IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 11 May 2024, accepted 5 June 2024, date of publication 11 June 2024, date of current version 27 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412807

RESEARCH ARTICLE

Output Forecasting for Multiple Geographically Distributed PVs Without Meteorological Data

HIROKI YAMAMOTO[1], TAIKI KURE[1], JUNJI KONDOH[1], AND DAISUKE KODAIRA[2], (Member, IEEE)

[1]Graduate School of Science and Technology, Tokyo University of Science, Noda 278-8510, Japan
[2]Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba 305-8573, Japan

Corresponding author: Hiroki Yamamoto (7322604@alumni.tus.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant JP23K13306.

ABSTRACT Photovoltaic (PV) output forecasting often uses meteorological and historical PV data, including cloud imagery and weather conditions. Access to such data can be limited for numerous dispersed PVs, particularly in remote areas, making accurate forecasting challenging. Recent advancements in distributed PVs and communication technologies, such as smart meters, have facilitated the collection of time-series data from numerous dispersed PV installations. This development has spurred research into new forecasting models that utilize these data to forecast the PV output across multiple locations. One notable technique is the optical flow algorithm, which estimates and forecasts PV power generation transitions by converting PV power generation data from various locations into images. This study introduces a hybrid model that combines optical flow with machine learning using historical PV generation, time, and location data from multiple installations. The proposed model has an 18.4% improvement in the mean absolute error (MAE) over traditional models that depend on weather data. It also exhibits a 5.8% improvement in MAE and a 10.8% improvement in the continuous ranking probability score compared to the optical flow alone.

INDEX TERMS Photovoltaic (PV) power forecast, multiple PV forecasting, ultra-short-term PV forecasting, prediction interval, optical flow, light gradient boosting machine (LGBM), hybrid model.

ABBREVIATIONS AND ACRONYMS

PV	Photovoltaic.
IEA	International Energy Agency.
NWP	Numerical weather prediction.
ConvLSTM	Convolutional long short-term memory.
GCLSTM	Graph convolutional long short-term memory.
GCTrafo	Graph convolutional transformer.
SVR	Support vector regression.
LSTM	Long short-term memory.
LGBM	Light gradient boosting machine.
MAPE	Mean absolute percent error.
ANN	Artificial neural networks.
MAE	Mean absolute error.
CRPS	Continuous ranked probability score.
NV	Normalized value of PV generation.
GOSS	Gradient-based one-side sampling.
EFB	Exclusive feature bundling.
CART	Classification and regression trees.
GBDT	Gradient boosted decision trees.
MSE	Mean squared error.
RMSE	Root mean squared error.
PICP	Prediction interval coverage probability.
PINAW	Prediction interval normalized averaged width.
XGB	Extreme gradient boosting.
AR	Autoregressive.

The associate editor coordinating the review of this manuscript and approving it for publication was Giovanni Angiulli.

SYMBOLS AND MATHEMATICAL NOTATIONS

OPTICAL FLOW

y_t	PV generation at time t.
$y_{2weeks(t)}$	Highest PV generation in the past two weeks at time t.

$u(x, y, t)$ Normalized velocity vector component of PV generation in the latitude direction at time t, where x and y denote longitude and latitude coordinates.

$v(x, y, t)$ Normalized velocity vector component of PV generation in the latitude direction at time t, where x and y denote longitude and latitude coordinates.

$f(x, y, t)$ Normalized PV generation of the mesh at coordinates (x, y) and time t, where x and y denote longitude and latitude coordinates.

E_D Data term representing the disparity between predicted and actual generation.

E_s Smoothness term, enforcing the smoothness constraint on the motion vector field.

λ Regularization parameter, balancing the impact of the smoothness term relative to the data term.

J Energy function combining the data and smoothness terms.

LIGHT GRADIENT BOOSTING MACHINE

$obj(s)$ Objective function at iteration s of LGBM.

$\hat{y}_t(s-1)$ Predicted PV generation at the $(s-1)$th iteration of LGBM.

$f_s(x_t)$ Mapping function of CART generated at iteration s, where x_t denotes the feature vector at time t.

$\Omega(f_s)$ Regularization term for the CART at iteration s.

S Total number of iterations in LGBM.

l Loss function used for calculating the error.

$\widehat{y}(x)$ Predicted PV generation for feature vector x.

HYBRID MODEL(FLOW-LGBM)

$\boldsymbol{E}_t$ Set of absolute errors at time t.

$e^i_{t_j,m}$ Forecasting error for the i-th day at time t for PV with ID j using Model m.

$y^i_{t_j}$ PV generation observed at time t for the i-th day at a PV with unique ID j.

$\widehat{y^i_{t_j,m}}$ PV power generation for the i-th day at time t at PV with a unique ID j, using model m.

$\boldsymbol{D}_{t_j}$ Set of forecast values considering error distribution for PV with ID $\ddot{j}$ at forecast time t.

METRIC

N Number of verification samples.

$\hat{y}_t$ Predicted PV generation at time t.

ε_t Indicator function: equals 1 if the observed value falls within the prediction interval, 0 otherwise.

L_t Lower bound of the prediction interval at time t.

U_t Upper bound of the prediction interval at time t.

$\hat{F}(y_t)$ Predicted cumulative distribution function at time t.

$\mathbf{1}(y-y_t)$ Heaviside function, which is 1 if $y-y_t$ is nonnegative and 0 otherwise.

I. INTRODUCTION

Distributed photovoltaic (PV) power generation enhances local energy productivity by adapting to diverse regional conditions and offers greater flexibility than traditional concentrated PV power generation. This lays the groundwork for sustainable energy supply [1]. The International Energy Agency (IEA) [2] reports that by 2022, distributed PV installations accounted for approximately 40% of the 1 terawatt (1 TW) of the PV capacity installed globally, with residential installations making up over one-third of that. Approximately 25 million PVs (PVs) are installed in homes worldwide and generate an output of approximately 130 gigawatts (130 GW). This number is expected to increase to 100 million units by 2030 units.

TABLE 1 presents the timescale and importance of PV output forecasting. In narrow areas, distributed PV power-generation systems are installed on individual rooftops and facilities, and their outputs are affected by local weather conditions. Thus, short-term power fluctuations are more extreme than concentrated PV power generation, requiring precise data analysis and accurate forecast models for output forecasting, especially within a 4-h time frame, known as ultra-short-term forecasting [3]. The fact that forecasting models are inherently uncertain has increased the demand for probabilistic forecasts [4], [5], [6]. These forecasts consider the uncertainty linked to errors as opposed to deterministic forecasts that rely on point estimates.

Many forecasting models have been developed to achieve higher accuracy in forecasting PV generation [7], [8], [9]. TABLE 2 summarizes relevant studies on ultra-short-term forecasts for PVs. These models use meteorological data, such as cloud images, satellite images, weather data, and numerical weather prediction (NWP), in addition to PV generation and location data. For ultra-short-term forecasts, in which the forecast period is less than 4 h, cloud generation, dissipation, and movement are the basic factors that generate PV power output variability. Consequently, studies have explored the use of machine learning [10], [11], [12]and image-processing models [8], [13] that use cloud images. These studies typically involve the acquisition of cloud images from ground-mounted cameras. However, the installation and maintenance of such cameras in many distributed PV installations are expensive. Using satellite imagery [14], [15] for forecasting is more suitable for concentrated PVs but may not be ideal for distributed ones because of the large scale and low resolution of satellite images. When using weather data [16], [17] instead of cloud images, distributed PVs often do not have a multisensor local weather monitoring system, unlike concentrated PVs. Therefore, data were collected from weather stations near the PV generation

TABLE 1. PV generation forecasting periods and their importance.

Forecasting period	Forecast lead time	Critical factor	Data	Mainstream algorithm
Ultra-short-term forecast	0–4 h	cloud generation, dissipation, and movement	cloud images, weather data, satellite images, historical PV generation	Machine learning, image processing, persistent forecast model, statistical model
Short-term forecast	4–72 h	cloud movement, weather front changes, atmospheric conditions	satellite images, NWP, weather data, historical PV generation	Machine learning, image processing, statistical model
Medium and long-term forecast	72 h to 1 year	seasonal variations, climatic changes, atmospheric trends	satellite images, NWP, weather data, historical PV generation	Machine learning, statistical model

TABLE 2. Ultra-short-term pv generation forecasting techniques.

Reference	Time interval	Methodology	Input data	Country
[10]	5 min	Radial basis function (RBF) neural network	Historical PV generation, cloud images, environmental temperature, surface radiation	China
[12]	5–30 min	Semi-supervised density-based spatial clustering of applications with noise (DBSCAN)	Historical PV generation, sky image	America, Australia
[14]	30–60 min	Graph neural network	Historical PV generation, satellite images, clear-sky irradiance, temperature, humidity, wind speed	Australia
[15]	1–5 h	Optical flow	Satellite images, global horizontal irradiance	America
[16]	1 h	Convolutional neural network	Historical PV generation, historical weather data	China
[17]	1, 5, and 50 min	Uncertain Basis Functions Method, Stochastic State-Space Models	Historical PV generation, solar radiation	America
[24]	1 h	ConvLSTM	Historical PV generation of 56 units,	America
[25]	15–180 min	GCLSTM	Historical PV generation of 304 units,	Switzerland
[26]	30 min	Optical flow	Historical PV generation of 5096 units,	Japan

station and used for output forecasting [18]. For instance, the Japan Meteorological Service Support Center which is affiliated with the Japan Meteorological Agency, a government agency that offers real-time, minute-by-minute weather data from 155 weather stations and special regional weather stations nationwide. These data include information on the location, rain, wind, temperature, sunshine, solar radiation, snow cover, air pressure, humidity, and visibility, which is available for approximately ten thousand yen per month (the exact amount varies depending on the communication facilities, data formats, and other conditions). In distributed PVs, weather observation points provided by weather stations are not always in close proximity to PV installations. This implies that accurate data on the actual weather conditions may not always be available, which could lead to issues with the forecast accuracy. NWP is often used when there is no meteorological observatory near a PV installation or when forecasting mid- to long-term periods. However, it is essential to note that the NWP [19], [20] is generally considered less reliable than field observations and is typically valid only for forecast steps longer than 4 h [21].

