Electric Power Systems Research 216 (2023) 109073

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Designing solar power generation output forecasting methods using time series algorithms

EunGyeong Kim [a], M. Shaheer Akhtar [a,*], O-Bong Yang [a,b,**]

[a] Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju, 54896, Republic of Korea
[b] School of Semiconductor and Chemical Engineering, Jeonbuk National University, Jeonju, 54896, Republic of Korea

ARTICLE INFO

Keywords:
Photovoltaic power output
Forecasting methods
Machine learning
Deep learning
Climate change
Time series algorithm

ABSTRACT

The present photovoltaic (PV) power generation systems are globally facing the irregularity problem in the distribution of PV generation. In particular, the exact PV power forecasting is critical for grid-connected photovoltaic (PV) systems under unwanted changes in environmental circumstances. The grid energy management, grid operation and scheduling are important factors to forecast the PV power output. Time series analysis is one of the most important aspects of PV output prediction, especially in places (in South Korea) where past solar radiation data or other weather parameters have not been recorded. In this paper, a variety of time-series methods including deep-learning algorithm and machine learning algorithms was used to predict the PV power generation output for quick respond to equipment and panel defects. For designing AI models, the input data were characterized by dividing seasons and choosing the multiple parameters from seasons. In this study, the photovoltaic power generation data was collected from Ansan city, South Korea during January 2017 to June 2021 and the weather data was collected from Suwon city, South Korea during January 2017 to June 2021. In this work, approx. 40,000 hours of operation data from 1.5 MW grid-connected PV system in South Korea was used. PV power generation forecasting was carried out on an hourly basis to test efficacy of various models. Among all models (Holt-Winters, Multivariate Linear Regression, ARIMA, SARIMA, ARIMAX, SARIMAX), LSTM model presented the lowest error rate as compared to other models for quick PV power generation forecasting.

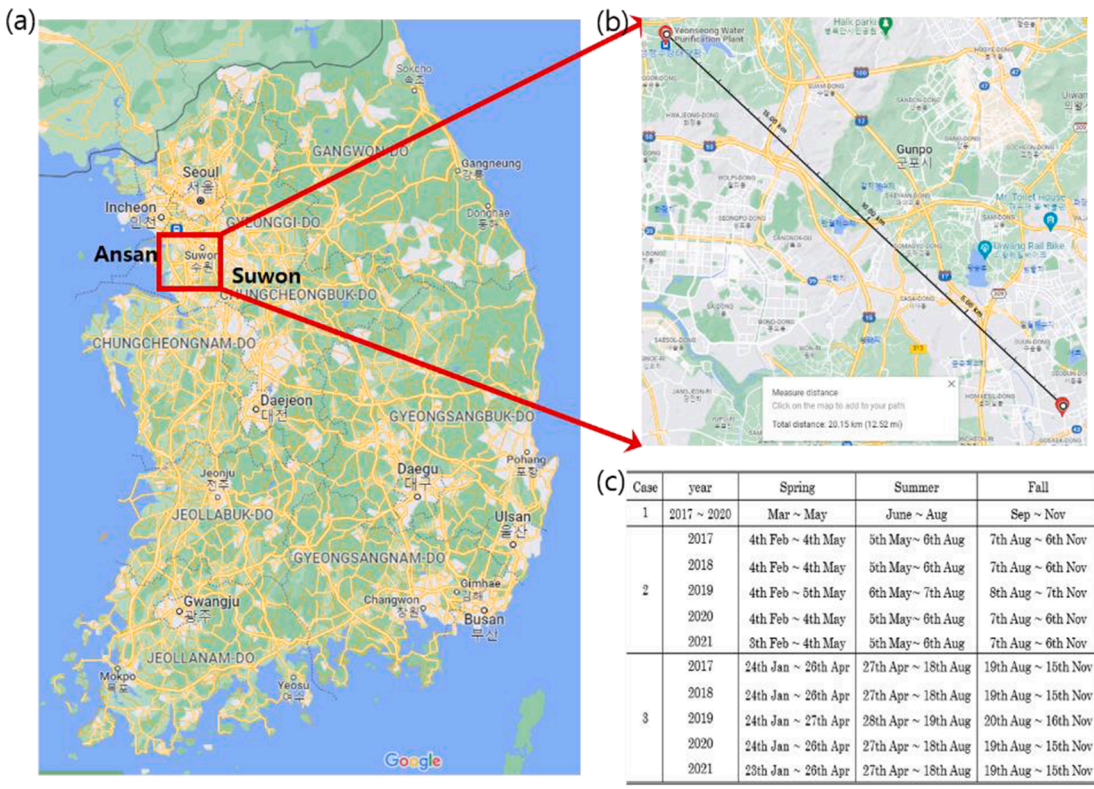
1. Introduction

Past few years, the climate change, global warming, and rising energy demands have prompted the Korean government to look for sustainable energy sources that are both economically and environmentally viable. Excessive industrialization and urbanization increase the power demand, so the power system must become more efficient and stable. When the supply of electricity exceeds the demand, surplus electricity is generated, which is uneconomic. Therefore, the insufficient power supply and management, resulting in power failure, which may lead the blackout. However, due to the irregular power generation problem of photovoltaic systems, there is a problem that engineers cannot quickly respond to equipment and panel defects [1]. Few studies have been done to PV power generation forecasting by Non-linear Autoregressive Exogenous Neural Network [2], LSTM-ARMA [3] and ARIMA [4] models. Therefore, predicting the amount of PV power generation in advance and stabilizing the power supply increases the efficiency of PV power generation.

The present PV power generation systems still shown numerous faults and dependencies which normally come from solar irradiance. The electrical power generated is influenced by a number of factors including the quality of the PV cells, the type of solar cells used, the electrical circuit of the module, the angle of incidence, weather conditions, and other parameters. Mainly the temperature of the solar cell in a PV system affects the amount of power produced [5]. Some reports are available to define PV output power forecasting based on the weather classification [6–9]. It is essential to predict PV power output in order to quickly respond to panel and equipment defects. It is demanded to develop model for improving the PV power generation using the artificial intelligence (AI) including machine learning, deep learning etc. Lee et al analyzed the high PV power generation forecasting model by considering meteorological factors [10]. Lee's group selected the factors such as date and time, temperature, wind speed, wind direction, humidity, total cloud cover, and solar insolation for forecasting the PV

* Corresponding author.
** Corresponding author at: Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju, 54896, Republic of Korea.
E-mail addresses: shaheerakhtar@jbnu.ac.kr (M.S. Akhtar), obyang@jbnu.ac.kr (O.-B. Yang).

https://doi.org/10.1016/j.epsr.2022.109073
Received 19 July 2022; Received in revised form 10 November 2022; Accepted 7 December 2022
Available online 18 December 2022



(c)

Case	year	Spring	Summer	Fall	Winter
1	2017 ~ 2020	Mar ~ May	June ~ Aug	Sep ~ Nov	Dec~ Feb
2	2017	4th Feb ~ 4th May	5th May~ 6th Aug	7th Aug ~ 6th Nov	7th Nov ~ 4th Feb
	2018	4th Feb ~ 4th May	5th May~ 6th Aug	7th Aug ~ 6th Nov	7th Nov ~ 4th Feb
	2019	4th Feb ~ 5th May	6th May~ 7th Aug	8th Aug ~ 7th Nov	8th Nov ~ 4th Feb
	2020	4th Feb ~ 4th May	5th May~ 6th Aug	7th Aug ~ 6th Nov	7th Nov ~ 4th Feb
	2021	3th Feb ~ 4th May	5th May~ 6th Aug	7th Aug ~ 6th Nov	7th Nov ~ 3th Feb
3	2017	24th Jan ~ 26th Apr	27th Apr ~ 18th Aug	19th Aug ~ 15th Nov	16th Nov ~ 23th Jan
	2018	24th Jan ~ 26th Apr	27th Apr ~ 18th Aug	19th Aug ~ 15th Nov	16th Nov ~ 23th Jan
	2019	24th Jan ~ 27th Apr	28th Apr ~ 19th Aug	20th Aug ~ 16th Nov	17th Nov 23th Jan
	2020	24th Jan ~ 26th Apr	27th Apr ~ 18th Aug	19th Aug ~ 15th Nov	16th Nov ~ 23th Jan
	2021	23th Jan ~ 26th Apr	27th Apr ~ 18th Aug	19th Aug ~ 15th Nov	16th Nov ~ 22th Jan

Fig. 1. (a) photograph of South Korea map and (b) selected sites where PV power generation data collected. (c) Summary of data selection in different seasons.

power generation and meteorological variables. Mellit et al. utilized more than a year of data to anticipate the electricity produced by a 50 W PV plant using two artificial neural networks [11]. Most of models are focused on the short-term (three-days-ahead) forecasting [12].

