12th International Conference on Electrical, Electronics and Computer Engineering (IcETRAN), Čačak, 9-12 June 2025

# Effect of Solar and Weather Parameters on LSTM-based Model Predictions of Solar Power Production

Novak Radivojević
*University of Niš, Faculty of Electronic Engineering*
Niš, Serbia
novak.radivojevic@elfak.ni.ac.rs
ORCID 0009-0000-4423-6224

Uroš Ilić
*University of Niš, Faculty of Electronic Engineering*
Niš, Serbia
uros.ilic@elfak.ni.ac.rs
ORCID 0009-0006-0414-6404

Andrija Petrušić
*University of Niš, Faculty of Electronic Engineering*
Niš, Serbia
andrija.petrusic@elfak.ni.ac.rs
ORCID 0009-0004-7615-6500

Miona Andrejević Stošović
*University of Niš, Faculty of Electronic Engineering*
Niš, Serbia
miona.andrejevic@elfak.ni.ac.rs
ORCID 0000-0002-2211-9024

***Abstract*—This paper discusses the process of training a neural network intended for predicting power generation in a solar power plant based on meteorological data and the historical data of power measurement readings. Initially, the neural network was trained on a wide array of different solar and weather quantities as inputs (26 in total), but since the prediction errors of this model were considerably large, the model was retrained, but on a much smaller subset of critical input variables, which resulted in an increase of accuracy where MAE dropped significantly.**

***Keywords—solar, weather, prediction, neural network, accuracy***

## I. INTRODUCTION

The exponential growth and scalability of renewable energy source (RES) power plants like wind, solar, and hydroelectric has given rise to the concept of distributed generation. Distributed generation has expanded with renewable energy sources at the distribution level, creating "Prosumers" – entities which are internally connected to a RES and have the ability to switch between the energy consumer and producer in real time by modifying energy flow at the exchange point [1].

In recent times, the electric power system has shifted from a natural monopoly to a market-based system with competition in generation and supply through spot markets. Electric power system liberalization has led to the emergence of independent power producers, traders, and suppliers, while the system infrastructure remains managed by transmission and distribution system operators. A problem which is often encountered in such systems is the balancing responsibility - equilibrium between generation and consumption must be maintained as any drastic deviation between the two could result in system failure. Balancing has become more challenging due to the intermittent nature of RES. Balance responsible parties must provide hourly forecasts of energy needs and planned generation to the transmission system operator, with the balancing market charging for deviations between forecasted and actual values [2]. Market mechanisms promoting predictability and controllability affect pricing for power producers, end consumers, prosumers, and active consumers through dynamic supply pricing and balancing service charges.

Forecasting energy needs and renewable generation potential is considered the greatest challenge in managing the electric power system, attracting significant academic research. This study proposes a neural network model based on weather parameters and historical data of generated power to estimate the power output of a prosumer-regime solar power plant in Vladičin Han, a town in southeastern Serbia, while aiming to use a minimal number of input variables. Predictions are made hourly, based on projected weather conditions and data from the power plant's previous measured output in the hours preceding the prediction

## II. INPUT DATA SELECTION AND DATASET OVERVIEW

Numerous studies have been published on forecasting electricity generation in solar power plants using meteorological conditions as input factors [3]. Forecast models are increasingly more often relying on data gathered from meteorological databases, rather than direct parameter measurement, which can be difficult to integrate into forecasting systems [4]. Artificial neural networks and other similar deep learning algorithms have been used in many studies for prediction of solar power plant generation, and their successful implementation requires the selection of appropriate input variables [5]. Meteorological variables connected with solar power production include solar irradiance, air temperature, cloud cover, the Sun's angular position (zenith and azimuth), wind speed and direction, air humidity, pressure, precipitation, snow depth, etc.

