A comparative study on time series prediction of photovoltaic-power production through classic statistical techniques and short-term memory networks

Juan F. Duran
Department of Electrical Engineering, Electronics, and Telecommunications
Universidad de Cuenca
Ave. 12 de Abril, 010101, Cuenca, Ecuador
juan.durans@ucuenca.edu.ec

Luis I. Minchala
Department of Electrical Engineering, Electronics, and Telecommunications
Universidad de Cuenca
Ave. 12 de Abril, 010101, Cuenca, Ecuador
ismael.minchala@ucuenca.edu.ec

***Abstract*—The inherent variability in the power production of renewable energy sources (RES) limits the effectiveness of energy management systems (EMS) since optimal dispatch on power networks highly depends on the accuracy of predictors associated with the energy output and load demand. Consequently, power prediction tools for variable time horizons allow for improving energy management decisions. In this context, this work presents a detailed methodology for the deployment of predictive models for the photovoltaic (PV) power output of a small solar farm. The prediction models process a PV power dataset's time series using statistical techniques and neural networks with long-short term memory (LSTM). Before the data fitting, we develop a data preprocessing system, which involves evaluating missing data in the time series and getting descriptive analysis of the data set to either complete portions or delete atypical data. The results strongly suggest that the LSTM network performs better than the statistical model in exchange for more considerable computation times for long-term predictions.**

Keywords— forecasting, LSTM, photovoltaic power generation, statistical methods

I. Introduction

Integrating renewable energy sources into power grids has experienced significant growth in the last two decades. Among these technologies, PV power systems have gained a significant market share, mainly due to the price drops associated with this technology. However, PV power generation is inherently associated with the variability of the solar irradiance due to the stochastic nature of this resource, which in weak networks could lead to the activation of the low-frequency protections, severe voltage fluctuations, and load damage, among other issues [1].

Renewable energy systems typically include energy management systems (EMS), which monitor and control power production to maximize the exploitation of renewable resources. A classic task integrated into an EMS is power dispatch. Nevertheless, the RES variability complicates achieving optimal dispatch due to its dependency on the accuracy of power and load demand predictions. Consequently, accurate power prediction systems in specific time horizons are highly desired within the structure of an EMS [2].

Literate related to time series PV power forecasting classifies prediction models into physical models, statistical methods, machine-learning algorithms, and hybrid approaches. Physical models use stochastic processes and meteorological information to perform power predictions; however, this approach is complex to implement in a real-world scenario due to the high amount of real-time data processing from pretty accurate sensors. The statistical approach establishes models from the correlation between the current and previous samples of the PV power generated. Machine learning techniques use artificial intelligence (AI) algorithms, *e.g.* artificial neural networks (ANN), fuzzy logic, etc. Finally, hybrid models combine two or more techniques to obtain more accurate predictions [3], [4], [5].

Several contributions to predicting power production in PV systems develop methodologies associated with statistical approaches and machine learning. For instance, authors of [6] compare the performance of power forecasting methods associated with an ARIMA model and a recurrent neural network (RNA). The ARIMA model offered better performance than the RNA; however, both models have an evident prediction deficiency when the clear-sky index is low. References [7], [8], present a similar approach and compare an ARIMA model and an LSTM network. These models are trained with time series extracted from different locations within the same territorial space; in this context, the results show a notable improvement in forecasts concerning those techniques trained with data from a single location. Furthermore, authors of [9] test an ARIMA model and an LSTM network and show how the prediction quality improves when historical data increases, especially in the LSTM network. Finally, in [10], several statistical models are tested, such as ARIMA, SARIMA, SARIMAX, ARIMAX, among others, in contrast to an LSTM network, where the winner turns out to be the AI technique; however, statistical models work well in situations where access to data is limited, or there is very little prior information.

This work presents a detailed methodology for designing PV power prediction systems using statistical approaches and an LSTM network. The training data set corresponds to real data from a small solar farm. Despite its greater processing time, the results show lower prediction errors for the LSTM approach in more extended time frames.

