Time Series Forecasting for Daily to Monthly Temporal Hourly-based Solar PV Output Power

Yusak Tanoto
Electrical Engineering Department
Petra Christian University
Surabaya, Indonesia
tanyusak@petra.ac.id

Gregorius Satia Budhi
Informatics Department
Petra Christian University
Surabaya, Indonesia
greg@petra.ac.id

Jimlee Christanto Widjaya
Informatics Department
Petra Christian University
Surabaya, Indonesia
c14190050@john.petra.ac.id

2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 979-8-3503-5834-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISRITI60336.2023.10467972

***Abstract*—This paper presents the application of the Auto-Regressive Integrated Moving Average Exogenous (ARIMAX) model and compares its performance with Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) in forecasting daily, weekly, and monthly average solar PV output power. This study considers long-term hourly temporal-based solar PV output power for the Java-Bali region of Indonesia, as obtained from the Renewables.ninja solar PV model web-based tool. Using the Dash framework and Python, the study develops a web-based dashboard application that allows users to explore and analyse daily to monthly forecasting using these three methods. The testing results show that the time series methods are best suited for predicting monthly average output power, with the ARIMAX outperforming all other methods when applied to all cities/regencies in Central Java. It achieved the RMSE values of 10.74, 25.36, and 60.27 for daily, weekly, and monthly forecasting, respectively.**

Keywords—solar PV, time series, renewable energy, forecasting, output power

I. INTRODUCTION

Solar photovoltaics (PV) is a green technology with a strategic role to aid in the energy transition [1]. While there are challenges in maximising the theoretical potential of solar PV in a specific area due to uncontrollable factors such as solar irradiation, temperature, humidity, and other weather-related parameters [2], the benefits of solar PV output forecasting are inevitable for both the macro and micro perspectives. From the macro perspective, forecasting PV output allows stakeholders to better understand a country's needs and energy potential, especially where solar PV can be used more effectively [3].

Solar PV output forecasting can be divided into two categories: short-term for operational needs and long-term for investment and planning decision-making. Short-term forecasting is useful for system operators in managing PV output fluctuations and estimating peak energy requirements during specific hours up to a few days and weeks. Long-term annual forecasting can assist stakeholders in better planning and anticipating the electricity industry generation mix future.

Solar energy has enormous potential for electricity generation, particularly in many tropical developing countries. However, progress on PV capacity deployment in some of these countries and jurisdictions, including Indonesia, remains slow [4]. Aside from the fact that regulations related to the use of solar PV have frequently changed and been uncertain in Indonesia, there is still a lack of use of solar historical data for operational or planning purposes, as well as studies regarding accurate estimates of PV output power based on long-term historical data, and across the spatial coverage of country or provinces.

Developing an accurate forecasting model in terms of long-term solar PV forecasting is especially difficult [5]. This difficulty is due in part to the uncertainty of future solar output power attributes such as irradiance and ambient temperature. Despite this, several studies have been conducted to forecast solar PV output power or solar radiation.

A study evaluated the Seasonal Auto-Regressive Integrated Moving Average Exogenous (SARIMAX) model for forecasting the PV output of a city in the Philippines [6]. The authors compared each season and one full year of forecasting to evaluate the SARIMAX model's performance and identified significant input parameters for each season. Another forecasting study used Seasonal Auto-Regressive Integrated Moving Average (SARIMA) and Auto-Regressive Integrated Moving Average (ARIMA) models to represent monthly and daily solar radiation, respectively, in Seoul, Korea [7]. The authors performed future trends of monthly solar irradiation based on 37-year data.

Another study has compared SARIMA, SARIMAX, modified SARIMA, and an artificial neural network method for grid-connected PV generation output forecasting [8]. The comparison concludes the necessity and benefits of using exogenous factors in a time series model. Other studies compared several time series methods including artificial intelligence algorithms to forecast PV output power in a city in South Korea using 4.5-year operation data on an hourly basis from a 1.5 MW grid-connected PV power plant [9].