The recent increase in distributed PV installations has led to the availability of geographically dispersed time-series data from numerous sites. This has sparked interest in enhancing the accuracy of variable-generation forecasts, also known as spatiotemporal forecasts [22], [23]. These forecasts leverage information from neighboring sites as additional features. Multi-point data from distributed PV installations are promising for tracking pseudo-cloud movements. For example, when a cloud moves into an area, it affects PV generation in that area and adjacent areas. Therefore, it is possible to simulate cloud movement in a pseudo-manner based on changes in the geographic distribution of power generation across a broad adjacent area. Machine-learning models that rely solely on multi-point data without meteorological data have been reported in the past. Chai et al. [24] conducted a multi-site forecast that captured the spatiotemporal characteristics of many PVs using convolutional long short-term memory (ConvLSTM). ConvLSTM combines a convolutional layer that captures spatially local translational invariance through convolution operations with LSTM. This combination allows the model to store long-term time dependence. It uses data from two-dimensional (2D) images of geographic PV distributions over time as features.

Simeunović et al. [25] conducted a multi-site forecast that captured spatio-temporal characteristics for numerous PVs. They employed two models, the graph convolutional long short-term memory (GCLSTM) and graph convolutional transformer (GCTrafo), which utilize data from a graph structure of PV power generation at multiple locations. These models relied solely on the generation of historical data. They outperformed the single-site forecast accuracy of support vector regression (SVR) and LSTM using NWP as the input for forecasts up to 4 h ahead. SVR and LSTM utilize NWP data, which typically have lower spatial and temporal resolutions than graphical data representing multi-point generation, a key feature of GCLSTM and GCTrafo. Thus, the accuracy of the ultra-short-term forecasts is lower. The main advantages of machine learning models, especially deep learning-based models, such as ConvLSTM, GCLSTM, and GCTrafo, are their flexibility and scalability to data. These models can effectively capture nonlinear and complex spatiotemporal patterns, enabling efficient modeling of the interactions and dependencies between different PV. The disadvantage is that complex machine-learning models have many parameters and require large datasets over a long period to properly adjust the parameters and learn the relationships between the PVs. For instance, the abovementioned study by Chai et al. [24] utilized multi-point PV generation at 15-min intervals, covering 10 h a day for approximately 10 months, to train the ConvLSTM. Additionally, the study by Simeunović et al. [25] used GCLSTM, GCTrafo models trained on data at 15 min, 24 h a day, for one year.

This study focuses on optical flow [26], [27] as a model for output forecasting that relies solely on multi-point data without meteorological data. Optical flow is an image-processing technique that estimates and forecasts the geographic distribution of PV generation based on multi-point data. Unlike traditional optical flow [13], [15], which used cloud or satellite imagery, the optical flow in this study estimates the power generation trends at each mesh and forecasts future power generation using images of the power generation distribution meshed by the latitude and longitude. In the context of smart grids, where the PV output power can be monitored, it is possible and cost-effective to forecast the PV output at multiple locations without the need for meteorological data or additional equipment. Optical flow motion estimation and prediction assume that the mesh distribution of the PV generation undergoes constant-velocity linear motion between two consecutive times. Therefore, unlike machine-learning models that require long-term data, optical flow can forecast the PV output at multiple locations using only multi-point data at two different times. In a previous study [26], the accuracy of 30-min-ahead forecasting for 96 PV installations was compared between artificial neural networks (ANNs) and optical flow during high output variability when large errors are likely to occur in PV output forecasting. Optical flow showed a 20.8% improvement over ANNs in the mean absolute percent error (MAPE). The ANN was trained with data collected every 30 min for 11 h a day for approximately ten months, using the month, day, time, solar radiation, and temperature as the inputs. Therefore, optical flow is an effective forecasting model that uses only multi-point data twice and does not require meteorological data. However, optical flow can cause approximation errors for two reasons. First, during the meshing of the geographic distribution of power generation and generation of image data, interpolation was performed on meshes without PV. The interpolation accuracy tends to be low in areas where the PV is not dense. Second, the transition of the power generation distribution (pseudo-cloud motion) is assumed to be invariant and in a constant-velocity linear motion when forecasting future power generation using an optical flow. However, actual clouds repeat sudden onsets and disappearances and exhibit complex nonlinear motions.

Owing to the unavailability of meteorological data for many distributed ultrashort-term PV forecasts, various forecasting models such as deep learning and optical flow have been studied. These models rely solely on multi-point data. However, each model has its limitations. Deep learning requires a large amount of distributed PV data over an extended period, whereas optical flow, which requires only two consecutive data periods, is prone to approximation errors. However, if available, optical flow can improve the existing forecast accuracy by utilizing accumulated data over time.

This study proposes a hybrid model that combines optical flow and a light-gradient boosting machine (LGBM) [28]. This model does not use meteorological data and utilizes multi-point data over a relatively short period (one month) to forecast the output of many geographically dispersed PV generation systems 30 min ahead. The proposed method aims to achieve two primary objectives. First, the output of many geographically dispersed PVs is forecasted in 30 min using relatively short-term (one month) multi-point data without relying on weather data and with higher accuracy than conventional forecast models using weather data. Second, the forecast accuracy can be improved by combining machine learning and optical flow to learn the spatiotemporal relationship of PVs and correct the approximation error of the optical flow. The proposed method first utilizes optical flow to forecast the output of distributed PV generation in Japan 30 min in advance. Next, by incorporating the predictions of the PV output by optical flow into the LGBM features, the method captures the spatiotemporal relationships of the PVs and forecasts their output. When forecasting the output, in addition to deterministic forecasting, probabilistic forecasting is employed, as proposed in [29], to provide quantitative information on the uncertainty of PV power generation concerning the forecasted value. This method constructs an error distribution each time based on forecast errors in the training data and uses it to construct a forecast distribution. Because the error distribution is established in advance, only the inference time for the deterministic forecast is required to construct the forecast distribution. This approach has the potential for ultra-short-term forecasting 30 min in advance for many distributed PVs.

The contributions of this study are as follows:

1) *Forecast 390 PVs without using meteorological data such as cloud images or weather data*: The authors proposed a hybrid model based on optical flow and LGBM that does not rely on meteorological data. The model utilizes multi-point data from historical generation over

one month. Our proposed method offers deterministic and probabilistic output forecasting 30 min ahead for 390 distributed PVs.

2) *More stable forecasts with shorter period training data than weather data models:* The accuracy of the deterministic forecasts of the proposed method is compared with the accuracy of forecasting models that require 4.5 months of weather data. Compared to the weather data model, the simulation results show that the proposed method, which learns the spatiotemporal relationship of the PVs by optical flow forecasting, improves the mean absolute error (MAE) by 18.4%, even with one month of training data.

3) *Most stable forecast accuracy among the multi-point data models:* The accuracy of the deterministic and probabilistic forecasts generated by the proposed method is compared to the forecasting accuracy of optical flow and an LGBM model trained on 390 distributed PVs without using optical flow. The simulation results indicated that the proposed method achieved the best forecast accuracy for 259 and 362 of 390 PVs in terms of MAE and the continuous ranked probability score (CRPS) [30], which are evaluation indices for deterministic and probabilistic forecasts, respectively, among the multi-point data models. Furthermore, compared with optical flow, the proposed method improved the average MAE and CRPS of all 390 PVs by 5.8% and 10.8%, respectively.

The remainder of this paper is organized as follows: Section I outlines the optical flow, LGBM, and hybrid model that combines the two. Sections II and III describe the evaluation metrics and datasets, respectively. Section IV compares the weather data model, which is a conventional method, with a multi-point data model that includes the proposed method. Section V compares the multi-point data models and demonstrates the superiority of the proposed method over other models. Finally, the conclusions are presented in Section VI.

II. METHODOLOGY

The authors proposed flow-LGBM, a hybrid model of optical flow and LGBM, which offers 30-min-ahead deterministic and probabilistic output forecasts for 390 distributed PVs. This section presents a brief overview of the optical flow and LGBM and concludes with a description of the flow-LGBM.

A. OPTICAL FLOW

Optical flow is an algorithm that converts distributed PV power generation into a mesh distribution based on the latitude and longitude of each time frame. The movement of the mesh distribution is then estimated. For forecasting purposes, the algorithm assumes that the motion of the mesh distribution between two consecutive time frames is invariant and follows a linear motion at a constant velocity. A flowchart of the optical flow is shown in FIGURE 1; the steps from (i) to (iv) in FIGURE 1 are described below.

