energies

MDPI

Article

Short Term Prediction of PV Power Output Generation Using Hierarchical Probabilistic Model

Dongkyu Lee [1], Jae-Weon Jeong [1] and Guebin Choi [2,*]

[1] Department of Architectural Engineering, Hanyang University, 222 Wangsimni-Ro, Seungdong-Gu, Seoul 04763, Korea; wavesg241@hanyang.ac.kr (D.L.); jjwarc@hanyang.ac.kr (J.-W.J.)
[2] Department of Statistics (Institute of Applied Statistics), Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: guebin@jbnu.ac.kr

Abstract: Photovoltaics are methods used to generate electricity by using solar cells, which convert natural energy from the sun. This generation makes use of unlimited natural energy. However, this generation is irregular because they depend on weather occurrences. For this reason, there is a need to improve their economic efficiency through accurate predictions and reducing their uncertainty. Most researches were conducted to predict photovoltaic generation with various machine learning and deep learning methods that have complicated structures and over-fitted performances. As improving the performance, this paper explores the probabilistic approach to improve the prediction of the photovoltaic rate of power output per hour. This research conducted a variable correlation analysis with output values and a specific EM algorithm (expectation and maximization) made from 6054 observations. A comparison was made between the performance of the EM algorithm with five different machine learning algorithms. The EM algorithm exhibited the best performance compared to other algorithms with an average of 0.75 accuracies. Notably, there is the benefit of performance, stability, the goodness of fit, lightness, and avoiding overfitting issues using the EM algorithm. According to the results, the EM algorithm improves photovoltaic power output prediction with simple weather forecasting services.

Keywords: photovoltaic power output prediction; expectation and maximization (EM) algorithm; probabilistic method; correlation analysis

Citation: Lee, D.; Jeong, J.-W.; Choi, G. Short Term Prediction of PV Power Output Generation Using Hierarchical Probabilistic Model. *Energies* **2021**, *14*, 2822. https://doi.org/10.3390/en14102822

Academic Editor: George Kosmadakis

Received: 6 April 2021
Accepted: 10 May 2021
Published: 14 May 2021

1. introduction

The adoption of the large-scale solar photovoltaic (PV) plants worldwide has been driven by the need to curb environmental pollution and through reduction of fossil fuel usage and the associated carbon emissions [1]. Although the development of PV systems is essential, of equal importance is PV resource prediction, since its availability limits the development of a PV system. To plan grid-connected PV systems efficiently with low fluctuations in the power output [2], PV power should be accurately predicted for at least up to 24 h. Having accurate forecasted PV output in an hour or more short-term time resolution helps the grid operator to determine the necessary backup generation capacity through the determination of the shortfall on the actual generation from the forecasted values to enhance grid reliability. Moreover, forecasting the expected PV capacity benefits the customer by lowering energy cost and reducing uncertainty [3]. A review of published literature reveals the proposals based on various methods and tools for forecasting expected PV power. Some of the approaches employ exogenous data from numerical weather prediction, data of images, and nearest PV system and local measurements as input, whereas other approaches model various previously acquired PV data from prospective sites [3]. Notably, the horizon for carrying out the forecast can range from a few seconds to days or years. For spatial forecasting, however, the range can either be a site or regional scale. The categories for forecasting horizons include very short-term forecast with a range from 1 min to several minutes, short-term forecast ranging from 1 h or several hours ahead to a day or a week

ahead, medium-term ranging from 1 month to 1 year, and finally long-term forecast which ranges from 1 to 10 years ahead [4]. To design an efficient energy management PV integrated system, that encompasses unit commitment, power scheduling, and dispatch, both short-term and very short-term forecast are very necessary to improve the overall grid efficiency. Most researches conducted not only prediction accuracy with fixed input parameters but also prediction models of machine learning and stochastic methodologies.

Regarding demonstrations of using probabilistic and machine learning methods with nonlinear systems, there have been several studies that applied these methods to industrial systems. Ponza et al. [5] established an algorithm combining rules and a database. In addition, Ahmed et al. [6] suggested six different machine learning models for neural network, support vector machine, decision tree, random forest, extra tree, and gradient boosted trees using an inertial measurement unit and global positioning system to classify different pedestrian events. Hedrea et al. [7] presented a tensor product-based model for tower crane modeling, which has good system identification performance with a nonlinear real system. Research indicates that the prediction accuracy of the model used depends on the forecast horizon irrespective of the application of similar model parameters for the forecast [3,4]. Pedro and Coimbra (2012) [8] studied five machine-learning and deep-learning models for PV forecasting. Their research indicated that an artificial neural network with a genetic algorithm gets the best performances compared to other models when applied to PV power forecasting respectively. Filipe et al. [9] proposed an ensemble model of the statistical method with application of the physical method. Dong et al. [10] applied filtered stochastic models recursively estimating PV forecasting. This method gets improvements in forecasting accuracy compared to other machine learning methods. Auto-regressive and auto-regressive exogenous models (ARX) were reported by Bacher et al. (2009) [11]. The outcome noted a 35% root mean squared errors by ARX model on the reference forecasting model. Furthermore, using weather data and sora irradiance Almonacid et al. (2014) [12] and Leva et al. (2017) [13] applied an ANN model to predict the hourly power output for a PV system based in Spain and Italy, respectively. The outcome showed that the proposed model could predict hourly PV power output for sunny, partially cloudy, and cloudy days having partially outperformed other models. On the other hand, De Giorgi et al. (2016) [14] proposed various machine learning models to predict PV power generation. The study used ambient and module temperatures, and irradiance data as the input for setting up and testing the models. Based on the outcome, the proposed support vector machine model was superior to the other models. Moreover, a support vector machine for regression (SVMR) model with a prediction ahead for forecasted PV generation was studied by Wolff et al. (2016) [15]. The model was based on the application of solar data attained from weather measurement for the prediction of PV generation. From the PV measurements for the SVMR model, good results were obtained for short forecast periods (1h ahead), while the NWP model predictions were superior for prediction ranges above 3 h. Through irradiance forecasts, SVMR predictions for cloud data of motions were suited for periods between 1 and 3 h.

Although various approaches of prediction with machine learning and stochastic algorithms are suggested, conventional prediction methods have limitations of performance over short-term PV power output with only consideration of forecasted weather data [16] and misses a few important variables. For example, a variable that records whether the sun has risen or not is a very important variable in predicting PV power output, but traditional models do not take this into account.

The proposed model proposes a hierarchical probability model that interprets unobserved (but important) variables as latent variables. We propose an algorithm that iteratively predicts latent variable and PV power out from given data. Figure 1 is a diagram devised to differentiate between the proposed method and the existing approaches.

