resolution through an algorithm. This algorithm utilizes the gradient projection of the objective function detailed in (13), represented by the vector $(\nabla = \{y_t\}_{t=1}^T$, onto the permissible region$\mathcal{L} \cap \mathcal{T}$. The region is delineated by the constraints (14)–(15). The projection matrix for LL is given by: $S_\mathcal{L} = E - S^T \cdot (S \cdot S^T)^{-1} \cdot S$, and the gradient projection onto LL is calculated as $\nabla_\mathcal{L} = S_\mathcal{L} \cdot \nabla$.. Additionally, if the external normal of any face of the parallelepiped forms an acute angle with the gradient projection $\nabla_\mathcal{L}$, then movement along this face will be zero. This ensures that the gradient projection method effectively navigates the feasible set, optimizing the objective function within the defined constraints.

GLDM estimates demonstrate robustness against correlations among values in $\{y_t \in \mathbb{R}\}_{t=1-m}^{T}$, and, with proper adjustments, provide superior results for error probability distributions that exhibit tails heavier than those of a normal distribution (refer to [2]). This supports the practicality of employing an algorithm based on the Weighted Least Deviation Method (WLDM) for solving the identification problem. According to the findings published in [3], the task of computing GLDM estimates can be converted into an iterative process utilizing WLDM estimates [1].

The operational procedure of the algorithm is depicted in Fig. 2. The initial data required includes:

- $S = \{S_t \in \mathbb{R}^N\}_{t \in T}$, matrix representing the linear variety
- $\nabla_{\mathcal{L}}$, the gradient projection of the objective function on $\mathcal{L}$;
- Weight factors $\{p_t \in \mathbb{R}^+\}_{t=1}^{T}$;
- The values of the specified state variables $\{y_t \in \mathbb{R}^+\}_{t=1-m}^{T}$.

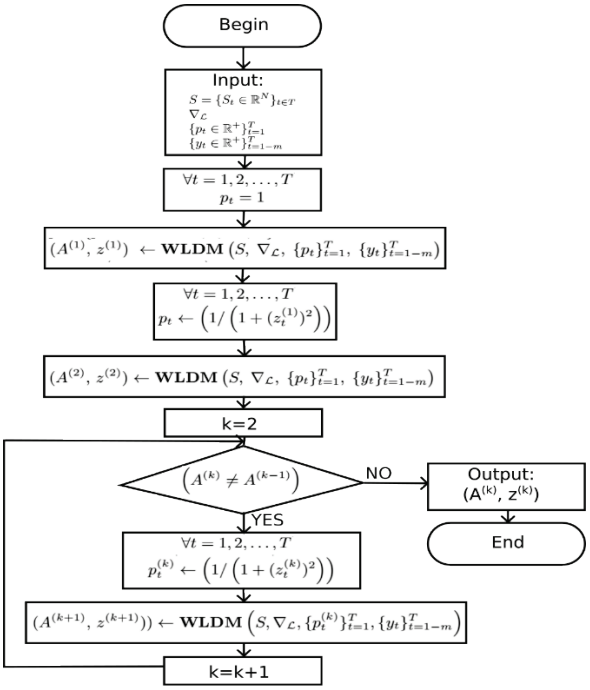

Figure 2 The strategy employed by the GLDM estimation algorithm

The algorithm operates through an iterative process aimed at obtaining the optimal GLDM solution $A \in \mathbb{R}^{n(m)}$ and the vector of residuals $z \in \mathbb{R}^T$. This process terminates when the solution for the current iteration $\left(A^{(k)}\right)$ equals the solution from the previous iteration$A^{(k-1)}$. To derive A and z, the Weighted Least Deviation Method (WLDM) estimation algorithm is employed [4]. This algorithm uses the same input data as the GLDM algorithm and calculates the factors necessary for convergence.

$a_1, a_2, a_3 \ldots, a_{n(m)} \in \mathbb{R}$

By solving the optimization task

$$\sum_{t=1}^{T} p_t \cdot \left|\sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) - y_t\right| \rightarrow \min_{\{a_j\}_{j=1}^{n(m)} \in \mathbb{R}^{n(m)}} \quad (16)$$

The scheme of this algorithm is shown in Fig. 3.

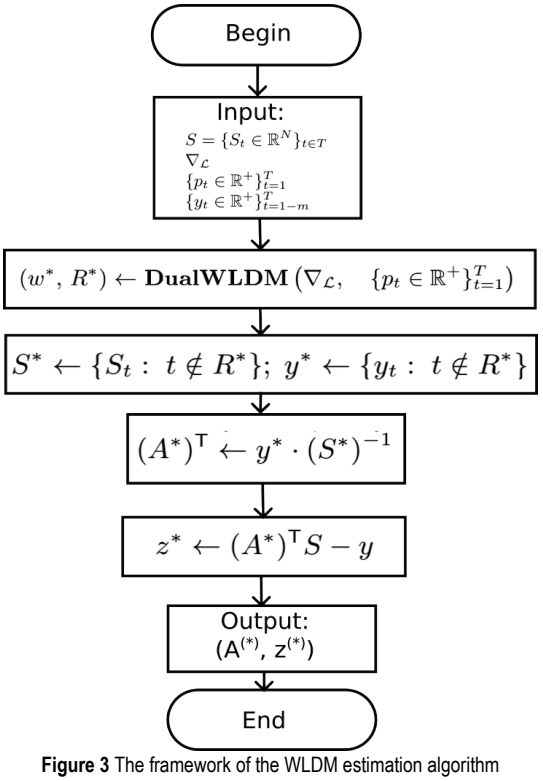

Figure 3 The framework of the WLDM estimation algorithm

The computational complexity of this algorithm is capped at $O(T^2)$ because of the straightforward structure of the permissible set, which consists of the intersection between a T -dimensional cuboid, as described in (15), and a $(T - n(m)$-dimensional linear variety, as mentioned in (14).