(i) Obtain the geographical distribution of normalized PV generation.

To account for the variations in the installation angles and power ratings among the PVs, the power generation of each PV system was normalized by its maximum power generation over two weeks. This normalization process converts the power generation values to a range from 0 to 1, unifying the different PVs uniformly. The conversion of power generation to the normalized value of NV is defined in Equation (1).

$$NV(t) = \frac{y_t}{y_{2weeks(t)}} \tag{1}$$

where y_t denotes the actual output power at a specific time t, and $y_{2weeks(t)}$ represents the highest output power among the 14 data points at a specific time t within the first two weeks (14 days). These two weeks were selected to ensure that at least one sunny day was included, and the seasonal variation in the sun elevation was considered negligible during this timeframe.

(ii) Meshing geographical distribution of PV generation.

The geographic distribution of the normalized generation was divided (meshed) into equally spaced distributions aligned in the latitude and longitude directions.

(iii) Impute missing values in the meshed distribution of PV generation.

An interpolation method using Delaunay triangulation [31] was applied to the missing regions in the meshed PV distribution.

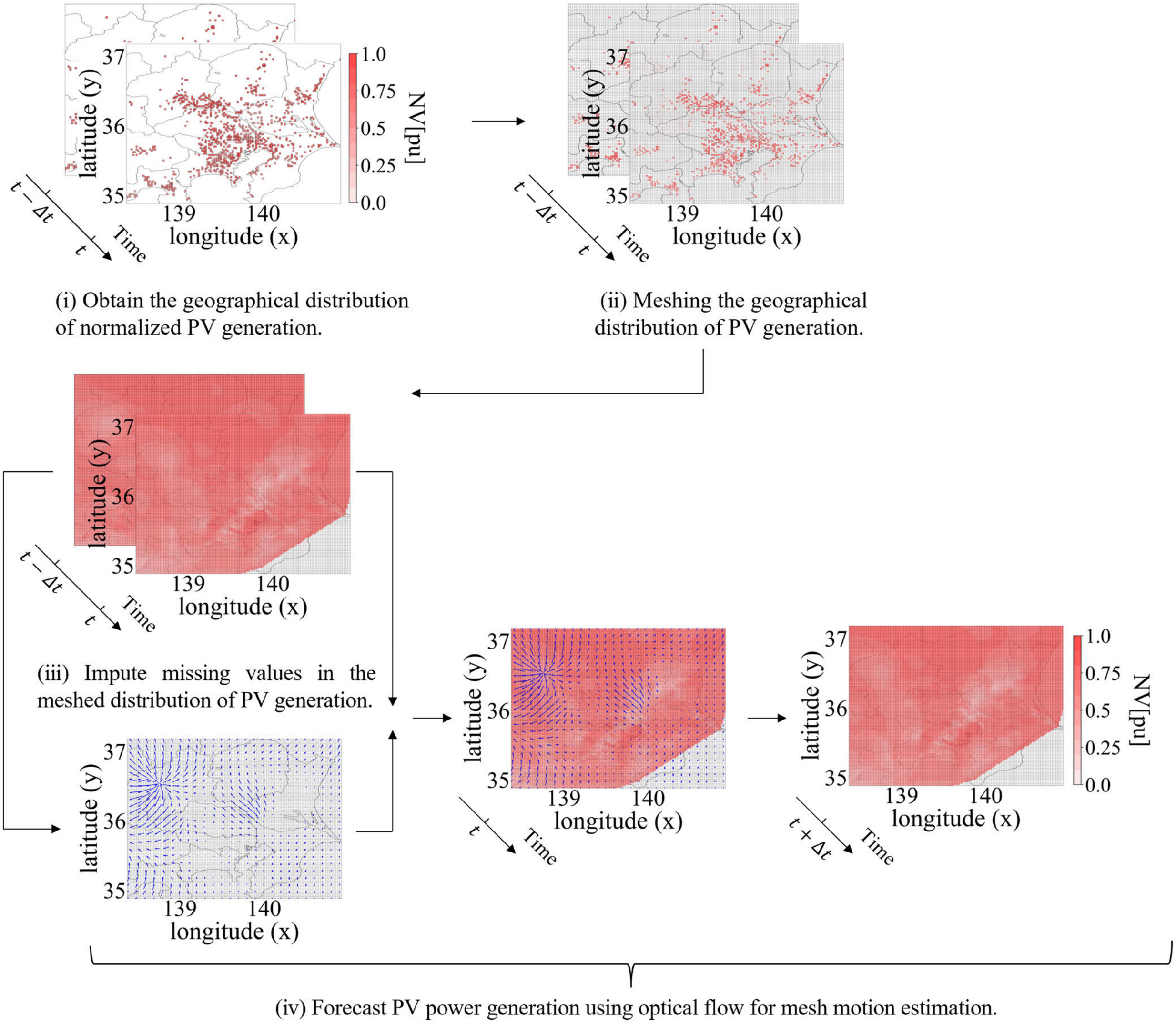
(iv) Forecast PV power generation using optical flow for mesh motion estimation.

This method assumes that the geographic distribution of normalized values changes over time and estimates the motion vector field of these values using an optical flow for prediction. Assuming that the normalized values remain constant and undergo linear motion at a constant velocity between two consecutive times and denoting the normalized value (NV) of each mesh at time t as $f(x, y, t)$, the relationship between the normalized values at time t and $t - \Delta t$ is expressed by Equation (2).

$$f(x - u(x, y, t)\,\Delta t - v(x, y, t)\,\Delta t, t - \Delta t) = f(x, y, t) \tag{2}$$

where x and y correspond to longitude and latitude values in each normalized coordinate, respectively, and $u(x, y, t)$ and $v(x, y, t)$ represent the normalized velocity of movement in the longitude and latitude directions at time t. In this context, the data term E_D, which quantifies the sum of the squared errors between the positions before and after movement for each mesh at two consecutive time steps, and the regularization term E_s, which enforces the smoothness constraint on adjacent movement vectors, are expressed in Equations (3) and (4), respectively.

$$E_D^2 = (f(x - u(x, y, t)\,\Delta t, y - v(x, y, t)\,\Delta t, t - \Delta t))^2 \tag{3}$$



(i) Obtain the geographical distribution of normalized PV generation.

(ii) Meshing the geographical distribution of PV generation.

(iii) Impute missing values in the meshed distribution of PV generation.

(iv) Forecast PV power generation using optical flow for mesh motion estimation.

FIGURE 1. **Optical flow flowchart.**

$$E_s = \lambda(|\nabla u(x, y, t)|^2 + |\nabla v(x, y, t)|^2) \tag{4}$$

The regularization term E_s is based on the sum of squares of gradients in the longitude and latitude directions, ∇u and ∇v, respectively. This characterizes the spatial rate of change in the velocity distributions $u(x, y, t)$ and $v(x, y, t)$. The parameter λ represents the weight of the regularization term, and a larger λ prioritizes the correlation between adjacent meshes, thereby increasing the uniformity of the predicted vectors among the neighboring meshes. The energy function is defined in Equation (5), where the data term E_D and regularization term E_s, are added‘.

$$J = \iint \left(E_D^2(x, y, t) + \lambda E_s(x, y, t)\right) dxdy \tag{5}$$

The solution (u, v) of the minimization problem in Equation (5) is the velocity vector that minimizes the error in the motion estimation and is derived using the Euler-Lagrange equations in Equations (6) and (7):

$$\lambda \nabla^{\mathrm{T}} \nabla u - \mathrm{E}_D \partial_x f = 0 \tag{6}$$

$$\lambda \nabla^{\mathrm{T}} \nabla v - \mathrm{E}_D \partial_y f = 0 \tag{7}$$

$\nabla^{\mathrm{T}}\nabla = (\partial_x^2 + \partial_y^2)$ is the Laplace operator. Further details of the equation solution and algorithm can be found in Kameda's study [32]. Using the derived velocity vectors (u, v), the normalized value of each mesh is predicted, assuming that the normalized value of each mesh will continue to move in a constant velocity linear motion after Δt. The predicted normalized values were converted to the output of each PV using Equation (1).

B. LIGHT GRADIENT BOOSTING MACHINE

The LGBM is a decision-tree-based gradient-boosting framework that facilitates efficient and scalable learning on large, high-dimensional datasets. The algorithm uses gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB). GOSS is a gradient-based sampling method that removes data instances with small gradients by utilizing only the remaining data to estimate the information gain. EFB bundles mutually exclusive features to reduce their number and employs a greedy algorithm to maintain the model accuracy while reducing the feature count. A series of basic classification and regression trees (CART) are iteratively constructed in the learning process using these methods. The weight parameters for each classifier were calculated to form a model that minimized the objective function. The objective function is expressed in Equation (8), and the final model is expressed in Equation 9).

$$obj\,(s) = \sum_{t=1}^{n} l\left(y_t, \hat{y}_t\,(s-1) + f_s(x_t)\right) + \Omega(f_s) \tag{8}$$

$$\hat{y}(x) = \sum_{s=1}^{S} f_s(x) \tag{9}$$

In these equations, x_t is the feature vector at time t, and y_t corresponds to the actual output at that same time. The term $\hat{y}_t\,(s-1)$ denotes the prediction result for x_t at iteration step $s-1$ within the additive training process of the LGBM. Furthermore, f_s is described as a new CART generated at the s-th iteration, which is responsible for mapping a particular training sample x_t to its corresponding leaves. The function l is utilized to calculate the squared error for each data sample, and $\Omega(f_s)$ acts as the regularization term to prevent over-fitting of the new CART. The predicted output $\hat{y}\,(x)$, emerges as the sum of the predictions from all iterations and is denoted as S in the LGBM. Each iteration contributes to a new prediction $f_s\,(x)$ based on feature x through a newly generated CART. This results in the final prediction being the aggregate of the individual predictions across all iterations. By employing these methods, the LGBM speeds up the gradient-boosted decision tree (GBDT) learning process by up to 20 times or more, providing superior performance on large, high-dimensional data.