For PV power generation forecasting, the common AI models are statistical models (ARIMA [4], hybrid model [5], holt winter [13], ARMA [14] and artificial neural network (ANN) [15,16]) and deep learning models (Non-linear Autoregressive Exogenous Neural Network [2], LSTM-ARMA [3], some Regression Neural Network [17,18], SARIMAX [18], SARIMA, modified SARIMA and ANN [19]). Several mathematical equations in statistical models are used to extract patterns from input data. In general, statistical techniques can be divided two groups: times-series and machine learning (ML) based models [20]. Time-series models such as ARMA [14–16,18], ARIMA [4,18,21], SARIMA [19,22] and holt-winters model [13,20] are popularly used because of their exponential smoothing. Nowadays, the machine learning and deep learning models are widely used to predict PV power output forecasting [3,23–26]. To overcome the aforementioned obstacles, fresh and sophisticated procedures must be used to achieve legitimate and reliable results. Several researchers have reported time series models for PV power generation forecasting using seasons such as 4 seasons [2,27–29], and the sunny day, cloud day and rainy day [7,10,17,30,31]. Last few years, deep learning (DL) model driven approaches have been experienced a great deal of interest in the time-series PV generation prediction fields because DL can enable to interpretate highly noisy and incorporate the irrelevant datasets [32]. Among several DL based model, long short-term memory (LSTM) approach has been used for the PV power prediction as LSTM delivers excellent result accuracy with time-series datasets due to remembering the previous information via a complex memory [33]. To advance the PV power forecast accuracy, the data was classified into four sub datasets based on the annual seasons.

In this work, a novel PV power generation forecast model using time series algorithms is developed by (i) six statistical and (ii) one deep learning time series models with non-stationary data. Herein, the main reason is to choose above time series models because they are applicable for PV management considering period, trend, and seasonality factors [28,34]. In order to make database in this work, we divide the data using four seasons (spring, summer, fall and winter) collected data from two cities of Korea rather than a method using a specific weather. In addition, the seasons are further divided into three methods and conduct the forecasting by seven AI models.

2. Methodology

2.1. Data description

In this work, we have chosen the data from PV power generation plant located in South Korea, as shown in Fig. 1. All PV power generation data have been collected from Ansan city, and the weather data was collected from Suwon city, South Korea during January 2017 to June 2021. In support, the Korea solar irradiance maps including provinces and cities have been studied by several groups [35,36]. These studies stated that the distribution of solar irradiance in terms of hours is similar all over Korea peninsula, except Jeju Island. Thus, the collecting PV data from Yeonseong Water Purification Plant Solar Power in Ansan city is used because this PV plant has the largest PV data provider among cities by public data portal (Government open portal systems managed by Ministry of the Interior). PV data from Ansan city and PV power forecasting can be suitable choice and would be useful for other cities or regions of Korea. Furthermore, the collected PV data on various variables (temperature, sunlight, insolation, clouds (total cloudiness, low-middle cloudiness, minimum cloud height), fine dust ([Particulate Matter (PM)] or PM10), precipitation, snow cover)) were used and analyzed in 3 methods (month, seasons, revised season affected by global warming). The seasons have divided into one dataset, the model's trend, seasonal, cycle change. In Fig. 1 (a)~(b), the PV power generation output data collected from Yeonseong Water Purification Plant Solar Power, Ansan city which have a capacity of 1.49 MW and consisted of a total 40,898 data.

For the weather data, there was no meteorological data provided by

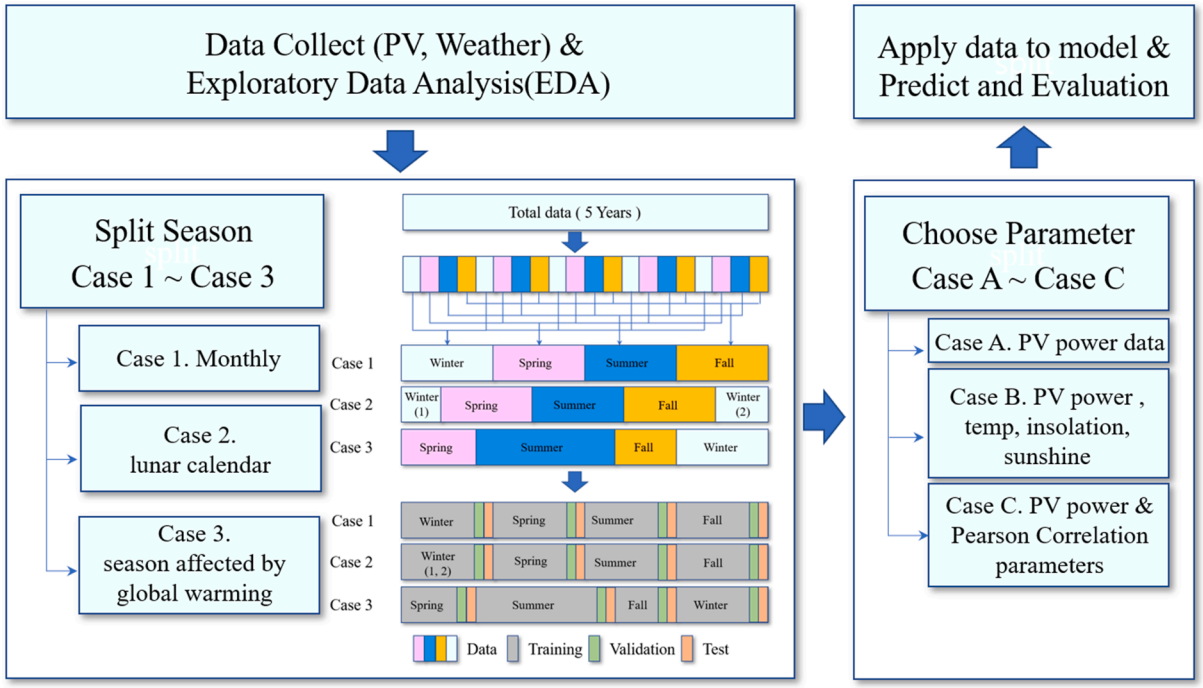


Fig. 2. Schematic of the time series procedure. [All collected data were taken from government open portal systems managed by Ministry of the Interior and safety, Korea meteorological administration (KMA), and Korea Environment Corporation]

Yeonseong Water Purification Plant Solar Power, Ansan city. We have chosen the weather data from Suwon city, which is just 20.15 km away in a straight line to Ansan city. From Suwon city, the collected 408,980 weather data were used on temperature, sunlight, insolation, clouds (total cloudiness, low-middle clouds, minimum cloud height), fine dust (PM10), snow, and rain. Before performing the model training, the collected season data (in Fig. 1. (c)) was divided into three cases. Case 1 was referred to seasons by months (like spring, summer, fall, and winter), Case 2 addressed the solar term-based subdivision of seasons (according to lunar calendar), and Case 3 described the season affected by global warming. The selection of Case 3 was based on the overall temperature increase due to global warming. According to weather forecasting department, due to global warming, the significant rise in temperature was recorded in Korean weather for examples, more than 5°C temperature rise in spring, more than 20°C rise in summer, in fall the average daily temperature fell below 20°C and temperature normal in winter. All collected data were taken from government open portal systems managed by Ministry of the Interior and safety, Korea meteorological administration (KMA), and Korea Environment Corporation.

The test dataset was first omitted the outliers for cleaning the unwanted values and thereafter the missing values were filled by using Python's fillna function. Afterward, the test dataset was normalized using Min-Max normalization method. Min-Max normalization is a preprocessing method for data scaling which is usually ranging from 0 to 1 [37]. The following expression (1) for Min-Max normalization is used for dataset normalization:

$$data = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

The obtained dataset is then trained, and the error rates are compared to propose a variable and season segmentation method for efficient model construction. In this work, the models such as Holt-winter, ARIMA, SARIMA, Multivariable Linear Regression models are programed by RStudio whereas LSTM model was programed by Python TensorFlow library.

2.2. Time series data procedure

Fig. 2 describes the entire time series modeling for PV power generation forecasting. For each modelling, 80% of the total data was used as training data and 18% was used as validation data. Afterward, only 2% test data was used to predict 5 days. Each dataset (Case 1-3) was resampled using the seasons for models. To predict highly accurate model, we prepare three cases based on several parameters. Only PV data is considered as Case A, and Case B include the PV data, temperature, insolation, sunshine and Case C describe the PV data with Pearson correlation parameters. For example, the multiple variables like temperature, fine dust, clouds, were used to models such as ARIMAX, Multivariate Linear Regression, ARIMAX, SAIRMAX, LTSM. By reflecting these seasonal characteristics, the variables to be used for each season were compared with the Pearson correlation coefficient for comparing the linear relationship and afterward applied to the model.