Solar irradiance, or the amount of solar radiation that reaches a specific surface, is commonly regarded as essential for such tasks [6]. Because of its relevance, weather databases include irradiance-related factors derived from other meteorological variables. Solar irradiance is usually expressed through different suitable components:

- Direct normal irradiance (DNI) – irradiance on a surface perpendicular to a beam coming directly (in a straight line) from the Sun
- Diffuse horizontal irradiance (DHI) - horizontal component of irradiance reaching a surface by diffusion (scattered by atmosphere)

ISBN 978-86-6200-031-6 ©2025 AM & ETRAN

https://doi.org/10.69994/12Ic25020

- Global horizontal irradiance (GHI) – Total irradiance represented by the sum of horizontal component of DNI and DHI
- Global tilted irradiance (GTI) - Total irradiance on an inclined plane

The databases also use certain methods to estimate the value of irradiance components under clear sky conditions, which can be useful in different cases, e.g. assessing a location's potential for energy production or solar plant fault detection [7] .

With the aim of reaching high prediction model accuracy, an assumption is made that including a large number of diverse meteorological variables as the model inputs should enhance the model's generalization capability by presenting the model to a set of different weather conditions under which a power plant can operate. However, some groups of variables provided by the database exhibit mutual dependencies (multicollinearity), e.g. the amount of irradiance is highly correlated to elevation of the Sun, and the effect of cloud cover can be observed through the difference between the regular and clear-sky irradiance, etc. Sets of input variables which are highly corelated carry similar information, so including only a single variable from such a set may be enough for model training. Reducing the number of input variable is beneficial since it reduces training time, memory usage, and the amount of data requested from the database. The goal of this study is to see how a neural network model performs when trained with different combinations of a subset of input variables compared to using all variables available.

Meteorological data were gathered using the *Solcast* platform, which collects and calculates weather variables suited for optimizing the operations of solar power plants. The weather estimates from this platform are created using advanced numerical models that analyze satellite imaging of cloud cover and combine data from various ground-based weather stations [8]. Solar power output data were collected directly from the power plant's measurement equipment. Power measurement was done at a 5-minute sampling period, and measurement data from December 2021 to November 2024 is provided for the research. The plant consists of three inverters, two of which have a power rating of 60 kW and one with 36 kW. Since hourly predictions are required, the power dataset is rearranged so that the model is trained on the sum of the power produced by all three inverters in each hour. Figure 1 depicts this dataset, while Table I. provides a statistical summary of the power dataset distribution after resampling to the 1-hour interval. All data and resulting metrics in Table I. are expressed in kW.

In Fig. 1, the seasonality of power production at the annual level can be clearly noticed, with production peaking in the summer and falling in the winter, indicating that solar irradiance is not the only parameter affecting the production and that various other weather conditions can influence production across seasons. Fig. 2 and Fig. 3 display how power production and GHI typically vary during a 24-hour period under clear sky conditions and with cloud presence, respectively, where the effect of irradiance components on power production can be observed.

TABLE I. POWER DATASET STATISTICS

| **Total** | **Mean** | **Std. dev.** | **Min.** | **25$^{th}$ perc.** | **50$^{th}$ perc.** | **75$^{th}$ perc.** | **Max.** |
|---|---|---|---|---|---|---|---|
| 9129 | 625.8 | 505.01 | 0.9 | 175.11 | 492.4 | 1025.55 | 2004.6 |

## III. TRAINING PROCESS AND MODEL SELECTION

For the purpose of this work, a neural network was implemented using two hidden layers: an LSTM and a regular fully connected (*Dense*) layer. The network was trained and validated using a standard train-test split approach - the weather and power datasets are combined into a single dataset of pairs

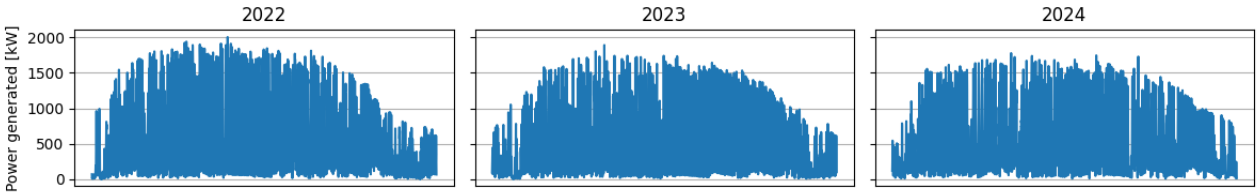


Fig. 1 Overview of the generated power dataset split per year.

