algorithms

MDPI

Article

Deep Feature Selection of Meteorological Variables for LSTM-Based PV Power Forecasting in High-Dimensional Time-Series Data

Husein Mauladdawilah [1,*], Mohammed Balfaqih [2], Zain Balfagih [3], María del Carmen Pegalajar [4] and Eulalia Jadraque Gago [1]

[1] School of Civil Engineering, University of Granada, 18001 Granada, Spain; ejadraque@ugr.es
[2] Department of Computer and Network Engineering, University of Jeddah, Jeddah 22254, Saudi Arabia; mabalfaqih@uj.edu.sa
[3] Effat College of Engineering, Effat Energy and Technology Research Center, Effat University, Jeddah 22332, Saudi Arabia; zbalfagih@effatuniversity.edu.sa
[4] Department of Computer Science and Artificial Intelligence, University of Granada, 18014 Granada, Spain
* Correspondence: mauladdawilah@correo.ugr.es

Abstract

Accurate photovoltaic (PV) power forecasting is essential for grid integration, particularly in maritime climates with dynamic weather patterns. This study addresses high-dimensional meteorological data challenges by systematically evaluating 32 variables across four categories (solar irradiance, temperature, atmospheric, hydrometeorological) for day-ahead PV forecasting using long short-term memory (LSTM) networks. Using six years of data from a 350 kWp solar farm in Scotland, we compare satellite-derived data and local weather station measurements. Surprisingly, downward thermal infrared flux—capturing persistent atmospheric moisture and cloud properties in maritime climates—emerged as the most influential predictor despite low correlation (1.93%). When paired with precipitation data, this two-variable combination achieved 99.81% R^2, outperforming complex multivariable models. Satellite data consistently surpassed ground measurements, with 9 of the top 10 predictors being satellite derived. Our approach reduces model complexity while improving forecasting accuracy, providing practical solutions for energy systems.

Keywords: deep learning; forecasting; long short-term memory; mean absolute; meteorological variables

Academic Editor: Muhammad Adnan Khan

Received: 7 July 2025
Revised: 31 July 2025
Accepted: 8 August 2025
Published: 10 August 2025

Citation: Mauladdawilah, H.; Balfaqih, M.; Balfagih, Z.; Pegalajar, M.d.C.; Gago, E.J. Deep Feature Selection of Meteorological Variables for LSTM-Based PV Power Forecasting in High-Dimensional Time-Series Data. *Algorithms* **2025**, *18*, 496. https://doi.org/10.3390/a18080496

1. Introduction

The global transition to renewable energy faces a critical challenge: integrating variable power sources into existing electrical grids. Photovoltaic (PV) systems, while promising, require sophisticated forecasting methods to ensure grid stability and efficient operation. The IEA projects PV capacity to reach 1277 GW by 2025 and 4815 GW by 2040 [1]. However, the variability of PV output, influenced by both internal factors (module types, inverter technology, array tilt) and external factors (weather conditions, shading, dust accumulation), presents significant challenges for grid integration [2].

The accuracy of PV power forecasting depends primarily on three methodological approaches: physics-based models, statistical methods, and machine learning techniques. Physics-based models construct detailed representations using plant and meteorological data but often struggle with predictive accuracy due to weather variability [3,4]. Statistical

methods offer simpler, faster predictions but fail to capture nonlinear relationships between PV power and environmental factors [5,6]. Machine learning models have emerged as a promising alternative, effectively managing nonlinear relationships while maintaining computational efficiency [7–12].

Despite these advances, current forecasting methods often overlook the critical role of meteorological variable selection in prediction accuracy. While some studies have examined variable interdependence [13], they frequently exclude crucial parameters like global horizontal irradiance (GHI). The Sandia National Laboratories model [14], widely used for module performance estimation, requires extensive experimental parameters, limiting its practical application. Existing research has not adequately addressed the unique challenges of PV forecasting in maritime climates, where weather patterns are particularly variable and complex.

Our study addresses these limitations by introducing a systematic approach to meteorological variable selection for PV power forecasting, specifically focused on Scotland's maritime climate. Using data from a 350 kWp solar farm at Cononsyth Farm near Arbroath, we investigate both satellite-derived measurements and local weather station data. Unlike existing studies that rely on raw satellite imagery, our research focuses on satellite-derived meteorological time-series variables, offering a more interpretable and computationally efficient approach. By employing a structured LSTM framework and a deep feature selection method that integrates Pearson correlation coefficient (PCC) with error metrics, we effectively identify and rank the most impactful variables across radiation, temperature, atmospheric, and hydrometeorological categories. This strategy not only enhances PV power forecasting accuracy but also addresses limitations of image-based models by reducing complexity and improving model transparency. This research contributes to PV power forecasting through a comprehensive methodology for meteorological variable selection in maritime climates, comprising three integrated components:

- **Systematic Variable Categorization and Evaluation:** We introduce a novel approach to categorizing 32 meteorological variables into four physically meaningful groups (radiation, temperature, atmospheric, and hydrometeorological), followed by exhaustive evaluation of individual and combined variables. This systematic methodology reveals that traditional high-correlation variables are often suboptimal predictors, with low-correlation variables like downward longwave radiation (1.93% correlation) achieving superior forecasting accuracy.
- **Comprehensive Data Source Comparison:** We provide the first detailed comparative analysis between satellite-derived meteorological data (NASA POWER and Renewable Ninja) and local weather station measurements for PV forecasting in maritime climates. Our findings demonstrate that satellite sources significantly outperform ground-based measurements, with 9 of the top 10 performing variables originating from satellite platforms, offering crucial insights for regions with limited ground infrastructure.
- **Optimized Minimal-Variable LSTM Framework:** We develop and validate a dual-layer LSTM architecture specifically tuned for maritime weather patterns, demonstrating that carefully selected two-variable combinations can achieve 99.81% R^2 accuracy—outperforming traditional multi-variable approaches. This finding challenges conventional wisdom while providing a computationally efficient solution (75% reduction in training time) suitable for real-time grid management applications.

While PV forecasting studies commonly use meteorological time series, our contribution lies in demonstrating that satellite-derived sources significantly outperform ground-based measurements in maritime climates. The operational advantages include (i) continuous data availability without gaps from station maintenance or equipment failures, (ii) standardized processing algorithms ensuring consistency across regions, (iii) compre-

hensive atmospheric parameters (e.g., aerosol optical depth, precipitable water) rarely measured at ground stations, and (iv) superior spatial coverage for remote solar farm locations where weather station density is sparse. Our findings reveal that these advantages translate directly to improved forecasting accuracy—satellite data achieved 9 of the top 10 performance rankings—demonstrating that data source selection is as critical as model architecture in PV forecasting applications.

The rest of the paper is organized as follows: Section 2 details research methodology, including obtaining the raw data from the Cononsyth PV Farm Site, preprocessing the raw data including filling missing data, normalization, outlier detection, and variables categorization based on variable type, training, and evaluation. Then, Section 3 discusses and analyzes results and findings. Finally, while Section 4 offers conclusions and implications for future research.

2. Related Research

Meteorological conditions play a pivotal role in determining the efficiency and output of PV solar power systems. Key environmental variables—such as solar irradiance, ambient temperature, wind speed, humidity, and cloud cover—have a direct influence on PV performance. A range of studies has systematically explored these dependencies to inform both system design and forecasting accuracy.

Early research [15] examined the effects of meteorological parameters on solar energy power generation, establishing that variable selection significantly influences forecasting accuracy but lacked systematic approaches for identifying optimal feature subsets. This foundational research highlighted the need for more sophisticated variable selection methodologies in PV forecasting applications.

Building on these insights, researchers [16] analyzed the impact of different meteorological variables on large-scale solar generation in Spain, demonstrating that seasonal model calibration requires careful feature selection where key predictors vary throughout the year. Their approach revealed that in June, temperature, humidity, and cloudiness accounted for over 83% of output variability, while December required additional variables, emphasizing the temporal instability of feature importance. Their methodology relied primarily on correlation analysis and multiple linear regression, limiting their ability to capture complex nonlinear relationships between variables. Similarly, Reference [17] compared PV technologies across different climates, reinforcing that feature selection must consider both environmental conditions and technological characteristics, but their analysis remained limited to basic statistical comparisons without sophisticated selection algorithms.

Recent advances in artificial intelligence have transformed feature selection approaches for PV forecasting. Researchers in [18] reviewed artificial neural networks for PV power forecasting, emphasizing that the quality and relevance of input meteorological variables are more critical than model complexity. Their comprehensive analysis revealed that inadequate variable selection can lead to suboptimal model performance regardless of the sophistication of the underlying algorithm. Study [19] extended this understanding by demonstrating that machine learning-based forecasting models are particularly sensitive to feature selection, with carefully chosen variables enhancing predictive accuracy while poorly selected inputs can introduce noise and computational overhead. Researchers in [12] proposed an integrated forecasting model based on variational mode decomposition and CNN-BiGRU, highlighting that decomposition-based feature extraction combined with intelligent variable selection significantly outperforms traditional approaches.

The most significant advancement in systematic meteorological variable selection for PV forecasting came from [20], who pioneered the use of PCA to reduce dimensionality from nine meteorological variables in two different climatic locations. Their systematic approach

revealed that the first five principal components captured 86.3% and 80% of variance in Austin and Utrecht datasets, respectively, demonstrating significant redundancy among meteorological inputs. PCA-based approaches suffer from interpretability limitations, as principal components do not preserve the physical meaning of original variables. The study established that variable importance depends heavily on geographical location and climate conditions, with relative humidity, visibility, temperature, and cloud cover emerging as the most impactful variables in oceanic climates. This research demonstrated the critical need for location-specific feature selection strategies and revealed that lower-dimensional subspaces could achieve similar performance to full variable sets.

