


Article

Modeling and Evaluation of Forecasting Models for Energy Production in Wind and Photovoltaic Systems

Imene Benrabia ^{*,†} and Dirk Söffker ^{*,†} 

Chair of Dynamics and Control, University of Duisburg-Essen, 47057 Duisburg, Germany

* Correspondence: imene.benrabia@uni-due.de (I.B.); soeffker@uni-due.de (D.S.)

† These authors contributed equally to this work.

Abstract: The comprehensive change from known, classical energy production methods to the increased use of renewable energy requires new methods in the field of efficient application and use of renewable energy. The urban energy supply presents complex challenges in improving efficiency; therefore, the prediction of the dynamical availability of energy is required. Several approaches have been explored, including statistical models and machine learning using historical data and numerical weather prediction models using mathematical models of the atmosphere and weather conditions. Accurately forecasting renewable energy production involves analyzing factors such as related weather conditions, conversion systems, and their locations, which influence both energy availability and yield. This study focuses on the short-term forecasting of wind and photovoltaic (PV) energy using historical data and machine learning approaches, aiming for accurate 8 h predictions. The goal is to develop models capable of producing accurate short-term forecasts of energy production from both resources (solar and wind), suitable for later use in a model predictive control scheme where generation and demand, as well as storage, must be considered together. Methods include regression trees, support vector regression, and regression neural networks. The main idea in this work is to use past and future information in the model. Inputs for the PV model are past PV generation and future solar irradiance, while the wind model uses past wind generation and future wind speed data. The performance of the model is evaluated over the entire year. Two scenarios are tested: one with perfect future predictions of wind speed and solar irradiance, and another considered realistic situation where perfect future prediction is not possible, and uncertain prediction is accounted for by incorporating noise models. The results of the second scenario were further improved using the output filtering method. This study shows the advantages and disadvantages of different methods, as well as the accuracy that can be expected in principle. The results show that the regression neural network has the best performance in predicting PV and wind generation compared to other methods, with an RMSE of 0.1809 for PV and 5.3154 for wind, and a Pearson coefficient of 0.9455 for PV and 0.9632 for wind.



Academic Editor: Fernando Sánchez Lasheras

Received: 15 December 2024

Revised: 22 January 2025

Accepted: 25 January 2025

Published: 29 January 2025

Citation: Benrabia, I.; Söffker, D. Modeling and Evaluation of Forecasting Models for Energy Production in Wind and Photovoltaic Systems. *Energies* **2025**, *18*, 625. <https://doi.org/10.3390/en18030625>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: wind energy; solar energy; forecasting; machine learning; regression neural network; random forest

1. Introduction and Literature Review

In the modern energy landscape, global electricity consumption in 2019 placed the residential sector second after industry with 26.6% of total consumption, which increased from 12.8 EJ in 2000 to 21.9 EJ in 2019 [1]. Consequently, renewable energy sources are becoming increasingly important as they can be provided from natural processes, dispatchable in different locations, and controllable in energy management schemes.

Energy management systems (EMSs) that utilize predictive methods, such as model predictive control (MPC), heavily rely on the accuracy of predicted variables. Predictive models, therefore, constitute a key element within these management systems. Among prediction models, renewable energy forecasting can anticipate production ramps, enabling the EMS to effectively balance energy during both flow and scheduling phases. This leads to reduced costs, improved system reliability, and minimized curtailment of renewable resources. In this context, wind and solar energy can be suitable selections for such applications, as their accurate forecasting can positively impact multiple operations, including scheduling, dispatching, and real-time balancing.

Traditional forecasting methods have limitations and are often not able to provide accurate predictions. In recent years, forecasting renewable energy production based on its resource characteristics (such as wind speed and solar irradiance) has become an active area of research. Various approaches and methods have been explored in the literature, including those based on numerical weather prediction (NWP) models, statistical methods using historical and real-time data, and hybrid approaches that combine multiple methods [2]. In [3], NWP models are further discussed. These physics-based models utilize mathematical equations such as governing equations, which describe the physical processes of the atmosphere, and data assimilation equations, which adjust the model's initial conditions and correct prediction errors using observational data. The equations can be further divided into global and regional models. Global models simulate the entire Earth's atmosphere, while regional models focus on specific areas. These models require significant computational effort to generate accurate predictions.

In addition to predicting the weather conditions used in energy generation forecasting, other methods directly target energy generation using historical data. These methods, known as statistical and intelligent methods, predict the future behavior of a system based on its past observations. A variety of statistical techniques analyze historical data (wind speed, solar irradiance, etc.) to learn the trend of the time series data (between the input and output) and develop a model for predicting future values. Multiple techniques are used, including moving average (MA), primarily used for smoothing historical data, autoregressive integrated moving average (ARIMA), which combines autoregressive (AR) and MA models while considering the irregular component of a time series, and machine learning algorithms, such as neural networks and support vector regression (SVR) [2,4].

Other approaches can also be combined to form a hybrid model. For example, in [2], seasonal autoregressive integrated moving average (SARIMA), random vector, and neural network approaches were combined for short-term photovoltaic (PV) energy forecasting. The results show that the combined models provide improved forecasting efficiency compared to individual ones. In [4], short-term wind speed was forecasted by applying particle-swarm-based optimization to a least square support vector machine. For short-term solar forecasting, an Elmann neural network was employed [4]. A preliminary examination of the connection between various input parameters was conducted through multivariate regression and the model was trained using different sets of input vectors. Moreover, a long short-term memory (LSTM) model was established to forecast the output of renewable energy in multiple regions, considering the correlation of time series and weather state. Then, particle swarm optimization (PSO) was adopted to optimize the parameters of the LSTM model and improve the forecast accuracy [2].

Furthermore, different works [5,6] have used machine learning methods to model wind and solar energy generation with good forecasting accuracy. In [5], SVR was used to forecast PV energy using real measured solar data from a 6.4 kWp plant and six input variables (day, month, hour, global normal irradiance, temperature, and wind speed). This method may handle nonlinearity in forecasting problems. However, SVR accuracy depends

on the choice of suitable parameters. In [6], short-term wind speed prediction was modeled using different regression tree models, including both linear and non-linear algorithms, and wind speed data from reference stations in wind farms. The methods were compared with multi-layer perceptron neural networks, extreme learning machines, and a support vector machine (SVM). The results showed that the used regression tree models were able to obtain good performances with a small computation time for wind speed prediction.