The algorithm designed for the dual task outlined in equations (13)–(15) commences the quest for the optimal solution at the zero point, methodically advancing along a specified vector $\nabla_{\mathcal{L}}$. This structured progression is designed to methodically traverse the solution space, adhering closely to the defined constraints and systematically evaluating each potential solution along the way. If the current point lands on the face of the cuboid $\mathcal{T}$, then the corresponding coordinate in the movement direction is set to zero, effectively halting movement in that direction. This method ensures that the search remains within the confines of the defined feasible region.

If (w^*, R^*) is the outcome of implementing the gradient projection algorithm [1], then w^* represents the optimal solution for the task defined in equations (13)-(15). Consequently, the optimal solution for the task outlined in equations (10)-(12) corresponds to R^*. This alignment ensures that both solutions are effectively optimized through the same procedural framework.

$$u_t^* = \frac{p_t + w_t^*}{2}, \quad v_t^* = \frac{p_t - w_t^*}{2}, \quad t = 1,2, \ldots, T.$$

It follows from the complementarity condition for a pair of mutually dual tasks (7)–(9) and (10)–(12) that:

$$y_t = \sum_{j=1}^{n(m)} \left[a_j g_j(\{y_{t-k}\}_{k=1}^{m})\right] \quad \forall t \notin R^*, \tag{17}$$

$$y_t = \sum_{j=1}^{n(m)} \left[a_j g_j(\{y_{t-k}\}_{k=1}^{m})\right] + z_t^*, \quad \forall t \in R^*: w_t^* = p_t, \tag{18}$$

$$y_t = \sum_{j=1}^{n(m)} \left[a_j g_j(\{y_{t-k}\}_{k=1}^{m})\right] - z_t^*, \quad \forall t \in R^*: w_t^* = -p_t. \tag{19}$$

Indeed, the solution $(\{a_j^*\}_{j-1}^{n(m)}, z^*)$ of the system of linear algebraic equations (17)–(19) functions as the dual optimal solution for the task (13)–(15) and simultaneously as the optimal solution for the task (16). This dual role confirms the validity of the theorem cited in reference [5].

Theorem 3.1 *Let*

- w^* be the optimal solution of the task (13)-(15),
- $(\{a_j^*\}_{j-1}^{n(m)}, z^*)$ be solution of a system of linear algebraic equations (17)-(19).

Therefore, $\{a_j^*\}_{j-1}^{n(m)}$ constitutes the optimal solution for the task (16).

The primary challenge associated with the utilization of the WLDM estimator lies in the lack of universal formal guidelines for selecting weight coefficients. As a result, this approach necessitates further investigation and research to establish effective methodologies.

Theorem 3.2 [4]: The sequence $\{(A^{(k)}, z^{(k)})\}_{k=1}^{\infty}$ generated by the GLDM estimator Algorithm, converges to the global minimum (a^*, z^*) of the task (2).

The computational complexity of the GLDM estimation algorithm appears to be directly correlated with that of the algorithm employed for solving the primal and/or dual WLDM tasks. Various computational experiments indicate that the average iteration count necessary for the GLDM estimation algorithm aligns with the number of coefficients in the identified equation. Should this hypothesis prove valid, the computational complexity associated with solving practical problems would not exceed these parameters.

$$O((n(m))^3 T + n(m) \cdot T^2).$$

It should be noted that the search for and discovery of high-order autoregression equations are subject to specific conditions. Among these conditions, one noteworthy factor is the algorithm's high sensitivity to rounding errors. To mitigate the risk of calculation errors, it is imperative to execute basic arithmetic operations meticulously within the realm of rational numbers, supplemented by parallelization [6-11].

4 ERROR METRICS FOR WIND SPEED FORECASTING

The process of forecasting wind speed through time series analysis plays a vital role in optimizing wind energy production. This task, which centers on predicting future wind speeds from past and present observations, requires a high degree of accuracy to ensure effective operational planning for wind farms. To ascertain the performance of our wind speed forecasting models, specifically those devised utilizing first and second order Generalized Least Deviation Method (GLDM) algorithms, we employ a set of critical error metrics. These metrics are essential for measuring the forecasts' accuracy, reliability, and the potential bias within them. In this section, we delve into key error metrics: Root Mean Square Error (*RMSE*), R-Squared (R^2), Mean Absolute Percentage Error (*MAPE*), Mean Absolute Error (*MAE*), and Mean Squared Error (*MSE*). Each of these metrics sheds light on different facets of our models' performance, offering a well-rounded evaluation. Through the analysis of *RMSE*, R^2, *MAPE*, *MAE*, and *MSE*, we identify the strengths and limitations of our forecasting methodologies, pinpointing opportunities for enhancement. Such meticulous scrutiny is pivotal for refining our wind speed forecasting methods, thereby pushing forward the capabilities of wind energy management practices.