C. HYBRID MODEL(FLOW-LGBM)

The flow-LGBM utilizes optical flow predictions as features for deterministic forecasting. It also generates error distributions for probabilistic forecasts based on error data from the learning process. This approach improves the forecasting accuracy of existing optical flows by correcting errors resulting from the interpolation of missing mesh values. Additionally, it addresses errors arising from the assumption that the mesh motion (pseudo-cloud motion) is invariant and follows a constant-velocity linear motion. This correction was achieved by learning the spatiotemporal relationships of multi-point PVs through optical flow predictions. The LGBM flowchart is shown in FIGURE 2, and steps (i) to (v) in FIGURE 2 are described below.

(i) Optical flow prediction

The power outputs 30 and 60 min before the distributed PV generation forecast time were extracted from the database. Subsequently, using the optical flow motion estimation, the PV generation output was forecasted for the next 30 min.

(ii) Train the models and search for parameters using a grid search.

The LGBM was trained using five-fold time series cross-validation. These features included the predicted optical flow values, past power generation, and variables related to the location and time. The target variable was the power generation at the forecast time. To ensure uniformity, the power generation data were normalized using the maximum power generation of each PV system for the training period, scaling it such that the maximum value was 1. A grid search was conducted to minimize the out-of-fold MAE for the parameter selection.

(iii) Derive error distribution from training result.

The absolute error set $\boldsymbol{E}_t$ for a specific time, t was obtained by comparing the predicted data with the observed data from the training results. The prediction error during the model training was calculated using Equation (10)

$$e^i_{t_j,m} = y^i_{t_j} - \widehat{y^i_{t_j,m}} \tag{10}$$

where $e^i_{t_j,m}$ denotes the forecasting error for the i-th day at time t in the training set; j represents the unique ID of the PV system; and m represents the model number. From step (ii), one model is obtained for each fold, resulting in a total of five models. This error quantifies the disparity between the actual observed value $y^i_{t_j}$ and predicted value $\widehat{y^i_{t_j,m}}$ for the corresponding day and time. From Equation (10), the absolute error set $\boldsymbol{E}_t$ is expressed as Equation (11).

$$\boldsymbol{E}_t := \left\{ e^i_{t_j,m} | \quad \text{for all } i, j, m \right\} \tag{11}$$

The set $\boldsymbol{E}_t$, as defined in Equation (11), provides a comprehensive overview of the forecasting errors throughout the dataset for each time instance. This set served as the basis for constructing a histogram at each time point, denoted as the error distribution in the context of this document.

(iv) Forecast deterministic PV generation by the trained models.

The five trained models from step (ii) were used as inputs, along with the predicted optical flow values, power generation in the past 30 min, latitude, longitude, and time, to forecast the power generation of the PV at multiple locations. To combine the predictions of each model, equal weights (weight factor of 1/5) were assigned to each model, and the predictions were averaged.

(v) Make prediction intervals from error distribution and deterministic forecasting.

To construct the prediction distribution, a set of predictions is calculated by adding the error distribution obtained in step (iii) to the predictions for each PV obtained in step (iv). For example, when the ID of the PV to be forecasted is $\ddot{J}$, the forecast time is t, and the forecast value is $\widehat{y_{t_{\ddot{j}}}}$, the set

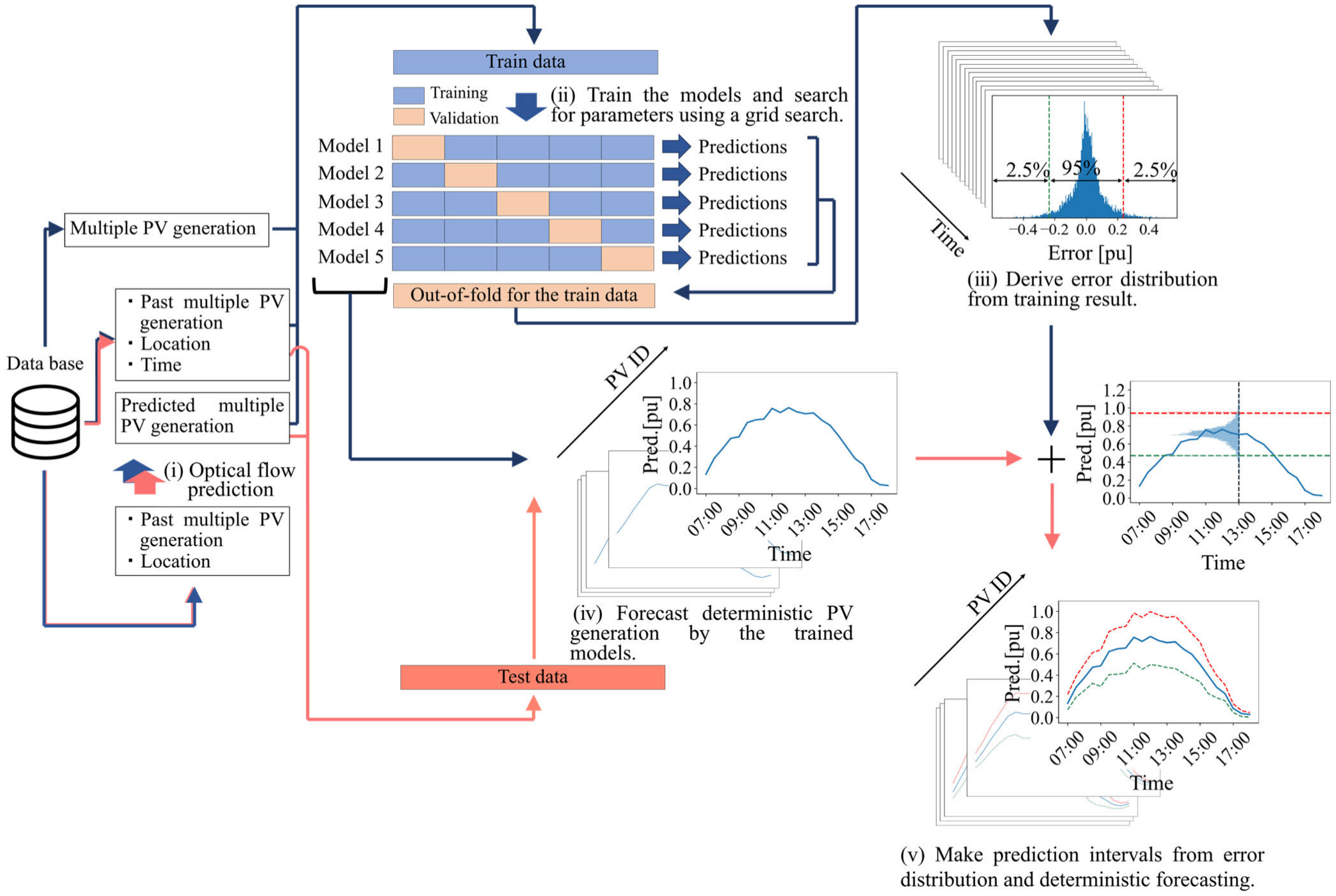


FIGURE 2. Flow-LGBM flowchart.

of forecast values $\boldsymbol{D}_{t_j}$ considering the error distribution is expressed in Equation (12).

$$\boldsymbol{D}_{t_j} := \left\{ \widehat{y_{t_j}} + e^{i}_{t_j,m} | \text{ for all } i, j, m \right\} \tag{12}$$

As a simple post-processing of set $\boldsymbol{D}_{t_j}$, the elements comprising $\boldsymbol{D}_{t_j}$ are corrected to 0 if they are below 0 and to 1 if they are above 1. These adjustments ensure that the generation output falls between zero and one. To output a 95% confidence interval, the upper and lower 2.5% points of the set $\boldsymbol{D}_{t_j}$ were defined as the upper and lower limits of the prediction intervals, respectively.

III. METRICS

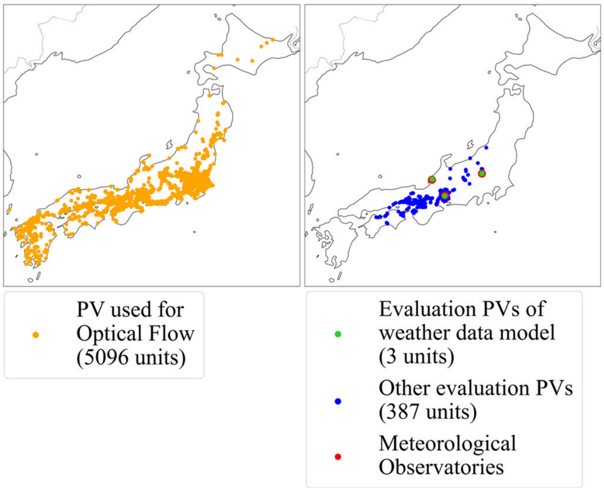
This section describes the metrics used to evaluate the deterministic and probabilistic forecasts. For deterministic forecasts, metrics such as the mean squared error (MSE), root mean squared error (RMSE), and MAE are commonly used to assess the error between the predicted values and observed data. In this study, the MAE, which is relatively robust against outliers, was employed as an evaluation metric. The MAE is defined by Equation (13).