Other regression models were explored in different studies: ref. [7] used a combined forecasting model, where multiple regression equations were established to forecast monthly wind and PV generation. The equations used monthly production data as independent variables and incorporated monthly average temperature as a correlating factor. A surface fitting–seasonal auto-regressive integrated moving average (SF-SARIMA) was used, and the prediction error of this model ranged from 4.6% to 8.3%. In [8], forecasting day-ahead PV generation was carried out by using a random forest as an ensemble learning method to combine forecasts from SVM models representing present and past PV forecasts, as well as meteorological data. The aggregated mean of the monthly root mean squared error (RMSE) demonstrated that the ensemble method achieved the most accurate forecasts, with a mean RMSE of 0.0725. Similarly, day-ahead forecasting was addressed in [9] using neural networks (ANNs), linear regression, an SVM, and weather factors such as wind speed, humidity, irradiance, and temperature. The results demonstrated that the ANN achieved the highest accuracy. Moreover, the accuracy of the power predictions could be further improved by incorporating real-time weather data from an on-site weather station and pyranometer, leading to enhanced model performance over time. Notably, the ANN achieved the best performance, with an RMSE of 468.23 and an R^2 value of 0.838. In [10], a multi-layer weather-classification-based regression model (MWCR) was developed to predict PV generation. Weather conditions were classified into clear sky, partly cloudy, mostly cloudy, and cloudy. For each class, a regression model was developed using local irradiance, temperature, wind speed, humidity, and a power ratio calculated using LSTM and a physical model. The MWCR model outperformed single regression, conventional neural networks, gated recurrent units, and LSTM models on both training and testing data. PV output power was forecasted in [5] using a modified SVR method where the parameters were optimized using cuckoo search (CS) and differential evolution (DE) algorithms. The SVR model was compared with a radial basis and linear kernels. The results showed that the SVR model with a radial basis function, optimized by CS and DE, achieved the highest accuracy, with an RMSE of 0.137 and an R^2 of 0.99.

Other studies have used more complicated structures, such as neural networks combined with other methods. In [11], forecasting based on quantile regression using gradient-boosted regression trees was approached. The model was trained using numerical weather forecasts for irradiance, total precipitation, and temperature. A deterministic model of the plant was trained using features such as the global horizontal irradiance, clear sky index, zenith and azimuth angles of the apparent position of the sun, and the weather forecast data, with a 24 h horizon assumed to be the closest to real conditions. The test of the PV system model of 1.3 MW was able to provide higher performance than those obtained with other methods. Similarly, a neural network method was used in [12] combined with grey wolf optimization (GWO) for short-term PV energy production forecasting. A general regression neural network (GRNN), which is expected to provide more accurate predictions with shorter computational times, was trained using GWO. The performance of the proposed model was investigated using short-term forecasting in different seasons and was compared to SVM and LSTM methods. The numerical results presented in [12] show that the proposed approach can significantly enhance the prediction accuracy of PV systems.

Summarizing the above discussion, PV and wind energy system predictions are dependent on many variables, mainly related to weather conditions. Traditional complex physical methods may not be well suited to addressing the problem. Instead, machine learning techniques can be used to better understand and address the variability of renewable energies. Previous research has explored various machine learning approaches for this purpose. However, many previous works predominantly relied on past information with multiple inputs along with combined layered methods to improve the forecast, which increased the complexity of the approach. In this study, a novel regression-based approach for forecasting hourly PV and wind energy generation is presented. Incorporating future information in predictive models has proven to generate better results, as in [13], where the prediction of models that consider future intentions performed better than models that do not account for this information. Therefore, the main idea in this work is to use the combination of past and future information as a training process for known machine learning methods, to reduce the reliance on different parameters as inputs, and to simplify the problem by accounting for one-layer methods.

The models are developed using a regression neural network, random forest, and SVR. This method considers future weather data, as well as historical data consisting of wind speed and solar irradiance, to train and validate different machine learning models for short-term predictions. Two different tests are conducted: the first test relates to using historical weather and energy data along with future weather data, and the second test relates to using historical weather and energy data along with non-accurate future weather data generated by noise models. The forecasting models' results are then further improved using a filtering method.

After the introduction and literature review, this paper is structured as follows: in Section 2, the theoretical background of the used methods is explained. The data collection and preprocessing are discussed in Section 3; additionally, the machine learning algorithms used in this paper are elaborated. In Section 4, the obtained results along with the used assessment parametrics are provided. Furthermore, the results are interpreted and discussed to conclude their implications and limitations.

2. Theoretical Background

Based on the literature, machine learning methods can be applied to model renewable energy systems. This study focuses on regression methods that use historical data to learn the relationship between input variables (e.g., wind speed, solar irradiance) and output variables (e.g., renewable generation). The obtained models can then be used to predict future hourly PV and wind energy production. In the following subsections, we provide a comprehensive explanation of regression tree, random forest, regression neural network, and SVR methods.

2.1. Regression Tree

A regression tree (RT) is a machine learning method for regression and classification, where leaf nodes represent regression models predicting numeric responses [11,14]. The tree splits data based on input features, forming parent and child nodes. Each parent node computes an equation to predict the target variable for its child nodes. The tree can then be pruned by removing branches that do not significantly reduce prediction errors, calculated as the weighted average absolute difference between predicted and actual target values. The process continues until all instances are covered by at least one rule in the tree [14].

Based on the regression tree, methods like the random forest (RF) emerged, which is an ensemble learning technique that combines multiple decision trees trained on random subsets of data to improve prediction accuracy and stability [15]. It is suitable for

handling high-dimensional data and non-linear relationships between features and target variables [6]. Given the large dataset in this work and the need to estimate production under stochastic weather changes, RFs are chosen in this study to address these challenges.

The RT elaborated in the RF of this study splits the data into smaller subsets based on the values of the predictor variables $x_i (i = 1 \dots n)$. At each node l , the tree splits the data into two or more branches based on a chosen split point. The goal of the tree is to identify the most important variables that can accurately predict the target variable. The tree computes the weighted mean squared error (MSE) of the responses in node l using

$$MSE_l = \sum_{j=1}^T w_j (y_j - \bar{y}_j)^2, \quad (1)$$

where w_j is the weight of observation j and T is the number of observations in node l . Accordingly, the probability that an observation is in node l is calculated by

$$P(T) = \sum_{j=1}^T w_j. \quad (2)$$

Moreover, the RT determines the best way to split node l using x_i by maximizing the reduction in MSE over all splitting candidates.

In this work, the RF algorithm is implemented using the regression tree method to construct each tree in the ensemble. Each regression tree is built with the classification and regression tree (CART) algorithm, which uses recursive partitioning [14]. At each internal node, the data are split based on an input feature value, chosen to minimize the mean squared error (MSE) among observations in the child nodes [16].

Since each tree is trained on a random subset of features and data, their predictions may vary. The RF algorithm combines these predictions by averaging them for regression tasks, resulting in the final prediction (Figure 1).