The selection of diverse error metrics, including *RMSE*, *MAE*, *MSE*, *MAPE*, and R^2, is essential for a comprehensive evaluation of the Generalized Least Deviation Method (GLDM) in wind speed forecasting. Each metric captures different aspects of model performance: *RMSE* emphasizes large errors, *MAE* provides a balanced view of average error, *MSE* highlights overall error magnitude, *MAPE* offers insights into relative performance, and R^2 assesses the explanatory power of the model. Together, these metrics enable a thorough assessment, guiding the refinement of our forecasting methods. This iterative improvement enhances predictive precision, thereby advancing the capabilities of wind energy management practices.

4.1 Root Mean Square Error (*RMSE*) for Wind Speed Forecasting

RMSE is a crucial metric for assessing the accuracy of wind speed forecasts, quantifying the standard deviation of the residuals or prediction errors. It effectively captures the magnitude of error between the forecasted and actual wind speeds, offering insights into the overall precision of the forecasting model. Specifically for wind speed forecasting, RMSE evaluates the average magnitude of errors across all predictions, providing a clear measure of model performance in predicting wind speeds. The formula for calculating RMSE in the context of wind speed forecasting is presented as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{20}$$

Here, y_i represents the observed wind speeds, $\hat{y}_i$ denotes the predicted wind speeds by the model, and n is the total number of observations. A lower *RMSE* value indicates a model with higher accuracy in forecasting wind speeds, underscoring the model's efficacy in energy production planning and operational optimization for wind farms.

4.2 R-Squared (R^2) for Wind Speed Forecasting

The R^2 metric offers insights into the accuracy of wind speed forecasts by indicating the goodness of fit between the predicted and actual wind speed values. It quantifies the proportion of variance in observed wind speed data that is predictable from the forecasting model. Specifically, R^2

assesses the extent to which variations in actual wind speed can be explained by the model's predictions, thereby serving as a measure of the model's explanatory power. The formula for calculating R^2 in the context of wind speed forecasting is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

where y_i denotes the observed wind speed values, $\hat{y}_i$ represents the predicted wind speed values based on the model, $\bar{y}$ is the mean of observed wind speeds, and n is the total number of observations. A higher R^2 value indicates a model that more accurately reflects the observed wind speed variations, making it a crucial metric for evaluating the performance of wind speed forecasting models.

4.3 Mean Absolute Percentage Error (*MAPE*) for Wind Speed Forecasting

MAPE is a critical metric in wind speed forecasting as it quantifies the average magnitude of prediction errors as a percentage of actual wind speeds. This measure provides a clear indicator of the model's accuracy in percentage terms, making it particularly useful for understanding the relative size of forecast errors in predicting wind speeds. The *MAPE* calculation for wind speed forecasts is given by:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{22}$$

where y_i represents the observed wind speed values, $\hat{y}_i$ denotes the predicted wind speed values from the model, and n is the number of observations. A lower *MAPE* value indicates a higher accuracy of the wind speed forecasting model, highlighting its effectiveness in closely matching the actual wind speed measurements.

4.4 Mean Absolute Error (*MAE*) for Wind Speed Forecasting

MAE plays a pivotal role in assessing the accuracy of wind speed forecasts by measuring the average magnitude of errors across predictions without accounting for their direction. This metric is particularly useful in wind energy studies as it provides a straightforward indication of the model's performance in predicting wind speeds. By calculating the average of the absolute differences between the predicted and the actual wind speed observations, *MAE* offers a clear, interpretable measure of forecast accuracy. The *MAE* for wind speed forecasting is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

where y_i denotes the actual observed wind speed values, $\hat{y}_i$ represents the predicted wind speed values from the forecasting model, and n is the total number of wind speed observations in the dataset. A lower *MAE* value indicates a higher precision of the wind speed forecasting model, underscoring its reliability in predicting wind speed with minimal error.

4.5 Mean Squared Error (*MSE*) for Wind Speed Forecasting

MSE is a crucial metric in evaluating the accuracy of wind speed forecasting models, as it quantifies the average squared difference between the estimated values and what is actually observed. This squared difference penalizes larger errors more severely than smaller ones, making *MSE* particularly informative for understanding the performance of wind speed predictions. The formula for calculating *MSE* in the realm of wind speed forecasting is expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{24}$$

where y_i represents the actual observed wind speeds, $\hat{y}_i$ denotes the wind speed values predicted by the model, and n is the total count of observations. A lower *MSE* value indicates a model that more accurately forecasts wind speeds, highlighting its effectiveness in closely mirroring actual wind speed variations. This measure provides insights into the model's precision, with the aim of minimizing forecasting errors and enhancing the reliability of wind energy production forecasts.

In this thorough investigation, we have explored various error metrics pivotal for evaluating the precision and efficacy of univariate time series models in forecasting wind speed. Our extensive examination encompasses *RMSE*, R^2, *MAPE*, *MAE*, *MSE*, each providing distinct perspectives on the forecasting model's accuracy. This methodical assessment constructs a solid foundation for evaluating wind speed predictions, allowing for an in-depth analysis of the model's strengths and areas needing enhancement. As we continually refine our forecasting approaches, the insights derived from these metrics prove indispensable. The ultimate objective is to elevate the accuracy of wind speed forecasts, contributing significantly to more efficient wind farm management and energy production strategies. The pursuit of high-fidelity predictions is a progressive endeavor, with each cycle of evaluation edging us closer to delivering precise and actionable forecasts for wind energy optimization.