The remainder of this document is organized as follows: section II presents the methodology associated with the data

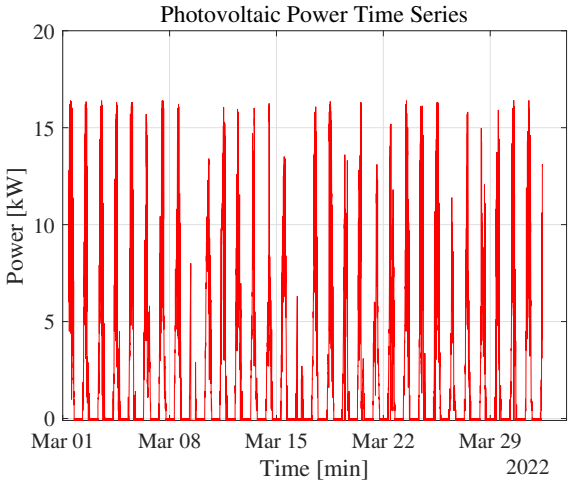


Fig. 1: PV power time series from the solar farm

fitting of the training set and the two proposed approaches. Section III presents the results of the proposed predictive approaches. Section IV presents the conclusions of this work.

II. Methodology

The data set used for training, testing, and validation corresponds to PV power from a small solar farm (15 monocrystalline panels) with a peak capacity of 20 kW. The observations were captured with a sampling time of 1 minute during 31 days of March 2022. Fig. 1 shows the PV power time series of this solar farm.

Fig. 2 shows the overview of the methodology we use throughout this work. There are four that briefly perform the following:

- Data imputation: replacing missing data with substituted values.
- Descriptive analysis: statistical analysis of the data.
- Model fit: data adjustment to SARIMA and LSTM models
- Evaluation: metrics evaluation to validate the models

A. Insertion of missing data

A preliminary step for analyzing a time series is to preprocess the dataset. Most real scenario datasets have missing data and atypical information, among other problems, which need to be handled beforehand [11]. For instance, as seen in Fig 1, there are portions of data lost in the time series, which can be associated with sensor's malfunction/disconnection.

There are several ways to complete the information within a time series. According to literature, traditional ways are: imputation by moving or fixed average, imputation by exponential filtering, and imputation through random values

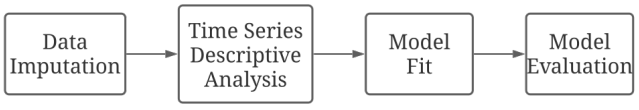


Fig. 2: The work process for the adjustment of time series prediction models

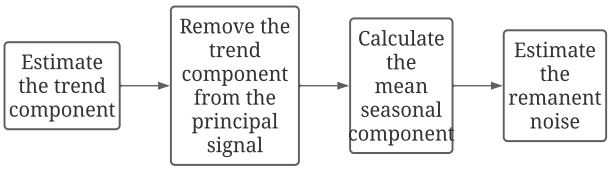


Fig. 3: Steps to decompose a time series

between the maximum limits and minimum of the time series [12], [11]. In this work, we use the imputation by fixed mean where missing data is the average of the observations of the series in similar time moments.

B. Descriptive analysis of the time series

To gain a better understanding of the behavior of the time series, a preliminary statistical analysis is performed based on a graphical approach and decomposition of the time series, which according to [11], [12], consists of four components:

1) Trend
2) Cyclic component
3) Seasonal component
4) Random component

Fig. 3 presents the steps we applied for decomposing the PV power time series. Fig. 4 shows three out of the four components of the time series from Fig. 1. These graphs allow concluding the following:

- The series' trend has variations over time, suggesting a non-constant mean and variance.
- The presence of a clear seasonality with an extension of 1440 samples (1 day), which is precisely the length of a day sampled at 1-minute intervals.
- The existence of high variability is undoubtedly associated with the stochastic movement of the clouds.