2.3. Pearson correlation coefficient (PCC)

The independent variable PV data are calculated by PCC to select the dataset of Case C. PCC generally calculated from the corr function of the Pandas library of Python. PCC is calculated between PV variable and other weather variables [38]. The following expression (2) is used for PCC:

$$\rho(a,\ b) = \frac{E(ab)}{\sigma_a \sigma_b} \tag{2}$$

Where $E(ab)$ is the cross-correlation between a and b, and $\sigma_a^2 = E(a^2)$ and $\sigma_b^2 = E(b^2)$ are the a and b variances, respectively [38]. PCC is usually ranging from -1 to 1. Positive PCC values imply positive correlation, negative values imply negative correlation. The obtained PCC results are displayed in Figs. 3–5 in terms of Case 1, Case 2 and Case 3. When comparing the strength of PCC, the absolute value is written first. It is known that the $\geq$0.7, $\geq$0.3 and $\leq$0.1 are referred to strong PCC, medium PCC and weak PCC, respectively. From Figs. 3–5, variables such as temperature, sunshine, insolation, cloud level, cloud amount, and cloud

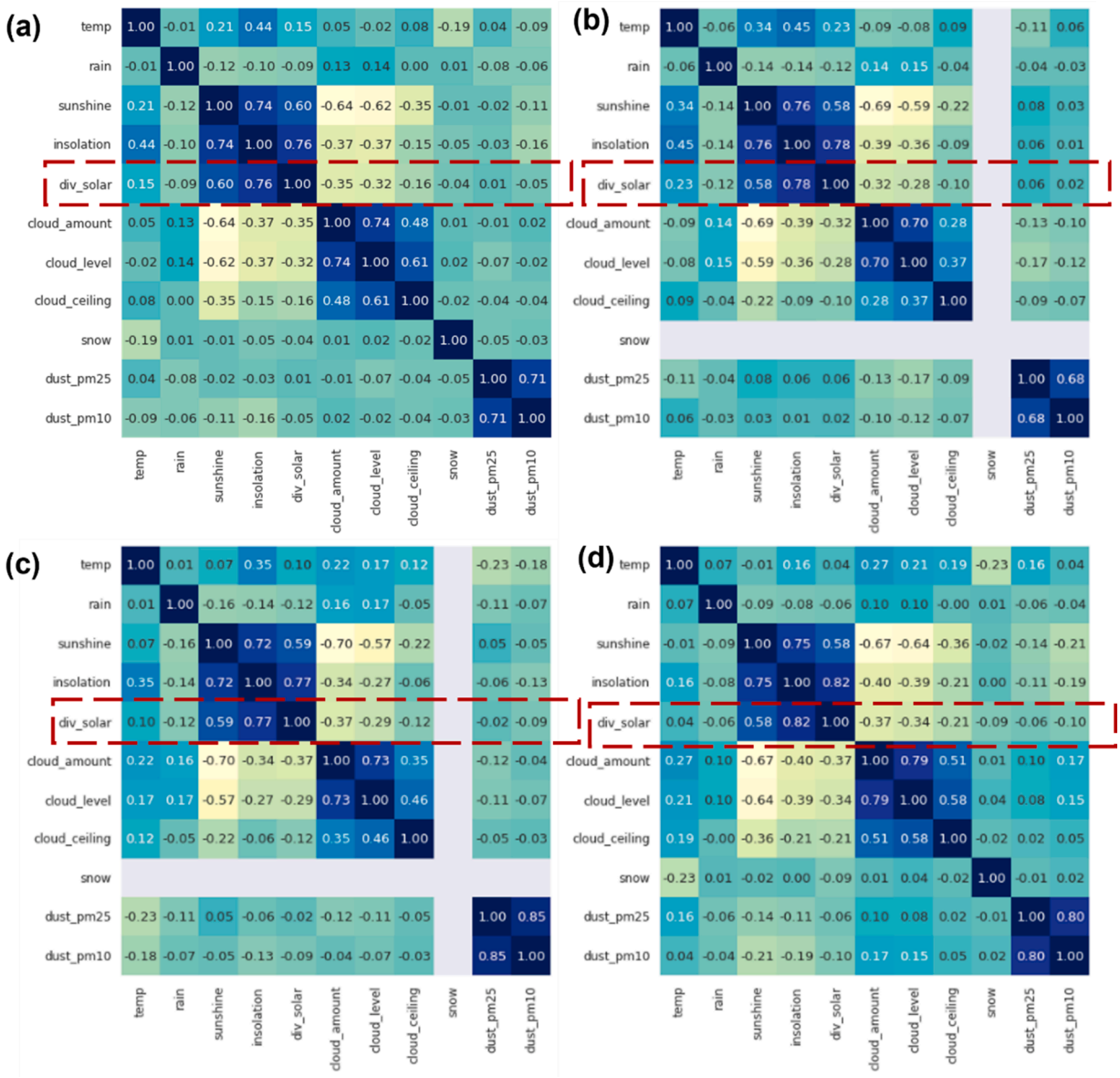


Fig. 3. Pearson correlation coefficients for Case 1 in different seasons as (a) spring, (b) summer, (c) fall and (d) winter.

ceiling are shown over 0.1 PCC values in spring and winter seasons. It is seen in Case 2 and Case 3 that the independent variables such as dust PM10 having the PCC values of 0.14. Based on PCC values, several variables are characterized into Case 1, Case 2, and Case 3, as shown in Table 1. Therefore, PCC analysis is helpful to choose the independent variables on PV power generation output.

2.4. Performance

In the performance evaluation of the model, the model has evaluated using the mean absolute error (MAE) [39], mean squared error (MSE) [40], root mean squared error (RMSE) [39,40] and maximum mean absolute percentage error (mMAPE) [41] values. MAE is the average absolute error value extracted from the difference between the actual value and the predicted value divided by n, as expressed in Eq. (3). MAE is scale dependent, if the error size is the same for each model, the error rate is not the same.

$$MAE = \frac{\sum |Actual - Predict|}{n} \tag{3}$$

MSE generally calculates the average error after squaring the difference between the actual value and the predicted value, as seen in Eq. (4).

$$MSE = \frac{\sum (Actual - Predict)^2}{n} \tag{4}$$

RMSE is the root of the MSE which is more robust to outlier than MAE, as described in Eq. (5).

$$RMSE = \sqrt{\frac{\sum (Actual - Predict)^2}{n}} \tag{5}$$

$$mMAPE = \begin{cases} \frac{100}{n} \sum_{t=1}^{n} |Actual_t - Predict|, \ if \ |Actual|_{max} < 1 \\ \frac{100}{n} \sum_{t=1}^{n} \frac{|Actual_t - Predict_t|}{|Actual|_{max}}, \ else \end{cases} \tag{6}$$

mMAPE is the index to prevent infinity (Inf) when the actual value was 0 [41]. Using the Eq. (1), the mMAPE is calculated after simulating two