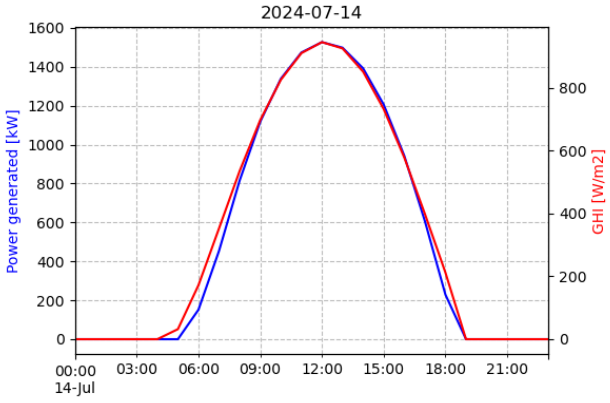


Fig. 2 Typical clear-sky daily production compared to GHI.

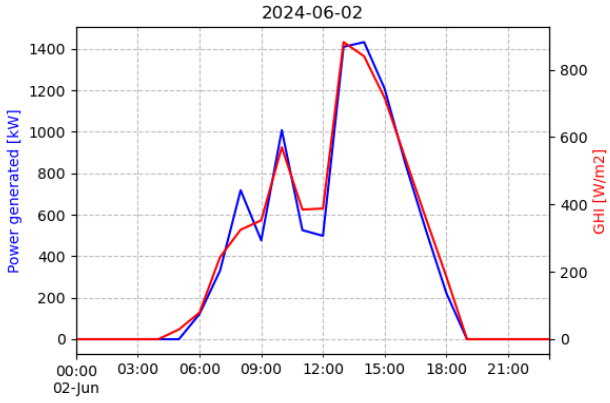


Fig. 3 Power production under cloud presence before noon, compared to GHI

of input and output vectors, which is split into train and test sets. The test set contains the last 30% of data from the original dataset, ordered by date. The input-output pairs for training and testing are formed from weather and power data, taking into account their values over a time interval of several hours prior to the prediction moment - for each prediction target (output), which is a value of power produced at some time instant *t*, the inputs include the values of each selected weather variable at times *t*, *t*-1, *t*-2, ... *t*-*h*, as well as the measured production at times *t*-1, *t*-2, ... *t*-*h*+1. Here, *h* represents the prediction window and can also be considered a hyperparameter for model training.

Given the seasonal characteristics of power production and with the goal of achieving high prediction model accuracy, the network was initially trained with all the weather variables available for download from the database. These variables include solar irradiance components (DNI, DHI, GHI, GTI), irradiance components under clear-sky conditions, cloud opacity, albedo, sun position (azimuth and zenith), amount of precipitation, relative air humidity, pressure, amount of snow, and finally wind speed and direction at 10m and 100m heights.

Aside from the choice of input variables, prediction accuracy may also depend on the network structure (for example, the number of neurons in the hidden layers) and the convergence flow of the learning algorithm. To address this, a cross-validation method was used on a set of structure and learning process parameters, which would yield both an accurate and quickly converging model. The list of selected hyperparameters selected and their chosen values are shown in Table II. Optimizer used for model training was *Adam*, a widely used stochastic gradient descent method that uses adaptive estimation of first- and second-order moments. During cross-validation, the model's accuracy was evaluated on the test set, where the error between predicted and actual values was computed for each test set output. Based on the errors calculated, mean absolute error (MAE) and coefficient of determination ($R^2$) were used to assess model accuracy.

After the cross-validation training with all inputs, a model with 120 neurons in both layers, trained for 300 epochs with learning rate of 0.002 and the batch size of 256 generated the best results based on the metrics chosen. Table III. shows metric values and a statistical summary of the model's prediction errors for each of the chosen prediction windows. In most cases, the MAE was above 110 kW which is within 17.57% of the mean of the power dataset (625.8 kW). Meanwhile, the $R^2$ value was lower than, or around 0.9 which suggests that the errors are considerably large and there is room for improvement in model training, under different variable or parameter selection.