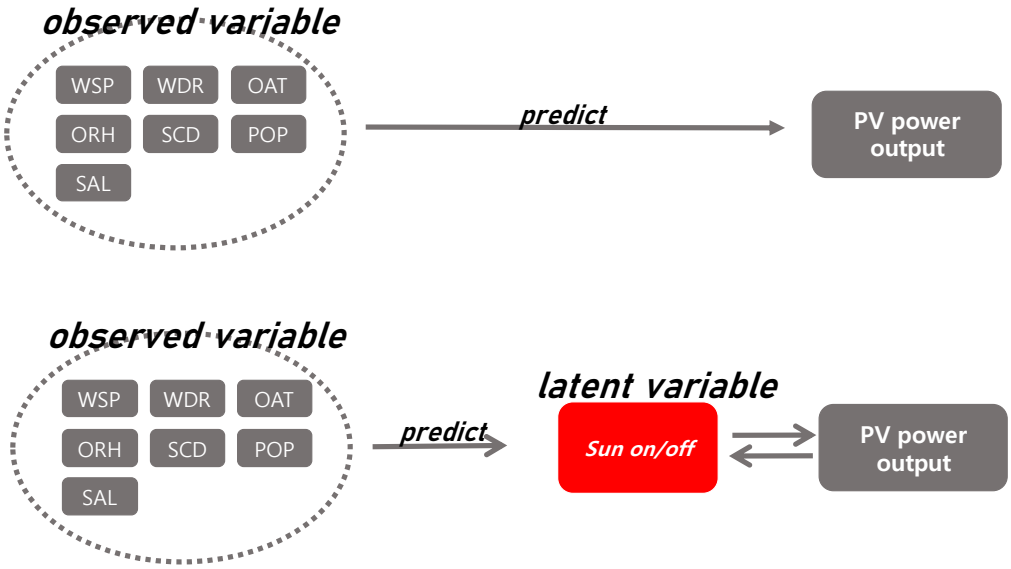

Figure 1. (**top**): Conventional prediction methods; (**bottom**): proposed method.

The paper's organization is illustrated as follows. Section 2 introduces the system description and the framework's prediction models. Section 3 will explore how the proposed model of the EM algorithm is characterized. Section 4 will be a discussion of the prediction performances of various machine learning models and proposed algorithms. Section 5 will include the conclusion.

2. Backgrounds

2.1. PV Generation System

The office buildings with rooftop PV systems in South Korea are operated in micro energy grid (MEG) with other renewable energy resources and energy storage systems (ESS). The geographical location of the buildings is latitude 37°31′ N and longitude 127°14′ E. The nominal power output capacity of the PV system is 50 kW which consists of 20 parallel strings of 10 series modules with a capacity of 250 W for each as shown in Figure 2. Each PV panel is positioned east of south-facing with an angle of 20 degrees from the roofline.

The grid-connected PV systems measure not only PV-related data but historical weather information with weather sensors. All measured parameters were stored in the building energy management system (BEMS) every one hour. The used data period was 1st of November 2016 to 31st of August 2017 (12 months) with hourly intervals. For instance, almost 70% of collected data are used for training and the model predicts hourly PV power output in a day-ahead (24 h) for the rest of the dataset.

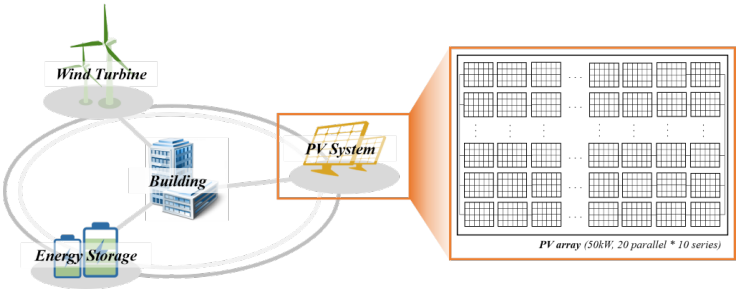

Figure 2. Schematic diagram of grid-connected PV system.

2.2. Online Weather Forecasting Dataset

The online weather forecasting data are used to suggest a day-ahead real prediction model obtained from Korea meteorological administration (KMA) and Korea astronomy

and space science institute (KASI). The weather forecasting information is provided in 3-h intervals by after tomorrow with online service in XML form and KASI offers solar altitude by the hour. However, there is no way to use historical data for PV forecasting. Thus, this paper gets the data by crawling and interpolates data in the one-hour interval.

The seven weather elements were used as predictors, which were pre-processed at 1-h intervals: wind speed (WSP), wind direction (WDR), outdoor air temperature (OAT), relative humidity (ORH), sky condition (SCD), probability of precipitation (POP), solar altitude (SAL). Table 1 describes the characteristics of the used weather forecasting datasets.

Table 1. Characteristics of online weather forecasting information.

Name	Specifications
Wind speed (WSP)	m/s
Wind direction (WDR)	North (0), North east (1), East (2), South east (3), South (4), South west (5), West (6), North west (7)
Outdoor air temperature (OAT)	°C
Relative humidity (ORH)	%
Sky condition (SCD)	Clear (1), Little cloudy (2), Cloudy (3), Overcast (4)
Probability of precipitation (POP)	%
Solar altitude (SAL)	°

2.3. Prediction Model Description

In this study, the prediction performance of PV power output was compared commonly used prediction models such expectation-maximization (EM), artificial neural network (ANN), support vector machine (SVM), random forest(RF), classification and regression tree (C&RT), Chi-squared automatic interaction detection (CHAID). Table 2 presents key hyperparameters of PV power output prediction models.

The ANN is a network composed of artificial neurons or nodes which emulate the biological neurons. This network is made up of input and output layers of processing units identified as nodes that are interconnected via one or more 'hidden' layers which are dictated by the numbers of independent and dependent variables sequentially. In this study, a typical feed-forward neural network (FFNN) is used which trained only forward direction with information.

The RF is an ensemble-learning algorithm that combines a large set of regression trees. Each of the classification trees is built with a bootstrap sample of the data. Thus, RF uses bootstrap aggregation, a successful approach for combining unstable learners, and random variable selection for tree building.

The CART is an effective binary recursive partitioning algorithm that can automatically search the optimal decision tree and feature selection with the classification effect of the features at each node. CART built the tree by recursively splitting the variable space based on the impurity of the variables to determine the split till the termination condition is met.

The CHAID is a single variable prediction one which is a dependent variable, the other variables are referred to as the predictor variables. CHAID is an iterative technique that determines the most effective method given the dataset. Child nodes are created by splitting the subsets of the space repeatedly. This is done to the whole dataset to construct CHAID. A merger is done between any allowable pairs of categories of the predictor variable to ascertain the best split at each node.