The graphic description of Fig. 4 provides accurate information about the behavior of the time series. However, it does not indicate the relationship between one sample and another, which is critical for adjusting a statistical model. Consequently, it is essential to use the simple autocorrelation function (ACF) and the partial autocorrelation function

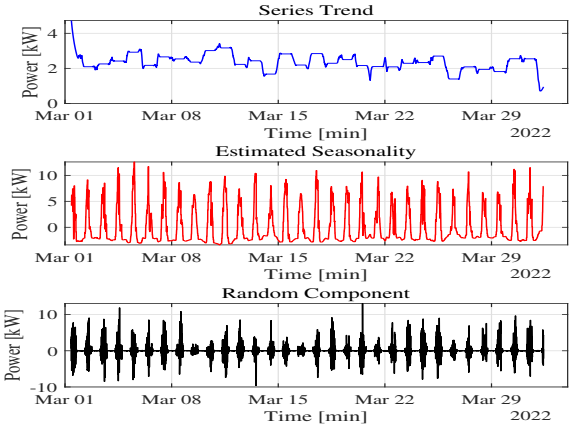


Fig. 4: Components of the time series of photovoltaic power

(PACF) to establish statistical relationships between instants of time [13], [12], [14].

The correlograms of the time series of Fig. 1 allow determing the following:

- A slow ACF decay is a sign that the mean and variance statistics are not constant. Consequently, the time series is not stationary.
- A significant lag in the ACF, in multiples of 1440, indicates the seasonality of the series for every 1440 samples.
- According to the PACF, the number of significant lags is eight samples.

C. Adjustment of the SARIMA and LSTM models

1) SARIMA model: In [14] is stated that the adjustment of this type follows the Box-Jenkins methodology, which encompasses three aspects: the identification of model components, the estimation of its parameters, and the evaluation.

To determine the parameters in a SARIMA model, the ACF and PACF have an exponential decay behavior with damped sinusoids. In addition, they include a differentiation operation that helps to find the stationary version of a series [14], [12], [13], [2]. Usually, a time series does not differ more than twice; if this happens, the time series requires some previous operation (smoothing). Additionally, the model considers a seasonal component where the order of the AR and MA components depends on the number of significant delays that appear in the time lag associated with the period.

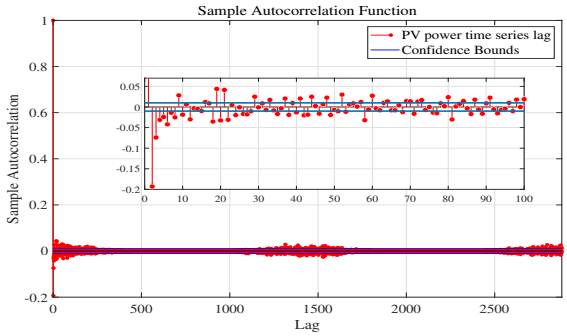
In the previous section, the ACF and PACF of the original time series describe an exponentially slow decay process, which is directly associated with a non-stationary process. Therefore, the need to apply differentiation at the level of the non-seasonal component is essential. The ACF and PACF of Fig. 5 were obtained by performing this first-order operation. As a result, these correlograms are more similar to the shapes described for a SARIMA model and allow us to deduce the information from Table I, where the maximum degrees of every component of this model are shown.

Values from Table I allow to find, iteratively, the best SARIMA model by varying the components order within these ranges, then , estimating its parameters, and evaluating the results of every iteration through the evaluation metrics. The best model is selected by evaluating the prediction error of every iterative model.

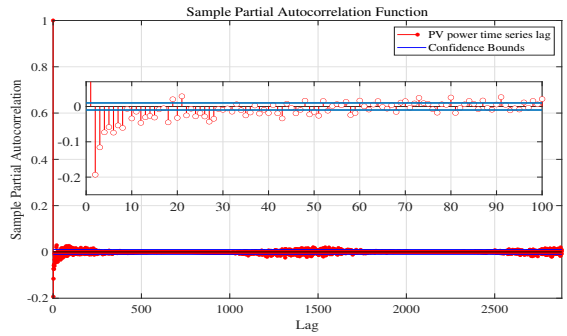
2) LSTM neural network: For the adjustment of the LSTM neural network, two architectures are considered. The first is mentioned in [15] and implies using a deep neural network for PV power forecast. This network comprises an input