TABLE II. HYPERPARAMETER COMBINATIONS

| **Hyperparameter name** | **Set of values** |
|---|---|
| Prediction window (*h*) | 0 (i.e. not taking previous values), 4, 8, 12, 16, 20, 24 |
| LSTM neurons | 60, 80, 100, 120, 140 |
| Dense layer neurons | 60, 80, 100, 120, 140 |
| Learning rate | 0.002, 0.004, 0.006, 0.008, 0.01 |
| Batch size | 256, 512, 1024 |
| Training epochs | 100, 200, 300, 400, 500 |

Figs. 4 and 5 show the predictions compared to the measured power values along with the absolute errors between the measured and predicted values, for the same days chosen as examples in Figs. 2 and 3. The two peaks which appear on Fig. 4 show that the model failed to capture the pattern of a clear-sky day production, suggesting that the model was affected by overfitting to a specific set of weather conditions. The model also gives more optimistic predictions under increased cloud presence, as seen on Fig. 5.

TABLE III. ALL-INPUTS TEST SET ERROR STATISTICS

| *h* | MAE | $R^2$ | Std. dev. | Min. | 25th per. | 50th per. | 75th per. | Max. |
|---|---|---|---|---|---|---|---|---|
| 0 | 106.54 | 0.906 | 102.71 | 0.01 | 36.76 | 80.25 | 139.69 | 714.68 |
| 4 | 95.39 | 0.909 | 110.23 | 0.02 | 25.92 | 58.80 | 119.51 | 1190.82 |
| 8 | 113.71 | 0.869 | 133.58 | 0.02 | 27.62 | 69.06 | 147.11 | 1188.09 |
| 12 | 117.77 | 0.866 | 132.27 | 0.01 | 30.28 | 74.96 | 150.15 | 1060.42 |
| 16 | 117.95 | 0.866 | 131.62 | 0.04 | 28.51 | 75.11 | 156.22 | 981.58 |
| 20 | 123.61 | 0.860 | 132.84 | 0.05 | 33.22 | 80.05 | 168.01 | 1234.2 |
| 24 | 121.83 | 0.858 | 135.5 | 0.03 | 29.35 | 77.13 | 163.42 | 981.99 |

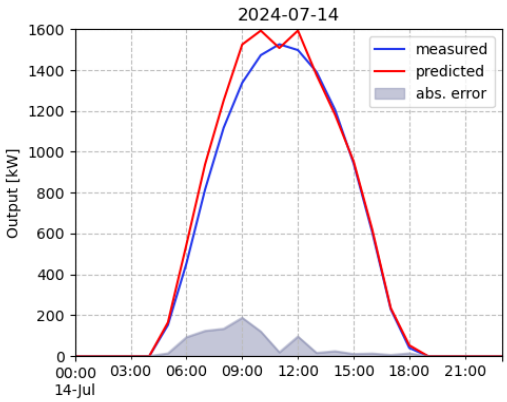


Fig. 4 All-input model predictions for a clear-sky day.

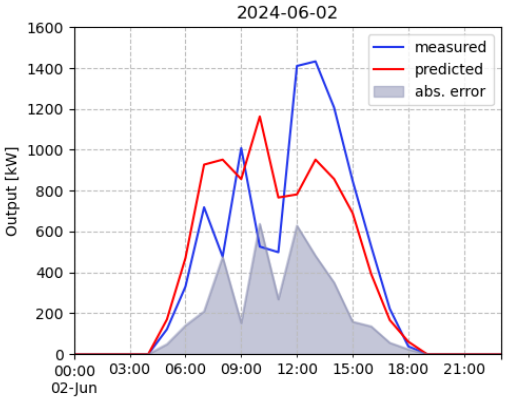


Fig. 5 All-input model predictions for a day with cloud presence.

In an attempt to improve the model's accuracy, some variables were removed from the list of inputs. Along with variables of major relevance, such as zenith, azimuth, and irradiance, only those with the greatest influence were selected, as determined by calculating the correlation coefficients between each weather input and the power output. In addition, the correlation among pairs of weather inputs was calculated in order to determine sets of similar inputs. For each set of multiple highly-correlated variables, only one of its variables is selected, which further reduces the list of necessary inputs. The ten inputs with highest absolute value of correlation factor with generated power measurement are as follows: GHI (0.871), GTI (0.835), clear-sky GHI (0.744), zenith (-0.733), clear-sky GTI (0.668), DNI (0.655), clear-sky DNI (0.587), cloud opacity (0.515), air temperature (0.444) and clear-sky DHI (0.338). As expected, the irradiance parameters (GHI, GTI, DNI, DHI and their clear-sky counterparts) have the highest correlation coefficients; however, the mutual correlations of all pairs of these parameters turned out to be greater than 0.8, so only GHI was chosen because its values take multiple other effects into account. The input set was reduced to GHI, air temperature, cloud opacity, zenith, and azimuth, with the remaining variables having little impact and hence being removed from the dataset.