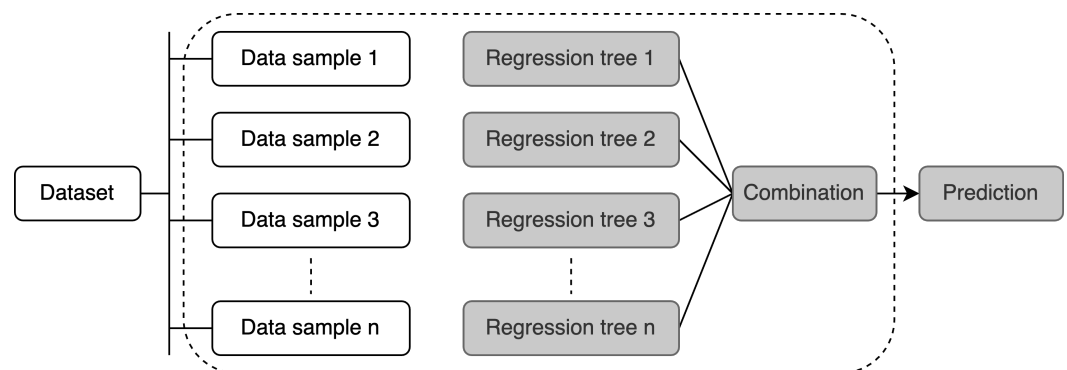


Figure 1. Random forest method.

2.2. Support Vector Regression

A support vector machine is a supervised learning algorithm typically used for classification problems by finding the best boundary that separates the data into different classes, maximizing the distance between the boundary and the closest data points from each class [4].

Support vector regression is a type of SVM used for regression problems. Instead of finding a boundary that separates the data into different classes, SVR aims to find the best line that fits the data and predicts the value of the target variable. Unlike SVMs, which classify data based on their class labels, SVR classifies data based on the regression errors by

identifying the errors that are greater or less than a specific threshold and then performing the classification of the data based on the identified errors [4,5].

2.3. Regression Neural Network

Regression neural networks are feedforward neural networks, also known as multi-layer perceptrons (MLPs), which are methods that aim to predict the value of a continuous target variable based on a set of input features. The basic architecture of a regression-NN is represented by

$$y_i = \sum_{m=1}^l f(W_{i,m}x_{i,m} + b_{i,m}), \quad (3)$$

where y_i denotes the predicted output in the layer i ($i = 1 \dots n$), $x_{i,m}$ is the input data features ($m = 1 \dots l$), $W_{i,m}$ and $b_{i,m}$ are the parameters (weights and biases) of the network, and f is the activation function (e.g., sigmoid, ReLU). This equation represents the output of a single layer of the neural network [12]. The weights and biases of the network are learned during the training process, using an optimization algorithm to minimize the difference between the predicted output and the actual output.

In this study, a feedforward neural network is elaborated. The information flows in one direction, from the input layer to the output layer, passing through hidden layers in between. There are no cycles or loops in the network, and the output of any layer is not fed back to the same layer or previous layers. Moreover, the layer structures of the used networks are indicated in Figure 2.

The network uses a limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm [17] to minimize the loss function of the network. The loss function in this case is the mean squared error (MSE). The LBFGS solver is an optimization algorithm that uses an approximation of the Hessian matrix (a matrix of second partial derivatives) to determine the direction of the next step in the optimization process. It also uses a standard line-search method, which is a method for finding the next point to evaluate in the optimization process.

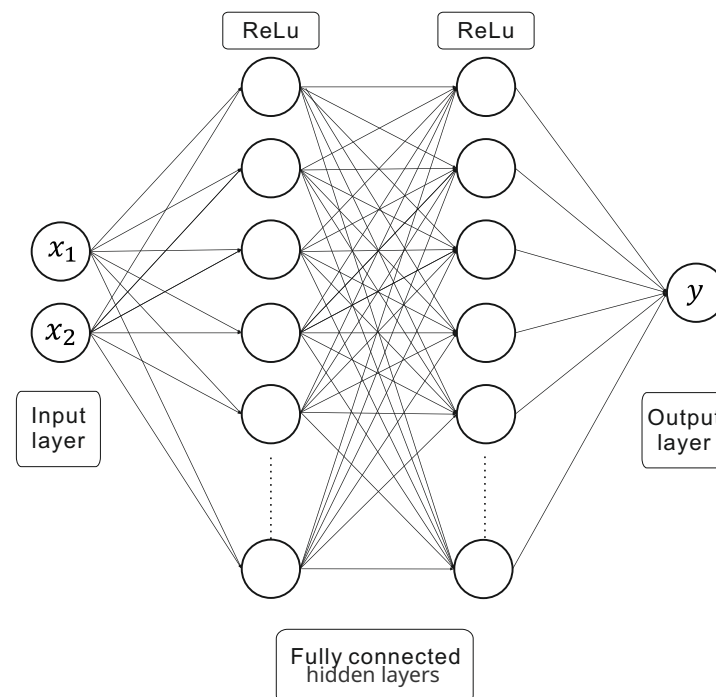


Figure 2. Regression neural network method.

3. Generated Models and Used Datasets

In this section, the process of developing the models is discussed. Furthermore, the used datasets and the modeling of the regression-NN, SVR, and RF are explained.

3.1. Data Source

For both PV and wind systems, weather data represented by wind speed, global irradiance, and energy production are integrated. Global irradiance is calculated as the sum of direct and diffuse irradiance:

$$I_{global} = I_{direct} + I_{diffuse}. \quad (4)$$

The database used to develop the forecasting models [18,19] uses reanalysis models for wind and PV prediction. Considering a 100 kWp capacity wind turbine system of type GE 1.5se and a 100 kWp capacity PV system with a tilt of 35 degree, energy production data are generated.

The PV and wind output data are separated with 1 h intervals for the period of four years from January to December, 2016 to 2019. Together with corresponding weather parameters, the entire dataset is divided into two subsets, namely the training dataset $Data_{train}$ (2016 and 2017 data) and the test dataset $Data_{test}$ (2018 and 2019 data), such that

$$Data = Data_{train} + Data_{test}. \quad (5)$$

In Figures 3 and 4, the wind and solar energy are shown along with the variations in wind speed and solar irradiance over time. All figures are plotted with respect to time, with the data returning to the year 2017, serving as an example for the analysis. The choice of a daily time frame (24 h) in this analysis allows one to observe the varying trends in the data over the course of the year from a daily perspective.

As can be seen from the figures, solar energy is strongly influenced by solar irradiance, while wind energy depends primarily on wind speed.

Given this observation for modeling and simulation, only past wind energy and future wind speed data are used to train the wind energy model. Similarly, past solar energy and future solar irradiance data are used to train the solar energy model.

In this study, two years of data are used for training to ensure that the model captures all relevant properties of the data, such as trends within a year and trends between two different years. The model's prediction accuracy is tested against data from the next two years. To verify the predictive capability for a short horizon (hourly forecast), data from a two-year run are used for hourly forecasts under various conditions throughout the entire annual run.