The SVM is a classification model based on statistical learning theory and structural risk minimization principle. The SVM, a classification model, was established on the statistical learning theory and the structural risk immunization principle. SVM operates as a non-linear mapping, renovating the new training data into a higher dimension. With the new dimension, the SVM examines the linear optimal separating hyper-plane.

Table 2. The hyperparameters of six PV power output prediction models.

Prediction Model	Hyperparameter	Value
ANN	Activation function	Sigmoid
	Hidden neurons	10
RF	Number of nodes	1000
	Maximum tree depth	100
C&RT	Impurity measure	0.0001
	Minimum change in impurity	Gini
CHAID	Tree depth4	
	Minimum change in frequencies	0.001
SVM	Kernel type RBF Gamma	Radial Basis Function(RBF)+0.1

3. Proposed Method

Machine learning models involve a heavy computational burden, which is not appropriate for the circumstances of Korean historical weather station data. In contrast to previous studies using machine learning models, this research uses the expectation-maximization (EM) algorithm [17]. There are numerous applications of the EM algorithm with the identification of missing data in aggressive models and hidden variables of each state [18]. Irregular performance with missing gathered data in industrial processes happens regularly. Kalyani et al. [19] suggested an EM algorithm setting for removal of outliers in irregular data. The EM algorithm has been applied in various ways to incomplete data with missing values. Also, Deng et al. [14] dealt with a Bayesian framework using the EM algorithm to reduce uncertainty errors in soft lab data. However, there is no research that applies the EM algorithm to the regression and prediction of specific values, and it is applicable to other industrial systems to handle various latent variables. Also, modeling of the EM algorithm has no hyper-parameters for tuning compared to other machine learning methods and consumes a lot of time and domain knowledge randomness. With similar circumstances as the previous research mentioned above, predicting the weather data of Korean weather stations is incomplete. So, this research applies the EM algorithm to conduct PV prediction. In figuring out the input features of data, the sky condition would have the lowest correlation with power generation. Also, this research considers the sky condition as a latent variable to estimate PV output. Although the sky condition is not an important variable, it is difficult to estimate in weather forecasting data due to the weather forecasting comprising four different levels of cloudiness. For this reason, there is a disadvantage that the performance of the previous model varies greatly depending on sky condition measurement accuracy. To avoid those drawbacks, the sky condition is considered a latent variable among other features and the EM algorithm is applied to jointly estimate the sky condition and PV output. In addition to correlation analysis of input data, the input of probabilities of precipitation and sky conditions are very correlated with the values shown as Table 3. The correlated values do not affect the prediction results. Thus, this research uses the POP input variable instead using both of POP and SCD input variables shown as Figure 3 .

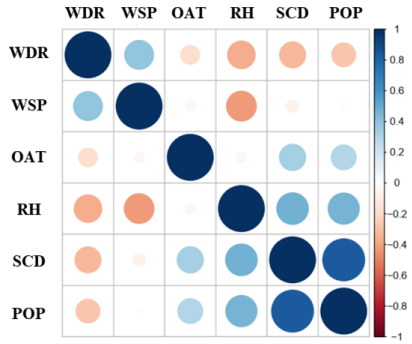

Figure 3. Visualization of Table 3.

Table 3. Correlation matrix.

	WDR	**WSP**	**OAT**	**RH**	**SCD**	**POP**
WDR	1	0.40	−0.17	−0.36	−0.32	−0.27
WSP	0.40	1	−0.05	−0.42	−0.07	0.01
OAT	−0.17	−0.05	1	0.05	0.33	0.30
RH	−0.36	−0.42	0.05	1	0.48	0.47
SCD	−0.32	−0.07	0.33	0.48	1	0.84
POP	−0.27	0.01	0.30	0.47	0.84	1

Our goal was to construct a model that can predict *pvoutput*. The *pvoutput* means the amount of power generated from sunlight, which is affected by various weather conditions. We observed an hourly rate of electricity production from 01:00 on 21st of November, 2016 to 23:00 on 26th of August, 2017, yielding 6504 observations. Let these values as $\mathbf{z}$, so $\mathbf{z}$ is a 6504×1 vector. Further, define the weather conditions that affect $\mathbf{z}$ as $\mathbf{X}$. The design matrix $\mathbf{X}$ is composed of observations of the following five weather conditions, thus $\mathbf{X}$ is a 6504×5 matrix.

The value of $\mathbf{X}$ was obtained from Korea Meteorological Administration (https://data.kma.go.kr). Predicting $\mathbf{z}$ based on $\mathbf{X}$ seems very reasonable at first glance, but the result is somewhat inaccurate. The reason for this is that the most important factor in predicting $\mathbf{z}$ is whether the sun is on or off. If the sun does not rise, the amount of power produced would only be zero, even though all of the values of $\mathbf{X}$ (*wind speed*, ... , *solar altitude*) would meet the perfect conditions for power generation.

For this reason, our study introduces a latent variable $\mathbf{w}$ which indicates the presence of the sun. Like $\mathbf{z}$, $\mathbf{w}$ is a vector of length 6504 and $w_i = 0$ means that there is no sun at time i and $w_i = 1$ means the sun exists at time i. Note that w_i is not a simple variable to distinguish day and night at the time i. That is, it is not true that the *solar altitude*> 0 at the time i implies $w_i = 1$. For example, during rainy daylight, the *solar altitude* is positive but $w_i = 0$.

To consider the effects of w_i, consider the following model:

$$z_i = W_i y_i + (1 - W_i)\tilde{y}_i, \quad i = 1, \ldots, n$$

and assume that (1) $W_i \sim$ *i.i.d. Bernoulli*(p_i) where $p_i = \frac{\exp(\gamma + \sum_{j=1}^{p} x_{ij}\beta_j)}{\exp(\gamma + \sum_{j=1}^{p} x_{ij}\beta_j) + 1}$, (2) $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon \sim$ *i.i.d.* $N(0, \sigma^2)$, and (3) $\tilde{y}_i = 0$. In here $n = 6504$, $p = 5$, and x_{ij} is the (i, j)th element of $\mathbf{X}$. The parameter γ controls the average value of p_i, which means the probability that the sun exists at the ith time. Above the model, y_i is the power output when $w_i = 1$, and we assume that it can be modeled by simple linear regression. The $\tilde{y}_i$ is the power output when $w_i = 0$, and we assume that it is always zero. For each i, we observe z_i rather than (w_i, z_i). Thus, $\mathbf{z}$ is the observed data and $(\mathbf{w}, \mathbf{z})$ is the complete data. To handle this missing problem, we adapt the EM algorithm.