After selecting the new subset of inputs, the model is retrained using the same structure and learning hyperparameters that generated the best results on the previous try. The model is trained with seven different input combinations, shown in Table IV. while using all previously selected prediction windows. All combinations include zenith and azimuth since these two parameters combined have the biggest impact on power production and assist the model in acquiring the time of day and year pattern. Table V. shows the resulting MAE for all models across previous time steps and input combinations after training.

TABLE IV. REDUCED INPUT COMBINATIONS

| | |
|---|---|
| 1. | Zenith, azimuth, GHI |
| 2. | Zenith, azimuth, air temperature |
| 3. | Zenith, azimuth, cloud opacity |
| 4. | Zenith, azimuth, GHI, air temperature |
| 5. | Zenith, azimuth, GHI, cloud opacity |
| 6. | Zenith, azimuth, air temperature, cloud opacity |
| 7. | Zenith, azimuth, GHI air temperature, cloud opacity |

TABLE V. RETRAINED MODELS MAE

| **Comb. No.** | **Prediction window (h)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **0** | **4** | **8** | **12** | **16** | **20** | **24** |
| 1. | 85.54 | 80.79 | 82.11 | 82.52 | 85.22 | 85.76 | 86.43 |
| 2. | 98.63 | 94.88 | 101.14 | 93.13 | 95.21 | 103.93 | 97.89 |
| 3. | 84.09 | 88.54 | 88.18 | 86.95 | 88.42 | 91.13 | 91.08 |
| 4. | 85.86 | 82.26 | 80.91 | 80.08 | 95.51 | 87.07 | 80.47 |
| 5. | 94.06 | 85.50 | 85.71 | 84.29 | 87.32 | 84.41 | 87.38 |
| 6. | 88.24 | 89.41 | 90.57 | 85.88 | 89.88 | 94.44 | 92.77 |
| 7. | 84.98 | 80.24 | 87.62 | 87.43 | 89.37 | 89.40 | 95.28 |

Metrics presented in Table V. suggest the following:

- In all cases, MAE is significantly lower (mostly ranging between around 80-90 kW) compared to the all-inputs model (mostly ranging between 110–120 kW) which clearly indicates an overall improvement in accuracy
- Especially for combination no. 4, MAE is considerably lower on majority of prediction window choices in comparison to other combinations
- Combination no. 3 (only using air temperature) gives less accurate predictions
- The lowest MAE in each combination tends to be achieved either for $h$=12 or $h$=4, while the lowest among all cases was achieved for combination no. 4 (80.08 kW).

Table VI. presents prediction error statistics for the model with lowest MAE ($h$=12, input combination no. 4) compared to the results achieved with a similar model trained on all inputs. The $R^2$ factor is now around 0.93 which suggests better predictions, while all the other metrics are significantly lower, with 25$^{th}$, 50$^{th}$, and 75$^{th}$ percentiles achieving a drop in about 35%. Resulting MAE is 80.08 kW which is within 12.79% of the mean of the power dataset, a significant drop compared to the model in the first pass. Figs. 6 and 7 show the predictions of this model for the two selected days. Predictions under both conditions show significantly lower errors.

TABLE VI. LOWEST-MAE MODEL STATISTICS

| | **MAE** | **$R^2$** | **Std. dev.** | **Min.** | **25$^{th}$ per.** | **50$^{th}$ per.** | **75$^{th}$ per.** | **Max.** |
|---|---|---|---|---|---|---|---|---|
| Comb. no. 4 | 80.08 | 0.932 | 97.36 | 0.03 | 19.46 | 45.23 | 100.77 | 790.24 |
| All inputs | 117.77 | 0.866 | 132.27 | 0.01 | 30.28 | 74.96 | 150.15 | 1060.42 |

## IV. CONCLUSION & FUTURE WORK

The subject of this paper is a neural network structure implementation in solar power plant production forecast and the problems encountered when choosing from a set of available input variables. The network was first trained on a set of more than 20 weather variables and the historic power measurements. Cross-validation was applied, including different network-related hyperparameters to determine the most accurate structure. Initially, the model made predictions with considerably large errors, with MAE in the range of 110-120 kW. In the second pass, the model was trained on different subsets of input variables selected based on correlation factors.