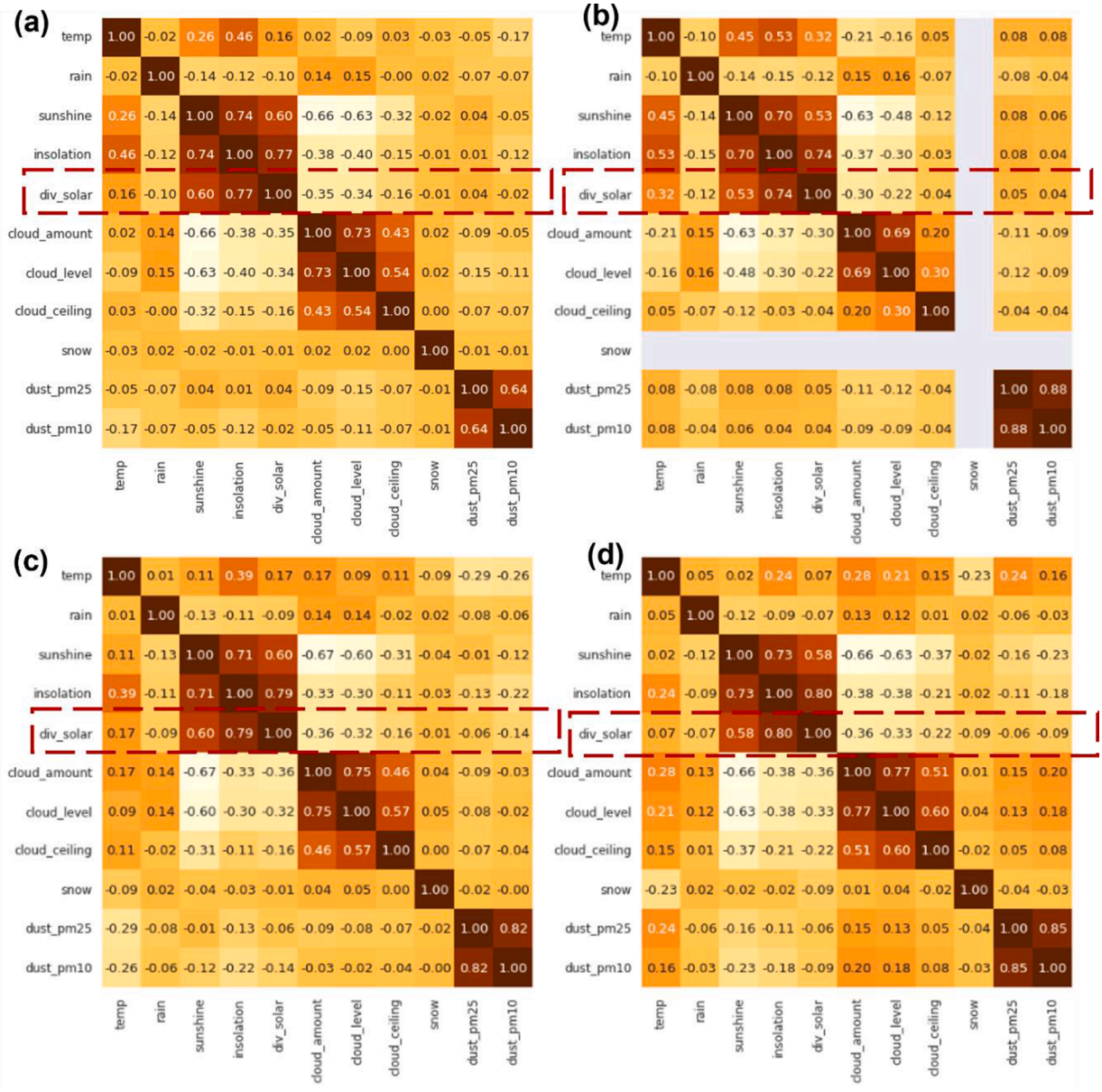


Fig. 4. Pearson correlation coefficients for Case 2 in different seasons as (a) spring, (b) summer, (c) fall and (d) winter.

cases. $|\ Actual\ |_{max}$ is the maximum absolute value of the actual data. When the value of $|Actual|_{max}$ is less than 1, the difference between the actual data and the predicted data is used as the evaluation value, which considerably avoid the slight difference between the actual value and the predicted data. Therefore, the predicted data is amplified by $|Actual|_{max}$. If $|Actual|_{max}$ is greater than 1, the error value is calculated by dividing the difference between the actual data and the predicted data by $|Actual|_{max}$. With MAPE, the magnitude of the error is affected by the difference between the actual and predicted values.

3. Models

3.1. Holt-Winter model

According to the studies in refs. [42] and [43], the Holt-Winters model is a simple model which presents excellent accuracy and good predictability for energy generation forecast. The Holt-Winters model is highly useful for a constant periodicity and seasonal component. The associated issues like exponential smoothing model make obstacle to estimate the trend and seasonality. However, the use of Holt-Winters model considerably overcome these issues [44]. Holt-Winters model is normally divided into two types according to the data characteristics such as multiplicative seasonal model and additive seasonal model. The multiplicative seasonal model deals the increase in variance of data with the passage of time. The additive seasonal model is referred to highly suitable when the data presents a linear and uniform increasing trend. In this work, PV output data from Case A was applied to suitable additive seasonal (Holt-Winters) model. It is found that PV power output data fluctuations are constant between 0 to 1400 kWh. The additive seasonal model can be expressed by following expressions;

$$L_t = \alpha(Y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1}), \tag{6}$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}, \tag{7}$$

$$S_t = \gamma(Y_t - L_t) + (1 - \gamma)S_{t-s}, \tag{8}$$

$$F_{t+m} = (L_t + b_t m) + S_{t-s} \tag{9}$$

Where, L_t = time series mean level at time t, b_t = time series trend

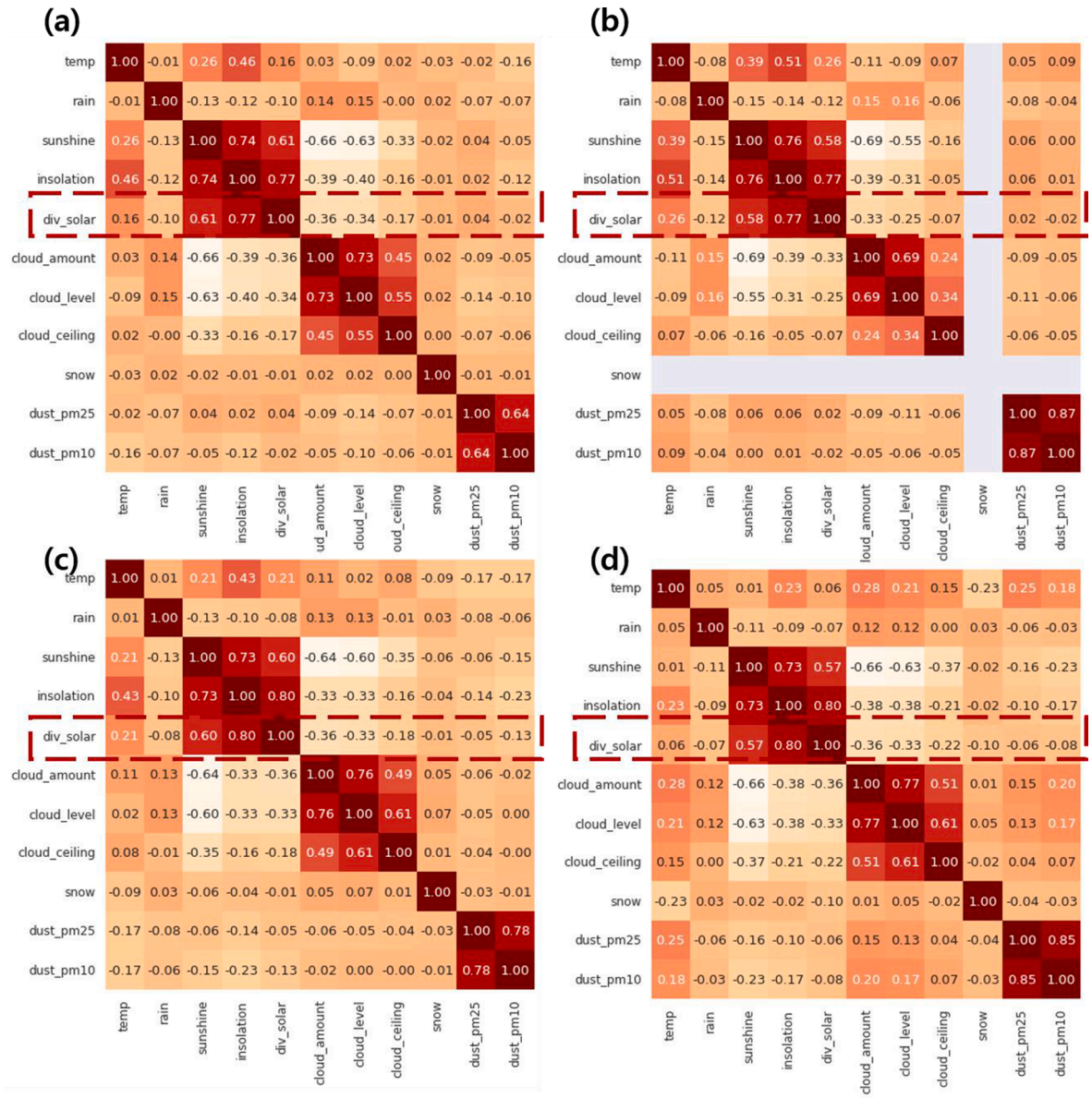


Fig. 5. Pearson correlation coefficients for Case 3 in different seasons as (a) spring, (b) summer, (c) fall and (d) winter.

Table 1
Summary of Pearson correlation coefficients for Case 1, Case 2, and Case 3.

Case	Season	Variables Temperature	Insolation	Sunshine	Cloud amount	Cloud level	Cloud ceiling	Dust PM10	Rain	Snow
Case 1	Spring	√	√	√	√	√	√			
	Summer	√	√	√	√	√	√		√	
	Fall	√	√	√	√	√	√		√	
	Winter		√	√	√	√	√	√		
Case 2	Spring	√	√	√	√	√	√		√	
	Summer	√	√	√	√	√			√	
	Fall	√	√	√	√	√	√	√		
	Winter		√	√	√	√	√		√	
Case 3	Spring	√	√	√	√	√	√			
	Summer	√	√	√	√	√				
	Fall	√	√	√	√	√	√			
	Winter		√	√	√	√	√			√

component at time t, S_t = time series seasonal component at time t, F_{t+m} = predicted value of time t+m predicted at time t, s = length of seasonal component, and α, β, γ = smoothing parameters.

The time series components such as L_t, b_t, and S_t. in Holt-Winters model are predicting PV generation for Case A i.e., seasonal, as presented in Table 5. This model is only use for seasonal not for weather factors (as error rate is very high). Therefore, it is necessary to search another model for predicting the PV power generation which is suitable

Table 2
p, d, q values from ARIMA model.

Case	Spring (p,d,q)	Summer (p,d,q)	Fall (p,d,q)	Winter (p,d,q)
1	2,1,5	4,1,4	3,0,5	5,1,1
2	4,1,3	5,1,3	3,0,5	4,0,5
3	2,1,5	4,1,5	3,0,3	4,0,5

Table 3
p, d, q and P, D, Q, S values from SARIMA model.