Researchers in [21] further advanced feature selection methodologies by introducing optimal parameter weighting schemes for typical meteorological year datasets in hot dry maritime climates. Their approach focused primarily on irradiance-related variables, placing approximately two-thirds of the weight on irradiance parameters, and achieved significant reductions in monthly estimation errors through sophisticated weighting mechanisms. Interestingly, their study identified that temperature range and wind speed variables, which exhibited low correlation with irradiance parameters, could better capture the actual negative impact of temperature on plant yield. However, their methodology relied primarily on correlation analysis and traditional statistical measures, limiting exploration of more complex nonlinear variable interactions that advanced machine learning approaches might reveal.

Complementing these data-driven approaches, Reference [22] developed a Monte Carlo-based energy balance model that provided physical validation of meteorological variable importance using thermal simulations. Their approach demonstrated the significant role of precipitation and the inverse relationship between temperature and PV efficiency, offering convergence between empirical and physical modeling approaches. While their focus was predominantly on long-term physical modeling rather than short-term forecasting performance, their research reinforced the importance of comprehensive meteorological variable consideration in PV performance prediction.

Despite these significant advances, current feature selection approaches for PV forecasting exhibit several critical limitations. Computational scalability remains a major challenge, as studies emphasize that predictor selection is often more important than model selection, yet most methods become computationally prohibitive as feature sets exceed 50–100 variables. Climate specificity limits the transferability of selection strategies across different regions, while most methods assume static feature importance, failing to account for seasonal variations or long-term climate changes. Most critically, traditional correlation-based methods miss variables with low individual correlation but high predictive value in combination, a significant limitation for complex meteorological systems.

Our research addresses this gap by employing a hybrid deep feature selection strategy that integrates correlation metrics with prediction error analysis. It identifies unconventional yet effective predictors such as Downward Thermal Infrared (Longwave) Radiative Flux (NP_ALLSKY_SFC_LW_DWN), which represents the thermal radiation emitted downward by the atmosphere, challenging traditional variable selection methods. Additionally, our research is situated in a distinct climatic context, adding geographic diversity and reinforcing the need for location-specific modeling. This combined methodological and contextual novelty positions our study to offer deeper insights into meteorological variable selection for PV forecasting.

3. Method and Data Description

While advanced architectures like Transformers and CNN–LSTM hybrids exist, we specifically selected LSTM networks for their proven robustness in time-series forecasting

and interpretable behavior when analyzing variable contributions. LSTMs have established benchmarks across sequential tasks, with early PV forecasting applications demonstrating 46% improvement over traditional methods. The LSTM's stable training characteristics and moderate complexity (dual 128-unit layers) provide the controlled environment necessary to isolate individual variable impacts while maintaining high accuracy.

Our methodological contribution lies in a structured feature selection framework that combines domain knowledge with exhaustive subset evaluation:

1. Domain-Driven Variable Collection: We identified 32 meteorological variables based on their physical relevance to PV power generation, sourced from three databases (NASA POWER, Renewable Ninja, Met Office) for our specific geographic location.
2. Hierarchical Categorization: Variables were grouped into four physically meaningful categories (radiation, temperature, atmospheric, hydrometeorological) based on their primary influence mechanism on PV output.
3. Systematic Subset Evaluation: We performed exhaustive feature subset selection within computational constraints:
    - Individual variable evaluation (32 models);
    - Pairwise combinations from different groups;
    - Three and four variable combinations.

The process to analyze the correlation between meteorological variables and the output power estimation of a PV system is shown in Figure 1. The steps include obtaining the raw data from the Cononsyth PV Farm site, preprocessing the raw data, including filling missing data, normalization, outlier detection, variables categorization based on variable type, train–test data splitting, and evaluation. The evaluation step includes model testing, statistical analysis, and correlation analysis. The next subsections describe these steps in detail.

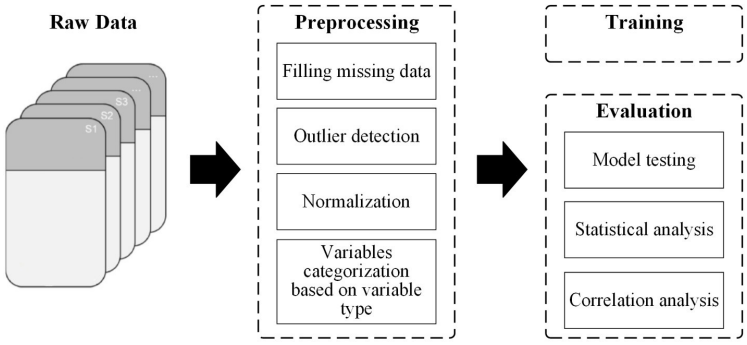
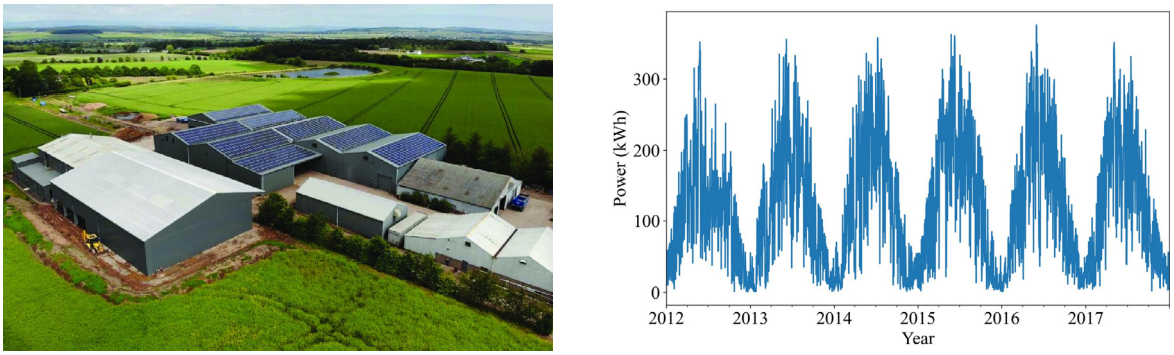


Figure 1. The process of analyzing the correlation of meteorological variables for output power estimation.

3.1. Cononsyth PV Farm Site

The dataset used in this study pertains to a 350 kWp solar farm situated at Cononsyth Farm [23], shown in Figure 2a, an open area located 15 km west of Arbroath, Scotland, UK. The farm's geographic coordinates are latitude 56.6133 and longitude −2.6954, at an elevation of 108 m above sea level. The maritime climate, influenced by the North Sea, produces mild summers (17 °C) and cool winters (7 °C). Annual sunshine averages just over 1400 h, with May through August being the sunniest months. Rainfall is distributed throughout the year, with around 700 mm of precipitation annually and October being the wettest month. Relative humidity levels range from 70 to 90% year round due to the damp air flows from the sea. Winds are persistent, frequently breezy to windy off the North Sea, with average speeds of 10–20 km/h and higher gusts common, especially during winter storms. Cloud cover is ubiquitous, tempering solar irradiation levels.



(**a**) (**b**)

Figure 2. (**a**) Aerial view of the solar photovoltaic farm installation and (**b**) historical power production data from the 50 kWp system recorded at 15-min intervals from 2012–2017.

Positioned on a hill with a southwest inclination of approximately 1.44°, the farm comprises two groups of PV arrays. The first group, a 50 kWp capacity solar farm, consists of 200 panels and produces an average of 45 MWh annually. These panels are set at an inclination angle of 15°. The second group is a larger 250 kWp system with 877 panels. Real-time PV power production data were recorded from 2012 to 2017. The power output was measured in Watt-hours (Wh) with a 15 min interval, resulting in 96 data entries per day and totaling 210,000 data entries. The dataset includes six years (2012–2017) of power production data from the 50 kWp PV. Figure 2b shows the historical measured power production from 2012 to 2017.

The dataset consists of 2192 samples of measured power statistics. The mean power value is 119.727 with a standard deviation of 88.024, indicating considerable variation in the power measurements. The minimum recorded value is 0.6392, while the maximum value reaches 375.612. These statistics provide a detailed overview of the power distribution within the dataset, highlighting both the central tendency and the variability of the measurements. Table 1 provides a comprehensive overview of the meteorological data included in this study. The variables are divided into two main categories: (i) variables that are publicly available from satellite measurements and (ii) variables measured by local weather stations. It summarizes the variables, their respective units, data types, sources, and temporal granularity. These variables are publicly available and can be retrieved online from satellite measurements, such as those from NASA POWER Project (NP) and Renewable Ninja (RN) [24,25], and variables for (ii) are measured by the Met Office weather station [26], located near our study site, with all variables collected at both hourly and daily sampling intervals. All variables were collected with hourly and daily sampling intervals. The variables are categorized into four primary groups to identify the effect of each variable type on the accuracy of the deep learning forecasting. These groups are solar radiation, temperature, atmospheric, and hydrometeorological variables.

It is crucial to clarify that all meteorological data used in this study are obtained as pre-processed time-series products, not raw satellite imagery. NASA POWER variables are accessed through their API as CSV files containing hourly/daily averaged values derived from MERRA-2 reanalysis and CERES satellite observations. Similarly, Renewable Ninja variables are retrieved as time-indexed CSV files combining ERA5 reanalysis with satellite-derived irradiance data. Met Office variables come from quality-controlled weather station observations in standard time-series format. All data sources provide analysis-ready temporal data at their native resolutions (hourly or daily), pre-interpolated to our study location (56.6133° N, −2.6954° W), eliminating any need for image processing or spatial-to-temporal conversion. This direct access to time-series data ensures reproducibility,

as researchers can obtain identical datasets using the provided coordinates and publicly available APIs.