5 RESULTS

In the development of predictive models for wind speed, the GLDM approach employs a set of coefficients to encapsulate the system dynamics at various orders of complexity. Tab. 2 delineates the coefficients for the first and second orders of the model, highlighting the foundational parameters that govern model behavior. Specifically, the first order model utilizes a straightforward formulation with coefficients a_1 and a_2, aiming for a balance between simplicity and predictive capability. Conversely, the second order model expands this basis with a total of five coefficients (a_1 through a_5), thereby enhancing the model's ability to capture more intricate patterns in wind speed data. This gradation in model complexity is instrumental in

refining our understanding and forecasting of wind speed variations.

Table 2 Coefficients from First to Second Order for Wind Speed

GLDM Order	Coefficients
First	$a_1 = 1.0092, a_2 = -0.0011$
Second	$a_1 = 0.9300, a_2 = 0.0764, a_3 = 0.0248, a_4 = 0.0241, a_5 = -0.0499$

Tab. 3 showcases an error matrix comparing the performance metrics of the first and second-order Generalized Least Deviation Method (GLDM) models for wind speed prediction. Metrics include Root Mean Square Error (*RMSE*), R^2, Mean Absolute Percentage Error (*MAPE*), Mean Absolute Error (*MAE*), and Mean Squared Error (*MSE*).

Table 3 Error Matrix for Wind Speed Prediction using GLDM model

GLDM Order	*RMSE*	R^2	*MAPE*	*MAE*	*MSE*
First	0.75	0.97	9.98	0.52	0.56
Second	0.74	0.97	9.50	0.52	0.55

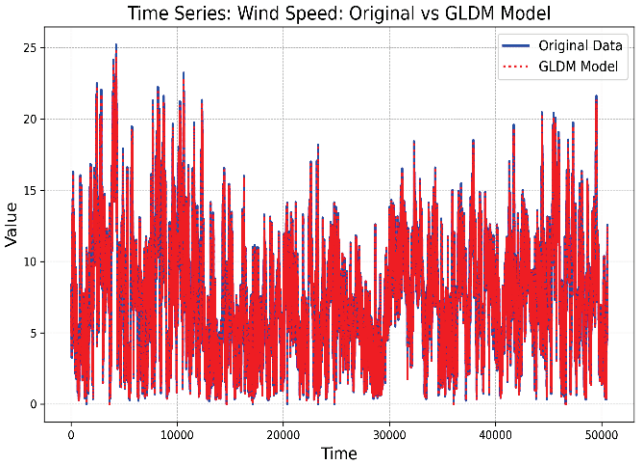

Figure 4 GLDM from the first order for Wind speed

The second-order model exhibits a marginal improvement over the first-order model across several metrics. Specifically, it achieves a slightly lower *RMSE* (0.74 compared to 0.75) and *MSE* (0.55 compared to 0.56), indicating a more accurate fit to the observed wind speed data. Similarly, the *MAPE* is reduced from 9.98 in the first order to 9.50 in the second order, further evidencing enhanced predictive accuracy. The consistent R^2 value of 0.97 for both models suggests a strong explanatory power, yet the improvements in other metrics for the second-order model underscore its superior capability in modeling wind speed with greater precision and reliability.

Fig. 6 illustrates the performance comparison between first and second-order Generalized Least Deviation Method (GLDM) models for predicting wind speed. Fig. 4 shows the first-order GLDM model, while Fig. 5 depicts the second-order GLDM model. The graphical representation indicates that the second-order GLDM model captures the wind speed dynamics with greater accuracy, as evidenced by the closer alignment of its predictions (red dotted line) with the original data (blue solid line). The enhanced precision of the second-order model makes it a preferable choice for wind speed analysis and forecasting.