Case	Spring (p,d,q) (P,D,Q,S)	Summer (p,d,q) (P,D,Q,S)	Fall (p,d,q) (P,D,Q,S)	Winter (p,d,q) (P,D,Q,S)
1	(2,1,5) (2,0,0,24)	(4,1,4) (2,0,1,24)	(3,0,5) (1,0,0,24)	(5,1,1) (0,0,1,24)
2	(4,1,3) (1,0,0,24)	(5,1,3) (1,0,2,24)	(3,0,5) (1,0,0,24)	(4,0,5) (0,0,1,24)
3	(2,1,5) (2,0,1,24)	(4,1,5) (1,0,0,24)	(3,0,3) (2,0,2,24)	(4,0,5) (0,0,0,24)

for both seasonal and weather conditions too.

3.2. Multivariate linear regression model (MLP)

The Multivariate Linear Regression Model (MLP) model is a statistical technique that predicts a independent variable using multiple dependent variables. The purpose of MLP model is to develop the linear correlationship between multiple dependent variables and target variables. The variables in MLP model are described as;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdot + \beta_p x_{ip} + \epsilon, \quad (10)$$

where, y_i = dependent variable, x_i = explanatory variables, β_0 = y-intercept (constant term), β_p = slope coefficients for each explanatory variable, ϵ = the model's error term (also known as the residuals)

The MLP model generally provide the information about the correlation between the dependent variable and the independent variable. In this work, the dataset of Case 1, Case 2, Case 3, Case B and Case C are used for this model, but Case A cannot use because it is used only for univariate models. From Table 5, the error rate is very high, which means not suitable for predict the PV power generation using seasonal. Thus, MLP model is suitable for weather factors.

3.3. ARIMA Model

ARIMA model is comprised of the autoregressive (AR), moving average (MA) and difference. The model parameters p, d, q are used in ARIMA model in which p is the AR model's time lag parameter, d is the difference parameter for time lag, and q is MA model's time lag parameter. The time series model is generally using the values at time (t) and time lag (t-h). In this case, it becomes difficult to apply a general regression operation, so we need to proceed with the difference. Difference refers to the process of replacing the sequence obtained by subtracting the data at (t-h) from the data at t with the subject of time series analysis [45].

Table 2 shows the p,d,q values of ARIMA using the dataset from Case 1, Case 2, and Case 3. To know the values of p, d, and q, the 'auto.arima' function of the forecast library of RStudio is used. ARIMA model is performed after setting the p, d, q ranging from 0 to 5, and the variable with the lowest AIC (Akaike's Information Criterion). In fall season, the p, d, q values in Case 1, Case 2 and Case 3 are similar because the weather factors are stable. However, the significant difference in p, d, q values for spring, summer and winter are observed, as seen in Table 2. ARIMA model is good for Case 1, Case 2, Case 3, and Case A. Even

Table 4
Window average values of daytime for Case 1, Case 2 and Case 3.

Case	Spring	Summer	Fall	Winter
1	13 hours	14 hours	11 hours	10 hours
2	12 hours	14 hours	12 hours	10 hours
3	12 hours	14 hours	12 hours	9 hours

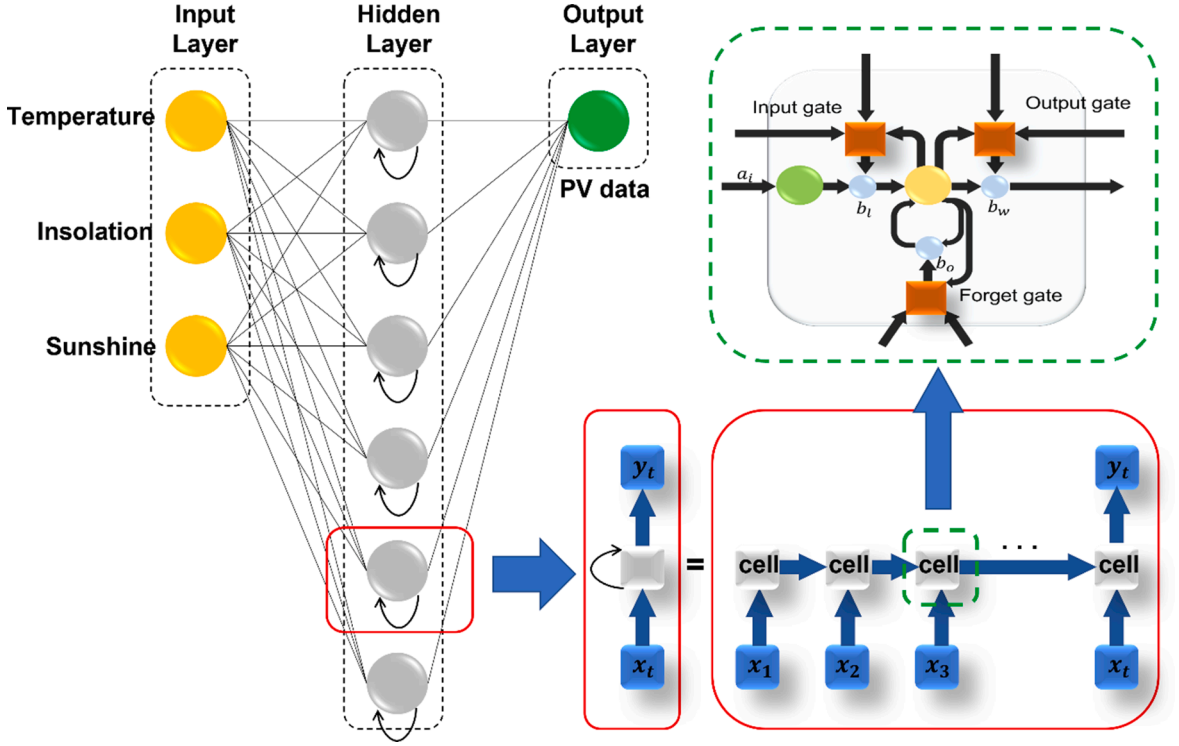


Fig. 6. LSTM model layout including Input, Output, forget Gate structure.

Table 5
Comparison of the error rate results Forecast with the actual data using 7 models (MLP, Holt-winters, ARIMA, SARIMA, ARIMAX, SARIMAX, LSTM).

Seasonal Case	Parameter Case	Model	Spring	Summer	Fall	Winter
1	A	Holt winters	12.53	10.96	18.18	13.30
		ARIMA	7.23	7.89	8.93	8.15
		SARIMA	7.23	7.89	8.93	8.16
	B	Multivariate	15.02	14.95	16.68	15.41
		ARIMAX	7.22	8.06	9.15	8.00
		SARIMAX	7.35	8.12	9.11	8.00
		LSTM	5.61	7.21	6.73	5.90
	C	Multivariate	15.11	15.56	16.89	14.91
		ARIMAX	6.74	15.95	11.23	9.57
		SARIMAX	7.29	9.85	9.19	8.20
		LSTM	5.45	7.00	6.83	5.46
2	A	Holt winters	12.45	11.12	13.47	15.05
		ARIMA	7.46	7.82	8.96	8.67
		SARIMA	7.54	7.83	8.95	8.68
	B	Multivariate	16.03	15.71	16.15	14.99
		ARIMAX	7.74	8.00	9.32	9.41
		SARIMAX	7.73	8.04	9.30	9.34
		LSTM	5.99	6.17	7.06	5.88
	C	Multivariate	15.93	15.61	16.06	14.50
		ARIMAX	7.62	8.07	9.40	9.50
		SARIMAX	7.62	8.07	9.40	9.50
		LSTM	7.16	6.61	8.07	6.62
3	A	Holt winters	11.86	11.48	14.53	13.01
		ARIMA	7.17	8.25	8.42	7.95
		SARIMA	7.32	8.22	8.46	7.95
	B	Multivariate	15.07	15.37	15.32	15.39
		ARIMAX	7.13	8.22	8.17	8.47
		SARIMAX	7.20	8.24	8.17	8.47
		LSTM	5.34	6.89	5.70	5.24
	C	Multivariate	15.20	15.36	15.54	14.87
		ARIMAX	7.13	8.24	8.22	8.40
		SARIMAX	7.12	8.24	8.22	8.40
		LSTM	5.55	6.70	5.90	5.57

Table 6
LSTM model error values in terms of MAE, MSE, RMSE, mMAPE for all cases.