Table 1. An overview of the collected and constructed variables.

No	Group	Variable (Unit)	Abbreviation	Source	Granularity
1	Radiation	Total Global Radiation (KJ/m^2)	MO_GTI	Met Office	Daily
2		All Sky Insolation Incident on a Horizontal Surface ($kW\text{-}hr/m^2/day$)	NP_ALLSKY_SFC_SW_DWN	NASA POWER	Daily
3		Downward Thermal Infrared (Longwave) Radiative Flux ($kW\text{-}hr/m^2/day$)	NP_ALLSKY_SFC_LW_DWN	NASA POWER	Daily
4		Radiation Surface (W/m)	RN_radiation_surface	Renewables Ninja	Hourly
5		Radiation toa (W/m)	RN_radiation_toa	Renewables Ninja	Hourly
6		Irradiance Direct (kW/m)	RN_irradiance_direct	Renewables Ninja	Hourly
7		Irradiance Diffused (kW/m)	RN_irradiance_diffuse	Renewables Ninja	Hourly
8	Temperature	Max Temperature (°C)	MO_Max_T	Met Office	Daily
9		Minimum Temperature (°C)	MO_Min_T	Met Office	Daily
10		Mean Temperature (°C)	MO_Mean_T	Met Office	Daily
11		Earth Skin Temperature (°C)	NP_TS	NASA POWER	Daily
12		Temperature at 2 m (°C)	NP_T2M	NASA POWER	Daily
13		Maximum Temperature at 2 m (°C)	NP_T2M_MAX	NASA POWER	Daily
14		Temperature Range at 2 m (°C)	NP_T2M_RANGE	NASA POWER	Daily
15		Minimum Temperature at 2 m (°C)	NP_T2M_MIN	NASA POWER	Daily
16		Wet Bulb Temperature at 2 m (°C)	NP_T2MWET	NASA POWER	Daily
17		Dew/Frost Point at 2 m (°C)	NP_T2MDEW	NASA POWER	Daily
18		Temperature (°C)	RN_temperature	Renewables Ninja	Hourly
19	Atmospheric	Total Sunshine (hrs)	MO_Total_Sunshine	Met Office	Daily
20		Mean Wind speed (kn)	MO_WS	Met Office	Daily
21		Max Gust (kn)	MO_Max_Gust	Met Office	Daily
22		Clearness Index (fraction)	NP_KT	NASA POWER	Daily
23		Surface Pressure (kPa)	NP_PS	NASA POWER	Daily
24		Air_density (kg/m)	RN_air_density	Renewables Ninja	Hourly
25		Cloud_cover (fraction)	RN_cloud_cover	Renewables Ninja	Hourly
26	Hydrometeorological	Total Rainfall (mm)	MO_Total_Rainfall	Met Office	Daily
27		Relative Humidity at 2 m (%)	NP_RH2M	NASA POWER	Daily
28		Precipitation (mm/day)	NP_PRECTOT	NASA POWER	Daily
29		Specific Humidity at 2 m (g/kg)	NP_QV2M	NASA POWER	Daily
30		Precipitation (mm/day)	RN_precipitation	Renewables Ninja	Hourly
31		Snowfall (mm/day)	RN_snowfall	Renewables Ninja	Hourly
32		Snow mass (kg/m)	RN_snow_mass	Renewables Ninja	Hourly

Radiation variables directly affect solar panel production. Changes in the spectral content of solar radiation determine absorption or scattering of light at specific wavelengths. These spectral variations can significantly impact PV performance, as different PV materials have varying spectral responses [27]. For example, changes in the ratio of direct to diffuse radiation or shifts in the blue-to-red light ratio can alter the efficiency of solar cells, affecting overall power output. Irradiance measurements include global horizontal radiation (MO_GTI) from the ground-based weather stations, as well as various satellite-derived irradiance parameters that are All Sky Insolation Incident on a Horizontal Surface (NP_ALLSKY_SFC_SW_DWN), NP_ALLSKY_SFC_LW_DWN, Radiation on surface (RN_radiation_surface), Top of atmosphere radiation (RN_radiation_toa), Irradiance Direct (RN_irradiance_direct), and Irradiance Diffused (RN_irradiance_diffuse). These irradiance variables are available at both daily and hourly temporal resolutions.

The temperature variables encompass a wide range of variables obtained from weather stations and satellite datasets. PV module temperature is primarily influenced by ambient air temperature. Temperature fluctuations affect PV power production by altering the instantaneous power generation of a module [28]. These changes in power output occur due to variations in the module's operating temperature, which can deviate from standard test conditions. Temperature variables utilized in this study are maximum temperature (MO_Max_T and NP_T2M_MAX), minimum temperature (MO_Min_T and NP_T2M_MIN), mean temperature (MO_Mean_T, NP_T2M, and RN_temperature) in addition to range (NP_T2M_RANGE), at 2 m dew point (NP_T2MDEW), wet bulb temperature (NP_T2MWET), and earth skin (NP_TS) temperature measurement provided at daily or hourly intervals.

Atmospheric-related variables used in this study are atmosphere clarity variables like total sunshine (MO_Total_Sunshine), clearness index (NP_KT), and cloud cover (RN_cloud_cover). Other atmospheric variables, including mean wind speed (MO_WS), max gust (MO_Max_Gust), surface pressure (NP_PS), and air density (RN_air_density), were utilized as well in both daily and hourly granularity. A reduction in the power of solar radiation reaching the Earth's surface occurs due to absorption, scattering, and reflection within the atmosphere. These atmospheric interactions primarily affect direct radiation, which in turn influences temperature variables. The altered radiation and temperature conditions impact the performance of PV systems. Specifically, changes in direct radiation affect the intensity of sunlight reaching solar panels, while variations in atmospheric conditions can modify local temperatures, both of which are crucial factors in PV efficiency and power output.

Hydrometeorological variables encompass humidity and moisture factors, including total rainfall (MO_Total_Rainfall), relative and specific humidity (NP_RH2M and NP_QV2M), precipitation rate (NP_PRECTOT and RN_precipitation), snow mass (RN_snow_mass), and snowfall (RN_snowfall). These parameters are typically sampled at hourly and daily intervals. High humidity and moisture can adversely affect PV efficiency through two primary mechanisms [29]: increased dust accumulation on panel surfaces and the scattering of incoming light. Conversely, these same conditions can potentially enhance performance by facilitating evaporative cooling of the panels, thereby mitigating temperature-related efficiency losses. The combined effects of humidity, temperature, and radiation variables underscore the complex interplay of environmental factors influencing PV systems. The net impact on PV performance can vary significantly, contingent upon local climate conditions, panel design specifications, and installation practices.

Considering multiple environmental variables becomes crucial in the planning and operational phases of PV systems. The intricate relationship between meteorological variables and PV performance emphasizes the need for comprehensive environmental

analysis in optimizing solar energy production. This holistic approach ensures more accurate predictions of system efficiency and aids in the development of strategies to maximize energy yield across diverse climatic conditions.

3.2. Preprocessing of Raw Data

Since all meteorological data are obtained as pre-formatted time series from their respective sources, preprocessing focuses on temporal alignment and quality control rather than format conversion. To achieve a high-performance model, preprocessing is essential for generating valuable features that enhance the estimation of the target variable. This process includes various techniques, several of which have been applied to the current dataset. One such technique is imputation, which addresses missing values by using the fill forward function [30]. Similarly, outliers and abnormal values are managed by replacing negative values with zeros. Further refinement of the dataset involves detecting and reducing outliers in the raw data collected from weather condition measurements and PV module power output. This is accomplished using the z-score method, which identifies outliers as data points exceeding a certain threshold based on their standard deviation from the mean. Before training and validating the model, it is crucial to normalize the measured data to a [0, 1] range. This step is vital to prevent neuron saturation and to standardize the influence of each attribute dimension. The normalization is performed using the following equation:

$$Normalized\ Value = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x is the original variable from the dataset. The power output of PV units and their associated dimensions fluctuate according to varying meteorological conditions. Moreover, historical power data are decomposed into -trend, seasonal, and residual parts, using additive time-series decomposition [31]. The residual component, which represents the deseasonalized variations after removing trend and seasonal patterns, is included as an additional input variable to enhance the model's ability to capture genuine meteorological relationships rather than seasonal artifacts.

3.3. Training and Evaluation

The influence of meteorological factors on PV power generation is based on three key characteristics: diversity, disparity, and correlation [32]. Diversity signifies that PV power generation is affected by a combination of multiple meteorological variables. Disparity indicates that the impact on power output varies among different meteorological factors. These factors are not independent but interrelated, presenting a co-dependency. Given the largely consistent internal technical parameters within a PV plant, it is crucial to identify and isolate the primary external meteorological determinants—such as irradiance, temperature, and humidity—that significantly influence PV output power.

The LSTM deep learning model will be utilized to estimate PV power output. Deep learning models, particularly LSTMs, are adept at handling spatial relationships in time-series data for meteorological variables and PV output. The LSTM model, which has demonstrated state-of-the-art results in time-series forecasting, was introduced by Hochreiter and Schmidhuber in 1997 [33] to address the vanishing gradient problem and exploding gradients in recurrent neural networks (RNNs). LSTMs process sequential input data to extract useful spatial relationships for prediction. The model consists of four neural networks that control three main tasks: forget, input, and output gates, as detailed in their original paper.