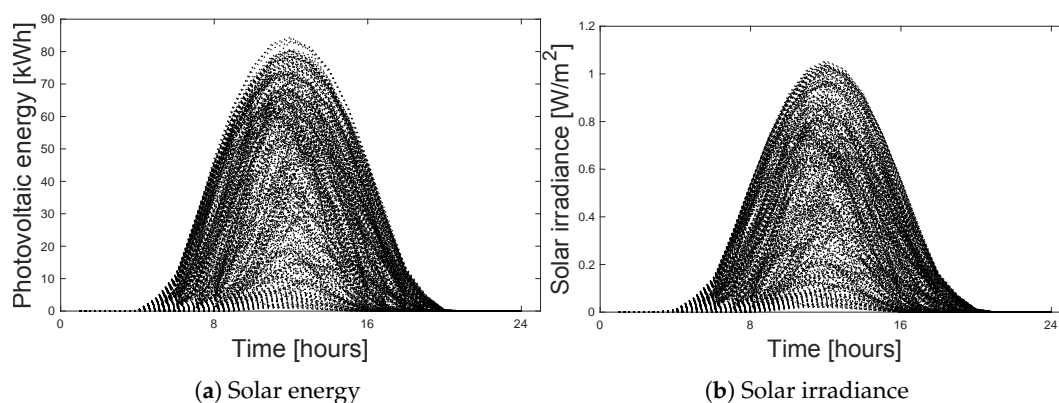


Figure 3. Daily solar energy and irradiance (year 2017).

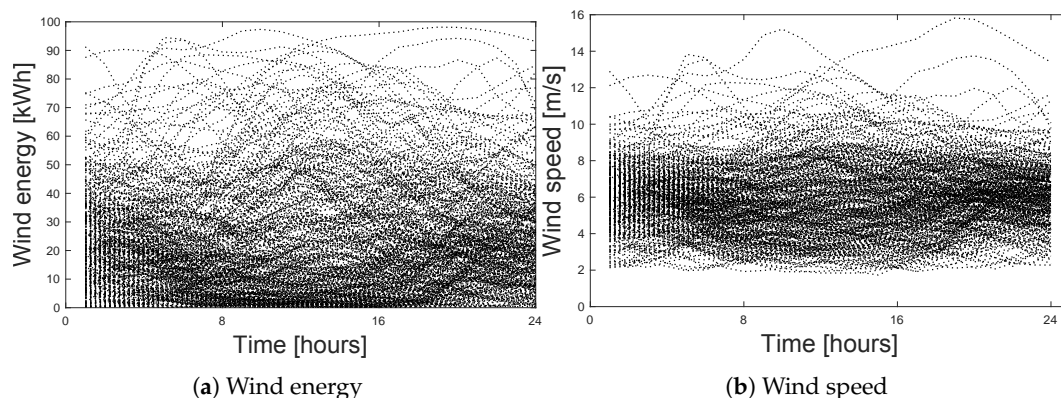


Figure 4. Daily wind energy and wind speed (year 2017).

3.2. Modeling Method

The main objective of this study is to develop accurate forecasts of the future 8 h generation output of both PV and wind energy systems. To achieve this goal, a two-part approach is applied for constructing the predictive models.

The first part of the approach involves using real data to train the models by utilizing information regarding the future wind speed and solar irradiance, along with past data on the produced energy (Figure 5). The second part is based on constructing a test method to assess the performance of the models. This involves modifying the input data by introducing noise to simulate non-accurate measurements, which allows one to evaluate the models' ability to handle disruptions.

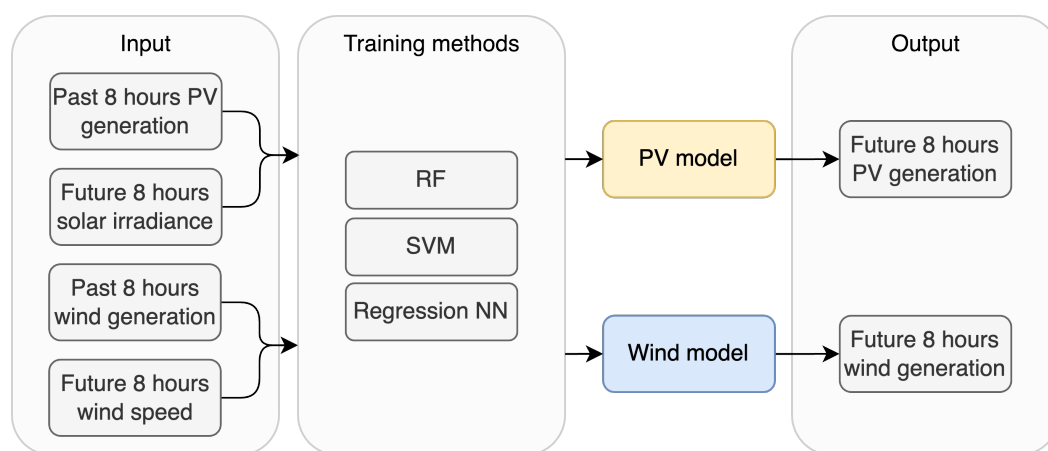


Figure 5. Training process of wind and PV models.

3.3. Noise Models

To model a noisy trend of real wind speed and solar irradiance data, a stochastic process with a known distribution is added to the data. A stochastic process is a mathematical model used to describe the evolution of a random variable over time. It is a collection of random variables that depend on time, and it is used to model various phenomena that involve randomness, such as weather patterns, as stated in [20,21].

As the models in this work account for future data related to weather conditions (solar irradiance and wind speed), the mentioned stochastic processes are used to emulate the forecasting errors in the data. Adding noise to a signal will introduce a stochastic process into the data. The noise itself is modeled as a random process with a known distribution, such as Gaussian or uniform noise. The noise properties can be defined by specifying the distribution and the parameters of the noise process.

In this work, different types of noise are generated and added to the core values of smoothed real data (wind speed and solar irradiance), which are obtained by using the moving average method [20]. The moving average method calculates the mean of the data over a fixed window, which will reduce the periodic trends in the data and their own real noise. The data modification is represented by

$$V_{new} = V_{smooth} + V_{noise}, \quad (6)$$

$$I_{new} = I_{smooth} + I_{noise}. \quad (7)$$

The noise models for the wind speed and solar irradiance are generated by using the standard normal distribution (SND) and the white Gaussian noise (WGN), given that these types of noise have been considered in past studies for performing simulations of weather noise patterns [20–22].

3.4. Statistical Metrics for Data Validation

In this subsection, the models' prediction accuracy is assessed by various indicators. The first statistical metric for evaluating prediction models is RMSE [5], which gives an indication about the overall measure of the errors magnitude as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}, \quad (8)$$

where N is the number of datasets involved in the analysis, y_i is the actual value to be estimated, and \hat{y} is the forecasted value.

From the literature, another often-used metric is called the mean absolute percentage error (MAPE) [5], calculated by

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}}{y_i} \right| * 100, \quad (9)$$

where N , y_i , and \hat{y} are the same as mentioned for the RMSE. The MAPE shows the absolute error in percentage [5]. However, the MAPE has a significant disadvantage: it produces infinite or undefined values when the actual values are zero or close to zero, which is common in some renewable energy productions, such as solar energy; when the panels do not receive sunlight, the irradiance values fall down gradually until they reach zero. Moreover, in order to measure the difference between the real data and the modified ones, the Pearson correlation coefficient is used to calculate how close two variables are to each other. The Pearson correlation coefficient [23] is described by

$$P_{coef} = \frac{cov(x, y)}{\sigma_x \sigma_y}, \quad (10)$$

where cov is the covariance between data x and y , σ_x is the standard deviation of data x , and σ_y is the standard deviation of data y . A correlation coefficient of 1 indicates a perfect correlation between the two datasets.