Let $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta})$. The likelihood of complete data is as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(z_i|w_i) f(w_i)$$

where

$$f(z_i|w_i) = \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{\left(z_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2}{2\sigma^2} \right) \right]^{w_i} \left[\delta(z_i) \right]^{1-w_i}$$

and

$$f(w_i) = \left[\frac{\exp(\gamma + \mathbf{x}_i'\boldsymbol{\beta})}{\exp(\gamma + \mathbf{x}_i'\boldsymbol{\beta}) + 1} \right]^{w_i} \left[1 - \frac{\exp(\gamma + \mathbf{x}_i'\boldsymbol{\beta})}{\exp(\gamma + \mathbf{x}_i'\boldsymbol{\beta}) + 1} \right]^{1-w_i}.$$

Here, $\delta(\cdot)$ is the Dirac delta function and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$.

[E-step]

In this step, we need to calculate the conditional expectation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\left[\log L(\boldsymbol{\theta}|\mathbf{z}, \mathbf{w}) \middle| \mathbf{z} \right] = \sum_{i=1}^{n} \left(E_{\boldsymbol{\theta}^{(k)}}\left[\log f(z_i|w_i) | z_i \right] + E_{\boldsymbol{\theta}^{(k)}}\left[\log f(w_i) | z_i \right] \right).$$

The most important part of the calculation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is the calculation of $E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i)$ which is a function of $\boldsymbol{\theta}^{(k)}$ where $\boldsymbol{\theta}^{(k)}$ is current estimate value of $\boldsymbol{\theta}$. By definition of conditional expectation, $E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i)$ can be expressed as

$$E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i) = \frac{p_i^{(k)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta}^{(k)})^2}{2\sigma^2} \right)}{p_i^{(k)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta}^{(k)})^2}{2\sigma^2} \right) + (1 - p_i^{(k)})\delta(z_i)}.$$

Note that $E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i)$ is always 0 when $z > 0$, otherwise $E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i)$ is always 1 due to the characteristics of the delta function. Thus,

$$E_{\boldsymbol{\theta}^{(k)}}(W_i|z_i) = I(z_i > 0)$$

where $I(\cdot)$ is the indicator function. This means that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is not a function of $\boldsymbol{\theta}^{(k)}$, so $\boldsymbol{\theta}$ can be obtained by repeating E-step and M-step only once. This ensures very fast convergence.

[M-step]

Let us differentiate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ by γ.

$$\frac{\partial}{\partial\gamma} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial\gamma} E\left[\log f(z_i|w_i) | z_i \right] + \frac{\partial}{\partial\gamma} E\left[\log f(w_i) | z_i \right] \right).$$

Note that $\frac{\partial}{\partial\gamma} E\left[\log f(z_i|w_i) | z_i \right] = 0$ since $f(z_i|w_i)$ is constant of γ. The term $\frac{\partial}{\partial\gamma} E\left[\log f(w_i) | z_i \right]$ can be expressed as follows:

$$\frac{\partial}{\partial\gamma} E\left[\log f(w_i) | z_i \right] = I(z_i > 0) \frac{1}{e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}} + 1} + I(z_i = 0) \frac{-e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}}}{e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}} + 1}.$$

From

$$p_i = \frac{\exp(\gamma + \sum_{j=1}^{p} x_{ij}\beta_j)}{\exp(\gamma + \sum_{j=1}^{p} x_{ij}\beta_j) + 1} = \frac{e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}}}{e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}} + 1}$$

and

$$1 - p_i = \frac{1}{\exp(\gamma + \sum_{j=1}^{p} x_{ij}\beta_j) + 1} = \frac{1}{e^{\gamma + \mathbf{x}_i'\boldsymbol{\beta}} + 1},$$

we can easily check that $\frac{\partial}{\partial \gamma} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = 0$ is equivalent to

$$m = \sum_{i=1}^{n} p_i \tag{1}$$

where $m = \sum_{i=1}^{n} I(z_i > 0)$. Note that Equation (1) is an interpretable and reasonable condition. Here, m can be interpreted as the number of observations when the sun rises and we define p_i as the probability of the sun rising at each observation. Therefore, Equation (1) means the sum of probabilities of sunrise equals to the number of observations that actually sun was up.

Now let us differentiate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ by $\boldsymbol{\beta}$.

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} E\big[\log f(z_i|w_i)|z_i\big] + \frac{\partial}{\partial \boldsymbol{\beta}} E\big[\log f(w_i)|z_i\big] \right).$$

The left one can be calculated as

$$\frac{\partial}{\partial \boldsymbol{\beta}} E\big[\log f(z_i|w_i)|z_i\big] = I(z_i > 0)\frac{(z_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{2\sigma^2} = I(z_i > 0)\frac{\mathbf{x}_i z_i - \mathbf{x}_i\mathbf{x}_i'\boldsymbol{\beta}}{2\sigma^2}.$$

Since $\mathbf{x}_i$ is a $p \times 1$ vector, $\mathbf{x}_i\mathbf{x}_i'$ is a $p \times p$ matrix such that

$$\mathbf{x}_i\mathbf{x}_i' = \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \dots & x_{i1}x_{ip} \\ x_{i2}x_{i1} & x_{i2}^2 & \dots & x_{i2}x_{ip} \\ \dots & \dots & \dots & \dots \\ x_{ip}x_{i1} & x_{ip}x_{i2} & \dots & x_{ip}^2 \end{bmatrix},$$

thus, $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$ can be expressed as follows:

$$\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \dots & \sum_{i=1}^{n} x_{i2}x_{ip} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \dots & \sum_{i=1}^{n} x_{ip}^2 \end{bmatrix} = \mathbf{X}'\mathbf{X}.$$