The errors have decreased according to most metrics, with MAE mostly between 80 and 90 kW. The results of this experiment illustrate the importance of choosing appropriate input variables for network training, apart from selecting a suitable network structure. Further research is pointed towards utilizing a standardized and more reliable approach in selecting the most appropriate input variables rather than using correlation factors. Another extension to this work can be training a model on different subsets of power datasets for different seasons of the year, as some weather conditions have a lot of impact only during a particular time of the year.

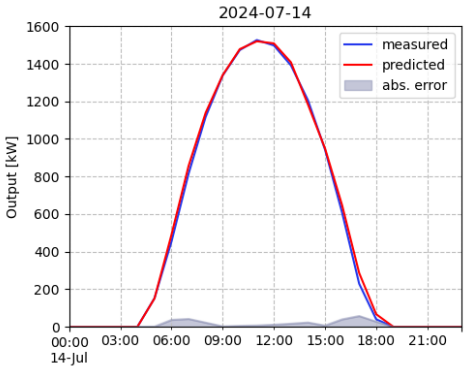


Fig. 6 Lowest-MAE model predictions for a clear-sky day.

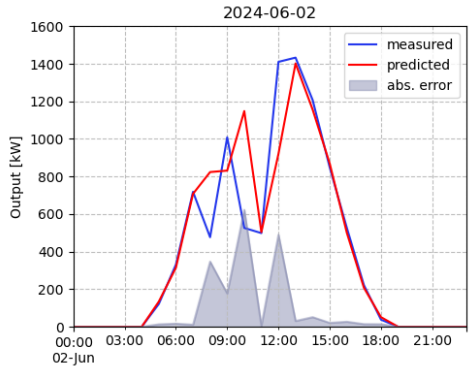


Fig. 7 Lowest-MAE model predictions for a day with significant cloud presence

## ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia. [Grant Number: 451-03-136/2025-03/200102]

## REFERENCES

[1] P. H. J. Nardelli, N. Rubido, C. Wang, M. S. Baptista, C. Pomalaza-Raez, P. Cardieri and M. Latva-aho, “Models for the modern power grid,” *European Physical Journal Special Topics*, vol. 223, Springer, 2014. doi:10.1140/epjst/e2014-02219-6.

[2] D. Huang and R. Billinton, "Effects of Load Sector Demand Side Management Applications in Generating Capacity Adequacy Assessment," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 335-343, 2012. doi: 10.1109/TPWRS.2011.2164425.

[3] G. de Freitas Viscondi and S. N. Alves-Souza, “A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting,” *Sustainable Energy Technologies and Assessments*, vol. 31, pp. 54–63, 2019.

[4] J. A. Ibañez, I. B. Benitez, J. M. Cañete, J. C. Magadia, and J. A. Principe, “Accuracy assessment of satellite-based and reanalysis solar irradiance data for solar PV output forecasting using SARIMAX”, *Journal of Renewable and Sustainable Energy*, vol. 15, no. 6, p. 066101, 12 2023.

[5] K. Anuradha, D. Erlapally, G. Karuna, V. Srilakshmi, and K. Adilakshmi, “Analysis of Solar Power Generation Forecasting Using Machine Learning Techniques”, *E3S Web of Conferences*, vol. 309, p. 01163, 01 2021.

[6] K. A. Ibrahim, G. Musa, and S. Aliyu, “The Effect of Solar Irradiation on Solar Cells”, *Science World Journal*, vol. 14, 01 2019.

[7] S. Peratikou and A. G. Charalambides, “Estimating clear-sky PV electricity production without exogenous data”, *Solar Energy Advances*, vol. 2, p. 100015, 2022.

[8] J. M. Bright, “Solcast: Validation of a satellite-derived solar irradiance dataset”, *Solar Energy*, vol. 189, p. 435-449., 2019, doi: 10.1016/j.solener.2019.07.086.