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^{N} |\widehat{y_t} - y_t| \tag{13}$$

where N is the number of verification samples, $\hat{y}_t$ and y_t are the point forecast and the corresponding actual value at time t, respectively.

In probabilistic forecasting, it is crucial to assess the reliability and sharpness of the forecast distribution simultaneously. Reliability assesses the proximity between the predicted and observed distributions, whereas sharpness assesses the sharpness of the peak within the predicted distribution. Both criteria should be evaluated simultaneously because excessively sharp predictions may lack reliability, making it impossible to characterize them uniquely as good prediction intervals without evaluating reliability. Probabilistic forecasting aims for a predictive distribution to be as sharp as possible if well-calibrated [33], [34]. In this study, the authors used the prediction interval coverage probability (PICP) as a measure of reliability, defined by Equation (14).

$$\text{PICP} = \frac{1}{N} \sum_{t=1}^{N} \varepsilon_t \tag{14}$$


PV used for Optical Flow (5096 units)
Evaluation PVs of weather data model (3 units)
Other evaluation PVs (387 units)
Meteorological Observatories

(a) Optical flow (b) Flow-LGBM

FIGURE 3. Locations of PVs used in the evaluation of each forecast model, and meteorological stations.

where ε_t is an indicator function, which takes the value of 1 if the observation at time t falls between the lower (L_t) and upper (U_t) bounds, and 0 otherwise. Additionally, to assess the sharpness, the authors utilized the prediction interval-normalized averaged width (PINAW). PINAW represents the average width of the prediction intervals at nominal probability and is defined by Equation (15). A smaller PINAW indicated a sharper prediction.

$$\text{PINAW} = \frac{1}{N}\sum_{t=1}^{N}(U_t - L_t) \tag{15}$$

As mentioned previously, a good prediction interval maximizes the sharpness of a prediction based on the assumption of high reliability, and the two indicators must be evaluated simultaneously. This study used the continuous ranked probability score (CRPS) to evaluate the two indices in a unified manner. CRPS is defined by Equation 16.

$$\text{CRPS} = \frac{1}{N}\sum_{t=1}^{N}\int_0^1 \left(\hat{F}\left(y_t\right) - \mathbf{1}\left(y - y_t\right)\right)^2 dy \tag{16}$$

where $\mathbf{1}$ represents the Heaviside function, which is 1 if its argument is nonnegative and 0 otherwise. CRPS is consistent with MAE at extremes, where the uncertainty in the probabilistic forecast decreases and approaches deterministic forecasting. Similar to MAE, CRPS is less influenced by outliers.

IV. DATASET DESCRIPTION

The PVs used for the optical flow in this study were located throughout Japan, totaling 5096 units, as shown in FIGURE 3(a). In the Flow-LGBM, 390 PVs, represented by blue and green dots, were evaluated for learning and testing, as shown in FIGURE 3(b). These 390 PVs were selected from 5096 PVs, and all data were available over the test period. The three green dots in FIGURE 3(b) indicate the three PVs evaluated in Section IV, whereas the three red dots indicate the locations of the weather stations. Each weather station was located within 4 km of the PVs denoted by green dots, and the weather data obtained from these stations were used in the weather data model as the comparison model for the Flow-LGBM in Section IV. The correlation coefficients between the solar radiation measured at each weather station and the PV generation of the respective green dots is greater than 0.85 during the test period. The test period spanned four months, from January 2014 to April 2014, with data collected every 30 min from 7:00 AM to 6:00 PM.

V. WEATHER DATA MODELS VS MULTI-POINT DATA MODELS

A. MODEL DESCRIPTION

This section presents a comparison of weather data models with multi-point data models that utilize generation data from multiple PVs instead of weather data. For the weather data model, the authors provided models trained for one month, four and a half months for machine learning, and one month for the multi-point data model. The authors aimed to determine whether the forecast accuracy of the multi-point data model was comparable to that of the conventional weather data model within a shorter training period. TABLE 3 and 4 outline each model and the data used. The three PVs indicated by the green dots in FIGURE 3 (b) are the evaluation targets, with the weather data for each PV system obtained from the meteorological stations. Notably, the weather data used here are real-time observational data, which tend to result in smaller forecasting errors than past weather data or NWP data. In TABLE 4, the forecast models that use weather data include XGB (1-month), XGB (4.5-month), LGBM (1-month), and LGBM (4.5-month), which utilize variables related to time, 30 min prior generation, and weather data such as solar radiation. Each forecasting model trains one model for each PV system. Specifically, XGB (1-month) and LGBM (1-month) use data up to one month before the test data as training data, whereas XGB (4.5-month) and LGBM (4.5-month) use data up to 4.5 months before the test data as training data. For example, when the test data are from January 1, 2014, to January 31, 2014, XGB (1-month) and LGBM (1-month) use data from December 1 to December 31, 2013, as training data, and XGB (4.5-month) and LGBM (4.5-month) use data from August 15, 2013, to December 31, 2013. In TABLE 4, the forecasting models using multi-point PV power generation are optical flow, multi-LGBM, and flow-LGBM. The optical flow forecasts the generation of the 5096 PVs shown in FIGURE 3 (a), using data meshed based on the latitude and longitude of the generation 30 and 60 min before the forecast time. The multi-LGBM uses variables related to time, historical power generation of the 390 PVs, unique PV features, and data specific to each PV, such as the location and rated power, as shown in TABLE 3 and 4. Flow-LGBM uses the variables used by the multi-LGBM and the predictions of each of the 390 PVs by

TABLE 3. Feature abbreviation.

Abbreviation	Features
Base features	Hour sine, hour cosine, daily sine, daily cosine, month sine, month cosine, annual sine, annual cosine, two weeks max generation, generation 30 min prior
PV unique features	PV ID, latitude, longitude, rated power, maximum generation
Weather features	Temperature, sunshine hours, humidity, precipitation, solar radiation
Optical flow features	Predicted generations for 390 PVs by optical flow

TABLE 4. Model description.

Model name	Model Type	Features	Data period	Forecasted PV installations
Persistence	baseline model	generation 30 min prior	30 min	1
AR	statistical model	generation 30 min prior	Four and a half months	1
XGB(1-month)	weather data model	base features, weather features	one month	1
LGBM(1-month)	weather data model	base features, weather features	one month	1
XGB(4.5-month)	weather data model	base features, weather features	Four and a half months	1
LGBM(4.5-month)	weather data model	base features, weather features	Four and a half months	1
Optical Flow	multi-point data model	latitude, longitude, generation 30 and 60 min prior (5096 PVs)	one hour	5096
Multi-LGBM	multi-point data model	base features, PV unique features (390 PVs)	one month	390
Flow-LGBM	multi-point data model	base features, PV unique features, optical flow features (390 PVs)	one month	390

TABLE 5. Hyperparameter.

Model	Hyperparameter
LGBM	objective : regression eval metric : mae learning rate : 0.05 num boost round : 100000 early topping rounds : 100 max depth values : [4, 6, 8] subsample values : [0.7, 0.8, 0.9] colsample bytree values : [0.8, 0.9, 1.0]
XGB	objective : regression eval metric : mae learning rate : 0.05 num boost round : 100000 early stopping rounds : 100 max depth values : [4, 6, 8] subsample values : [0.7, 0.8, 0.9] colsample bytree values : [0.8, 0.9, 1.0]

optical flow. The multi-LGBM and flow-LGBM have one month of training data and forecast 390 PVs using a single model. The hyperparameters of the machine learning model were grid-searched and updated monthly, as presented in TABLE 5. In addition to weather data models and multi-point data models, the authors added a persistent forecasting model that uses the power generation 30 min prior as the forecast value and an autoregressive (AR) model, which is a statistical model.

B. RESULT

TABLE 6 and FIGURE 4 present the MAE for each forecast model when evaluating the three PV systems. Flow-LGBM demonstrated the highest forecast accuracy in PV 1 and PV 3, whereas, in PV 2, it ranked second after optical flow. The average MAE across the three PVs for the flow-LGBM improved by 37.7%, 16.9%, 25.7%, 18.4%, 17.0%, and 5.6% compared with persistence, AR, XGB (4.5-month), LGBM (4.5-month), multi-LGBM, and optical flow, respectively. The statistical significance was confirmed using a t-test, with p-values of less than 1% in all comparisons. Optical flow shows the second-best forecast accuracy in the average MAE across the three PVs but performs less accurately than AR, multi-LGBM, and flow-LGBM at PV1. The multi-LGBM ranks fourth in terms of average MAE forecast accuracy, but its performance is significantly lower in PV2. The flow-LGBM enhances the forecasting accuracy of both the optical flow and multi-LGBM by combining the optical flow and LGBM. Regarding weather data models, LGBM outperformed XGB when comparing models trained for the same period. Both models exhibited improved forecast accuracy with a training period of 4.5 months compared with a training period of one month.