Seasonal case	Parameter case	Season	MAE	MSE	RMSE	mMAPE (%)
1	B	Spring	12.02	10334.11	101.66	5.61
		Summer	11.85	16472.7	128.35	7.21
		Fall	12.24	14641.1	121	6.73
		Winter	16.24	17528.55	132.4	5.90
	C	Spring	18.03	10788.39	103.87	5.45
		Summer	17.72	14622.45	120.92	7.00
		Fall	20.69	12553.03	112.04	6.83
		Winter	21.72	14622.01	120.92	5.46
2	B	Spring	14.39	21946.01	148.14	5.99
		Summer	12.22	11264.24	106.13	6.17
		Fall	15.92	27467.21	165.73	7.06
		Winter	14.99	15251.83	123.5	5.88
	C	Spring	25.11	19459.46	139.5	7.16
		Summer	23.02	12500.68	111.81	6.61
		Fall	22.66	20442.45	142.98	8.07
		Winter	20.32	11072.05	105.22	6.62
3	B	Spring	5.61	4371.33	66.12	5.34
		Summer	6.18	7637.41	87.39	6.89
		Fall	6.71	8402.67	91.67	5.70
		Winter	6.26	6665.44	81.64	5.24
	C	Spring	8.43	3754.56	61.27	5.55
		Summer	10.07	8779.57	93.7	6.70
		Fall	7.77	5403.88	73.51	5.90
		Winter	9.64	5938.16	77.06	5.57

though, it shows good results in different seasons and weather but still presents high error rate. However, the error rate in ARIMA model is lower than those of above models.

3.4. SARIMA Model

SARIMA model integrates the AR model and the MA model including the seasonal variables. SARIMA model allows to measure the p, d, q and P, D, Q, S values. P, D, Q, S values in SARIMA are more appropriate wherein P is the AR model's seasonal parameter, Q is MA model's seasonal parameter, D is difference seasonal parameter and S is seasonal parameter in dataset. SARIMA model is also a univariate model, thus only Case 1, Case 2, Case 3 and Case A are used.

Table 7
Comparison designed LSTM model error rate with reported similar forecasting models.

Year	Location	Horizon	Model	Error rate	Ref.
2021	Northern region India	15 mins 30 mins	ARIMA	9.5% 12.6%	[4]
2020	South China	3 hours	BI-LSTM	6.1% ~ 10.2%	[46]
2020	Germany	1 day	LSTM	12%	[47]
2019	Alice Springs (Australia)	1 day	LSTM, CNN, C-LSTM	8% ~ 11.2%	[48]
2020	Nevada (USA)	1 day	LSTM-RNN	6.29%	[49]
2020	Alice Springs (Australia)	1 day	CNN (ResNet and DenseNet)	18%, 15%	[50]
2022	Korea	1 hour	LSTM	5.24 ~ 6.70%	This work

3.5. ARIMAX Model

ARIMAX is an extension of the ARIMA model to predict the PV power output value using multiple parameters and p, d, q values, as shown in Table 2. In this model, the multiple input parameters are based on PCC results in Table 1. ARIMAX model is shown the good suitability to season data of Case 1, Case 2, Case 3 and parameter data of Case B and Case C. The error rate in ARIMAX model is similar to SARIMA model.

3.6. SARIMAX Model

SARIMAX model is an extension of the SARIMA/ARIMAX to predict the PV power data using seasonal and exogenous parameters. SARIMAX p, d, q and P, D, Q, S values are found same as in SARIMA (Table 3). Multiple parameters in Table 1 were used as exogenous parameters in SARIMAX model. This model is good for season data like Case 1, Case 2, Case 3 and parameter data of Case B and Case C. It can be seen that all above models are not completely suitable for seasonal (Case 1, Case 2, Case 3) and weather factors (Case A, Case B and Case C).

3.7. Long-short term memory (LSTM) Model

LSTM model is a Neural Network (NN) model in which the whole data can process and predict the future data. The most representative time series NN model is recursive neural network (RNN), but it has gradient vanishing problem. Gradient vanishing is a problem in which old data cannot affect the forecasting values. LSTM model is solving the gradient vanishing problem using three gates such as forget, input, and output gates, as illustrated in Fig. 6. From Fig. 6, the designed LSTM model is composed of three layers including input layer, one hidden layer and output layer. This hidden layer consists of several nodes which are normally exporting the result through the activation function in the hidden layer. Each node size is the twice of input parameters, that can be considered as cell. As described in Fig. 6, the forget gate is located in the cell, suggesting that forget gate is the part of one hidden layer in this model. In other word, one hidden layer is necessary for this model, in which the forget gate does not have any impact on increasing the hidden

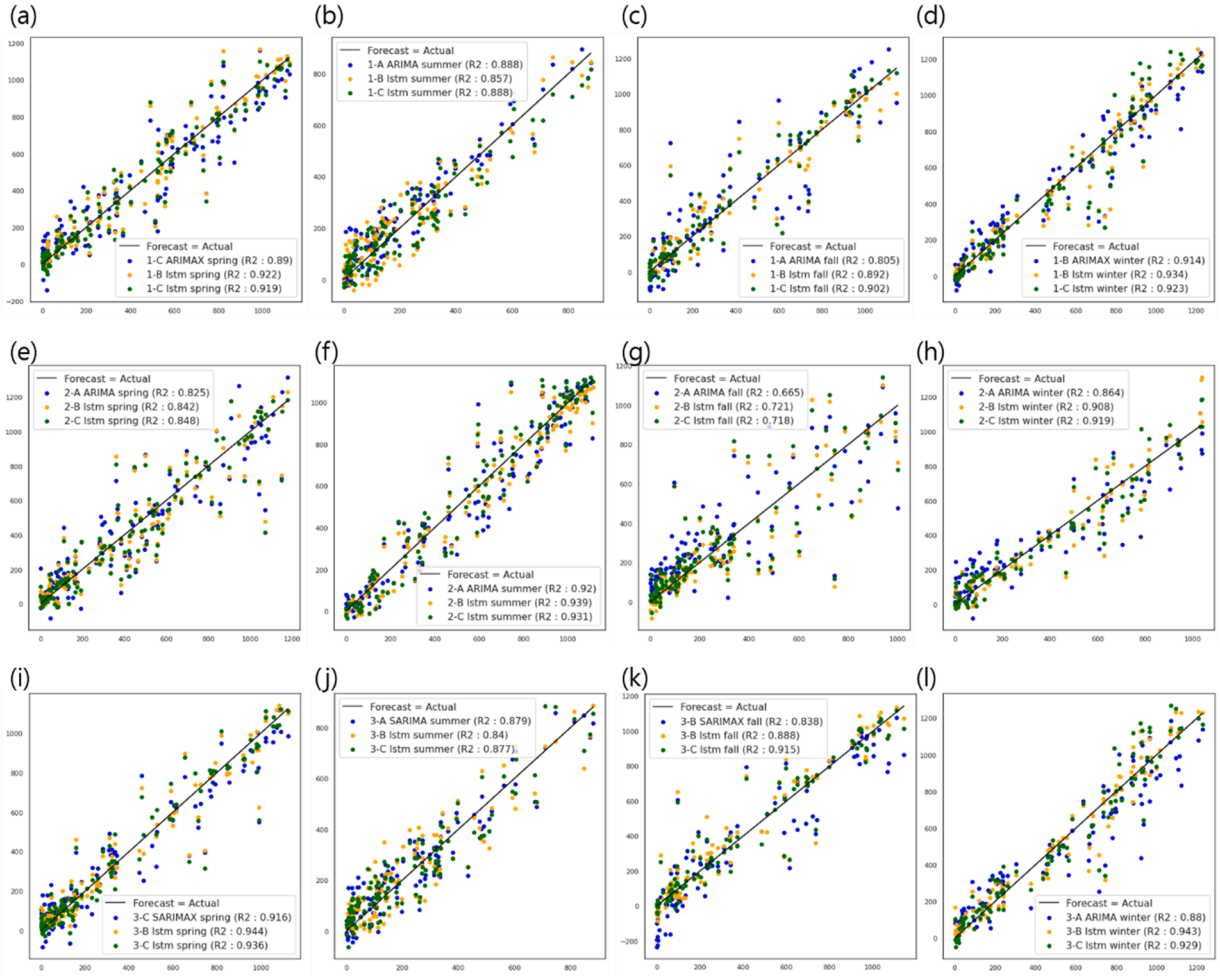


Fig. 7. R^2 graph between actual PV power output data and predict data. (a-d) are for Case 1, (e-h) are for Case 2 and (i-l) are for Case 3.