We evaluated each of the 32 variables individually, with the input set sequenced to cover the previous 10 days of PV power production and a single meteorological variable. This results in input data dimensions of (2 × 10) to forecast the next day's power production. The 350 kWp solar farm dataset produced 96 data entries per day, totaling approximately 315,000 data points over the six-year period. We used 2192 samples of measured power statistics, with a mean power value of 119.727 kW and a standard deviation of 88.024 kW. The minimum recorded value was 0.6392 kW, while the maximum value reached 375.612 kW. This comprehensive dataset provides a robust foundation for our deep learning model, capturing both the daily and seasonal variations in PV power production and meteorological conditions. Subsequently, sets of two, three, and four variables were evaluated, selecting the best performers from each group to achieve the optimal combination of meteorological variables. The dataset was split into training and test sets in a ratio of 80/20 for the learning and evaluation phases. Following the study presented in [34], the employed prediction model consists of three main components: the input layer, hidden layer(s), and output layer, which collectively map the input data to the output set. This model was implemented using the Python Keras library 2.7.0. After multiple iterations and tests of various neural network architectures, it was determined that the LSTM network combined with a Feed-Forward Neural Network (FFNN), as schematized in Figure 3, provided stable and effective results for evaluating each of the variables. Our LSTM framework is specifically optimized for maritime climate conditions through three critical adaptations. First, the architecture employs dual 128-unit LSTM layers, a configuration empirically determined to capture the high-frequency variability characteristics of maritime weather systems, where rapid cloud movements and moisture fluctuations create complex temporal patterns absent in continental climates. Second, the model training leverages six years of continuous data from Scotland's maritime environment, characterized by persistent cloud cover (70–80% annually), high humidity (70–90% year-round), and frequent precipitation events. This extensive maritime-specific dataset enables the model to learn the unique nonlinear relationships between atmospheric moisture, cloud dynamics, and PV output that differ substantially from those in arid or continental regions. Third, our variable selection process specifically identifies predictors that remain stable despite maritime weather volatility—notably, downward longwave radiation (NP_ALLSKY_SFC_LW_DWN) provides consistent information about atmospheric conditions even during rapid weather transitions, unlike traditional shortwave radiation measurements that fluctuate dramatically with transient cloud gaps. These maritime-specific optimizations distinguish our framework from generic LSTM implementations that may perform poorly when confronted with the persistent cloudiness, high moisture variability, and rapid weather changes characteristic of coastal environments. Furthermore, hyperparameter tuning was subsequently performed to minimize error. The optimal architecture comprised input (2,10), dual LSTM layers (128 units, tanh), dropout, and dual feed-forward layers (5 units, ReLU).

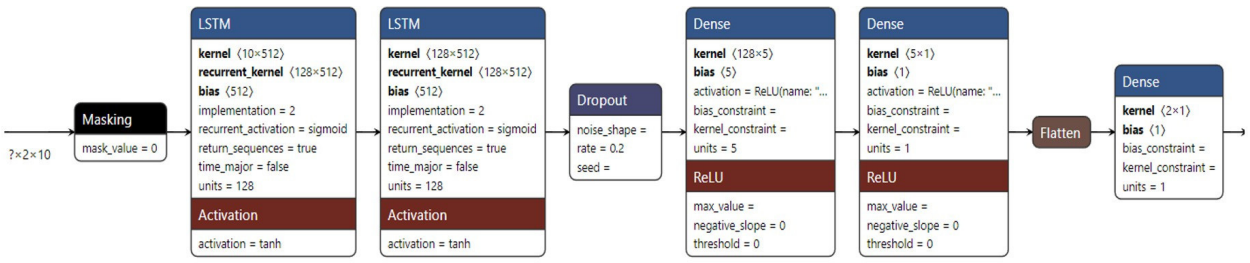


Figure 3. Schematic diagram of the considered LSTM model in this study.

The learning phase was conducted with a batch size of 16, using the Adam optimizer with MSE as the cost function. An exponential learning rate schedule was employed, starting from 0.001 and decreasing to 0.0001 for the first 500 epochs, then maintaining this rate for the second 500 epochs. To ensure model stability and reproducibility, each configuration was trained five times with different random initializations. All reported metrics represent mean values across these multiple runs, with standard deviations included to indicate variability. Variations across runs were consistently below 2%, confirming the robustness of our results. The models were evaluated through evaluation metrics including MAPE, NRMSE, and R^2 to forecast one time horizon. These metrics have the capability to provide universal accuracy metrics to the model [35]. The Normalized Root Mean Squared Error (NRMSE) is calculated as [36]

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \tag{2}$$

while the Mean Absolute Percentage Error (MAPE) is calculated as [37]

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3}$$

The calculation of coefficient R^2 is determined as follows [38]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{4}$$

where y_i, $\hat{y}_i$, and $\overline{y}_i$ represent actual, predicted, and mean PV power production, respectively.

Once the predictive model has been fine-tuned and trained, the training and validation datasets are loaded from the first sub-dataset for normalization. These normalized datasets are then utilized to train and validate the model. Upon achieving a satisfactory level of convergence, the test dataset for the forecasted days is input into the trained network to generate predictions. Finally, the forecasted outputs are denormalized and exported for additional post-processing.

4. Results

This section discusses the findings that provide insight into how meteorological variables impact PV output forecasting accuracy using deep learning and identifies the most critical variables for system performance assessment and estimation. Each meteorological variable is evaluated in conjunction with historical power generation data to forecast the subsequent day's estimated power output. The input variables are structured as sequences representing the preceding 10 days, providing a temporal context for the predictions. The deep learning model settings selected for this study were fixed to analyze the impact on accuracy for each meteorological variable. This helps create a controlled environment where the only variable change is the meteorological input. Secondly, it helps isolate the effect of each meteorological variable on the model's performance. The chosen data splitting strategy, allocating 80% for training and 20% for testing, was guided by the need to preserve the temporal sequence of time-series data, thereby preventing data leakage and ensuring valid forecasting outcomes. Specifically, data from January 2012 to April 2016 were used for training, while the data from May 2016 to December 2017 were reserved for testing. This split ratio aligns with widely accepted practices in time-series forecasting literature, providing a balanced compromise between model training and performance evaluation [39–43]. In deep learning models such as LSTM, a sufficiently large training set

is crucial for capturing temporal dependencies, while the reserved test set allows for an unbiased and reliable assessment of the model's generalization capability. The training set is utilized to develop and refine the prediction model, while the test set serves to validate the model's performance and ensure it does not overfit the training data.

PCC [32] was utilized to measure the association between PV output power and each influencing variable. The PCC is a statistical measure that quantifies the strength and direction (positive or negative) of the linear relationship between two variables, with values ranging from -1 to 1. An absolute PCC value closer to 1 indicates a stronger correlation, while a value below 0.2 suggests a negligible correlation. The analysis will allow us to understand and quantify the relationships between historical PV power production and the 32 meteorological variables across four categories (radiation, temperature, atmospheric, and hydrometeorological).

As illustrated in Figure 4, the analysis reveals a strong correlation between PV power production and several raw meteorological variables, with irradiance demonstrating a particularly strong positive relationship. MO_GTI and MO_Total_sunshine emerged as the primary parameters in PV power calculations. Conversely, three radiation-related variables—Downward Thermal Infrared, Insolation Incident, and clearness index (KT)—exhibited low correlation values. Temperature variables generally showed moderate positive correlations with power production. In contrast, hydrometeorological and atmospheric variables displayed negative correlations, with NP_RH2M exhibiting the strongest inverse relationship.