While previous research on solar PV output forecasting using time series models provided planners and policymakers with useful insights into the potential impact of possible future PV output power on, for example, system security and possible generation mixes, there is no one-size-fits-all for different goals and forecasting problems.

This study therefore evaluates the application of time series-based models of Auto-Regressive Integrated Moving Average Exogenous (ARIMAX) and compares the performance with ARIMA and SARIMA in forecasting PV output power based on the long-term hourly temporal-based solar PV model applied for the Java-Bali, Indonesia. Moreover, this study creates a web-based dashboard application using the Dash framework and Python that allows users to explore and analyse the forecasting results for all cities/regencies in the Java-Bali region. This study aims to compare the forecasting performance of the three models in terms of daily, weekly, and monthly average PV output power as well as the best-suited forecasting period.

The rest of this paper is structured as follows. Section 2 presents the methodology, followed by a brief description of the system and implementation in Section 3. Section 4 presents results and discussions, and finally, Section 5 concludes this paper.

II. METHODOLOGY

A. The Dataset

This study considers 2005–2016 long-term hourly-temporal data of modelled solar PV output power across the gridded latitude and longitude of 0.05° x 0.05° or 5 km x 5 km across the Java-Bali region, Indonesia, as obtained from the Renewables.ninja (RN) solar PV model web-based tool introduced in [10]. Considering a 1 MW peak PV capacity, the tool calculated PV output power for all locations while accounting for the effects of direct and diffuse irradiations, and ambient temperature. The main processes in this study are data preprocessing, modelling, and forecasting, and evaluation and visualisation.

B. Data Preprocessing

The first goal of conducting the data preprocessing is to obtain the area of all cities/regencies, as well as all data points within these boundaries. This study iterates all locations by conducting API requests at OpenRoute service API under the Reverse Geocoding. The required parameters include API key, and latitude and longitude. This procedure includes the removal of all data points that do not belong to any city/regency, such as those in the sea. This process ends up with 4,150 data points representing 4,150 locations.

Subsequently, this study assigns one location or data point for every city/regency across the Java-Bali region. The assigned location is selected based on the highest capacity factor (CF) among all locations in each city/regency in each year. The CF is a ratio of the electrical energy produced by a PV capacity to the electrical energy that could have been produced in one year. Table 1 shows an example of data preprocessing results from 2005 to 2014 for Bogor Regency in West Java, highlighting one location with the highest yearly average CF for the corresponding year.

TABLE I. EXAMPLE OF DATA PREPROCESSING RESULTS FOR BOGOR, WEST JAVA

Date	Lat Lon	Regency	Province	Output Power	Year	CF
2005-01-01, 11:00	-6.75 106.9	Bogor	West Java	626.566	2005	16.19
2005-01-02, 11:00	-6.75 106.9	Bogor	West Java	642.625	2005	16.19
2005-01-03, 12:00	-6.75 106.9	Bogor	West Java	500.715	2005	16.19
2005-01-04, 11:00	-6.75 106.9	Bogor	West Java	487.778	2005	16.19
2005-01-05, 12:00	-6.75 106.9	Bogor	West Java	369.951	2005	16.11
....	….	….	….	….	….	….
2014-12-27, 12:00	-6.6 107.2	Bogor	West Java	463.848	2014	16.54
2014-12-27, 12:00	-6.6 107.2	Bogor	West Java	494.045	2014	16.54
2014-12-27, 11:00	-6.6 107.2	Bogor	West Java	627.448	2014	16.54
2014-12-27, 12:00	-6.6 107.2	Bogor	West Java	686.127	2014	16.54
2014-12-27, 10:00	-6.6 107.2	Bogor	West Java	485.593	2014	16.54

C. Modelling, Forecasting, and Evaluation

Fig. 1 shows the modelling flowchart to determine the best parameters for all the time series models and forecasting frequency using the grid search method.