TABLE I: Maximum degrees of each of the components of the SARIMA model

Component	Seasonal component	Non-seasonal component
AR	8	1
MA	4	1
Differentiation	1	1



(a) Simple autocorrelation function



(b) Partial autocorrelation function

Fig. 5: Time Series Correlation Functions with First Order Differentiation

layer, 2 LSTM layers of 469 hidden units, and a third layer of 338 neurons, which are separated by dropout layers with probability 0.2, 0.15, and 0.41. In addition, this architecture also has a fully-connection layer and, finally, a regression layer, which calculates the loss of the mean square error of the predicted value versus the actual value.

A second architecture is proposed by [16], that recommends an ANN with an input layer, an LSTM layer of 200 hidden units, a fully connected layer with an output of 50 sequential data serving as input to a dropout layer of probability 0.5, which connects to a fully connected second layer that delivers the predicted output. Finally, the prediction quality is measured with a regression layer.

The MATLAB *Deep Learning Toolbox* tool was used to train each architecture. Table II summarizes the training options we used. Some key points to take into consideration for the network training process are:

- Separate the data in a proportion of 70% for the training phase, 10% will serve as model validation data, and 20% will be test data for future predictions.
- The form of test for the network will be in a closed loop. The only known data the network will have for future predictions are the first 1440 data of the test time series.
- The reason for initializing the network with 1440 data is directly associated with the ACF of the original series since the first 1440 data have a strong relationship with each other.

TABLE II: LSTM Network Training Options

Option	Value	Description
Training algorithm	adam	Adaptive moment estimation algorithm. Adjusts learning rates by optimizing the loss function, with consideration of a moment term
MaxEpochs	100	Maximum number of epochs
GradientThreshold	1	Limit gradient burst to avoid training divergence
InitialLearnRate	0.005	Specifies the initial learning rate of the training
LearnRateSchedule	piecewise	Update the learning rate every certain number of epochs
LearnRateDropPeriod	50	Indicates the number of epochs in which the previous parameter modifies the learning rate
LearnRateDropFactor	0.2	Learning rate reduction factor
SequencePaddingDirection	right	Right truncation direction, to prevent later time units from influencing earlier time predictions
ValidationFrequency	10	The frequency with which the algorithm performs validation tests

D. Models evaluation

There are several ways to validate the performance of a prediction model. The indicators we used for evaluation in this work are described below:

- For statistical models, an evaluation metric is the Bayesian Schwarz Information Criterion (BIC) which is calculated by Eq. (1), where T corresponds to the observations used for the estimation, k is the number of predictors, the term $k+2$ considers the predictor coefficients k, the intercept and the variance of the residuals. The idea of this criterion is to penalize the fit of the model (SSE) depending on the number of parameters that need to be estimated and the amount of total data of the time series [14].

$$BIC = Tlog\left(\frac{SSE}{T}\right) + 2(K+2)log(T) \quad (1)$$

- Statistical metrics, such as root mean square error (RMSE) and normalized root mean square error (NRMSE), are the leading measures of model evaluation [3], [4], [5]. These metrics are calculated through equations (2) and (3), respectively.

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

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}}{\sum_{i=1}^{N} {y_i}^2} * 100 \quad (3)$$

where y_i is the actual value, $\hat{y_i}$ is the prediction value, and N is the total number of data.

III. Results and Discussion

The models' evaluation consists in applying the predictive approaches to the dataset (see Fig. 1) corresponding to the solar farm under study. Discussion about the predictions is based on the analysis of the evaluation metrics discussed in subsection II.C. For instance, Fig. 6 shows the evolution of the BIC versus the NRMSE for different SARIMA models. If the fit quality were evaluated solely by the root mean square error, every model would have performed excellently since the estimation error is less than 1%. However, when analyzing the BIC, it can be noticed that models with either excessively high orders (e.g$(7,1,3)\times(0,1,0)_{1440}$) or excessively low orders (e.g$(1,1,1)\times(0,0,1)_{1440}$) have relatively poor performances. This means that although the prediction error is low, it is inadequate for the number of terms in the model. Consequently, the optimal model that was obtained through this criterion is a SARIMA $(3,1,3)\times(1,0,1)_{1440}$. Table III show the SARIMA model parameters.