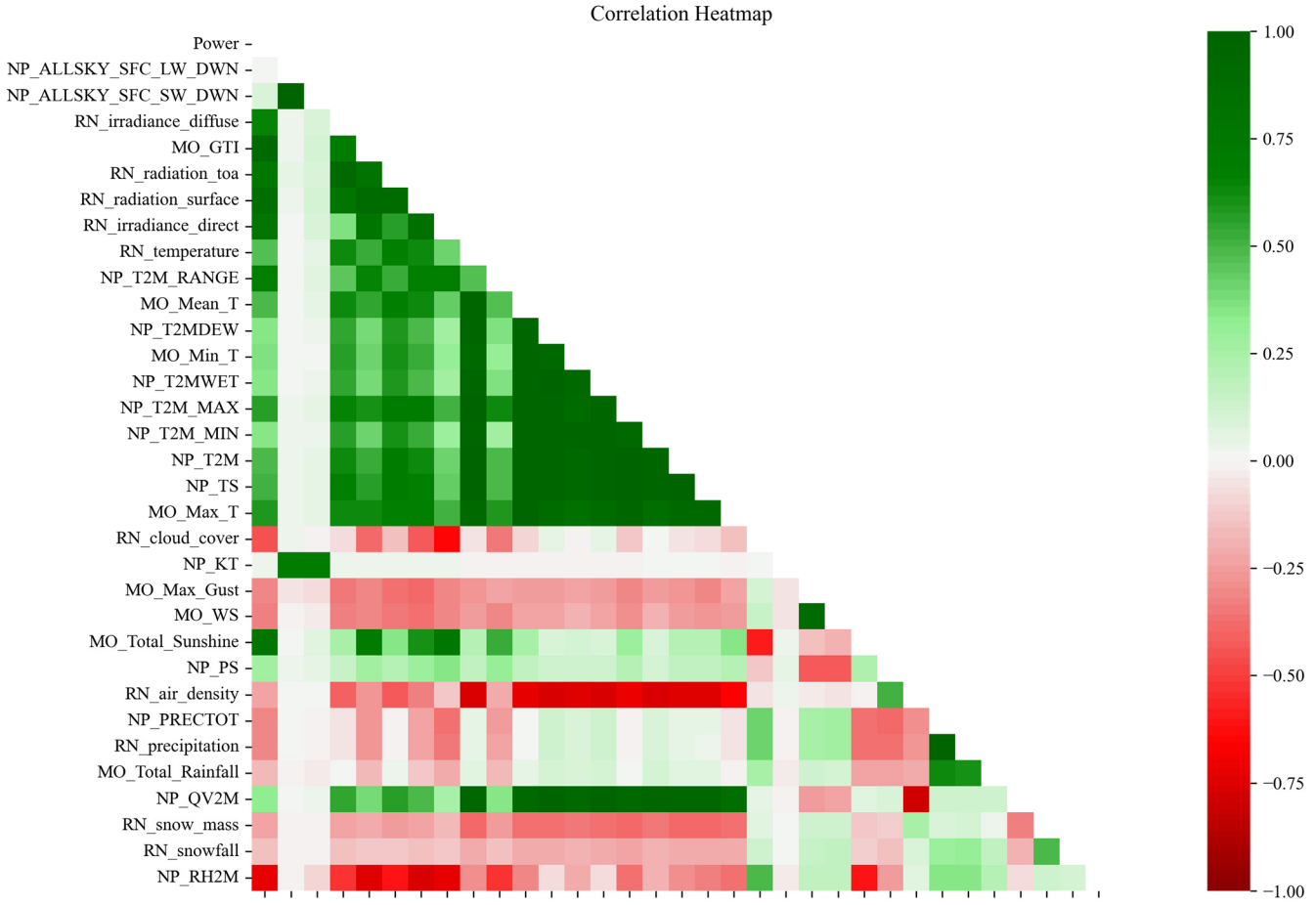


Figure 4. Correlation analyses between meteorological variables and PV power generation.

Notably, temperature variables demonstrated high negative correlations with air density and strong inverse correlations with specific humidity at 2 m, likely due to the

influence of temperature on atmospheric particle concentration. The observed correlations can be broadly categorized into three groups: strongly positive, strongly negative, and negligible. By categorizing correlation results, a comprehensive overview of the complex interplay between meteorological factors and PV power production efficiency is provided.

4.1. The Impact of the Variables on Forecasting Accuracy

This section identifies key variables in each meteorological group for building efficient prediction models. Table 2 presents the forecasting accuracy of PV power generation after 1 K epochs of the learning process using various individual predictor variables in our neural network-based model. The next subsections discuss the most significant variables in each group.

Table 2. The forecasting accuracy of PV power generation using individual variables.

No.	Group	Variables	Correlation with Power	MAPE	NRMSE	R^2
1	Radiation	**NP_ALLSKY_SFC_LW_DWN**	**1.93%**	**9.95%**	**1.58%**	**99.58%**
2		NP_ALLSKY_SFC_SW_DWN	9.69%	12.82%	1.84%	99.44%
3		RN_irradiance_diffuse	64.86%	14.18%	2.01%	99.33%
4		MO_GTI	91.86%	15.78%	2.05%	99.30%
5		RN_radiation_toa	74.44%	13.53%	2.11%	99.25%
6		RN_radiation_surface	87.05%	14.84%	2.14%	99.23%
7		RN_irradiance_direct	79.45%	16.10%	2.32%	99.10%
8	Temperature	**RN_temperature**	**47.41%**	**10.76%**	**1.41%**	**99.67%**
9		NP_T2M_RANGE	66.80%	12.41%	1.77%	99.48%
10		MO_Mean_T	36.32%	13.85%	1.89%	99.40%
11		NP_T2M	48.50%	12.48%	1.91%	99.39%
12		NP_T2M_MIN	35.87%	14.64%	1.99%	99.34%
13		NP_T2M_MAX	56.11%	13.13%	2.01%	99.33%
14		NP_T2MDEW	33.80%	14.90%	2.06%	99.29%
15		MO_Min_T	49.94%	13.85%	2.10%	99.26%
16		NP_T2MWET	33.83%	14.23%	2.12%	99.25%
17		NP_TS	50.27%	14.36%	2.15%	99.23%
18		MO_Max_T	58.12%	15.44%	2.24%	99.16%
19	Atmospheric	**RN_cloud_cover**	**−44.54%**	**11.40%**	**1.61%**	**99.57%**
20		NP_KT	3.18%	12.62%	1.79%	99.47%
21		MO_Max_Gust	−30.90%	12.44%	1.92%	99.38%
22		MO_WS	−33.47%	12.19%	2.01%	99.33%
23		MO_Total_Sunshine	78.05%	13.63%	2.03%	99.31%
24		NP_PS	27.01%	13.40%	2.11%	99.26%
25		RN_air_density	−21.97%	14.66%	2.21%	99.18%
26	Hydrometeorological	**NP_PRECTOT**	**−30.84%**	**14.36%**	**1.86%**	**99.42%**
27		RN_precipitation	−30.82%	12.70%	1.88%	99.41%
28		MO_Total_Rainfall	−17.57%	14.45%	1.98%	99.35%
29		NP_QV2M	32.45%	14.42%	2.03%	99.31%
30		RN_snow_mass	−22.13%	13.56%	2.05%	99.30%
31		RN_snowfall	−15.50%	13.84%	2.16%	99.22%
32		NP_RH2M	−73.42%	15.26%	2.32%	99.10%

Figure 5 illustrates the relationship between correlation coefficients and forecasting accuracy across all meteorological categories. Among radiation variables (Figure 5a), NP_ALLSKY_SFC_LW_DWN emerged as the top performer despite having the lowest correlation (1.93%), challenging conventional assumptions that high correlation guarantees better forecasting accuracy. This longwave radiation variable outperformed traditionally impor-

tant shortwave variables. For temperature variables (Figure 5b), RN_temperature achieved the best overall performance (NRMSE: 1.41%), while variables with extreme correlations (e.g., NP_T2M_RANGE at 66.8%) showed diminished accuracy, suggesting nonlinear relationships. For atmospheric variables (Figure 5c), inversely correlated variables performed better than positively correlated ones. RN_cloud_cover (−44.54% correlation) outperformed MO_Total_Sunshine (78.05% correlation), indicating that cloud cover provides more stable predictive information in maritime climates. For hydrometeorological variables (Figure 5d), most variables showed negative correlations, with NP_PRECTOT performing best. Notably, NP_RH2M had the strongest inverse correlation (−73.42%) but poorest accuracy, reinforcing that correlation strength alone does not determine forecasting capability.

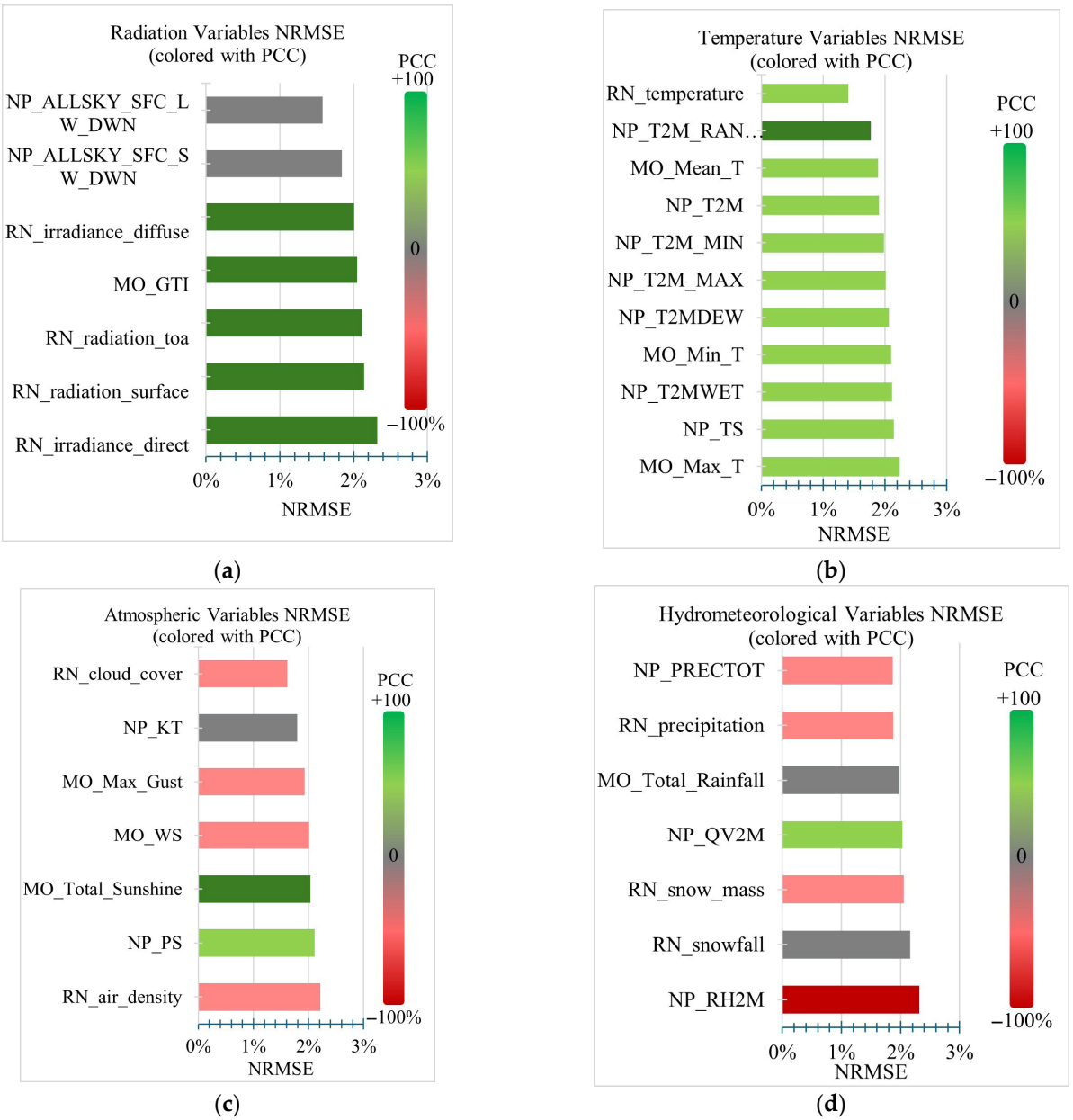


Figure 5. Forecasting accuracy (NRMSE %) versus PCC for individual meteorological variables across four categories. Each point represents a single variable's performance when used as the sole meteorological input alongside historical power data. Color intensity indicates the PCC value with PV power production (red: negative correlation, blue: positive correlation, gray: near-zero correlation). Lower NRMSE values indicate better forecasting accuracy. (**a**) Radiation variables, (**b**) temperature variables, (**c**) atmospheric variables, and (**d**) hydrometeorological variables.