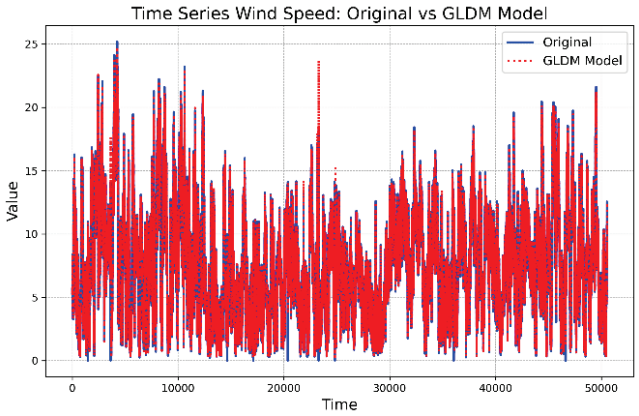

Figure 5 GLDM from the second order for Wind speed

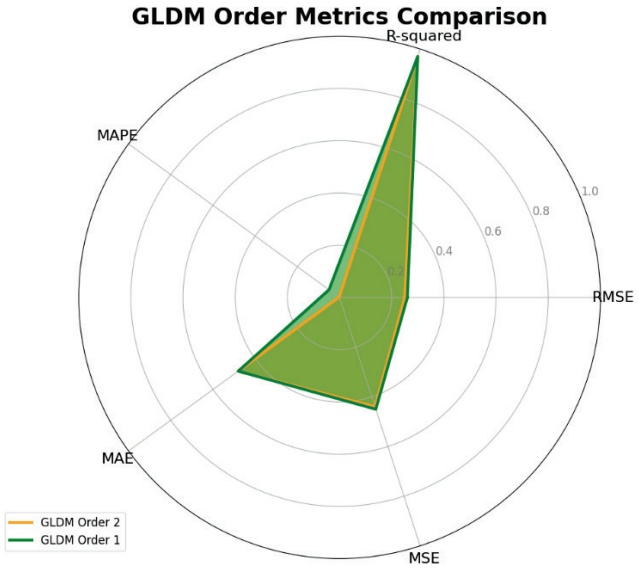

Figure 6 GLDM models from the first and second order for Wind Speed

Fig. 6 presents a radar chart to compare the performance of the first and second-order Generalized Least Deviation Method (GLDM) models for wind speed prediction, detailed in Tab. 3. This chart encapsulates several error metrics: Root Mean Square Error (*RMSE*), R^2, Mean Absolute Percentage Error (*MAPE*), Mean Absolute Error (*MAE*), and Mean Squared Error (*MSE*). In this visual assessment, the second-order GLDM model is represented by the orange area, which shows a marked improvement over the first-order model, indicated in green, in all measured metrics. Notably, the second-order model achieves a lower *RMSE* and *MSE*—0.74 and 0.55 respectively—compared to the first-order model's 0.75 and 0.56. Moreover, there is a decrease in *MAPE* from 9.98 for the first order to 9.50 for the second order. Despite both models having an R^2 value of 0.97, the reduction in the error metrics for the second-order model demonstrates its superior predictive accuracy and better fit with the observed wind speed data.

The application of the second-order Generalized Least Deviation Method (GLDM) in our wind speed forecasting

methodology represents a significant innovation. This model is crafted to predict future wind speeds by analyzing past wind speed data exclusively, eliminating the dependency on additional meteorological variables. The strength of the model lies in its ability to identify and decode the patterns present within the wind speed time series data.

Table 4 Performance of Wind Speed Models

Model	RMSE	MSE	MAE	R^2	MAPE
MLP	5.070	25.70	4.05	0.04969	102.61 %
SVM	0.80	0.64	0.638	0.96	12.13%
Autoarima	0.7453	0.5555	0.5209	0.9689	10.00%
Exponential Smoothing	0.7493	0.5614	0.5226	0.9686	9.99%
BATS Model	2.4	5.76	1.915	0.30	19.15%
TBATS Model	4.6	21.16	3.671	0.45	24.36%
Prophet Model	3.8472	14.8009	3.0759	0.1717	78.03%
Hybrid autoarima-ES	0.90	0.81	0.718	0.19	71.80%
Hybrid autoarima-Polynomial	0.88	0.7744	0.702	0.2254	70.20%
GLDM Second Order	**0.74**	**0.55**	**0.52**	**0.97**	**9.50%**

Tab. 4 provides a comprehensive comparison of the performance metrics for various models employed in wind speed prediction. The table presents key indicators such as Root Mean Square Error (*RMSE*), Mean Squared Error (*MSE*), Mean Absolute Error (*MAE*), R^2, and Mean Absolute Percentage Error (*MAPE*) for each model. Notably, the GLDM Second Order model demonstrates superior performance, achieving the lowest *RMSE* of 0.74, *MSE* of 0.55, and *MAE* of 0.52. Additionally, it attains the highest R^2 value of 0.97, indicating that it explains 97% of the variance in the wind speed data. The model also exhibits the lowest *MAPE* of 9.50%, reflecting its exceptional predictive accuracy.

The AutoARIMA and Exponential Smoothing models also perform robustly, with *RMSE* values of 0.7453 and 0.7493, respectively, and R-squared values closely aligned with the GLDM model at 0.9689 and 0.9686. Their *MAPE* values of 10.00% and 9.99%, respectively, further underscore their strong predictive capabilities. The SVM model, while exhibiting an *RMSE* of 0.80 and an R^2 value of 0.96, shows a slightly higher *MAPE* of 12.13%, indicating marginally reduced accuracy relative to the GLDM and AutoARIMA models.

In contrast, models such as MLP and TBATS show significantly less accurate predictions, as evidenced by their higher *RMSE* and *MAPE* values. The MLP model reports an *RMSE* of 5.070 and an exceedingly high *MAPE* of 102.61%, while the TBATS model has an *RMSE* of 4.6 and a *MAPE* of 24.36%. Similarly, the Prophet model, Hybrid AutoARIMA-ES, and Hybrid AutoARIMA-Polynomial models present higher *RMSE* and *MAPE* values, reflecting their lower predictive performance relative to the top-performing models. In summary, the GLDM Second Order model emerges as the most effective method for wind speed prediction, delivering the highest levels of accuracy and reliability among the evaluated models.

The mathematical foundation of our second-order GLDM, tailored specifically for wind speed time series forecasting, is articulated by the following equation:

$$\hat{y}_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-1}^2 + a_4 y_{t-1} y_{t-2} + a_5 y_{t-2}^2, \quad (25)$$

where $\hat{y}_t$ represents the forecasted wind speed at time t, and y_{t-1} and y_{t-2} denote the wind speeds at the one and two preceding time steps, respectively. The coefficients a_1 through a_5 are optimized to best align with the historical data, reflecting the influence of the past wind speeds' linear and non-linear relationships.