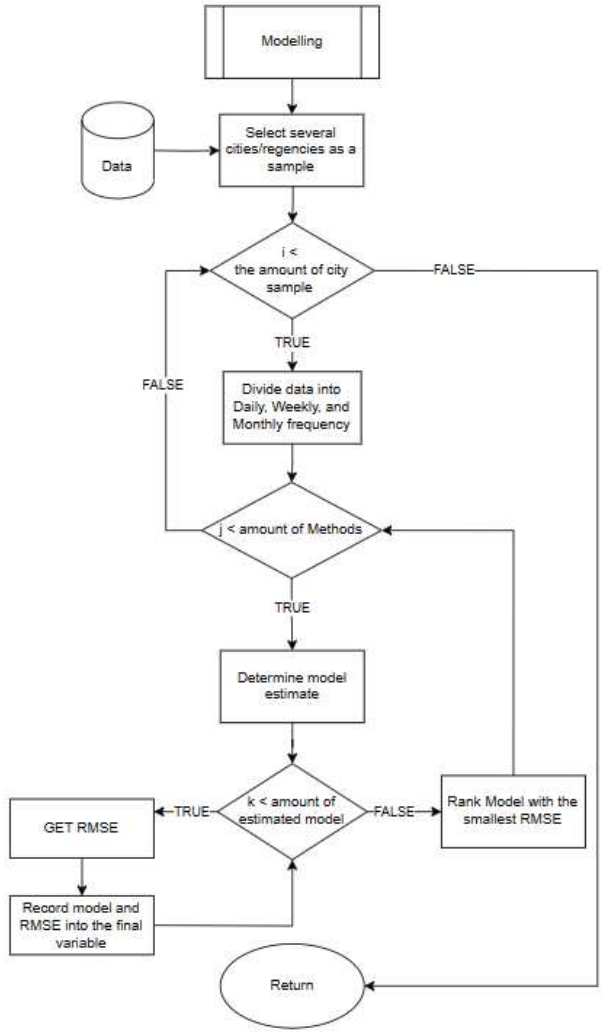

Fig. 1. Modelling flowchart.

This modelling takes one city/regency from each province in the Java-Bali region as a sample. In comparison to other cities/regencies in that province, each selected city/regency has the highest average CF. Data on daily output power, like those shown in Table 1, are further grouped into weekly and monthly average PV output power for this regency. The process for the selected sample of city/regency continues by conducting a grid search method for ARIMAX, SARIMA, and ARIMA with respect to daily, weekly, and monthly output power data to obtain the best parameters for each model, i.e., the model with the lowest RMSE.

The parameters used in ARIMAX are p (autoregressive), d (integrated), q (moving average), and exogenous variable. Meanwhile, the parameters used in SARIMA are p (autoregressive), d (order of differencing), q (moving average), P (seasonal autoregressive), D (order of seasonal differencing), Q (seasonal moving average), S (seasonal period length), and for ARIMA are p (autoregressive), d (integrated), and q (moving average).

Subsequently, the forecasting parameters obtained for the sample city/regency can be used to forecast solar PV output power along with the desired forecasting periods for other cities/regencies within the same province as the sample city/regency. Alternatively, users can also proceed directly with the modelling and forecasting for any city. This study applies RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) to measure the accuracy of prediction results. The correlation between variables is determined using the coefficient of determination R-squared (R^2).

D. Visualisation

The Solar PV Data Visualization dashboard is created using the web-based Dash framework. It has four main pages, namely: 'about' page, 'explanatory' page, 'time series analysis' page, and 'prediction' page.

Every page designed in this study has its own items and functionalities. The explanatory page contains the general overview of the data along with related items, including the data frequency. The time series analysis page contains a time series analysis of selected data, displays patterns of data movement, and stationary tests, including the autocorrelation and partial autocorrelation function plot during the selected period. Meanwhile, the following lists are available items and functionality on the prediction page.