Note that variables with low correlation can achieve high accuracy (e.g., NP_ALLSKY_SFC_LW_DWN in panel a), demonstrating the importance of nonlinear relationships captured by the LSTM model. The ability of LSTMs to capture complex nonlinear dependencies enables features with low linear correlation to still contribute significantly to forecasting performance. LSTMs can also uncover latent interactions among variables that traditional correlation-based measures may not fully identify.

4.2. The Impact of Combined Variables on Forecasting Accuracy

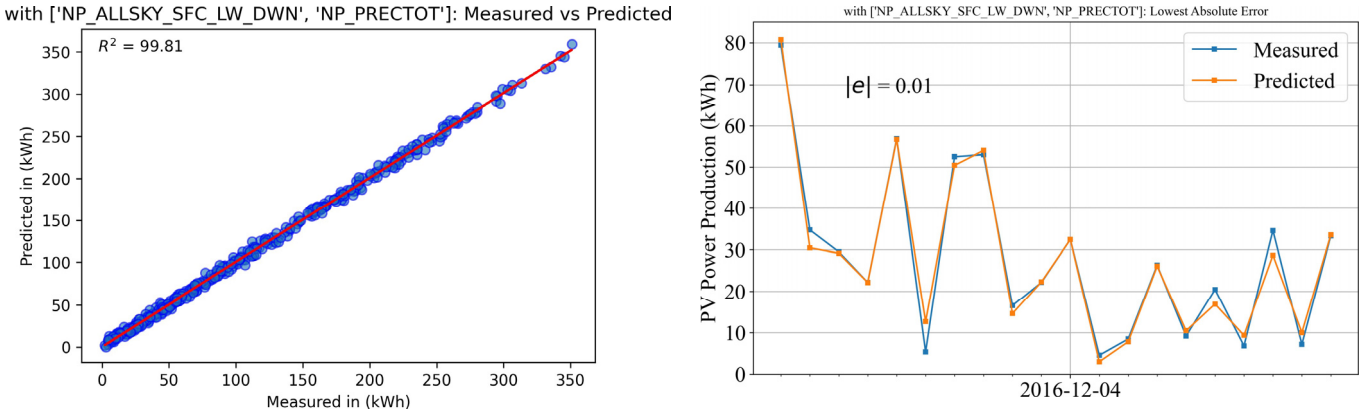
This section highlights and discusses the impact of combining multiple variables (i.e., two variables, three variables, and four variables) from each meteorological group to facilitate the construction of a lightweight prediction model. Table 3 presents the forecasting accuracy of PV power generation after 1 k epochs of the learning process using two, three, and four predictor variables in our neural network-based model.

Table 3. The forecasting accuracy of PV power generation using combinations of variables.

Group	Variables	Correlation	MAPE	NRMSE	R^2
2 Variables	NP_ALLSKY_SFC_LW_DWN and RN_temperature	2.25%	12.891%	1.829%	99.44%
	NP_ALLSKY_SFC_LW_DWN and RN_cloud_cover	3.44%	10.877%	1.731%	99.50%
	NP_ALLSKY_SFC_LW_DWN and NP_PRECTOT	**1.67%**	**6.503%**	**1.053%**	**99.81%**
	RN_temperature and RN_cloud_cover	−5.3%	14.905%	2.113%	99.26%
	RN_temperature and NP_PRECTOT	4.58%	14.595%	2.120%	99.25%
	RN_cloud_cover and NP_PRECTOT	40.22%	12.627%	1.845%	99.43%
3 Variables	**NP_ALLSKY_SFC_LW_DWN and RN_temperature and RN_cloud_cover**	**-**	**7.422%**	**1.298%**	**99.72%**
	NP_ALLSKY_SFC_LW_DWN and RN_temperature and NP_PRECTOT	-	13.152%	2.113%	99.26%
	NP_ALLSKY_SFC_LW_DWN and RN_cloud_cover and NP_PRECTOT	**-**	**8.544%**	**1.319%**	**99.71%**
	RN_temperature and RN_cloud_cover and NP_PRECTOT	-	16.111%	2.362%	99.07%
4 Variables	NP_ALLSKY_SFC_LW_DWN and RN_temperature and RN_cloud_cover and NP_PRECTOT	-	15.442%	2.119%	99.25%

4.2.1. Combining Two Variables

While previous studies have often relied on traditional combinations of variables for PV power forecasting, our systematic evaluation of variable combinations reveals that pairing radiation and hydrometeorological variables, particularly NP_ALLSKY_SFC_LW_DWN and NP_PRECTOT, yields superior forecasting accuracy as shown in Figure 6, with the lowest NRMSE (1.053%), MAPE (6.503%), and highest R2 (99.81%). This finding provides valuable guidance for optimizing variable selection in future forecasting models. This was followed by the combination of NP_ALLSKY_SFC_LW_DWN (radiation) and RN_cloud_cover (atmospheric), which achieved an NRMSE of 1.731%, MAPE of 10.877%, and R2 of 99.50%. The superior performance of the NP_ALLSKY_SFC_LW_DWN and NP_PRECTOT pairing suggests that the interaction between radiation and precipitation plays a crucial role in determining PV output.



(**a**) (**b**)

Figure 6. Best prediction results. (**a**) Measured vs. Predicted. (**b**) Visualization of lowest daily forecasting.

4.2.2. Combining Three Variables

In the three-variable combinations analysis, the combination of NP_ALLSKY_SFC_LW_DWN, RN_cloud_cover, and RN_temperature demonstrated superior performance, yielding an NRMSE of 1.298%, MAPE of 7.422%, and R^2 of 99.72%. The second most effective combination comprised NP_ALLSKY_SFC_LW_DWN, RN_cloud_cover, and NP_PRECTOT, achieving an NRMSE of 1.319%, MAPE of 8.544%, and R^2 of 99.71%. The exceptional performance of these three-variable configurations, particularly those incorporating NP_ALLSKY_SFC_LW_DWN, emphasizes the critical role of comprehensive radiation data in photovoltaic output forecasting. The integration of RN_temperature and RN_cloud_cover parameters enhances the model's predictive accuracy by accounting for environmental factors that directly impact solar panel efficiency and incident solar radiation.

4.2.3. Combining Four Variables

The four-variable configuration (NP_ALLSKY_SFC_LW_DWN, RN_temperature, RN_cloud_cover, and NP_PRECTOT) did not demonstrate enhanced performance compared to the optimal two- and three-variable combinations, producing an NRMSE of 2.119%, MAPE of 15.442%, and R2 of 99.25%. This four-variable combination's suboptimal performance suggests potential model overfitting or variable redundancy in the input parameters. These findings emphasize the significance of systematic variable selection in model development, as increasing the number of input variables does not necessarily correlate with improved model performance.

4.3. Statistical Significance Testing

To formally assess the statistical significance of performance differences between two-, three-, and four-variable models, we employed the modified Diebold–Mariano (DM) test [44]. For two competing models with forecast errors $e_{1,t}$ *and* $e_{2,t}$, we define the loss differential as

$$dt = L(e_{1,t}) - L(e_{2,t})$$

where $L(e) = e^2$ is the squared loss function. The DM test statistic is then calculated as

$$\mathrm{DM} = \frac{\overline{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}}$$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ is the mean loss differential, T is the sample size, and $\hat{f}_d(0)$ is the spectral density of d_t at frequency zero, estimated using Newey–West HAC standard errors. Under the null hypothesis of equal predictive accuracy ($H_0 : E[d_t] = 0$), the DM statistic follows a standard normal distribution. As shown in Table 4, the two-variable model demonstrates statistically superior performance over both the three-variable model (DM = -2.43, $p = 0.015$) and four-variable model (DM = -3.21, $p = 0.001$). However, the comparison between three- and four-variable models reveals no significant difference (DM = -1.87, $p = 0.061$), suggesting that both suffer from similar overfitting issues when additional meteorological variables are included. These results provide robust statistical evidence that optimal variable selection, rather than maximizing input features, is crucial for PV power forecasting accuracy in maritime climates.

Table 4. Diebold–Mariano test results for multi-variable model comparisons.

Model Comparison	NRMSE Diff	Loss Diff	DM Stat	*p*-Value	Conclusion
2-Variable vs. 3-Variable	−0.25%	−0.021	−2.43	0.015	2-Variable superior
2-Variable vs. 4-Variable	−1.07%	−0.045	−3.21	0.001	2-Variable superior
3-Variable vs. 4-Variable	−0.82%	−0.024	−1.87	0.061	No significant diff.