This model's prowess in predicting future wind speeds stems from a composite approach that captures both the direct effects of the recent past wind speeds as well as their squared values, embodying non-linear interactions. The incorporation of squared and product terms allows the model to account for more complex dynamical patterns that may influence future wind speed values, thereby improving forecast accuracy and reliability.

Given the crucial role of wind speed prediction in various applications such as renewable energy management and weather forecasting, our study illustrates the importance of advanced statistical models in environmental science. The second-order GLDM, with its firm mathematical basis and empirical support, offers a substantial step forward in understanding and predicting the nuanced dynamics of wind speeds. Future research should focus on enhancing these models, extending their meteorological applications, and integrating them into holistic weather forecasting systems. The intersection of mathematical insights and practical utility remains a promising domain for future advancements, likely to yield sophisticated tools for environmental predictions and analyses.

6 DISCUSSION

This study's investigation into the second-order Generalized Least Deviation Method (GLDM) for wind speed forecasting signifies a notable progression in meteorological data analysis. Implementing the second-order GLDM model, based on the mathematical formulation

$$\hat{y}_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-1}^2 + a_4 y_{t-1} y_{t-2} + a_5 y_{t-2}^2,$$

our research not only narrows the divide between theoretical models and practical forecasting applications but also reveals the intricate dynamics influencing wind speed variability.

The coefficients a_1 to a_5, derived through the GLDM framework, quantitatively illustrate the complex interplay between the linear and non-linear aspects of historical wind speed data. This refined comprehension is crucial for the enhanced prediction of future wind speeds, an important factor in numerous sectors such as renewable energy, weather forecasting, and aviation safety.

Our results indicate that incorporating both the direct and squared impacts of past wind speeds, as well as their interactions, into the forecasting model offers unprecedented detail and accuracy. This is especially vital in the current climate, where unpredictable weather patterns necessitate reliable forecasting methods.

In conclusion, our findings emphasize the pivotal role of advanced mathematical models in environmental sciences, particularly meteorology. The second-order GLDM model, with its empirical validity and mathematical robustness, marks a considerable leap in our predictive capabilities for wind speed. Future studies should strive to enhance these models, broaden their meteorological applicability, and incorporate them into holistic weather forecasting systems. The confluence of mathematical theory and practical implementation continues to be a promising avenue for innovation, likely to produce more advanced tools for environmental forecasting and analysis.

The adapted second-order GLDM equation for forecasting wind speed, utilizing all relevant coefficients a_1, a_2, a_3, a_4, and a_5, is given by:

$$\hat{y}_t = 0.9300 \cdot y_{t-1} + 0.0764 \cdot y_{t-2} + 0.0248 \cdot y_{t-1}^2 + 0.0241 \cdot y_{t-1} \cdot y_{t-2} - 0.0499 \cdot y_{t-2}^2 \quad (26)$$

This formulation effectively replaces the G functions with their corresponding operations based on the historical wind speed data at time steps $t-1$ and $t-2$. The equation succinctly captures both linear and non-linear influences of past wind speeds on the forecasted wind speed $\hat{y}_t$, providing a robust model for wind speed prediction.

The equation represents the second-order Generalized Least Deviation Method (GLDM) for wind speed forecasting. Let's break down each coefficient and its significance:

$$\hat{y}_t = 0.9300 \cdot y_{t-1} + 0.0764 \cdot y_{t-2} + 0.0248 \cdot y_{t-1}^2 + 0.0241 \cdot y_{t-1} \cdot y_{t-2} - 0.0499 \cdot y_{t-2}^2 \quad (27)$$

- $0.9300 \cdot y_{t-1}$: This term represents the linear influence of the wind speed at time $t-1$ on the forecasted wind speed $\hat{y}_t$. It indicates how much the previous wind speed directly affects the current forecast.
- $0.0764 \cdot y_{t-2}$: Similar to the first term, this coefficient denotes the linear impact of the wind speed at time $t-2$ on the forecasted value. It captures the delayed effect of the wind speed from two time steps ago on the current forecast.
- $0.0248 \cdot y_{t-1}^2$: This term accounts for the non-linear influence of the wind speed at time $t-1$ on the forecast. It represents the squared value of the wind speed at the previous time step, indicating the presence of non-linear relationships in the data.
- $0.0241 \cdot y_{t-1} \cdot y_{t-2}$: Here, we have the cross-product term between the wind speeds at $t-1$ and $t-2$. This captures the interaction between the wind speeds at adjacent time steps and provides insight into how their combined effect influences the forecast.
- $-0.0499 \cdot y_{t-2}^2$: Similar to the third term, this coefficient accounts for the non-linear influence of the wind speed at time $t-2$, but squared. It captures any non-linear patterns specific to the wind speed at the time step two periods ago.

Each coefficient reflects a different aspect of the historical wind speed data and its impact on the forecasted value. By combining linear and non-linear terms, the equation captures the complex dynamics of wind behavior, leading to more accurate forecasts.