4. Simulation and Results

Using the previously described methods, three distinct forecasting models are developed: a random forest, SVR, and a regression neural network. All simulations in this work are conducted using Matlab R2023b software.

The development of these models follows a basic workflow that involves several key steps. The first step is about gathering and organizing the data. The used dataset is

obtained from [18,19] and explained in detail in Section 3. In Figure 6, taking wind energy as an example, the graphical expression of the data inputs and outputs for the wind energy model is shown. The same procedure is carried out for training the solar energy model. Here, the used data for illustration are a 16 h set from the 1 January 2016. The model is trained by introducing past 8 h energy data (solar and wind) and future 8 h weather data (irradiance and wind speed) as inputs to forecast the future 8 h of energy production as the output. The next step concerns selecting the method from the already mentioned list of methods for each model, taking into consideration the specific characteristics of the data and the desired outcomes.

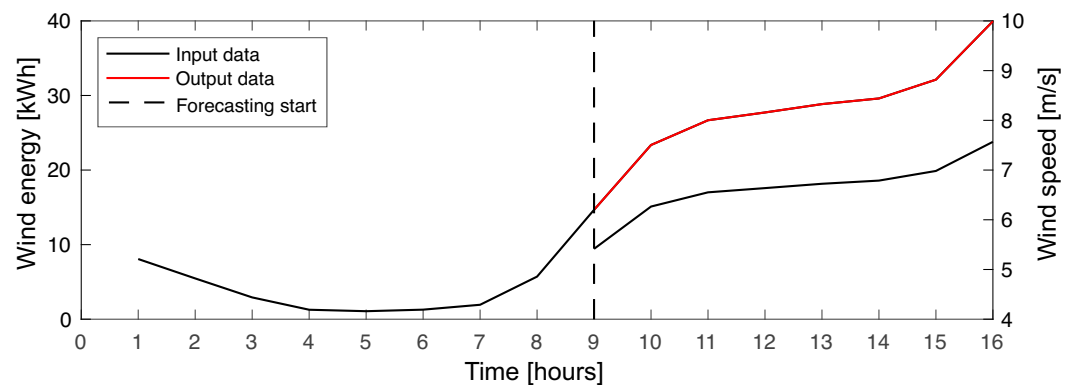


Figure 6. Illustration of the wind energy model training.

To ensure that the model is suitable for the given data, the fit is examined and any necessary updates are made until the accuracy and reliability of the model are improved.

Finally, the fitted model is used for predictions. The results are compared to the real data using the assessment factors mentioned in Section 3.

4.1. Models Training

Firstly, training occurs using the hourly data from 2016 to 2017, which amount to 17,520 data points. The structure of the chosen models is as follows.

- First model: The RF is built by the combination of 10 regression trees for the wind energy model and 13 for the solar energy model.
- Second model: The SVR model is built considering the standardization of the data by centering and scaling each variable by the corresponding mean and standard deviation.
- Third model: The regression-NN is built considering the same standardization as the second model, along with using two fully connected layers of 13 nodes each for the wind energy model and 20 nodes each for the solar energy model, both with ReLu as an activation function.

4.2. Wind Speed and Irradiance Noise Models

As already mentioned, white Gaussian noise with a standard normal distribution is used to simulate the noise trend in the modified data. The aim of these data is to test the energy prediction models' robustness toward the errors in the future weather data. To apply this modification, a moving average is used on the real data with a fixed window of four points. The used wind speed and solar irradiance data in the simulations of the results are the whole data for the period from 2018 to 2019. The correlation coefficients and RMSE results are presented in Table 1.

The notations w and s in the coefficients correspond to wind speed and solar irradiance, respectively. The correlation metric provides an indication of the degree to which the modified wind data and solar irradiance aligns with the overall trend of the original data.

The RMSE measures the level of deviation between the modified data and the actual wind data, taking into account the impact of the added noise and fluctuations. In this case, a high RMSE value indicates a significant level of differences between the modified data and the actual wind data.

Table 1. Correlation coefficient and RMSE of noisy wind speed and solar irradiance from 2018 to 2019.

Method	P_{coefw}	P_{coefs}	$RMSE_w$ (kWh)	$RMSE_s$ (kWh)
SND	0.9506	0.9216	0.6164	0.1748
WGN	0.9398	0.9031	0.6868	0.1961

The observed high RMSE value in wind speed data can be attributed to the inherent characteristics of wind, which tend to exhibit more fluctuations throughout most hours of the day compared to solar irradiance data. This is due to the fact that solar irradiance is directly limited by the number of sunlight hours, whereas wind patterns can be influenced by a variety of complex factors that contribute to variability and unpredictability.

4.3. Test and Discussion

The test process is summarized in Figure 7. The initial test involves using real data. In terms of metrics, both the Pearson correlation coefficient and the RMSE error are considered. In Table 2, the results obtained from the entire dataset spanning from 2018 to 2019 are presented.

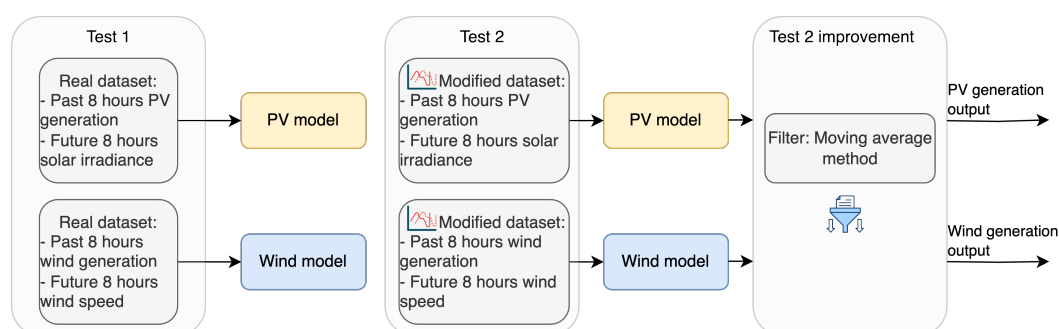


Figure 7. Test process for wind and PV models.

Table 2. Test results using real data from 2018 to 2019.

Method	$RMSE$ (kWh)	P_{coef}
Wind energy: Regression-NN	0.0959	1
Wind energy: RF	0.4651	0.9997
Wind energy: SVR	4.0564	0.9796
Solar energy: Regression-NN	0.0009	1
Solar energy: RF	0.0035	0.9999
Solar energy: SVR	0.0150	1

For the predicted energy evaluation, MAPE is expressed as a percentage of the relative error between actual and forecasted values, while RMSE is expressed in kWh. Regarding wind energy prediction, the regression-NN model has the lowest RMSE value of 0.0959, suggesting that its predictions are closest to the actual data. Moreover, it has a perfect Pearson correlation coefficient of 1, indicating a strong linear relationship between predicted and actual values. The RF model has a higher RMSE of 0.4651, but still has a very high Pearson correlation coefficient of 0.9997. The SVR model has the highest RMSE value of 4.0564, indicating that its predictions have a larger deviation from the actual data, but it still has a strong Pearson correlation coefficient of 0.9796. The obtained results suggest that regression-NN and RF models perform better than SVR for wind energy prediction.