- Button–Info: Button to display the description of the time series analysis page and available items.
- Slider–Range Year: Provides 10-year selection options from the dataset that filter the forecasting results that will be displayed.
- Dropdown–Select Regency: Provides a selection of 108 cities/regencies for forecasting.
- Dropdown–Select Attribute: Provides 4 attributes to be selected for predictions, i.e., direct irradiation, diffuse irradiation, temperature, and PV output power.
- Dropdown–Select Frequency: Provides a selection of daily, weekly, and monthly forecasting periods.
- Dropdown–Select Method: Provides a selection of time series models that can be selected to conduct forecasting.
- Button–Predict: Button for conducting forecasting.
- Line Chart–Actual vs Predicted Data: Displays forecasted data and actual data with a trend line chart.
- Table–Evaluation Metrics: Displays forecasting evaluation results based on MAE, MSE, RMSE, and R^2.
- Card–Average Predicted: Displays the average value of forecasted data.
- Card–Average Real Data: Displays the average value of actual data.
- Table–Location and Capacity Factor per Year: Displays list of points of location used every year for the selected city/regency with the CF.
- Table–Location per Year: Displays points of locations used each year for a selected city/regency in a map form.

Fig. 2 depicts the visualisation design of the prediction page on the Solar PV Data Visualization dashboard.

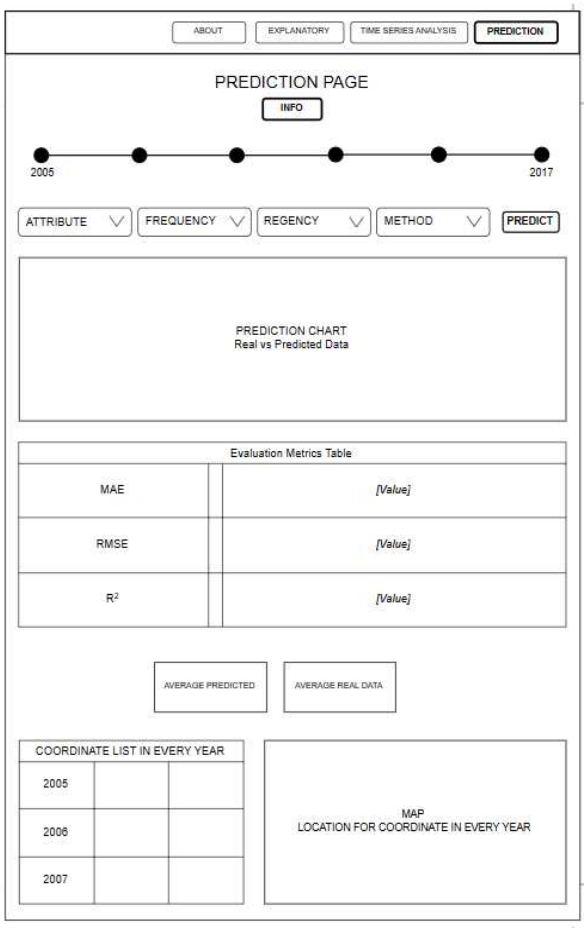

Fig. 2. Visualisation design of the prediction page.

III. System Implementation

The forecasting system is completely built with Python 3.10.9 and Microsoft Visual Studio Code as the software code editor. The Dash framework is used to create a website-based dashboard application. It enables the development of web-based applications entirely in Python. This study also uses the Bootstrap framework to enhance the user interface. Several

Python libraries are used in the programming codes, including scikit-learn, pandas, matplotlib, seaborn, plotly dash, and statmodels. Scikit-learn provides functions for data preprocessing, such as normalisation, category encoding, and data grouping, other than functions for dividing data into training and testing sets, as well as metrics for evaluating results. Pandas provides functions for efficient data manipulation and analysis using a dataframe, which aids in data preprocessing such as cleaning, sorting, merging, and handling duplicate data. Pandas also supports reading and writing data from a variety of file formats, including CSV, Excel, and JSON. It can also be combined with Matplotlib, a visualisation library that generates graphs to aid in hyperparameter tuning, and with Seaborn, which

offers the ability to view the relationship between attributes. Plotly dash library offers the ability to generate visually appealing graphs and data visualisations. Statmodels provides a time series algorithm and functions for analysing time series data, such as determining data stationarity.

The system implementation uses the following hardware configuration for coding and testing the application: Processor: Intel(R) Core (TM) i7-8750H CPU @ 2.20GHz; Memory: 24 GB; Graphic Card: Nvidia GeForce GTX 1050 Ti-4GB; Hard Disk: 1TB; Solid-state Drive: 128GB; Operating system: Windows 10, 64-bit operating system, x64-based processor.