Similarly, one can easily check that

$$\sum_{i=1}^{n} \mathbf{x}_i z_i = \mathbf{X}'\mathbf{z}.$$

Note that $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$ and $\sum_{i=1}^{n} \mathbf{x}_i z_i$ can be expressed as

$$\begin{cases} \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' = \sum_{i\in\{i:z_i>0\}} \mathbf{x}_i\mathbf{x}_i' + \sum_{i\in\{i:z_i=0\}} \mathbf{x}_i\mathbf{x}_i' = \tilde{\mathbf{X}}'\tilde{\mathbf{X}} + (\mathbf{X}'\mathbf{X} - \tilde{\mathbf{X}}'\tilde{\mathbf{X}}) \\ \sum_{i=1}^{n} \mathbf{x}_i z_i = \sum_{i\in\{i:z_i>0\}} \mathbf{x}_i z_i + \sum_{i\in\{i:z_i=0\}} \mathbf{x}_i z_i = \tilde{\mathbf{X}}'\mathbf{z} + (\mathbf{X}'\mathbf{z} - \tilde{\mathbf{X}}'\mathbf{z}) \end{cases}$$

where $\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \sum_{i\in\{i:z_i>0\}} \mathbf{x}_i\mathbf{x}_i'$ and $\tilde{\mathbf{X}}'\mathbf{z} = \sum_{i\in\{i:z_i>0\}} \mathbf{x}_i z_i$. Here, $\tilde{\mathbf{X}}$ can be considered as another design matrix defined except for the observation where z_i is zero. By introducing the $\tilde{\mathbf{X}}$, we can express $\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(z_i|w_i)|z_i\right]\right)$ as

$$\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(z_i|w_i)|z_i\right]\right) = \frac{1}{2\sigma^2}(\tilde{\mathbf{X}}'\mathbf{z} - \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta}). \tag{2}$$

Now, let us focus on the term $\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(w_i)|z_i\right]$. This term can be calculated as

$$\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(w_i)|z_i\right] = \frac{1}{2\sigma^2}\left(I(z_i>0)\frac{\mathbf{x}_i'}{e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}+1} + I(z_i=0)\frac{-e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x_i}'}{e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}+1}\right).$$

Using $p_i = \frac{e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}}{e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}+1}$ and $1-p_i = \frac{1}{e^{\gamma+\mathbf{x}_i'\boldsymbol{\beta}}+1}$, we get

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(w_i)|z_i\right] &= \frac{1}{2\sigma^2}\left(I(z_i>0)\mathbf{x}_i'(1-p_i) - I(z_i=0)\mathbf{x}_i'p_i\right)\\ &= \frac{1}{2\sigma^2}\left(I(z_i>0)\mathbf{x}_i' - I(z_i>0)\mathbf{x}_i'p_i - I(z_i=0)\mathbf{x}_i'p_i\right) = \frac{1}{2\sigma^2}\left(I(z_i>0)\mathbf{x}_i' - \mathbf{x}_i'p_i\right).\end{aligned}$$

Observe

$$\sum_{i=1}^{n}\mathbf{x}_i' = \left(\sum_{i=1}^{n}x_{i1},\ldots,\sum_{i=1}^{n}x_{ip}\right)' = \mathbf{X}'\mathbf{1}$$

where $\mathbf{1}$ is an $n\times 1$ vector such that $\mathbf{1} = (1,1,\ldots,1)'$. Similarly, one can check that

$$\sum_{i=1}^{n}\mathbf{x}_i p_i = \mathbf{X}'\boldsymbol{p}$$

where $\boldsymbol{p}$ is an $n\times 1$ vector with $\boldsymbol{p} = (p_1,\ldots,p_n)'$. Thus,

$$\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log f(w_i)|z_i\right]\right) = -\mathbf{X}'\boldsymbol{p} + \tilde{\mathbf{X}}'\mathbf{1}. \tag{3}$$

Combining Equations (2) and (3), we obtain

$$\frac{\partial}{\partial\boldsymbol{\beta}}E\left[\log L(\boldsymbol{\theta})|\mathbf{z}\right] = \tilde{\mathbf{X}}'\mathbf{z} - \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{p} + \tilde{\mathbf{X}}'\mathbf{1}.$$

Therefore the last step is to solve the following equations.

$$\begin{cases} m - \sum_{i=1}^{n}p_i = 0 \\ \tilde{\mathbf{X}}'\mathbf{z} - \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{p} + \tilde{\mathbf{X}}'\mathbf{1} = \mathbf{0} \end{cases} \tag{4}$$

To solve (4), we propose the following Algorithm 1.

Algorithm 1 Proposed algorithm for prediction of PV power output

1: Get m and $\tilde{\mathbf{X}}$.
2: Initialize $\boldsymbol{p}$ and $\boldsymbol{\beta}$ as $\boldsymbol{p}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ and set $\ell = 0$.
3: $\boldsymbol{\beta}^{(\ell+1)} \leftarrow (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{z} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}'\mathbf{1} - \mathbf{X}'\boldsymbol{p}^{(\ell)})$
4: $\boldsymbol{p}^{(\ell+1)} \leftarrow m\left(\frac{e^{\mathbf{X}\boldsymbol{\beta}^{(\ell+1)}}}{1+e^{\mathbf{X}\boldsymbol{\beta}^{(\ell+1)}}}\right)\left(\mathbf{1}'\frac{e^{\mathbf{X}\boldsymbol{\beta}^{(\ell+1)}}}{1+e^{\mathbf{X}\boldsymbol{\beta}^{(\ell+1)}}}\right)^{-1}$
5: $\ell \leftarrow \ell+1$
6: Repeat 3–5 until convergence.

4. Results

This research conducts various machine learning models and probabilistic models to predict PV generation included in Sections 2 and 3. The analysis was conducted from largely the following three viewpoints.

4.1. Overall Performance Comparison

In this section, the overall performances of five machine learning models and the proposed method are compared. Figure 4 shows the goodness-of-fit through R^2 when each of 5 months-8 months was set as a training set. The x-axis of each figure is the actual data, and the y-axis is the estimated values. Therefore, it can be interpreted that the more the points are concentrated near the straight line, the more excellent the models are. The degree to which the points are concentrated near the straight line is indicated through R^2, and the higher the value of R^2, the better the fit of the model.