For solar energy prediction, the regression-NN model has the lowest RMSE value of 0.0009 along with a perfect Pearson correlation coefficient of 1, indicating that this model fits the actual data most closely and has a strong linear relationship between predicted and actual values. The RF model has a higher RMSE of 0.0035, but still has a very high Pearson correlation coefficient of 0.9999. Finally, the SVR model has an RMSE value of 0.0150, suggesting a larger deviation from the actual data, but it still has a perfect Pearson correlation coefficient of 1. As a conclusion, regression-NN and RF also perform better than SVR for solar energy prediction.

The observed behavior of all the models presenting a high accuracy can be interpreted considering the nature of the case. The high accuracy result is expected as, at every moment, wind energy is highly correlated with wind speed, and the same applies to PV energy with solar irradiance. Using perfect future wind speed and solar irradiance will eventually lead to very good results.

In general, a model with a lower RMSE and a higher Pearson correlation coefficient is considered as more useful because it allows for a better prediction of energy production. The regression-NN model, for example, performed well in both wind and solar energy prediction, suggesting that it may have been able to capture the underlying patterns in the data more effectively than the other models. However, each model has its own strengths and weaknesses, and the choice of the best model for a given energy prediction task depends on several factors. Accordingly, the models' robustness is tested using the generated noisy data for future wind speed and solar irradiance.

To perform this, noisy data using the WGN and the SND noise models are implemented in the already trained models. The results regarding the whole dataset from 2018 to 2019 are shown in Table 3.

Table 3. Test results using modified data from 2018 to 2019.

Metrics	$RMSE_{wgn}$ (kWh)	P_{wgn}	$RMSE_{snd}$ (kWh)	P_{snd}
Wind: Regression-NN	6.8570	0.9408	6.1007	0.9524
Wind: RT	6.8729	0.9404	6.1137	0.9522
Wind: SVR	7.7939	0.9241	7.2319	0.9339
Solar: Regression-NN	0.1958	0.9033	0.1745	0.9218
Solar: RT	0.1936	0.9026	0.1732	0.9212
Solar: SVR	0.1999	0.9031	0.1794	0.9216

In this case, the noise models have a significant role in modifying the results of the prediction models. From Table 3, it can be seen that the results show that all models have a relatively higher RMSE and lower Pearson correlation coefficients when the data are corrupted with the used standard normal deviation noise compared to white Gaussian noise.

In terms of the specific models, the regression-NN and RF models perform similarly for both types of noise, while the SVR model has a larger RMSE and lower Pearson correlation coefficients. This indicates that the regression-NN and RF models may be more robust to different types of noise compared to the SVR model.

For wind energy production, the regression-NN and RF models have similar performance in both types of noise, with RMSE values between 6.1 and 6.9 and Pearson correlation coefficients between 0.9404 and 0.9524. The SVR model performs worse in both types of noise, with RMSE values around 7.2 to 7.8 and Pearson correlation coefficients between 0.9241 and 0.9339.

For solar energy production, the regression-NN and RF models again have similar performance in both types of noise, with very low RMSE values of around 0.17 to 0.20 and Pearson correlation coefficients around 0.9026 to 0.9218. The SVR model performs slightly

worse, with RMSE values around 0.18 and Pearson correlation coefficients around 0.9031 to 0.9216.

The performance of the regression-NN and RF models appears to be similarly good for both wind and solar energy production, while the SVR model performs worse. The difference in performance between the two types of noise is not very significant. Additionally, solar energy models have a lower RMSE and higher Pearson correlation coefficients compared to wind energy prediction, likely due to the higher predictability and stability of solar energy compared to wind energy.

Moreover, in order to visualize the results, the solar energy prediction using WGN noise and wind energy prediction using SND noise are plotted in Figures 8 and 9. The first week of June 2018 is chosen as an example. It can be seen that the prediction follows the trend of the energy production: due to noisy spikes in the wind speed and irradiance inputs, the predicted energy has many disturbances compared to the real energy production data.

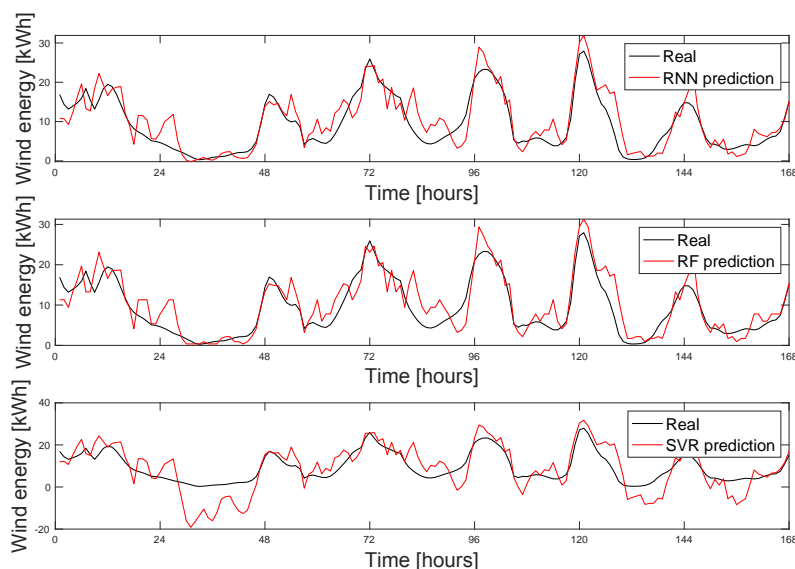


Figure 8. Wind energy prediction using the SND noise for a period of one week.

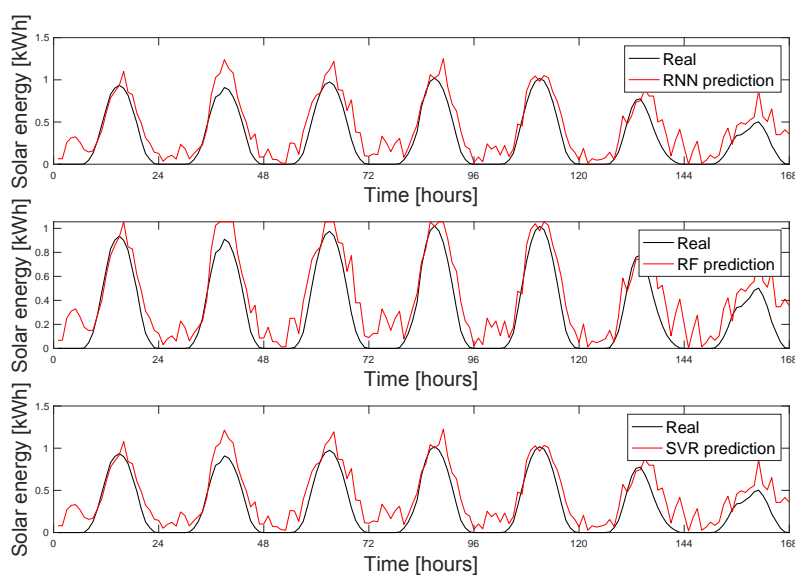


Figure 9. Solar energy prediction using the WGN noise for a period of one week.