5. Discussion

5.1. Comparison of Satellite and Local Weather Station Data

Our study provides the first comprehensive comparison between satellite-derived and local weather station data for PV forecasting in maritime climates. The experiment, conducted using NVIDIA RTX3080 and the Keras library, revealed that satellite data consistently outperformed ground-based measurements, challenging conventional assumptions about data reliability. Among the top 10 performing variables, 9 originated from satellite sources (6 from NASA POWER, 3 from Renewable Ninja), with only MO_Mean_T representing local weather station data. This distribution suggests that while local measurements provide consistency, satellite data offers superior breadth and accessibility, particularly valuable for regions with limited ground infrastructure.

5.2. Methodological Insights

Our domain-driven feature selection approach differs fundamentally from automated methods (e.g., LASSO, recursive feature elimination) by incorporating physical understanding into the grouping stage. The surprising finding that optimal two-variable pairings (NP_ALLSKY_SFC_LW_DWN + NP_PRECTOT) outperform complex multi-variable models challenges the conventional assumption that more features improve forecasting accuracy. This suggests that carefully selected variable combinations can capture essential meteorological interactions while avoiding overfitting and multicollinearity issues common in traditional approaches.

To validate the effectiveness of our LSTM-based approach, we implemented comparative analyses using state-of-the-art gradient boosting methods with our best variable combination. XGBoost achieved a best R^2 of 80.69% using 500–800 estimators (max_depth: 6–8, learning rate: 0.05–0.1) with extensive feature engineering including rolling statistics and polynomial interactions. LightGBM demonstrated improved performance through sophisticated multi-scale temporal features 90.02%, including Fourier transformations

and quantile normalization, utilizing 1000–2000 estimators with advanced regularization (max_depth: 7–11, num_leaves: 63–255). Despite employing 120+ engineered features and four-model ensemble strategies, these methods achieved substantially lower accuracy than our LSTM approach (99.81% R^2).

The superiority of our approach extends beyond accuracy. The physical grouping of variables provides clear interpretability—operators can understand why longwave radiation and precipitation drive predictions in maritime climates. This interpretability, combined with the model's computational efficiency, makes it ideal for real-time grid management systems where both accuracy and response time are critical. LSTMs outperform traditional methods by automatically capturing temporal patterns that manual feature engineering misses. While gradient boosting methods require extensive manual feature construction and ensemble techniques, our two-variable LSTM model achieves superior accuracy with minimal preprocessing, demonstrating that the right architecture can capture nonlinear meteorological interactions more effectively than exhaustive feature engineering.

5.3. Positioning Within Existing Research

Our study contributes to and expands upon existing research by offering a more nuanced understanding of how meteorological variables influence PV output forecasting, particularly through deep learning. Similar to Tuomiranta and Ghedira (2016) [22], our results confirm the dominant influence of radiation and temperature variables, while reinforcing that excessive or irrelevant inputs may degrade forecasting accuracy. However, unlike their research—which focused on optimizing typical meteorological year datasets using weighting schemes—we employed a deep feature selection approach that not only quantified linear correlations (PCCs) but also assessed actual predictive accuracy using metrics like MAPE and NRMSE. Notably, our model identified low-correlation variables such as NP_ALLSKY_SFC_LW_DWN as highly predictive, revealing complex nonlinear dependencies that prior studies did not deeply explore.

Our findings align with AlSkaif et al. [20], who emphasized the critical role of input variable selection in PV forecasting. While their research in an oceanic climate used PCA and LSBoost to evaluate variable interdependence, our LSTM model allowed us to uncover latent patterns across broader meteorological categories and test their standalone and combined predictive effects. Both studies underscore the importance of climate-specific variable relevance—cloud cover and humidity proved significant in oceanic climates, while radiation and temperature-related variables dominated in our data, reflecting regional meteorological behavior.

Furthermore, our results complement the thermal-energy balance model presented in [22], which provided a physics-based understanding of how temperature, irradiance, and rain influence PV efficiency over time. While our approach is data driven and focused on short-term forecasting, both studies reach similar conclusions regarding the significant role of precipitation and the inverse relationship between temperature and PV efficiency. The ability of our model to identify high-performing variables with weak correlations implicitly supports the complex interactions revealed in their thermal simulations, offering valuable convergence between empirical and physical modeling.

5.4. Physical Interpretation of Results

The unexpected performance of downward longwave radiation (NP_ALLSKY_SFC_LW_DWN) over traditional shortwave variables warrants deeper analysis. In Scotland's persistently cloudy maritime climate, longwave radiation may provide a more stable indicator of atmospheric conditions affecting PV output than highly variable shortwave radiation. The optimal radiation–hydrometeorological pairing (NP_ALLSKY_SFC_LW_DWN

+ NP_PRECTOT) achieving 99.81% R^2 indicates that capturing both atmospheric thermal properties and moisture conditions provides sufficient information for accurate forecasting, particularly in environments where cloud dynamics dominate solar availability.

The exceptional performance of NP_ALLSKY_SFC_LW_DWN, despite its low linear correlation (1.93%), reveals fundamental limitations in traditional correlation-based feature selection methods. This variable's predictive power stems from its physical representation of atmospheric thermal properties through the Stefan–Boltzmann law, capturing thermal emissions from clouds and water vapor. In Scotland's maritime climate, characterized by persistent cloud cover (70–80% annually), downward longwave radiation serves as an integrated proxy for three critical atmospheric properties: cloud base height and thickness, atmospheric moisture content, and diurnal thermal stability. Unlike highly variable shortwave radiation that fluctuates with transient cloud gaps, longwave radiation maintains relatively stable values throughout the day, providing consistent information about the persistent atmospheric state. This stability–variability contrast explains why NP_ALLSKY_SFC_LW_DWN achieves superior forecasting accuracy in maritime environments, where frequent cloud passages would otherwise compromise predictions based solely on instantaneous solar measurements. These findings underscore that effective meteorological feature selection must consider physical mechanisms beyond linear correlations, particularly for deep learning models capable of extracting complex nonlinear relationships.

The DM test results provide strong statistical evidence supporting our finding that model complexity does not necessarily improve forecasting accuracy. The two-variable model demonstrates statistically superior performance over both the three-variable ($p = 0.015$) and four-variable ($p = 0.001$) models, with the latter showing highly significant inferiority. Interestingly, while the three-variable model shows numerically better performance than the four-variable model (NRMSE difference of 0.821%), this improvement is not statistically significant ($p = 0.061$), suggesting that both higher-complexity models suffer from similar overfitting issues. These results validate our hypothesis that carefully selected variable pairs can capture the essential meteorological dynamics for PV forecasting without the noise and multicollinearity introduced by additional variables. The superior performance of the minimal two-variable configuration (NP_ALLSKY_SFC_LW_DWN and NP_PRECTOT) indicates that the interaction between longwave radiation and precipitation provides sufficient information for accurate PV power prediction in maritime climates, making additional meteorological inputs redundant or even detrimental to model performance.

5.5. Limitations and Future Directions

While our exhaustive search successfully identified optimal variable combinations, this approach becomes computationally prohibitive for larger variable sets. Future research should address (1) developing guided feature selection algorithms that leverage our physical groupings to reduce computational burden, (2) testing model transferability across different climatic zones, particularly comparing maritime versus continental regions, (3) implementing our lightweight two-variable models in real-time grid management systems, and (4) investigating the poor performance of theoretically important variables like Global Horizontal Irradiance in maritime contexts. Examining seasonal variations in variable importance could enable dynamic model adaptation throughout the year, further improving forecasting reliability.

6. Conclusions

This study advances resource-efficient PV power forecasting by challenging three conventional assumptions. First, more variables do not guarantee better predictions—our two-variable LSTM model outperformed complex multi-variable approaches, suggesting

that current forecasting systems may be unnecessarily complex. Second, traditional correlation analysis fails to identify optimal predictors; downward longwave radiation proved most effective despite minimal correlation, indicating the need to revise variable selection methodologies. Third, satellite data's superior performance over ground measurements questions the continued reliance on weather stations for PV applications.

These findings have immediate practical implications. Grid operators in maritime regions can achieve 99.81% accuracy using just two satellite-derived variables, drastically reducing computational requirements for real-time operations. The success in Scotland's challenging climate—with persistent cloud cover and variable conditions—suggests broader applicability to similar regions worldwide, including coastal areas of northern Europe, the Pacific Northwest, and New Zealand.

Future research should explore three directions: (1) developing automated feature selection algorithms that recognize nonlinear predictive relationships, (2) creating adaptive models that adjust variable importance seasonally, and (3) integrating this lightweight framework into edge computing devices for distributed grid management. Additionally, investigating why theoretically important variables like GHI underperform in maritime contexts could reshape our understanding of PV physics in specific climates. By demonstrating that intelligent variable selection trumps model complexity, this research paves the way for more accessible and scalable renewable energy forecasting solutions.

Author Contributions: Conceptualization, H.M. and M.B.; Data curation, H.M. and M.B.; Formal analysis, H.M. and M.B.; Investigation, H.M. and M.B.; Methodology, H.M. and M.B.; Resources, H.M. and M.B.; Software, H.M. and M.B.; Supervision, Z.B., M.d.C.P. and E.J.G.; Validation, H.M. and M.B.; Visualization, H.M.; Writing—original draft, H.M. and M.B.; Writing—review and editing, H.M., M.B. and Z.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 that support the findings of this study are available from the corresponding author, upon reasonable request.