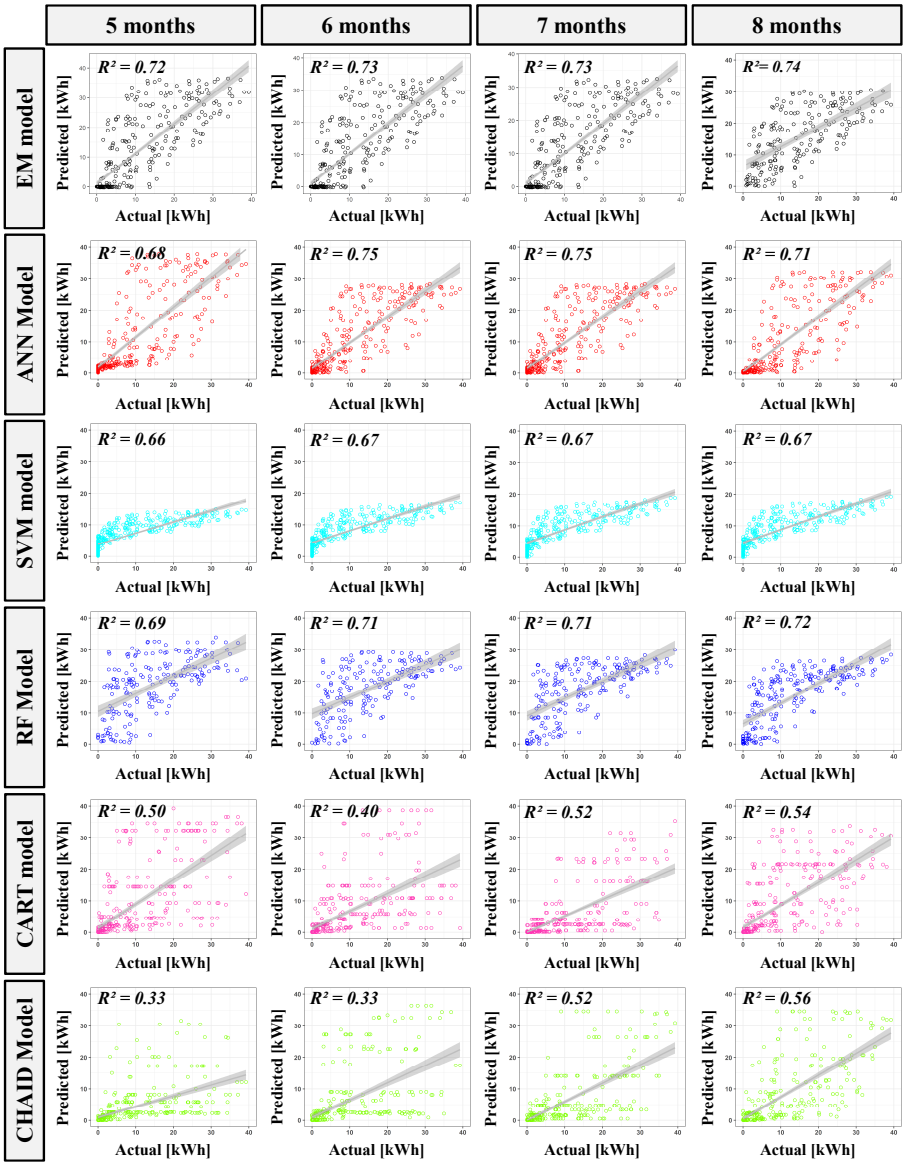

Figure 4. The R^2 values of the six models and plots of actual values vs. predicted values when the training sets were set to 5–8 months.

When 6 months and 7 months were set as training sets, respectively, the ANN model showed the highest fits, and the R^2 value at this time was 0.75. However, when trained

on data for 5 months and 8 months, respectively, the ANN model did not show the same high fits as with 6 and 7 months. In addition to the ANN method, the proposed method EM algorithm showed high fits. EM showed evenly good performances for 5–8 months and has the advantage of showing relatively uniform performances compared to the ANN method. Other methods generally have poorer fits compared to EM and ANN.

In analyses thereafter, all training sets were set to 7 months. The reasons why the training sets were set to 7 months are (1) that our target model, ANN, is the setting that shows the best performance, and (2) that if the methods learn for 7 months, the methods will learn until June and given that the rainy season in Korea is June-August, learning June, which is one month in the rainy season, was thought to be appropriate. Figure 5 is a box plot of errors when 7 months was set as the training set. The box plot is meaningful for comparing the overall performances of the models, but it is not informative. When evaluating the prediction performance of solar electric energy, the point that is worthy of attention is the fit between 8:00 and 17:00 because this is the time zone in which electric energy is mainly produced. The analysis in this section gives no information about the fit for the time we are interested in.

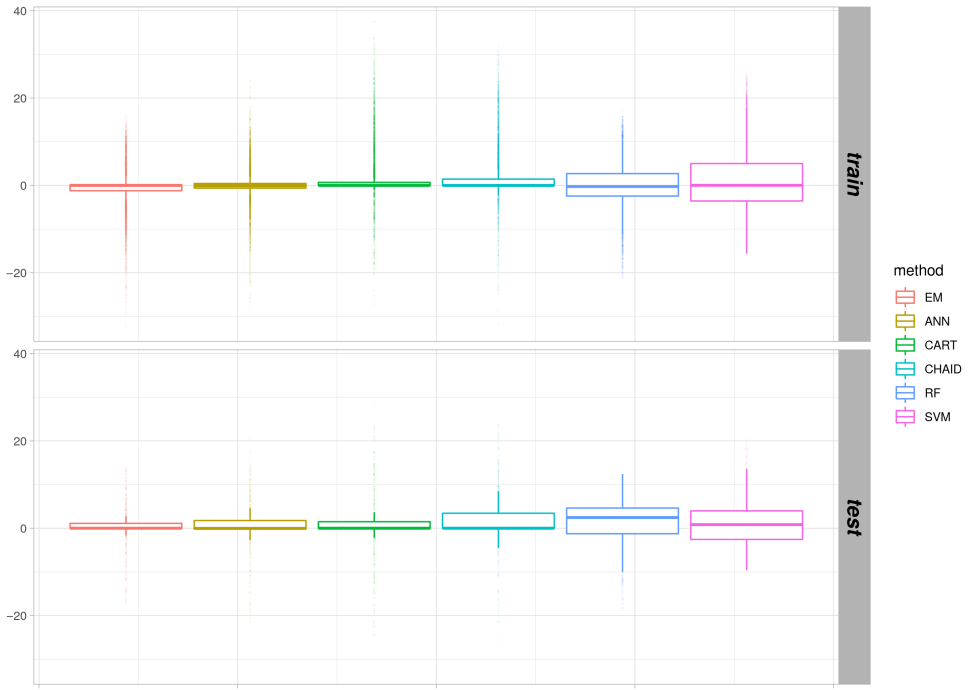

Figure 5. The train and test errors of the six models when the training was set to 7 months.

4.2. Analysis of Model Performances by Time

To overcome the limitations of the analysis described in Section 4.1, the performances of the model should be reanalyzed by time zone. Figure 6 shows the comparison of the fits of individual models by time.

An overall feature is the outliers of the errors from 7:00 to 17:00. This coincides with the time when the sun is up. The differences in errors between the time when the sun is up and the time when the sun is not up are very intuitive because when the sun is not up, the electric energy production can be set to zero. Except for the SVM model, it can be seen that most of the models well-adjusted the electric energy production to 0 during the time when the sun was up. All models generally have large errors during the time. In particular, such errors are outstanding between 10:00 and 14:00. Since this is the time when the sunlight is the strongest in a day, it is more difficult to predict accurate electric energy production.