To enhance the outcomes of the models, a filtering technique is proposed. Filtering is a commonly used approach for handling noisy data. To ensure the simplicity and

reproducibility of this model with different datasets, a moving average method with a fixed window of 3 h is selected. The simulation evaluation results are presented in Table 4.

Table 4. Test results using filtered and modified data from 2018 to 2019.

Metrics	$RMSE_{wgn}$ (kWh)	P_{wgn}	$RMSE_{snd}$ (kWh)	P_{snd}
Wind: Regression-NN	5.9664	0.9540	5.3154	0.9632
Wind: RT	5.9736	0.9538	5.3195	0.9631
Wind: SVR	7.0248	0.9370	6.5938	0.9443
Solar: Regression-NN	0.1809	0.9345	0.1623	0.9455
Solar: RT	0.1794	0.9342	0.1615	0.9450
Solar: SVR	0.1867	0.9344	0.1688	0.9453

For wind energy, and taking the regression-NN model as an example, the RMSE for the non-filtered data is 6.8570, while the RMSE for the filtered data is 5.9664. This means that the filtering technique results in an RMSE reduction of 0.8906. Similarly, the Pearson correlation has an improvement of 0.0132. In general, the filtered data result in a lower RMSE and higher Pearson correlation compared to non-filtered data for all models, indicating that the filtering technique helps in improving the accuracy of the models' predictions.

Considering the RF model for solar energy, the RMSE for the non-filtered data is 0.1936 while the RMSE for the filtered data is 0.1794. This means that the filtering technique results in an RMSE reduction of 0.0142. Similarly, the Pearson correlation has an improvement of 0.0316. In general, the filtered data result in a slightly lower RMSE and higher Pearson correlation compared to non-filtered data for all models, indicating that the filtering technique helps in improving the accuracy of the models' predictions, although the improvement is not as significant as in the wind energy case.

Furthermore, in Figures 10 and 11, the results are shown for illustration; here, the effect of using the filter on the output is demonstrated. It is evident that the filtered output significantly reduces the disturbances and tracks the trend of the energy production, which shows that using the smoothing method has a good impact on both wind and solar energy models.

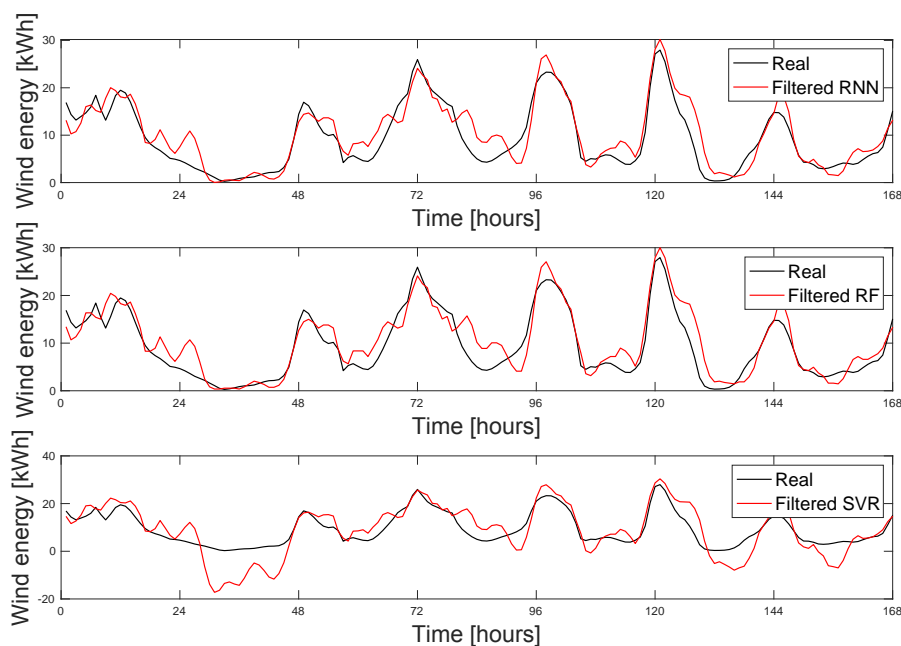


Figure 10. Filtered wind energy prediction using the SND noise for a period of one week.

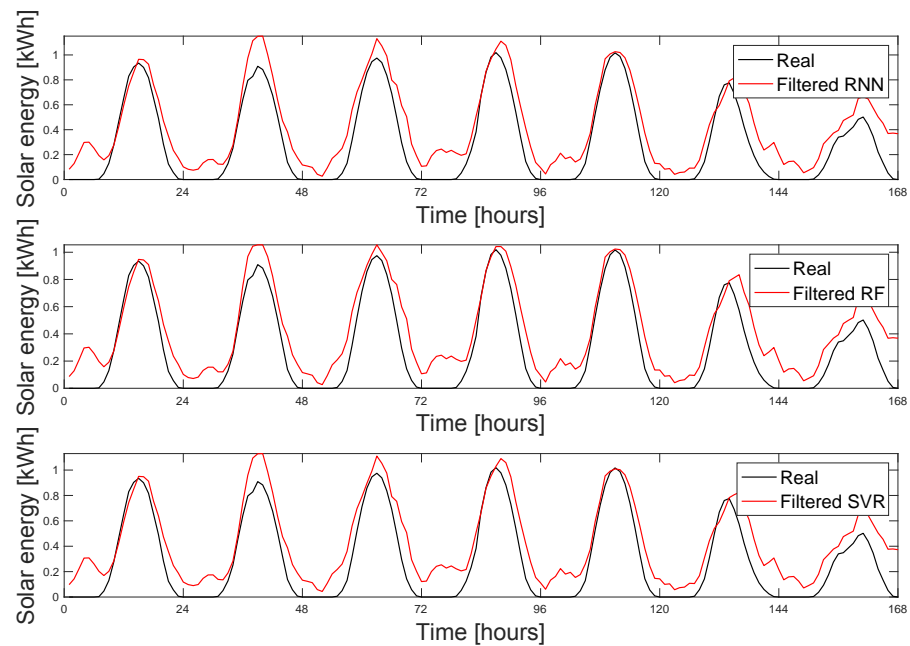


Figure 11. Filtered solar energy prediction using the WGN noise for a period of one week.