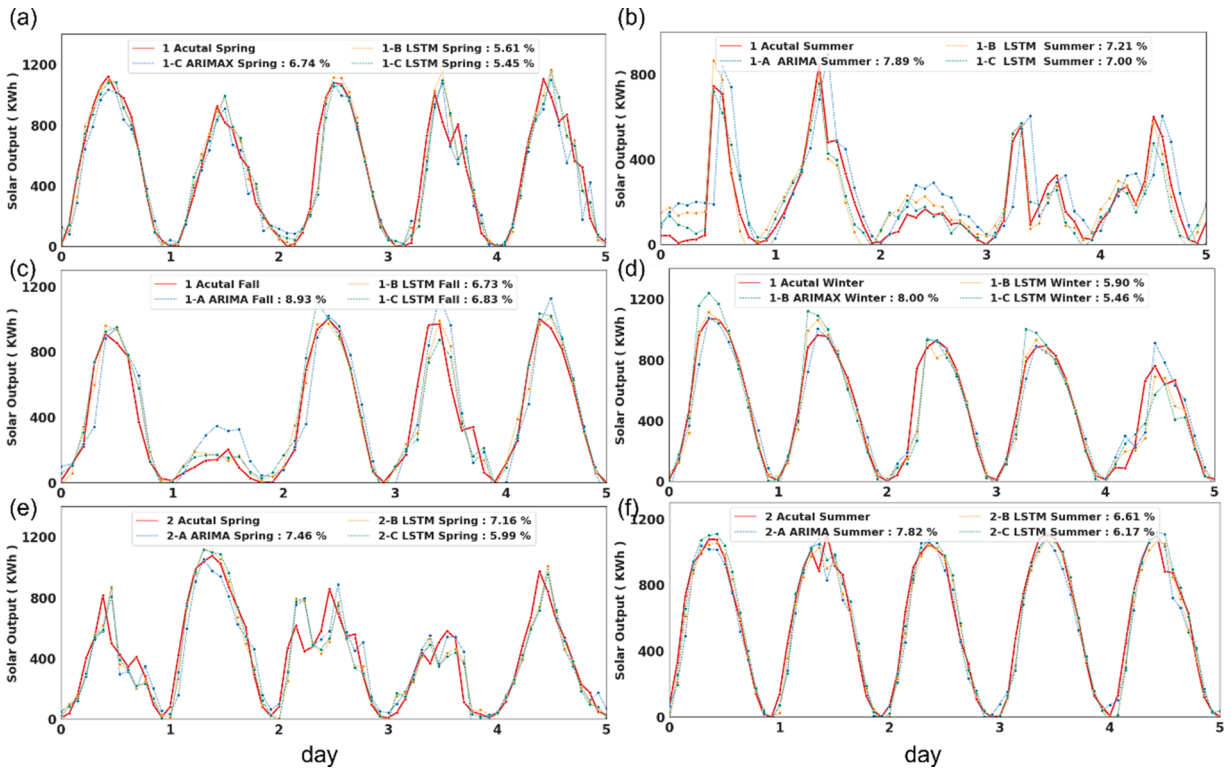


Fig. 8. Regression plot extracted from results of top three models for PV power prediction, 5 days ahead for the validation. (a-d) for Case 1, and (e-f) for Case 2.

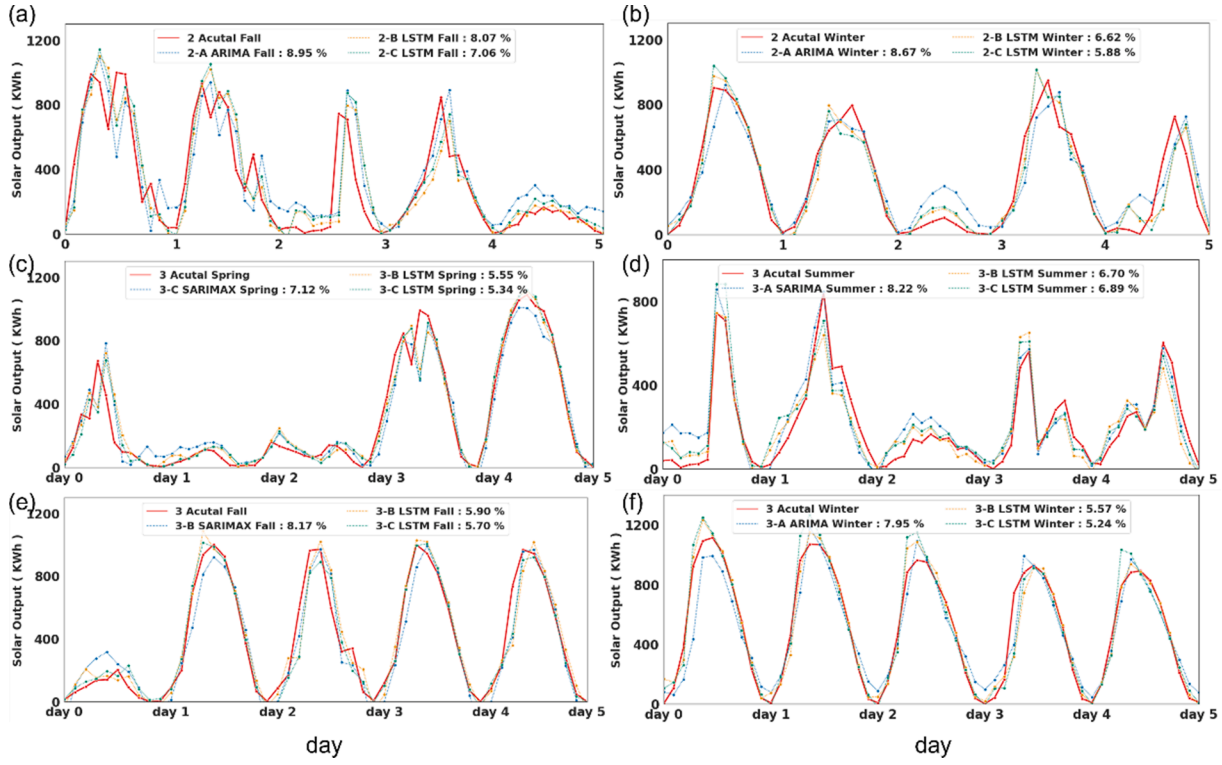


Fig. 9. Regression plot extracted from results of top three models for PV power prediction, 5 days ahead for the validation. (a-b) for Case 2, and (c-f) for Case 3.

layers. However, each node or cell is responsible for deciding the average window size and values. The window value and horizon value are required for performing the LSTM model. The window value is a variable that sets number of previous data to predict. Horizon is a variable that sets how many lags to predict the value. Herein, the horizon value is set to 1. Window values are calculated from the average daytimes. The average daytime for each season is calculated using the API for sunset and sunrise times of the Korea Astronomy and Space Science Institute. The window values are summarized in Table 4. The error rate in this model is the lowest as compared to other models (MLP, Holt-

winters, ARIMA, SARIMA, ARIMAX, SARIMAX) used for all cases. Table 6 summarizes the LSTM model performance evaluation in terms of MAE, MSE, RMSE and mMAPE values. From the Table 6, LSTM model performances are in the order of Case 3 > Case 1 > Case 2. Case 3 presents the lowest values of MAE, MSE, RMSE and mMAPE as compared to Case 1 and Case 2, suggesting the good model performance and excellent accuracy. By considering the parameter cases, the Case B expresses the better performance, as estimated the lower values of MAE, MSE, RMSE and mMAPE. Moreover, our model presents the lowest error rate as compared to reported AI models used for forecasting the PV generation in various regions (worldwide), as seen in Table 7.

4. Results and discussions

Among all used models, LSTM model has exhibited the good accuracy along with an excellent correlation coefficient (R^2) of 0.943 and average mMAPE of 5.79 for seasonal and weather factors. Based on data, Fig. 7 depicts the top 3 models R^2 graphs between actual PV power output data and predict PV power data for seasonal and weather factors considering 5 days. From Fig. 7, the R^2 values are getting bigger from Case 1 to Case 3. In this study, the highest R^2 value of 0.922 records in Case 1 Spring. In Case 3 spring and Case 3 winter, the highest R^2 of 0.944 and 0.943 are obtained respectively. However, the fall and winter in Case 1 present the similar R^2 values. This observation clearly demonstrates the low error rates as the observed high R^2 values. It is well known that R^2 value is ranging from 0 to 1. The high and low error rate is decided by evaluating the R^2 value. The high error rate means R^2 value close to 0, and low error rate means R^2 value nearly 1. For all Cases, herein LSTM model displays over 0.9 R^2 value. The high R^2 values in our results are comprehensively reflected the lower error rate of the system.

Figs. 8 and 9 displays the graph between actual PV power output data and top 3 model's forecasting data for 5 days. The data gap between actual data and predict data are smaller when we compare Case 1 to Case 3. From these figures, the lowest error rate is observed by LSTM model as compared to other models. This explains the seasonal and cycle must be considered rather than simple datasets. Case B presents the lower error rate than that of Case C. It is seen that when large number of variables are used in Case C, the complexity of the model increases, and the error rate rises. Thus, Case B with three variables result in the lowest error rate. The reduction in error rate of graph for actual and predict value is needed to find the good accuracy model. In this regard, the hyperparameter (activation, optimization method) for NN models, optimization number of parameters, study of ensemble model (bagging, stacking, tune model) would be considered in future to predict highly reliable and quick model for PV power generation forecasting.