IV. Results and Discussions

The search for the best parameters for ARIMAX models for daily, weekly, and monthly forecasting periods while considering *p* (autoregressive), *d* (integrated), *q* (moving average), and one exogenous variable as parameters. Direct irradiation is chosen as the exogenous variable for ARIMAX modelling because it has the highest correlation with PV output power, i.e., 0.94, compared to other solar attributes.

Applying the hyperparameter tuning and model fitting to Wonogiri Regency in Central Java, Table 2 and Table 3 summarises the results of the best parameters and evaluation metrics obtained from the hyperparameter tuning for the ARIMAX models with direct irradiation as the exogenous variable and ARIMA models, respectively.

TABLE II. Summary of The Results of The Best Parameters Obtained from The Hyperparameter Tuning for ARIMAX (with Direct) Models in Wonogiri Regency

Data Frequency Period	Parameter (*p, d, q*)	Error Evaluation		
		RMSE	MAE	R^2
Daily	(18, 0, 20)	61.12	6.29	0.79
Weekly	(15, 0, 16)	22.97	3.84	0.86
Monthly	(12, 0, 12)	9.21	2.52	0.93

TABLE III. Summary of The Results of The Best Parameters Obtained from The Hyperparameter Tuning for ARIMA Models in Wonogiri Regency

Data Frequency Period	Parameter (*p, d, q*)	Error Evaluation		
		RMSE	MAE	R^2
Daily	(20, 0, 19)	106.85	9.07	0.38
Weekly	(15, 0, 14)	51.34	6.32	0.31
Monthly	(12, 0, 1)	23.17	4.43	0.41

As in Table 2, given the monthly forecasting period as an example, the model employs 12 autoregressive lags, as indicated by the *p* parameter value. The model considers the value of the current data's linear relationship with the values in the previous 12 periods or the previous 12 months. This model performs no differencing, as evidenced by the parameter value *d*, which is 0. Because the data used is stationary, there is no need to repeat the differencing process. The *q* parameter value indicates that this model employs a 12-lag moving average. To forecast the current value, the model considers the influence of 12 residual values or previous forecasting errors. A similar analysis of the model description can be carried out for weekly and daily forecasting periods in the ARIMAX modelling, as well as in the ARIMA modelling.

Table 4 summarises the best parameters and evaluation metrics obtained from the hyperparameter tuning for the SARIMA models considering daily, weekly, and monthly forecasting. The monthly forecasting period provides the SARIMA with the model's seasonal component of *S, P, D,* and *Q*. The *S* parameter represents the seasonal period in the time series data used. The seasonal pattern therefore can be interpreted as repeating itself every 12 months. The *P* value indicates that the model considers the impact of values from the previous 3 seasons. The *D* parameter indicates that no seasonal differencing was performed because the data used is seasonally stationary. From the *Q* value, it is revealed that the model considers the influence of 7 previous seasonal residual values when forecasting the current value.

TABLE IV. Summary of The Results of The Best Parameters Obtained from The Hyperparameter Tuning for SARIMA Models in Wonogiri Regency

Data Frequency Period	Parameter (*p, d, q*) x (P, D, Q, S)	Error Evaluation		
		RMSE	MAE	R^2
Daily	(17, 0, 19) x (2, 0, 2, 20)	106.80	81.95	0.38
Weekly	(15, 0, 15) x (0, 0, 2, 20)	50.97	39.73	0.32
Monthly	(11, 0, 10) x (3, 0, 6, 12)	24.17	18.63	0.49

Fig. 3 shows the PV monthly average output power versus the monthly forecasted output power plot of an ARIMAX model (12, 0, 12) for Wonogiri Regency, as presented in Table 2. Meanwhile, Fig. 4 shows the PV weekly average output power versus the weekly forecasted output plot of an ARIMAX model (15, 0, 16) from Table 2.