The authors analyzed the relationship between the importance of each feature and the forecast accuracy for the LGBM (1-month), LGBM (4.5-month), multi-LGBM, and flow-LGBM using the LGBM feature importance. TABLE 7 lists the correlation coefficients between the features used in the weather data models and generation, representing the average correlation across the three PVs. FIGURE 5 illustrates the relative importance of each feature, as evaluated by the LGBM feature importance assessment. As shown in FIGURE 5(a), the importance of each feature is nearly identical for LGBM (1-month) and LGBM (4.5-month). Both models assigned high importance to features such as the two-week maximum generation, 30 min prior generation, and solar radiation, which exhibited high correlation coefficients,

TABLE 6. MAE of each model for three PVs.

PV ID	Persistence	AR	XGB(1-month)	LGBM(1-month)	XGB(4.5-month)	LGBM(4.5-month)	Optical Flow	Multi-LGBM	Flow-LGBM
1	0.0634	0.0454	0.0761	0.0682	0.0621	0.0568	0.0484	0.0439	**0.0393**
2	0.0711	0.0518	0.0545	0.0466	0.0461	0.0427	**0.0363**	0.0524	0.0387
3	0.0562	0.0459	0.0531	0.0473	0.0519	0.0462	0.0413	0.0470	**0.0409**
All	0.0636	0.0477	0.0613	0.0540	0.0534	0.0486	0.0420	0.0478	**0.0396**

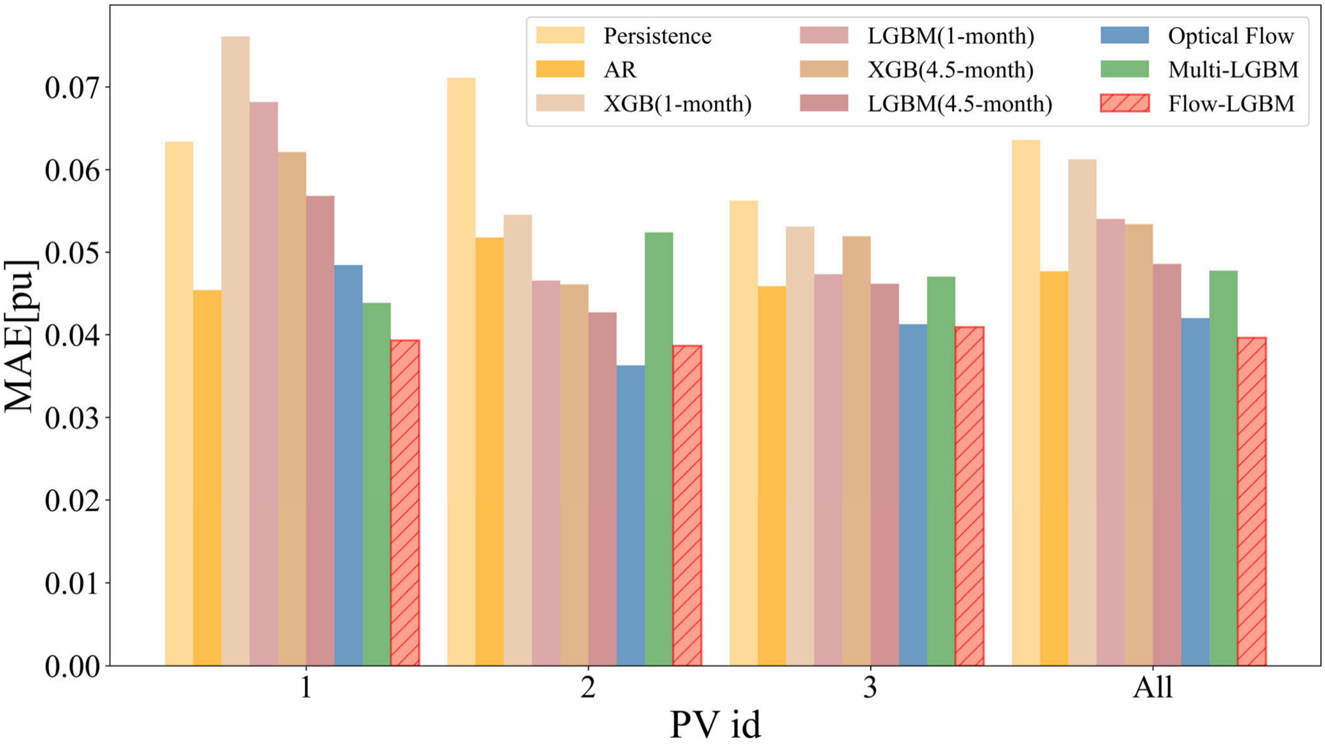
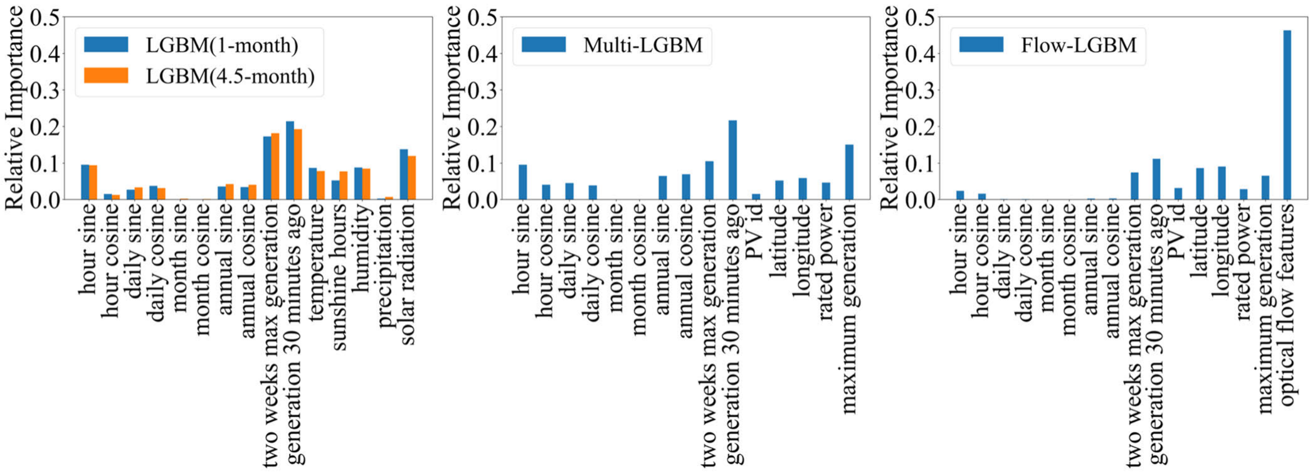


FIGURE 4. MAE [pu] of three PVs for each forecast model.

as shown in TABLE 7. Whereas the correlation coefficient assesses the linear relationship of each feature with the target variable, the LGBM feature importance evaluates the linear and nonlinear relationships between the target variable and features and the interaction between features. Therefore, there may be a slight discrepancy between the correlation coefficient and feature importance, as evaluated by the LGBM. FIGURE 5(b) shows that the multi-LGBM assigns high importance to features such as generation 30 min prior and maximum generation. Conversely, FIGURE 5(c) shows that Flow-LGBM is highly important to the optical flow features, representing the predicted generations for 390 PVs using optical flow. Furthermore, the Flow-LGBM rates the importance of time-related variables as low. This assessment may stem from the inclusion of time information in optical flow features. Hence, the 17.0% MAE improvement of the flow-LGBM over the multi-LGBM can be attributed to learning spatial-temporal PV relationships through optical flow features. The 25.7% MAE improvement of Flow-LGBM over LGBM (4.5-month) could be attributed to factors related to the weather data quality and their correspondence with the generation, in addition to the utilization of optical flow features in Flow-LGBM.

The authors analyzed the error factors for PV1, which exhibited the lowest forecast accuracy among all the weather data models, based on the relationship between solar radiation (the most important weather data), generation, and the forecast accuracy of the LGBM (4.5-month) and Flow-LGBM. TABLE 8 presents the correlation coefficient between generation and solar radiation and the MAE of the flow-LGBM and LGBM (4.5-month) for PV1. FIGURE 6 shows the relationship between generation and solar radiation and the MAE difference between Flow-LGBM and LGBM (4.5-month). TABLE 8 and FIGURE 6 demonstrate that the LGBM (4.5-month) exhibits better forecast accuracy and less variation in the forecast accuracy than the Flow-LGBM when the correlation between power generation and solar radiation is strong. Conversely, when the correlation is weak, the LGBM (4.5-month) exhibits a lower forecast accuracy and higher variability than Flow-LGBM. Specifically,



(a) LGBM(1-month), LGBM(4.5-month) (b) Multi-LGBM (c) Flow-LGBM

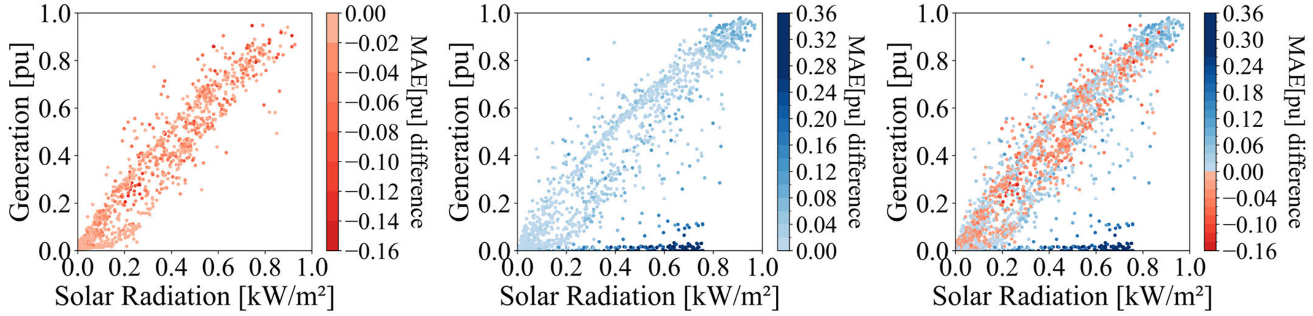
FIGURE 5. LGBM feature importance.