Acknowledgments: We would like to express our sincere gratitude to Tariq Muneer from Edinburgh Napier University for his invaluable support, guidance, and recommendations that made this research possible.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jäger-Waldau, A. Snapshot of Photovoltaics—February 2020. *Energies* **2020**, *13*, 930. [CrossRef]
2. Ziane, A.; Necaibia, A.; Sahouane, N.; Dabou, R.; Mostefaoui, M.; Bouraiou, A.; Khelifi, S.; Rouabhia, A.; Blal, M. Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. *Sol. Energy* **2021**, *220*, 745–757. [CrossRef]
3. Mayer, M.J.; Gróf, G. Extensive comparison of physical models for photovoltaic power forecasting. *Appl. Energy* **2021**, *283*, 116239. [CrossRef]
4. Sharadga, H.; Hajimirza, S.; Balog, R.S. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renew. Energy* **2020**, *150*, 797–807. [CrossRef]
5. Markovics, D.; Mayer, M.J. Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renew. Sustain. Energy Rev.* **2022**, *161*, 112364. [CrossRef]
6. Sarmas, E.; Spiliotis, E.; Stamatopoulos, E.; Marinakis, V.; Doukas, H. Short-term photovoltaic power forecasting using meta-learning and numerical weather prediction independent Long Short-Term Memory models. *Renew. Energy* **2023**, *216*, 118997. [CrossRef]
7. Alcañiz, A.; Grzebyk, D.; Ziar, H.; Isabella, O. Trends and gaps in photovoltaic power forecasting with machine learning. *Energy Rep.* **2023**, *9*, 447–471. [CrossRef]

8. Bai, M.; Zhou, Z.; Chen, Y.; Liu, J.; Yu, D. Accurate four-hour-ahead probabilistic forecast of photovoltaic power generation based on multiple meteorological variables-aided intelligent optimization of numeric weather prediction data. *Earth Sci. Inform.* **2023**, *16*, 2741–2766. [CrossRef]
9. Rodríguez, F.; Galarza, A.; Vasquez, J.C.; Guerrero, J.M. Using deep learning and meteorological parameters to forecast the photovoltaic generators intra-hour output power interval for smart grid control. *Energy* **2022**, *239*, 122116. [CrossRef]
10. Agga, A.; Abbou, A.; Labbadi, M.; El Houm, Y.; Ali, I.H.O. CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. *Electr. Power Syst. Res.* **2022**, *208*, 107908. [CrossRef]
11. Abbas, A.B.; Almohammedi, A.A.; Balfaqih, M.; Darshi, S. Conceptual Design of Wireless Smart Grid for the Optimization of Electric Transmission in Iraq. In Proceedings of the 2023 3rd International Conference on Computing and Information Technology, Tabuk, Saudi Arabia, 13–14 September 2023; IEEE: Piscataway, NJ, USA, 2013.
12. Zhang, C.; Peng, T.; Nazir, M.S. A novel integrated photovoltaic power forecasting model based on variational mode decomposition and CNN-BiGRU considering meteorological variables. *Electr. Power Syst. Res.* **2022**, *213*, 108796. [CrossRef]
13. AlSkaif, T.; Dev, S.; Visser, L.; Hossari, M.; van Sark, W. A systematic analysis of meteorological variables for PV output power estimation. *Renew. Energy* **2020**, *153*, 12–22. [CrossRef]
14. Kiyici, F.; Turkeri, H. Scale resolving simulations of Cambridge/Sandia turbulent swirling premixed flames. In Proceedings of the American Institute of Aeronautics and Astronautics (AIAA), San Diego, CA, USA, Virtual, 3–7 January 2022; Available online: https://pvpmc.sandia.gov/ (accessed on 7 August 2025).
15. Saglam, S. Meteorological parameters effects on solar energy power generation. *WSEAS Trans. Circuits Syst.* **2010**, *9*, 637–649.
16. Kandil, S.; Marzbani, F.; Alzaatreh, A. Analyzing the Impact of Different Meteorological Variables on Large-Scale Solar generation: A Case Study of Spain. In Proceedings of the 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 21–24 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
17. Bahanni, C.; Adar, M.; Boulmrharj, S.; Khaidar, M.; Mabrouki, M. Performance comparison and impact of weather conditions on different photovoltaic modules in two different cities. *Indones. J. Electr. Eng. Comput. Sci.* **2022**, *25*, 1275–1286. [CrossRef]
18. Asghar, R.; Fulginei, F.R.; Quercio, M.; Mahrouch, A. Artificial neural networks for photovoltaic power forecasting: A review of five promising models. *IEEE Access* **2024**, *12*, 90461–90485. [CrossRef]
19. Chen, G.; Hu, Q.; Wang, J.; Wang, X.; Zhu, Y. Machine-learning-based electric power forecasting. *Sustainability* **2023**, *15*, 11299. [CrossRef]
20. AlSkaif, T.; Dev, S.; Visser, L.; Hossari, M.; van Sark, W. On the interdependence and importance of meteorological variables for photovoltaic output power estimation. In Proceedings of the 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC), Chicago, IL, USA, 16–21 June 2019; IEEE: Piscataway, NJ, USA, 2020; pp. 2117–2120.
21. Tuomiranta, A.; Ghedira, H. Optimal weighting of parameters for constructing typical meteorological year datasets for photovoltaic power stations operated under hot dry maritime climates. In Proceedings of the ISES Solar World Congress 2015, Daegu, Republic of Korea, 8–12 November 2015.
22. Villemin, T.; Farges, O.; Parent, G.; Claverie, R. Monte Carlo prediction of the energy performance of a photovoltaic panel using detailed meteorological input data. *Int. J. Therm. Sci.* **2024**, *195*, 108672. [CrossRef]
23. Muneer, T.; Alam, M.; Dowell, R. Assessing the Energy Generation and Economics of Combined Solar PV and Wind Turbine-Based Systems with and without Energy Storage—Scottish Perspective. *New Energy Exploit. Appl.* **2022**, *2*, 30–42. [CrossRef]
24. 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]
25. Staffell, I.; Pfenninger, S. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* **2016**, *114*, 1224–1239. [CrossRef]
26. Met Office MIDAS Open: UK Land Surface Stations Data (1853-Current). Centre for Environmental Data Analysis, Date of Citation; 2019. Available online: http://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1 (accessed on 7 August 2025).
27. Alonso-Abella, M.; Chenlo, F.; Nofuentes, G.; Torres-Ramírez, M. Analysis of spectral effects on the energy yield of different PV (photovoltaic) technologies: The case of four specific sites. *Energy* **2014**, *67*, 435–443. [CrossRef]
28. Dubey, S.; Sarvaiya, J.N.; Seshadri, B. Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World–A Review. *Energy Procedia* **2013**, *33*, 311–321. [CrossRef]
29. Sher, A.A.; Ahmad, N.; Sattar, M.; Ghafoor, U.; Shah, U.H. Effect of Various Dusts and Humidity on the Performance of Renewable Energy Modules. *Energies* **2023**, *16*, 4857. [CrossRef]
30. Aljuaid, T.; Sasi, S. Proper imputation techniques for missing values in data sets. In Proceedings of the 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, 23–25 August 2016; IEEE: Piscataway, NJ, USA, 2016.
31. Mbuli, N.; Mathonsi, M.; Seitshiro, M.; Pretorius, J.H.C. Decomposition forecasting methods: A review of applications in power systems. *Energy Rep.* **2020**, *6*, 298–306. [CrossRef]

32. Liu, W.; Mao, Z. Short-term photovoltaic power forecasting with feature extraction and attention mechanisms. *Renew. Energy* **2024**, *226*, 120437. [CrossRef]
33. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
34. Garip, Z.; Ekinci, E.; Alan, A. Day-ahead solar photovoltaic energy forecasting based on weather data using LSTM networks: A comparative study for photovoltaic (PV) panels in Turkey. *Electr. Eng.* **2023**, *105*, 3329–3345. [CrossRef]
35. Husein, M.; Gago, E.; Hasan, B.; Pegalajar, M. Towards energy efficiency: A comprehensive review of deep learning-based photovoltaic power forecasting strategies. *Heliyon* **2024**, *10*, e33419. [CrossRef]
36. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
37. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
38. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
39. Son, N.; Jung, M. Analysis of meteorological factor multivariate models for medium-and long-term photovoltaic solar power forecasting using long short-term memory. *Appl. Sci.* **2020**, *11*, 316. [CrossRef]
40. Qu, J.; Qian, Z.; Pei, Y. Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy* **2021**, *232*, 120996. [CrossRef]
41. Konstantinou, M.; Peratikou, S.; Charalambides, A.G. Solar photovoltaic forecasting of power output using lstm networks. *Atmosphere* **2021**, *12*, 124. [CrossRef]
42. Mauladdawilah, H.; Gago, E.; Pegalajar, M.; Balfaqih, M. An Evaluation of Meteorological Variables Impact on Photovoltaic Power Generation Estimation Based on Deep Learning Model. In Proceedings of the 2025 4th International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 13–14 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 638–642.
43. Mauladdawilah, H.; Balfaqih, M.; Balfagih, Z.; Gago, E.; Pegalajar, M. Optimization of Photovoltaic Power Forecasting: A Comparative Study of Deep Learning Architectures, Optimization Techniques, and Evaluation Metrics. In Proceedings of the 2025 22nd International Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 15–16 January 2025; IEEE: Piscataway, NJ, USA, 2025; Volume 22, pp. 109–114.
44. Harvey, D.; Leybourne, S.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [CrossRef]