5. Conclusion

In summary, the PV power output forecast with high accuracy is studied by using seven AI models. In this work, the input data were characterized by dividing seasons and choosing the multiple parameters from seasons for the designing time series models. PV power generation data and weather data were collected from Ansan city and Suwon city during January 2017 to June 2021, respectively. PV power generation forecasting was evaluated on an hourly basis to test the efficacy of various models. As compared to all models (Holt-Winters, Multivariate Linear Regression, ARIMA, SARIMA, ARIMAX, SARIMAX). LSTM model showed the lowest error rate for quick PV power generation forecasting in seasonal and weather factors. In future, the tuning of current LSTM model would carry out by using finetuning and ensemble model to find best efficiency. In order to advance the current study, the more seasonal and weather-related factors incorporating AI systems would be considered for PV generation forecasting with more accuracy. Moreover, more accurate solar forecasting models can be used to advance weather-sensitive energy systems to achieve high industrial and environmental efficiencies.

CRediT authorship contribution statement

EunGyeong Kim: Conceptualization, Methodology, Writing – original draft. **M. Shaheer Akhtar:** Visualization, Investigation, Data curation, Validation, Writing – review & editing. **O-Bong Yang:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the "Human Resources Program in Energy Technology" of the Institute of Energy Technology Evaluation and Planning (KETEP) and granted financial resources from the Ministry of Trade, Industry & Energy, Republic of Korea (Project No.: 20204010600470). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (Project No. 2022M3J7A1066428).

References

[1] S.K. Chow, E.W. Lee, D.H. Li, Short-term prediction of photovoltaic energy generation by intelligent approach, Energy Build. 55 (2012) 660–667.
[2] C.A.F. Frederiksen, Z. Cai, Novel machine learning approach for solar photovoltaic energy output forecast using extra-terrestrial solar irradiance, Appl. Energy 306 (2022), 118152.
[3] Y. Jiang, L. Zheng, X. Ding, Ultra-short-term prediction of photovoltaic output based on an LSTM-ARMA combined model driven by EEMD, J. Renew. Sustain. Energy 13 (4) (2021), 046103.
[4] S. Das, Short term forecasting of solar radiation and power output of 89.6 kWp solar PV power plant, Mater. Today. 39 (2021) 1959–1969.
[5] M.d.A. Al-Nimr, S. Kiwan, H. Sharadga, Simulation of a novel hybrid solar photovoltaic/wind system to maintain the cell surface temperature and to generate electricity, Int. J. Energy Res. 42 (3) (2018) 985–998.
[6] C. Chen, et al., Online 24-h solar power forecasting based on weather type classification using artificial neural network, Sol. Energy 85 (11) (2011) 2856–2870.
[7] J. Shi, et al., Forecasting power output of photovoltaic systems based on weather classification and support vector machines, IEEE Trans. Ind. Appl. 48 (3) (2012) 1064–1069.
[8] Y. Jung, et al., Long short-term memory recurrent neural network for modeling temporal patterns in long-term power forecasting for solar PV facilities: case study of South Korea, J. Clean. Prod. 250 (2020), 119476.
[9] M. Konstantinou, S. Peratikou, A.G. Charalambides, Solar photovoltaic forecasting of power output using LSTM networks, Atmosphere 12 (1) (2021) 124.
[10] S. LEE, J. KIM, SPV Dominant Parameter Estimation Exploration and Visualization from Weather and Prediction Model, The Korean Institute of Electrical Engineers, 2021, pp. 464–465.
[11] A. Mellit, S. Sağlam, S.A. Kalogirou, Artificial neural network-based model for estimating the produced power of a photovoltaic module, Renew. Energy 60 (2013) 71–78.
[12] D. Su, E. Batzelis, B. Pal, Machine learning algorithms in forecasting of photovoltaic power generation, in: 2019 International Conference on Smart Energy Systems and Technologies (SEST), IEEE, 2019.
[13] W. Kanchana, S. Sirisukprasert, PV Power Forecasting with Holt-Winters Method, in: 2020 8th International Electrical Engineering Congress (iEECON), IEEE, 2020.
[14] R. Huang, et al., Solar generation prediction using the ARMA model in a laboratory-level micro-grid, in: 2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2012.
[15] M. Hassanzadeh, M. Etezadi-Amoli, M. Fadali, Practical approach for sub-hourly and hourly prediction of PV power output, in: North American Power Symposium 2010, IEEE, 2010.
[16] Y. Chu, et al., Short-term reforecasting of power output from a 48 MWe solar PV plant, Sol. Energy 112 (2015) 68–77.

[17] L. Liu, et al., Prediction of short-term PV power output and uncertainty analysis, Appl. Energy 228 (2018) 700–711.
[18] M.Q. Raza, M. Nadarajah, C. Ekanayake, On recent advances in PV output power forecast, Sol. Energy 136 (2016) 125–144.
[19] S.I. Vagropoulos, et al., Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting, in: 2016 IEEE International Energy Conference (ENERGYCON), IEEE, 2016.
[20] R. Ahmed, et al., A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization, Renew. Sustain. Energy Rev. 124 (2020), 109792.
[21] Y. Li, Y. Su, L. Shu, An ARMAX model for forecasting the power output of a grid connected photovoltaic system, Renew. Energy 66 (2014) 78–89.
[22] A.K. Dubey, et al., Study and analysis of SARIMA and LSTM in forecasting time series data, Sustain. Energy Technol. Assessm. 47 (2021), 101474.
[23] E. Lorenz, et al., Qualified forecast of ensemble power production by spatially dispersed grid-connected PV systems, Measurement (2007) 1–7.
[24] F. Mirzapour, et al., A new prediction model of battery and wind-solar output in hybrid power system, J. Ambient Intell. Hum. Comput. 10 (1) (2019) 77–87.
[25] S. Chen, H. Gooi, M. Wang, Solar radiation forecast based on fuzzy logic and neural networks, Renew. Energy 60 (2013) 195–201.
[26] C. Shang, P. Wei, Enhanced support vector regression based forecast engine to predict solar power output, Renew. Energy 127 (2018) 269–283.
[27] D.A.R. de Jesús, et al., Solar pv power prediction using a new approach based on hybrid deep neural network, in: 2019 IEEE Power & Energy Society General Meeting (PESGM), IEEE, 2019.
[28] M. Gao, et al., Short-term forecasting of power production in a large-scale photovoltaic plant based on LSTM, Appl. Sci. 9 (15) (2019) 3192.
[29] Y. Zhou, et al., Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine, Energy 204 (2020), 117894.
[30] S.Z. Islam, et al., Photovoltaic modules evaluation and dry-season energy yield prediction model for NEM in Malaysia, PLoS One 15 (11) (2020), e0241927.
[31] P. Mandal, et al., Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques, Procedia Comput. Sci. 12 (2012) 332–337.
[32] M. Mishra, et al., Deep learning in electrical utility industry: a comprehensive review of a decade of research, Eng. Appl. Artif. Intell. 96 (2020), 104000.
[33] Y. Yu, J. Cao, J. Zhu, An LSTM short-term solar irradiance forecasting under complicated weather conditions, IEEE Access 7 (2019) 145651–145666.
[34] A.C. Harvey, N. Shephard, 10 Structural Time Series Models, 1993.
[35] S.-M. Kim, M. Oh, H.-D. Park, Analysis and prioritization of the floating photovoltaic system potential for reservoirs in Korea, Appl. Sci. 9 (3) (2019) 395.
[36] O. Nematollahi, K.C. Kim, A feasibility study of solar energy in South Korea, Renew. Sustain. Energy Rev. 77 (2017) 566–579.
[37] Patro, S. and K.K. Sahu, *Normalization: A preprocessing stage.* arXiv preprint arXiv: 1503.06462, 2015.
[38] J. Benesty, et al., *Pearson correlation coefficient*, in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.
[39] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res. 30 (1) (2005) 79–82.
[40] C.J. Willmott, On the validation of models, Phys. Geogr. 2 (2) (1981) 184–194.
[41] K.-H. Shin, et al., Estimation method of predicted time series data based on absolute maximum value, J. Energy Eng. 27 (4) (2018) 103–110.
[42] C. Cho, Forecasting the Cement Traffic Volume at the Port of Donghae, Korea Logist. Rev. 18 (1) (2008) 33–53.
[43] B. Billah, et al., Exponential smoothing model selection for forecasting, Int J Forecast 22 (2) (2006) 239–247.
[44] J.-T. Kim, Forecasting number of student by Holt-Winters additive model, J. Korean Data Inf. Sci. Soc. 20 (4) (2009) 685–694.
[45] A. Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning, O'Reilly Media, 2019.
[46] H. Sharadga, S. Hajimirza, R.S. Balog, Time series forecasting of solar power generation for large-scale photovoltaic plants, Renew. Energy 150 (2020) 797–807.
[47] N. Maitanova, et al., A machine learning approach to low-cost photovoltaic power prediction based on publicly available weather reports, Energies 13 (3) (2020) 735.
[48] K. Wang, X. Qi, H. Liu, A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network, Appl. Energy 251 (2019), 113315.
[49] F. Wang, et al., A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework, Energy Convers. Manage. 212 (2020), 112766.
[50] H. Zang, et al., Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta learning, Int. J. Electr. Power Energy Syst. 118 (2020), 105790.