TABLE 7. Correlation coefficients between features used in the weather data model and power generation (average of three PV).

Hour sine	Hour cosine	Daily sine	Daily cosine	Month sine	Month cosine	Annual sine	Annual cosine	Two weeks max generation	Generation 30 minutes prior	Temperature	Sunshine hours	Humidity	Precipitation	Solar radiation
0.208	-0.612	-0.041	-0.021	0.265	-0.292	0.259	-0.291	0.705	0.906	0.380	0.613	-0.531	-0.182	0.894

TABLE 8. Correlation coefficient between generation and solar radiation, and MAE of Flow-LGBM and LGBM(4.5-month) for PV1.

Type	Solar radiation vs generation correlation	Mean MAE (Standard Deviation) for LGBM(4.5-month)	Mean MAE (Standard Deviation) for Flow-LGBM
LGBM(4.5-month)>Flow-LGBM	0.949	0.0302 (0.0431)	0.0515 (0.0534)
LGBM(4.5-month)<Flow-LGBM	0.828	0.0741 (0.0889)	0.0314 (0.0449)
All	0.860	0.0568 (0.0773)	0.0393 (0.0494)



(a) LGBM(4.5-month)> flow-LGBM (b) LGBM(4.5-month)< flow-LGBM (c) All

FIGURE 6. Relationship between generation, solar radiation, and difference in MAE between Flow-LGBM and LGBM (4.5-month).

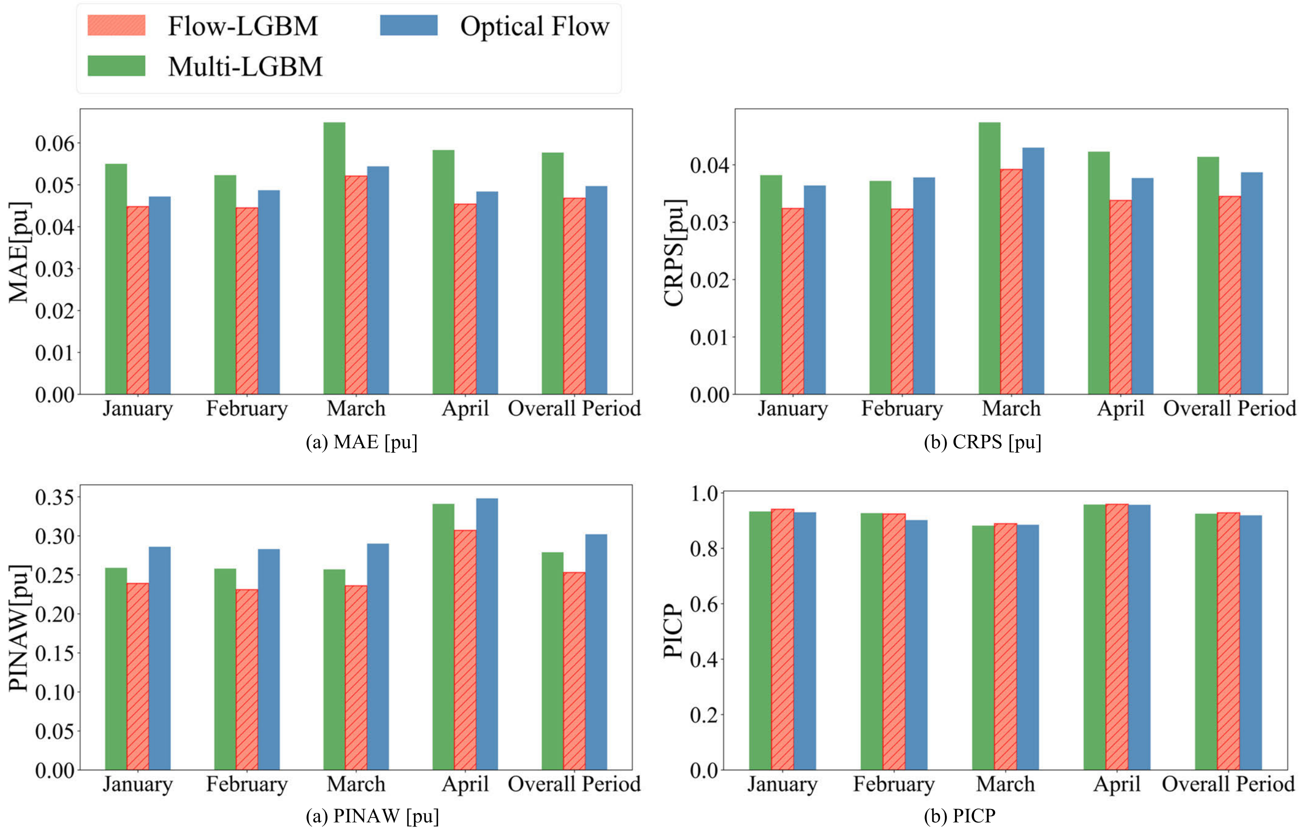
FIGURE 6 (b) depicts conditions where the solar radiation exceeds 0.33 kW/m^2 and power generation is below 0.1 pu; the forecast accuracy of LGBM (4.5-month) significantly lags behind that of the Flow-LGBM. This notable decrease in forecast accuracy is attributed to PV1's location in Maebashi City, Gunma Prefecture, an area with heavy snowfall, where generation can decrease owing to snow covering the PV panels. Thus, the alignment between weather data, such as solar radiation and generation, is crucial for forecasting the accuracy of weather data models. Addressing factors such as data loss and quality due to sensor failure and environmental elements such as snow cover are essential for improving the accuracy of forecast models. In contrast, the Flow-LGBM, a multi-point data model, maintains a stable forecast accuracy even during short learning periods by learning the relationship between spatiotemporal PV patterns, akin to pseudo-cloud motion, using optical flow features.

VI. COMPARISON OF MULTI-POINT DATA MODELS

The authors compared the deterministic and probabilistic forecast accuracies of the 390 PVs evaluated in this study for multi-LGBM, flow-LGBM, and optical flow, all of which are multi-point data models. This study aimed to assess the superiority of the Flow-LGBM over other models. TABLE 9 evaluates the deterministic and probabilistic forecast accuracies of the 390 PVs of the multi-point data model during the test period, as shown in FIGURE 7. TABLE 9 and FIGURE 7 demonstrate that the deterministic and probabilistic evaluation indices MAE and CRPS are superior for the Flow-LGBM compared with the other models for each month. Over the entire period, Flow-LGBM improved the multi-LGBM and optical flow by 18.9% and 5.8% for MAE, and 16.7% and 10.8% for CRPS, respectively. A T-test confirmed these improvements were statistically significant, with p-values of less than 1% for the MAE and CRPS comparisons. FIGURE 8 shows the average forecast accuracy of each PV system for the MAE and CRPS for the entire period. This indicates that for each of the 390 PVs, the flow-LGBM had the best forecast accuracy relative to the multi-LGBM and optical flow, with the MAE and CRPS at 259 and 362 PVs, respectively. Considering the feature importance in FIGURE 5(b) and 5(c),

TABLE 9. Average forecast accuracy for 390 PVs in the multi-point data models.