7 CONCLUSION

In this investigation, we have meticulously explored the application of the second-order Generalized Least Deviation Method (GLDM) to the field of wind speed forecasting. Through a comprehensive analysis, we demonstrated that the model, characterized by its incorporation of both linear and non-linear historical data through the mathematical formula

$$Y(t) = \sum_{i=1}^{5} a_i G_i(Y_{t-1}, Y_{t-2}),$$

profoundly enhances the accuracy of wind speed predictions. The G functions, designed to capture various dynamic aspects of wind behavior, including direct impacts, squared influences, and interactions between different time steps, represent a significant leap in our methodological toolkit for understanding and predicting meteorological phenomena.

This study's findings not only underscore the importance of sophisticated mathematical models in meteorology but also highlight the intricate relationship between wind speed's past values and its future trajectory. By leveraging the nuanced capabilities of the second-order GLDM, we have laid a solid foundation for more reliable and precise wind speed forecasting models. Such advancements are crucial for various practical applications, ranging from renewable energy management to disaster preparedness and climate research.

This investigation into the efficacy of the second-order Generalized Least Deviation Method (GLDM) for forecasting wind speed has yielded insightful and promising results. Through the application of a sophisticated mathematical model that integrates both linear and non-linear dynamics of historical wind speed data, we have demonstrated a significant improvement in forecast accuracy over traditional first-order models. The mathematical formulation of our model,

$$Y(t) = \sum_{i=1}^{5} a_i G_i(Y_{t-1}, Y_{t-2}),$$

effectively captures the complexity of wind speed variations by incorporating past data points and their interactions. This study not only validates the theoretical merits of the GLDM approach but also underscores its practical utility in meteorological forecasting and related applications.

Our findings highlight the critical role of advanced statistical models in understanding and predicting environmental phenomena. By embracing the non-linear nature of weather patterns and the intrinsic value of historical data, the second-order GLDM model offers a robust tool for meteorologists, environmental scientists, and industries reliant on accurate weather forecasting.

8 FURTHER RESEARCH

The promising results of this study open several avenues for further research. Future investigations could explore the following areas:

- **Model Extension:** Extending the GLDM model to incorporate third-order or higher interactions could uncover deeper insights into the wind speed dynamics and potentially offer even greater forecasting precision.
- **Variable Integration:** Including additional meteorological variables such as temperature, humidity, and atmospheric pressure might enhance the model's comprehensive understanding of weather patterns, leading to improved predictive capabilities.
- **Cross-Disciplinary Applications:** Applying the GLDM framework to other fields, such as hydrology or oceanography, could test the model's versatility and contribute to a broader spectrum of environmental science.
- **Real-time Forecasting Systems:** Developing real-time forecasting systems based on the GLDM model could have significant implications for disaster preparedness, renewable energy management, and aviation safety, among other sectors.
- **Machine Learning Integration:** Investigating the integration of GLDM with machine learning algorithms could yield advanced models that learn from and adapt to changing weather patterns dynamically.

In conclusion, the second-order GLDM model represents a significant step forward in the predictive modeling of wind speed. Its success paves the way for further research aimed at enhancing our understanding of meteorological phenomena and our ability to forecast them accurately. As we continue to refine and expand upon this model, the potential for impactful discoveries and practical applications in weather prediction and beyond remains vast and compelling.

9 REFERENCES

[1] Panyukov, A., Makarovskikh, T., & Abotaleb, M. (2022). Forecasting with using quasilinear recurrence equation. In *International Conference on Optimization and Applications* (pp. 183–195). Springer. https://doi.org/10.1007/978-3-031-22990-9_13

[2] Makarovskikh, T., Panyukov, A., & Abotaleb, M. (2023). Using general least deviations method for forecasting of crops yields. In *International Conference on Mathematical Optimization Theory and Operations Research* (pp. 376–390). Springer. https://doi.org/10.1007/978-3-031-43257-6_28

[3] Panyukov, A. V., & Mezaal, Y. A. (2018). Stable estimation of autoregressive model parameters with exogenous variables on the basis of the generalized least absolute deviation method. *IFAC-PapersOnLine, 51*(11), 1666–1669. https://doi.org/10.1016/j.ifacol.2018.08.217

[4] Panyukov, A. V., & Mezaal, Y. A. (2020). Improving of the identification algorithm for a quasilinear recurrence equation. In *Advances in Optimization and Applications: 11th International Conference, OPTIMA 2020, Moscow, Russia, September 28–October 2, 2020, Revised Selected Papers 11* (pp. 15–26). Springer. https://doi.org/10.1007/978-3-030-65739-0_2

[5] Panyukov, A., Makarovskikh, T., & Abotaleb, M. (2022). Forecasting with using quasilinear recurrence equation. In *International Conference on Optimization and Applications* (pp. 183–195). Springer. https://doi.org/10.1007/978-3-031-22990-9_13

[6] Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*(4), 609–637. https://doi.org/10.1177/0013164415596420

[7] LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*(3), 418–435. https://doi.org/10.1177/1094428107308984

[8] Anderson, M. J., & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation, 62*(3), 271–303. https://doi.org/10.1080/00949659908811936

[9] Ståhl, L., Wold, S., et al. (1989). Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems, 6*(4), 259–272. https://doi.org/10.1016/0169-7439(89)80095-4

[10] Rouder, J. N., Engelhardt, C. R., McCabe, S., et al. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review, 23*, 1779–1786. https://doi.org/10.3758/s13423-016-1026-5

[11] Graham, M. H., & Edwards, M. S. (2001). Statistical significance versus fit: Estimating the importance of individual factors in ecological analysis of variance. *Oikos, 93*(3), 505–513. https://doi.org/10.1034/j.1600-0706.2001.930317.x