We proceed similarly with the evaluation of LSTM networks. Table IV shows the metrics obtained during adjustment. It can be considered that both networks had an

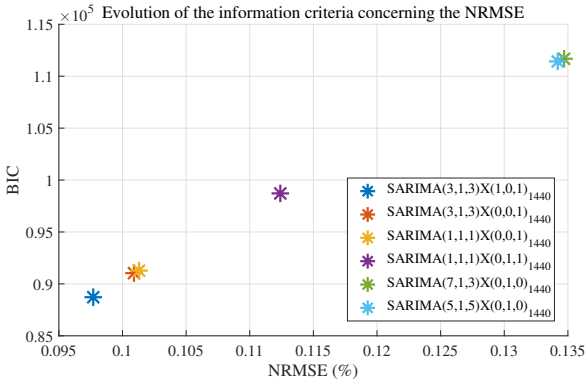


Fig. 6: Evolution of the BIC of the adjusted SARIMA models

TABLE III: Value of the adjusted SARIMA model coefficients

Parameter	Value	Standard error
c	-2.08E-05	6.84E-05
AR{1}	0.3611	0.019017
AR{2}	0.85954	0.0082965
AR{3}	-0.3479	0.01172
SAR{1440}	0.65942	0.0018554
MA{1}	-0.59855	0.018773
MA{2}	-0.87783	0.010533
MA{1}	0.51207	0.013237
SMA{1440}	-0.64896	0.0033403
Variance	0.70009	0.0013656

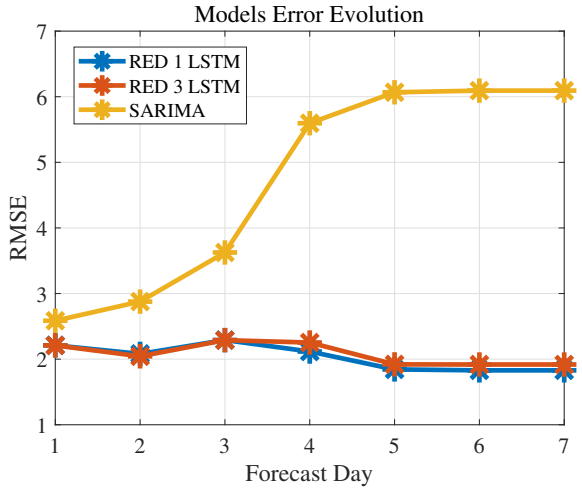


Fig. 7: Evolution of the RMSE depending on the number of days forecast

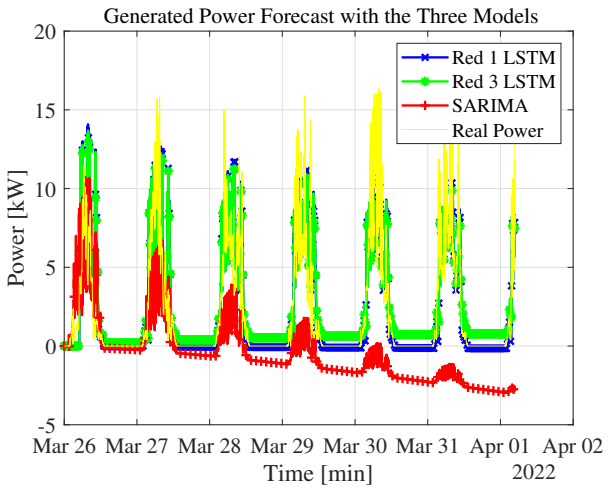


Fig. 8: Prediction results with trained models

acceptable fit, with an NRMSE below 1%. However, unlike the SARIMA models, it is impossible to determine which network to use clearly. That is why the final selection between the statistical model and the neural networks is based on the quality of the closed-loop predictions and the computation time. Figures 7, 8, and 9 show the quality of the predictions of each model graphically.