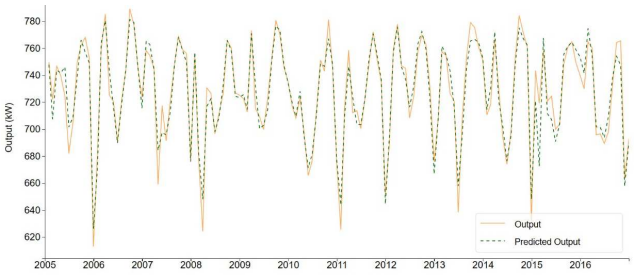

Fig. 3. PV monthly average output vs forecasted results of ARIMAX model (12, 0, 12) for Wonogiri Regency

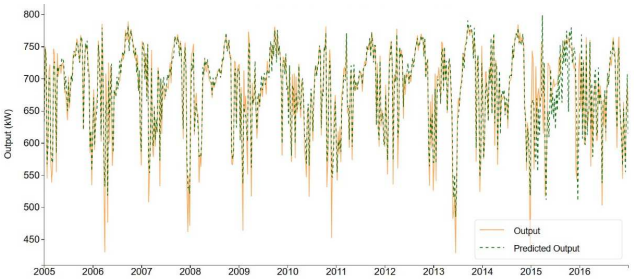

Fig. 4. PV weekly average output vs forecasted results of ARIMAX model (15, 0, 16) for Wonogiri Regency.

From Table 2 to Table 4, it can be seen that the ARIMAX outperform the ARIMA and SARIMA across all forecasting periods. This is evidenced by ARIMAX producing the lowest RMSE and MAE than the other two, as well as R^2 values that are closer to one, particularly for the monthly forecasting.

Based on forecasting results for all cities/regencies in Central Java, the ARIMAX model has shown the best performance for all forecasting periods. The model has an average RMSE of 10.74, an average MAE of 7.21, and an average R^2 of 0.9 for the monthly forecasting period.

Table 5 presents the results of monthly forecasting for other sample cities across the Java-Bali region using ARIMAX. All models obtain considerably low RMSE like that obtained in Table 2 for monthly forecasting. The results indicate that the monthly forecasting using the ARIMAX model is the best among other time series models. Meanwhile, Fig. 5 shows the visualisation of the prediction page, as described in Section 2, for Badung Regency's ARIMAX (10, 0, 12) monthly forecasting from 2005–2016.

TABLE V. RESULTS OF ARIMAX MODELS OF MONTHLY FORECASTING IN OTHER SAMPLE CITIES IN THE JAVA-BALI REGION

Province	Sample City/Regency	Parameter	RMSE
Jakarta	East Jakarta	(9, 0, 12)	10.15
West Java	Garut	(11, 0, 11)	10.82
Banten	Lebak	(8, 0, 12)	8.99
Yogyakarta	Gunung Kidul	(11, 0, 12)	12.01
East Java	Banyuwangi	(12, 0, 12)	10.65
Bali	Badung	(10, 0, 12)	6.94

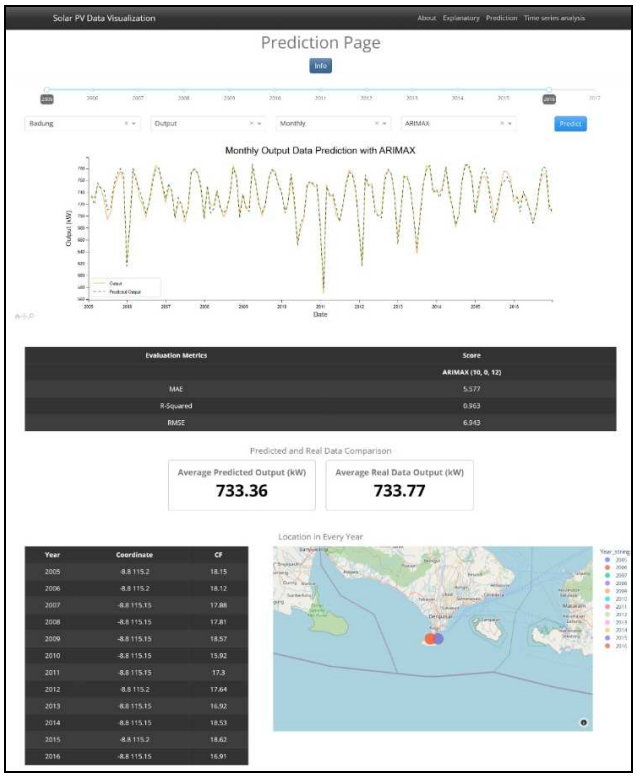

Fig. 5. Visualisation of the prediction page for Badung Regency's ARIMAX (10, 0, 12) monthly forecasting from 2005–2016.