5. Summary and Conclusions

Photovoltaic and wind energy systems are highly reliant on weather conditions, which can result in fluctuations in energy output. To overcome this challenge, machine learning techniques have emerged as a promising approach for energy generation from these sources. In this study, a novel regression-based approach is developed for the hourly prediction of PV and wind energy generation, utilizing future weather data (wind speed and solar irradiance) and historical energy data. The approach is tested using real wind and solar energy data. Optimal performance with the lowest RMSE and highest correlation is demonstrated for both PV and wind energy. Furthermore, the approach is tested using only past historical weather and energy data, along with non-accurate future weather data. This is important as non-accurate data can be a problem in data collection and sensor readings. Based on the results, it can be concluded that incorporating a filtering technique, specifically a moving average method with a fixed window, might enhance the accuracy and reproducibility of energy prediction models when using non-accurate wind and solar irradiance data. The results show that the filtering technique improves both RMSE and Pearson correlation metrics for all models, with more improvements observed for wind energy.

In the future, the models developed in this work can be extended to incorporate independent forecasting models for wind speed and solar irradiance using advanced methods such as machine learning. This extension will address further the limitation of using exact future data by integrating realistic, data-driven prediction models where the stochasticity error is generated from the prediction itself, differently from the noise modeling method used in this contribution. Furthermore, this work will be applied to applications in energy management.

Author Contributions: Conceptualization, I.B. and D.S.; methodology, I.B. and D.S.; software, I.B.; validation, I.B. and D.S.; formal analysis, I.B. and D.S.; investigation, I.B. and D.S.; resources, I.B.; data curation, I.B.; writing—original draft preparation, I.B. and D.S.; writing—review and editing, I.B. and D.S.; visualization, I.B.; supervision, D.S.; funding acquisition, I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study were obtained from publicly available open sources, which are appropriately cited in the work and listed in the references section.

Acknowledgments: This work is supported through a scholarship awarded to the first author by the German Academic Exchange Service (DAAD) for her Ph.D. study at the Chair of Dynamics and Control, UDE, Germany.

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

References

1. IEA. *Key World Energy Statistics 2021*; IEA: Paris, France, 2021; Licence: CC BY 4.0. Available online: <https://www.iea.org/reports/key-world-energy-statistics-2021> (accessed on 10 January 2025).
2. Harrou, F.; Sun, Y. *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*; IntechOpen: Rijeka, Croatia, 2020. [CrossRef]
3. Pu, Z.; Kalnay, E. Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation. In *Handbook of Hydrometeorological Ensemble Forecasting*; Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H.L., Schaake, J.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–31. [CrossRef]
4. Prema, V.; Bhaskar, M.S.; Almakhlles, D.; Gowtham, N.; Rao, K.U. Critical Review of Data, Models and Performance Metrics for Wind and Solar Power Forecast. *IEEE Access* **2022**, *10*, 667–688. [CrossRef]
5. Alrashidi, M.; Alrashidi, M.; Pipattanasomporn, M.; Rahman, S. Short-Term PV Output Forecasts with Support Vector Regression Optimized by Cuckoo Search and Differential Evolution Algorithms. In Proceedings of the IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, 16–19 September 2018; pp. 1–8. [CrossRef]
6. Troncoso, A.; Salcedo-Sanz, S.; Casanova-Mateo, C.; Riquelme, J.C.; Prieto, L. Local models-based regression trees for very short-term wind speed prediction. *Renew. Energy* **2015**, *81*, 589–598. [CrossRef]
7. Du, X.; Lang, Z.; Liu, M.; Wu, J. Regression Analysis and Prediction of Monthly Wind and Solar Power Generation in China. *Energy Rep.* **2024**, *12*, 1385–1402. [CrossRef]
8. Abuella, M.; Chowdhury, B. Random forest ensemble of support vector regression models for solar power forecasting. In Proceedings of the 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 23–26 April 2017; pp. 1–5. [CrossRef]
9. Kuriakose, A.M.; Kariyalil, D.P.; Augusthy, M.; Sarath, S.; Jacob, J.; Antony, N.R. Comparison of Artificial Neural Network, Linear Regression and Support Vector Machine for Prediction of Solar PV Power. In Proceedings of the 2020 IEEE Pune Section International Conference (PuneCon), Pune, India, 16–18 December 2020; pp. 1–6. [CrossRef]
10. Bahij, Z.; Yan, B. A Multi-layer Weather Classification-based Regression Model for PV Power Prediction. In Proceedings of the 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), Bari, Italy, 28 August–1 September 2024; pp. 3231–3236. [CrossRef]
11. Massidda, L.; Marrocu, M. Quantile Regression Post-Processing of Weather Forecast for Short-Term Solar Power Probabilistic Forecasting. *Energies* **2018**, *11*, 1763. [CrossRef]
12. Tu, C.S.; Tsai, W.C.; Hong, C.M.; Lin, W.M. Short-Term Solar Power Forecasting via General Regression Neural Network with Grey Wolf Optimization. *Energies* **2022**, *15*, 6624. [CrossRef]
13. Thind, N.S.; Soffker, D. Probabilistic Ship Behavior Prediction Using Generic Models. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 250–255. [CrossRef]
14. Breiman, L. *Classification and Regression Trees*, 1st ed.; Routledge: New York, NY, USA, 1984. [CrossRef]
15. Zhou, Z.; Li, X.; Wu, H. Wind Power Prediction based on Random Forests. In Proceedings of the 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEECS), Jinan, China, 15–16 October 2016; pp. 352–356. [CrossRef]
16. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
17. Nocedal, J.; Wright, S.J. *Numerical Optimization*, 2nd ed.; Springer: New York, NY, USA, 2006; ISSN 1431-8598. [CrossRef]
18. Pfenninger, S.; Staffell, I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy* **2016**, *114*, 1251–1265. [CrossRef]
19. Staffell, I.; Pfenninger, S. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* **2016**, *114*, 1224–1239. [CrossRef]
20. Bright, J.M.; Smith, C.J.; Taylor, P.G.; Crook, R. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. *Sol. Energy* **2015**, *115*, 229–242. [CrossRef]
21. Junlakarn, S.; Diewvilai, R.; Audomvongseeree, K. Stochastic Modeling of Renewable Energy Sources for Capacity Credit Evaluation. *Energies* **2022**, *15*, 5103. [CrossRef]

22. Spiliotis, E.; Petropoulos, F.; Nikolopoulos, K. The Impact of Imperfect Weather Forecasts on Wind Power Forecasting Performance: Evidence from Two Wind Farms in Greece. *Energies* **2020**, *13*, 1880. [[CrossRef](#)]
23. Jebli, I.; Belouadha, F.Z.; Kabbaj, M.I.; Tilioua, A. Prediction of solar energy guided by pearson correlation using machine learning. *Energy* **2021**, *224*, 120109. [[CrossRef](#)]

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

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

ub | universitäts
bibliothek

This text is made available via DuEPublico, the institutional repository of the University of Duisburg-Essen. This version may eventually differ from another version distributed by a commercial publisher.

DOI: 10.3390/en18030625

URN: urn:nbn:de:hbz:465-20250730-090445-5



This work may be used under a Creative Commons Attribution 4.0 License (CC BY 4.0).