A salient feature is the outliers of the errors between 7:00 and 17:00. The outliers are expressed as points that deviated from the box plot. SVM has the least number of outliers followed by the EM algorithm. The fact that a model has many outliers means that there

are many points where the prediction by the model deviates greatly because the variance of the model is large. Therefore, models with many outliers are suspected of overfitting. Other models' outliers are very large from 15:00 to 17:00.

In terms of performance stability, the EM algorithm and SVM are excellent. However, SVM has a disadvantage in that the overall fitting is poor. In addition, the fitting of SVM is much poorer compared to other models when electric energy production is zero. ANN has a good overall fit but has a weakness of unstable performances in certain time zones.

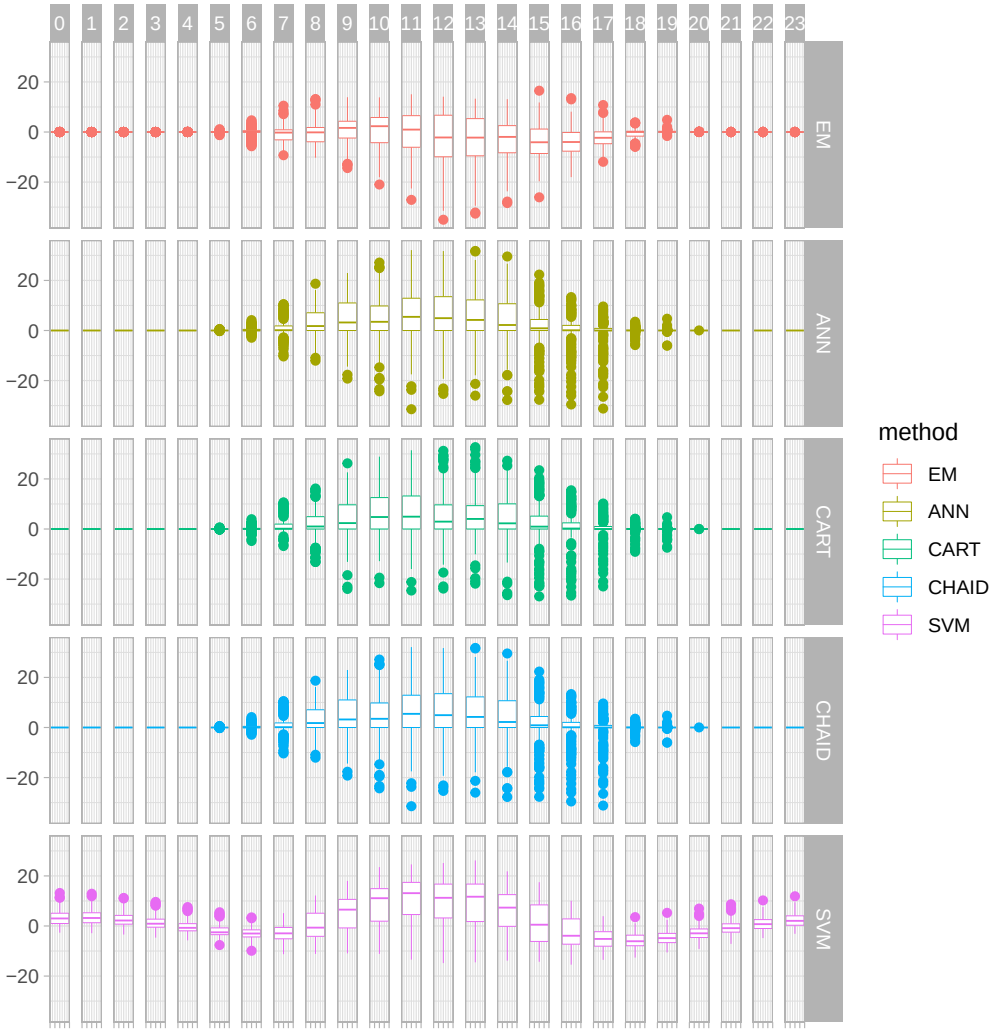

Figure 6. The goodness-of-fit of each model is analyzed by time. Each column represents hours, and each row represents a model. Each figure is a representation of the errors of the relevant method during the relevant time.

4.3. Analysis of a Certain Day after Magnifying and Refocusing

In this section, we compare the performances of EM and ANN-based on a certain day. In addition, we refocus on the overfitting issue of ANN more clearly. ANN has an issue of overfitting with various training datasets. The top row of Figure 7 shows that both EM and ANN predicting PV outputs. The bottom row of Figure 7 enlarges among EM, ANN, and real values of PV from 28th of June to 3rd of July.

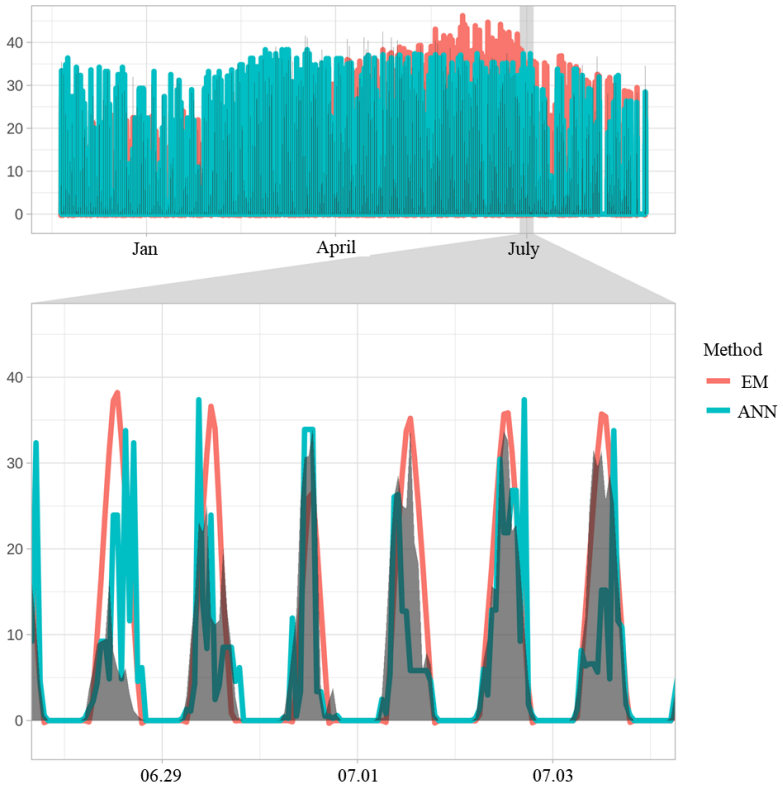

Figure 7. Comparison performances between EM and ANN.