[12] Shen, W., Chen, X., Qiu, J., et al. (2020). A comprehensive review of variable renewable energy levelized cost of electricity. *Renewable and Sustainable Energy Reviews, 133*, 110301. https://doi.org/10.1016/j.rser.2020.110301

[13] Widén, J. (2011). Correlations between large-scale solar and wind power in a future scenario for Sweden. *IEEE Transactions on Sustainable Energy, 2*(2), 177–184. https://doi.org/10.1109/TSTE.2010.2101620

[14] Ren, G., Liu, J., Wan, J., et al. (2017). Overview of wind power intermittency: Impacts, measurements, and mitigation solutions. *Applied Energy, 204*, 47–65. https://doi.org/10.1016/j.apenergy.2017.06.098

[15] Ruiz, S. A. G., Barriga, J. E. C., Martínez, J. A. M., et al. (2021). Wind power assessment in the Caribbean region of Colombia, using ten-minute wind observations and ERA5 data. *Renewable Energy, 172*, 158–176. https://doi.org/10.1016/j.renene.2021.03.033

[16] Augustyn, A., & Kamiński, J. (2018). A review of methods applied for wind power generation forecasting. *Polityka Energetyczna, 21*(2), 139–150. https://doi.org/10.33223/epj/96214

[17] Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. In *IEEE North American Power Symposium 2010* (pp. 1–8). https://doi.org/10.1109/NAPS.2010.5619586

[18] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

[19] do Nascimento Camelo, H., Sérgio Lucio, P., Verçosa Leal Junior, J. B., et al. (2018). Innovative hybrid modeling of wind speed prediction involving time-series models and artificial neural networks. *Atmosphere, 9*(2), 77. https://doi.org/10.3390/atmos9020077

[20] Makarovskikh, T., Panyukov, A., & Abotaleb, M. (2023). Using general least deviations method for forecasting of crops yields. In *International Conference on Mathematical Optimization Theory and Operations Research* (pp. 376–390). Springer. https://doi.org/10.1007/978-3-031-43257-6_28

[21] Ochoa, G. V., Alvarez, J. N., & Vanegas Chamorro, M. N. (2019). Data set on wind speed, wind direction and wind probability distributions in Puerto Bolivar – Colombia. *Data in Brief, 27*, 104753. https://doi.org/10.1016/j.dib.2019.104753
[22] Bebi, E., Stermasi, A., Alcani, M., & Cenameri, M. (2023). Assessment of wind potential: The case of Puka region in Albania. *International Journal of Innovative Technology and Interdisciplinary Sciences, 6*(1), 1112–1120. https://doi.org/10.35741/issn.0258-2724.57.6.97
[23] Baballëku, M., Verzivolli, A., Luka, R., & Zgjanolli, R. (2023). Fundamental basic wind speed in Albania: An adoption in accordance with Eurocodes. *Journal of Transactions in Systems Engineering, 1*(2), 56–72. https://doi.org/10.15157/JTSE.2023.1.2.56-72
[24] Qosja, S., Rolle, R., & Gebremedhin, A. (2022). Solving the bottleneck issue of energy supply: Case study of a wind power plant. *International Journal of Innovative Technology and Interdisciplinary Sciences, 5*(2), 874–891. https://doi.org/10.15157/IJITIS.2022.5.2.874-891
[25] Dhoska, K., Bebi, E., Markja, I., et al. (2024). Modelling the wind potential energy for metallurgical sector in Albania. *Scientific Reports, 14*, 1302. https://doi.org/10.1038/s41598-024-51841-x
[26] Jabbar, R. I. (2021). Statistical analysis of wind speed data and assessment of wind power density using Weibull distribution function (Case study: Four regions in Iraq). *Journal of Physics: Conference Series, 1804*, 012010. https://doi.org/10.1088/1742-6596/1804/1/012010
[27] Manusov, V., Matrenin, P., Nazarov, M., Beryozkina, S., Safaraliev, M., & Zicmane, I., Ghulomzoda, A. (2023). Short-term prediction of wind speed based on a learning process control algorithm in isolated power systems. *Sustainability, 15*, 1730. https://doi.org/10.3390/su15021730
[28] Simankov, V., Buchatskiy, P., Teploukhov, S., Onishchenko, S., Kazak, A., & Chetyrbok, P. (2023). Review of estimating and predicting models of the wind energy amount. *Energies, 16*(16), 5926. https://doi.org/10.3390/en16165926
[29] Kadhem, A. A., Abdul Wahab, N. I., Aris, I., Jasni, J., & Abdalla, A. N. (2017). Advanced wind speed prediction model based on a combination of Weibull distribution and an artificial neural network. *Energies, 10*, 1744. https://doi.org/10.3390/en10111744
[30] Hocaoglu, F. O., Gerek, O. N., & Kurban, M. (2010). A novel wind speed modeling approach using atmospheric pressure observations and hidden Markov models. *Journal of Wind Engineering and Industrial Aerodynamics, 98*, 472–481. https://doi.org/10.1016/j.jweia.2010.02.003

Author's contacts:

Mostafa Abotaleb, Assoc. Prof.
Engineering School of Digital Technologies,
Yugra State University,
Khanty-Mansiysk, 628012, Russia
E-mail: abotalebmostafa@bk.ru