The three models present a reasonably good prediction quality at the metric level. Nevertheless, it is crucial to highlight essential points such as the statistical model gradually loses precision, as samples are forecasted beyond a seasonal period, and this is directly reflected in the evolution of the RMSE, as seen in Fig. 7, from the second forecast day, this parameter increases and tends to keep increasing as the forecast advances in time. This behavior is directly associated with the nature of the model since it considers non-seasonal relationships of up to 3 past samples and seasonal ones with a period of 1440. However, something important to highlight is its computational speed, which is significantly lower than the machine learning techniques.

On the other hand, neural networks present a better performance when predicting time instants beyond a seasonal period, as shown in Fig. 7, the RMSE tends to remain constant over time, and this was to be expected thanks to the goodness of the memory of the LSTM layers. However, the computational cost increases significantly, and for this time series, an architecture with a single LSTM layer is more accurate and less expensive.

Fig. 9 shows the dispersion of the residuals of the models.

TABLE IV: Prediction model evaluation metrics

Metric	SARIMA		Network with an LSTM layer (net1)		Network with three LSTM layers (net3)	
	Fit	Prediction	Fit	Prediction	Fit	Prediction
RMSE	0.837	4.126	0.199	2.885	0.204	2.828
NMRSE [%]	0.098	1.169	0.112	0.767	0.114	0.752
Time [s]	7200	18	900	209	1200	1032

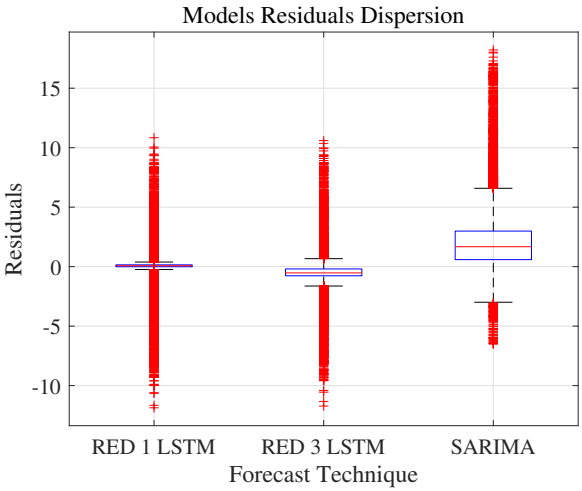


Fig. 9: Dispersion of the residuals of each of the predictive models

It can be seen how machine learning techniques present much more precise prediction results with slight variability since its box is located centered at the origin, and the whiskers are not at an alarming distance apart from the median. In simpler words, the predictions made by the neural networks are pretty similar to the real ones and have little variability, something that does not happen with the statistical model since its box is centered on a value different from zero, approximately 2kW, and a distance separates their whiskers concerning the median, this indicates that the average prediction error is around 2kW with high variability. An important feature that can be seen in this graph is the number of values that are outside the upper and lower limits in the 3 models. This behavior can be considered normal since the real-power value can be affected by the stochastic movement. of the clouds at that instant in time, causing an unexpected increase or decrease in power and thus generating an outlier.

IV. Conclusions

This work presents the adjustment of predictive models based on traditional statistical techniques and machine learning. Both techniques operate adequately depending on the approach that is intended to achieve. For instance, we are looking for long-term predictions where we do not have continuous access to the data. The LSTM network performs better than a statistical predictor, in addition to the potential advantage of evolving and adjusting to eventual changes in the time series at the cost of increasing the computation time. On the other hand, if the prediction horizon is not long and prediction speeds are needed, implying decision-making in short periods, the SARIMA model is a potential candidate due to its low computational cost.