In Figure 7, the results of ANN and EM are represented in blue and red, respectively. Actual values are expressed in gray shades. During the periods of rainy days, EM has more stable predicting results compared to ANN. ANN has no mountain-shaped fit results like the overfitting results. The results of EM are explainable to predict the PV outputs. Especially, from 28th of June to 29th of June, and 2nd of July, the predicted values of ANN are very volatile, and if a real-time prediction result was required, the ANN's results are useless. The results of EM are also the same in that it is not suitable for the periods, but EM yields explainable mountain-shape results. Without the sudden rain, EM would have been a very good fit.

Considering the overall aspects, the advantages of EM can be summarized as follows.

- **Goodness of Fit**: Overall results of the fit are excellent with ANN (deep neural network).
- **Stability**: There is little difference in performance between performances based on R^2 values.
- **Explainable model**: It is an interpretable model as described before.
- **Lightness**: The model is light and estimates $\theta^{(k)}$ very fast. The reason is that $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ is not the function of $\theta^{(k)}$.
- **Avoid overfitting issue**: It is more free from the overfitting issue than other complex models (ANN).
- **No hyperparameter**: There are no separate hyperparameters to be set to train the model. This means that it is easy to use overall aspects with no expertise.

5. Concluding Remarks

5.1. Summary

This research conducted PV prediction with Korea Meteorological predicting data. To analyze predicting data, latent variables were introduced and an EM process that can optimize both PV and latent variables was designed. For performance comparison, various

methods and comparative experiments were performed, and the proposed method proved to be excellent in performance while having several advantages, such as providing an interpretable model that is free from overfitting issues.

The proposed model had many advantages over the existing methods. There are many advantages, such as performance, stability of the model, lightness, and no hyperparameters, but the main advantage is that a new hierarchical model including latent variables is proposed. In this study, a model including one latent variable has been studied, but various latent variables that were not observed may be included depending on the situation. Our methods could easily be extended to these studies.

5.2. Future Research

In our model, the value of m is assumed to be a constant. This means that the rate at which the sun shines is always the same. From a macroscopic point of view, this is correct, but it may not be the case if you observe the data more in high resolution. For example, the rate of sunshine may vary depending on the season. In some seasons the sun rises long, in some seasons it does not. This seasonality is not reflected in our study. In addition, m may vary by region. Some areas may have a longer sun than others. In the end, m is not a constant, but a variable that changes subtly according to time and space. If we model these subtle local fluctuations, the performance of the prediction model can be improved. In the future, we plan to study a prediction model that reflects seasonality.

Author Contributions: D.L. and J.-W.J. performed the experiments, analyzed the data, and interpreted the results. G.C. contributed in developing the proposed method. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sahu, B.K. A study on global solar PV energy developments and policies with special focus on the top ten solar PV power producing countries. *Renew. Sustain. Energy Rev.* **2015**, *43*, 621–634. [CrossRef]
2. Elsinga, B.; van Sark, W.G. Short-term peer-to-peer solar forecasting in a network of photovoltaic systems. *Appl. Energy* **2017**, *206*, 1464–1483. [CrossRef]
3. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Martinez-de-Pison, F.J.; Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Sol. Energy* **2016**, *136*, 78–111. [CrossRef]
4. Raza, M.Q.; Nadarajah, M.; Ekanayake, C. On recent advances in PV output power forecast. *Sol. Energy* **2016**, *136*, 125–144.
5. Pozna, C.; Precup, R.E. Applications of signatures to expert systems modelling. *Acta Polytech. Hung.* **2014**, *11*, 21–39.
6. Ahmed, M.U.; Brickman, S.; Dengg, A.; Fasth, N.; Mihajlovic, M.; Norman, J. A machine learning approach to classify pedestrians' events based on IMU and GPS. *Int. J. Artif. Intell.* **2019**, *17*, 154–167.
7. Hedrea, E.L.; Precup, R.E.; Roman, R.C.; Petriu, E.M. Tensor product-based model transformation approach to tower crane systems modeling. *Asian J. Control* **2021**. [CrossRef]
8. Pedro, H.T.; Coimbra, C.F. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Sol. Energy* **2012**, *86*, 2017–2028.
9. Filipe, J.M.; Bessa, R.J.; Sumaili, J.; Tomé, R.; Sousa, J.N. A hybrid short-term solar power forecasting tool. In Proceedings of the 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP), Porto, Portugal, 11–16 September 2015; pp. 1–6. [CrossRef]
10. Dong, J.; Olama, M.M.; Kuruganti, T.; Melin, A.M.; Djouadi, S.M.; Zhang, Y.; Xue, Y. Novel stochastic methods to predict short-term solar radiation and photovoltaic power. *Renew. Energy* **2020**, *145*, 333–346. [CrossRef]
11. Bacher, P.; Madsen, H.; Nielsen, H.A. Online short-term solar power forecasting. *Sol. Energy* **2009**, *83*, 1772–1783. [CrossRef]
12. Almonacid, F.; Pérez-Higueras, P.J.; Fernández, E.F.; Hontoria, L. A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator. *Energy Convers. Manag.* **2014**, *85*, 389–398. [CrossRef]

13. Leva, S.; Dolara, A.; Grimaccia, F.; Mussetta, M.; Ogliari, E. Analysis and validation of 24 h ahead neural network forecasting of photovoltaic output power. *Math. Comput. Simul.* **2017**, *131*, 88–100. [CrossRef]
14. Deng, J.; Xie, L.; Chen, L.; Khatibisepehr, S.; Huang, B.; Xu, F.; Espejo, A. Development and industrial application of soft sensors with on-line Bayesian model updating strategy. *J. Process. Control* **2013**, *23*, 317–325. [CrossRef]
15. Wolff, B.; Kühnert, J.; Lorenz, E.; Kramer, O.; Heinemann, D. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. *Sol. Energy* **2016**, *135*, 197–208. [CrossRef]
16. Lee, D.; Jeong, J.; Yoon, S.H.; Chae, Y.T. Improvement of Short-Term BIPV Power Predictions Using Feature Engineering and a Recurrent Neural Network. *Energies* **2019**, *12*, 3247.
17. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22. [CrossRef]
18. Sammaknejad, N.; Zhao, Y.; Huang, B. A review of the expectation maximization algorithm in data-driven process identification. *J. Process. Control* **2019**, *73*, 123–136.
19. Kalyani, S.; Giridhar, K. Robust statistics based expectation-maximization algorithm for channel tracking in OFDM systems. In Proceedings of the 2007 IEEE International Conference on Communications, Glasgow, Scotland, 24–28 June 2007; pp. 3051–3056.