V. CONCLUSIONS

This paper presents the use of time series methods to forecast daily to monthly average values of solar PV output power models and predictions based on the 2005–2016 long-term hourly temporal-based PV output model obtained from Renewables.ninja. Given the Java-Bali region as a case study, it is revealed that ARIMAX outperforms ARIMA and SARIMA in all forecasting windows. Furthermore, given the lowest RMSE and MAE and the highest R-squared values, the ARIMAX model is best suited for predicting monthly average data. While this study considers the Java-Bali region of Indonesia as a case study, the methodology presented in this paper has a broader relevance to stakeholders conducting similar studies in other countries or jurisdictions. Future work could improve the dashboard and expand its capability by including the demand dataset as well as the use of machine learning or deep learning algorithms to enhance the analysis and improve results.

ACKNOWLEDGMENT

This paper is part of the research funded by The Ministry of Education, Culture, Research, and Technology of The Republic of Indonesia through the Fundamental Research Scheme 2023, contract: 002/SP2H/PT/LL7/2023, 11/SP2H/PT/LPPM-UKP/2023.

REFERENCES

[1] C. Breyer, D. Bogdanov, A. Gulagi, A. Aghahosseini, L.S.N.S. Barbosa, O. Koskinen, M. Barasa, U. Caldera, S. Afanasyeva, M. Child, J. Farfan, and P. Vainikka, "On the role of solar photovoltaics in global energy transition scenarios,"Progress in Photovoltaics, vol. 25, issue 8, pp. 727-745, March 2017.

[2] S.C. Lim, J.H. Huh, S.H. Hong, C.Y. Park and J.C. Kim, "Solar power forecasting using CNN-LSTM hybrid model," Energies, vol. 15, issue 21, pp. 8233, November 2022.

[3] K.J. Iheanetu, "Solar photovoltaic power forecasting: A review," Sustainability, vol. 14, issue 24, pp. 17005, December 2022.

[4] J. Langer, J. Quist, and K. Blok, "Review of renewable energy potentials in Indonesia and their contribution to a 100% renewable electricity system," Energies, vol. 14, issues 21, pp. 7033, October 2021.

[5] P. Gupta and R. Singh, "Forecasting hourly day-ahead solar photovoltaic power generation by assembling a new adaptive multivariate data analysis with a long short-term memory network," Sustainable Energy, Grids, and Networks, vol. 35, pp. 101133, September 2023.

[6] I. Benitez, L. Gerna, J. Ibañez, J. Principe and F.D.L. Reyes, "Use of SARIMAX model for solar PV power output forecasting in Baguio City, Philippines," in *2022 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE)*, Pattaya, Thailand, IEEE Press Piscataway, November 26-28 2022.

[7] M. Alsharif, M. Younes and J. Kim, "Time series Arima model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea, Symmetry, vol. 11, issue 2, pp. 240, February 2019.

[8] S.I. Vagropoulos, G.I. Chouliaras, E.G. Kardakos, C.K. Simoglou and A.G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," in *2016 IEEE International Energy Conference (ENERGYCON)*, Leuven, Belgium, IEEE Press Piscataway, April 4–8 2016.

[9] E.G. Kim, M.S. Akhtar and O.B. Yang, "Designing solar power generation output forecasting methods using time series algorithms," Electric Power System Research, vol. 216, pp. 109073, March 2023.

[10] S. Pfenninger and I. Staffell, "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data," Energy, vol. 114, pp. 1251–1265, September 2016.