Projecting the results to the area of interest, which is the optimal energy management in systems with high penetration of renewable sources, the LSTM network can be implemented for techniques that involve the commitment of the generation unit in such a way that the available power is predicted in a time horizon of 1 or 2 days to program an optimal dispatch of power. On the other hand, the SARIMA model can be beneficial in shorter time horizons, in the order of minutes, to take corrective actions, through a battery bank, in the variability of the renewable resource.

In conclusion, the choice of one technique over another depends on the purpose of the prediction. LSTM networks are preferred for long-term predictions, while SARIMA models are more suitable for short-term predictions. The ability to quickly adjust to changes in the time series as well as the low computational cost make both techniques ideal for energy management in systems with high penetration of renewable sources. As such, they can be used together to create a robust system that is capable of accurately predicting and responding to changes in energy demand and supply.

References

[1] M. A. Syed and M. Khalid, "Neural network predictive control for smoothing of solar power fluctuations with battery energy storage," *Journal of Energy Storage*, vol. 42, p. 103014, 10 2021.

[2] B. Singh and D. Pozo, "A guide to solar power forecasting using arma models," in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 2019, pp. 1–4.

[3] Y.-K. Wu, C.-L. Huang, Q.-T. Phan, and Y.-Y. Li, "Completed review of various solar power forecasting techniques considering different viewpoints," *Energies*, vol. 15, no. 9, 2022, cited by: 3; All Open Access, Gold Open Access.

[4] T. C. Carneiro, P. C. M. De Carvalho, H. A. Dos Santos, M. A. F. B. Lima, and A. P. De Souza Braga, "Review on photovoltaic power and solar resource forecasting: Current status and trends," *Journal of Solar Energy Engineering, Transactions of the ASME*, vol. 144, no. 1, 2022, cited by: 16.

[5] A. Mellit, "An overview on the application of machine learning and deep learning for photovoltaic output power forecasting," *Lecture Notes in Electrical Engineering*, vol. 681, p. 55 – 68, 2021, cited by: 2.

[6] L. Fara, A. Diaconu, D. Craciunescu, and S. Fara, "Forecasting of energy production for photovoltaic systems based on arima and ann advanced models," *International Journal of Photoenergy*, vol. 2021, 2021, cited by: 12; All Open Access, Gold Open Access.

[7] S. De Jongh, T. Riedel, F. Mueller, A. E. Yacoub, M. Suriyah, and T. Leibfried, "Spatio-temporal short term photovoltaic generation forecasting with uncertainty estimates using machine learning methods," 2020, Conference paper, cited by: 2.

[8] O. El Alani, C. Hajjaj, H. Ghennioui, A. Ghennioui, P. Blanc, Y.-M. Saint-Drenan, and M. El Monady, "Performance assessment of sarima, mlp and lstm models for short-term solar irradiance prediction under different climates in morocco," *International Journal of Ambient Energy*, 2022, cited by: 0.

[9] E. J. Santana, R. P. Silva, B. B. Zarpelão, and S. Barbon Junior, "Photovoltaic generation forecast: Model training and adversarial attack aspects," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12320 LNAI, p. 634 – 649, 2020, cited by: 3.

[10] E. Kim, M. S. Akhtar, and O.-B. Yang, "Designing solar power generation output forecasting methods using time series algorithms," *Electric Power Systems Research*, vol. 216, 2023, cited by: 0.

[11] W. A. Woodward, B. Sadler, and S. Robertson, *Time Series for Data Science: Analysis and Forecasting*, 1st ed. A Chapman & Hall Book, 2022.

[12] J. García, *Predicción en el dominio del tiempo: análisis de series temporales para ingenieros.*, 1st ed. Editorial de la Universidad Politécnica de Valencia, 2016.

[13] Álvaro Montenegro, *Análisis de series de tiempo*, 1st ed. Pontificia Universidad Javeriana, 2010.

[14] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Australia: OTexts, 2018.

[15] R. Costa, "Convolutional-lstm networks and generalization in forecasting of household photovoltaic generation," *Engineering Applications of Artificial Intelligence*, vol. 116, 2022.

[16] J. Little and C. Moler, "Sequence-to-one regression using deep learning," 2023.