	MAE [pu]			CRPS [pu]			PICP			PINAW [pu]		
Model Month	Optical Flow	Multi-LGBM	Flow-LGBM	Optical Flow	Multi-LGBM	Flow-LGBM	Optical Flow	Multi-LGBM	Flow-LGBM	Optical Flow	Multi-LGBM	Flow-LGBM
January	0.0472	0.0550	**0.0448**	0.0364	0.0382	**0.0324**	0.930	0.933	**0.941**	0.286	0.259	**0.239**
February	0.0487	0.0523	**0.0445**	0.0378	0.0372	**0.0323**	0.902	**0.927**	0.924	0.283	0.258	**0.231**
March	0.0544	0.0649	**0.0521**	0.0430	0.0474	**0.0392**	0.885	0.882	**0.889**	0.290	0.257	**0.236**
April	0.0484	0.0583	**0.0454**	0.0377	0.0423	**0.0338**	0.957	0.958	**0.959**	0.348	0.341	**0.307**
Overall Period	0.0497	0.0577	**0.0468**	0.0387	0.0414	**0.0345**	0.919	0.925	**0.928**	0.302	0.279	**0.253**



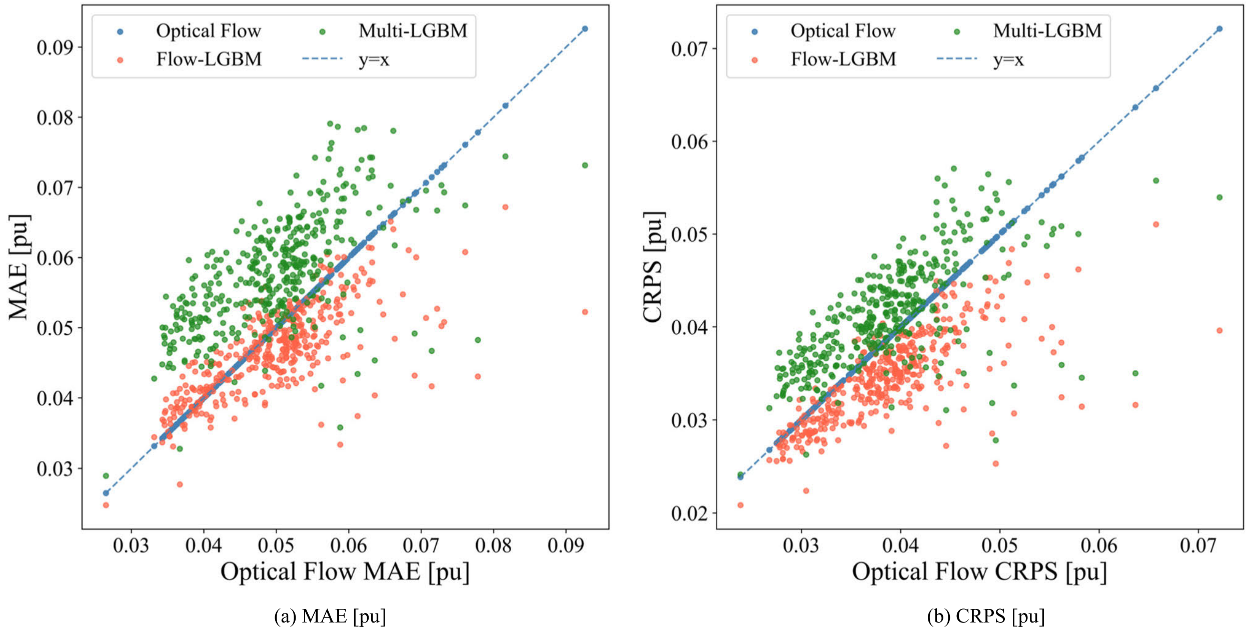
(a) MAE [pu]

(b) CRPS [pu]

(a) PINAW [pu]

(b) PICP

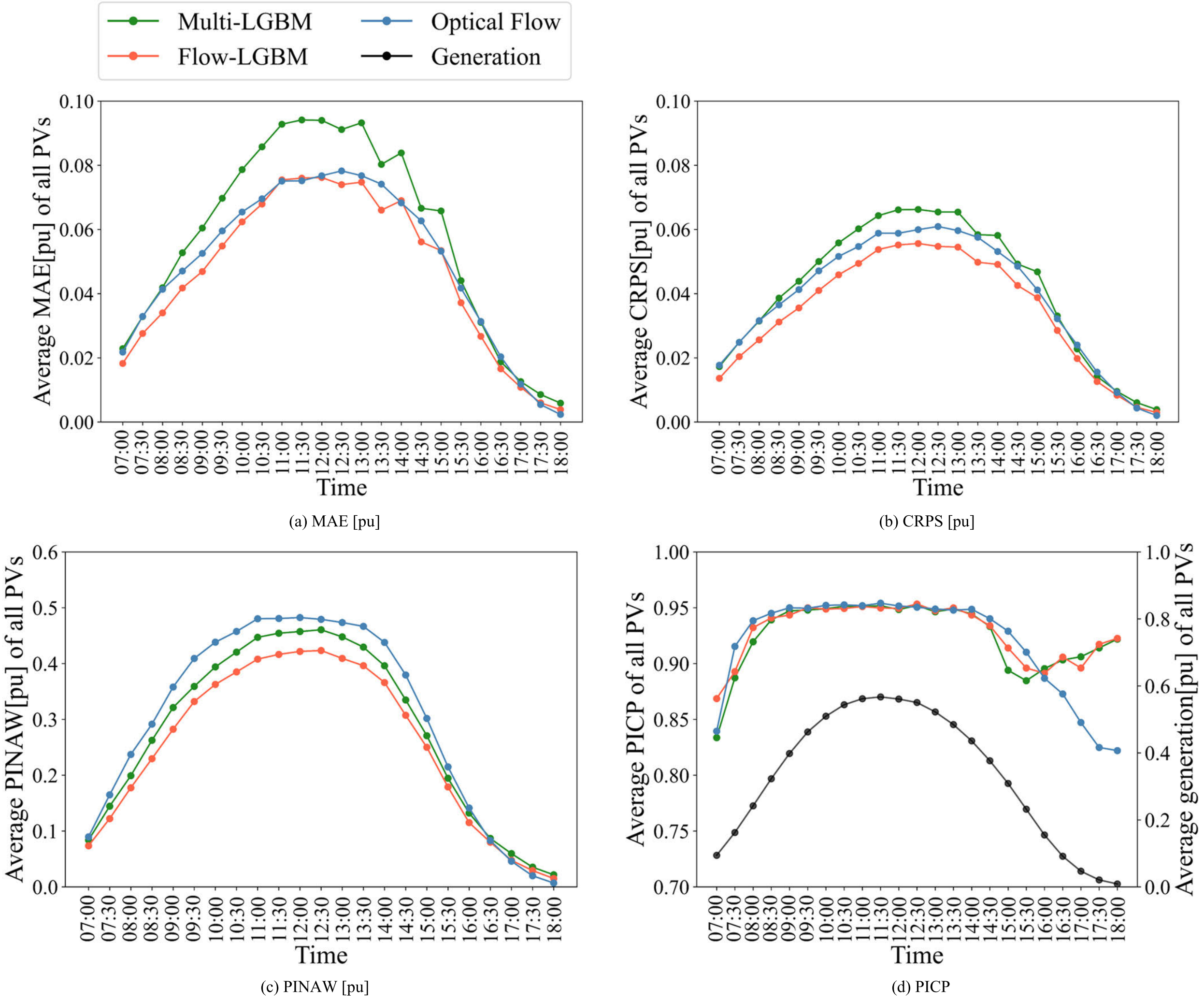
FIGURE 7. Average forecast accuracy for 390 PVs.



(a) MAE [pu]

(b) CRPS [pu]

FIGURE 8. Average forecast accuracy for each PV over the entire period.



(a) MAE [pu]

(b) CRPS [pu]

(c) PINAW [pu]

(d) PICP

FIGURE 9. Average forecast accuracy by time for all PVs.

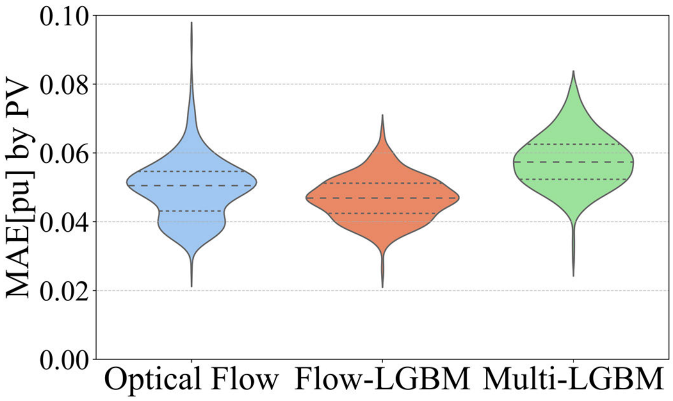


FIGURE 10. Violin diagram of average MAE over the entire period of 390 PVs.

the Flow-LGBM evaluates the importance of the optical flow features. This suggests that learning the spatiotemporal PV relationship through optical flow features can improve forecasting accuracy compared to other models. TABLE 9 and FIGURE 7 also illustrate that for PICP and PINAW, Flow-LGBM is superior in all months except February for PICP and each month for PINAW. The flow-LGBM improves the multi-LGBM and optical flow for the entire period by 0.3% and 0.9% for PICP and 9.3% and 16.2% for PINAW, respectively. The PICP falls below 0.9 in March in all the multi-point data models. This decrease in the PICP can be attributed to the high average power generation and large output fluctuations during this month, resulting in increased prediction errors.

The authors analyzed the accuracy of the prediction intervals for each time because the multi-point data models obtained the error distribution for each time and generated prediction intervals. FIGURE 9 shows the average forecast accuracy over time for all PVs, and FIGURE 10 presents a violin diagram of the average MAE over the entire period for 390 PVs. FIGURES 9(a) and 9(b) show that Flow-LGBM has the best forecast accuracy in terms of MAE and CRPS for many periods. FIGURE 9(c) demonstrates that Flow-LGBM has the narrowest PINAW at almost all times, whereas the optical flow exhibits the widest PINAW at almost all times. The wide PINAW of the optical flow may be due to errors in the interpolation of missing values in the image of the power generation distribution and errors resulting from the assumption that the power generation distribution is invariant and in constant velocity linear motion, leading to a larger variation in the forecast, as shown in FIGURE 10. FIGURE 9(d) indicates that the PICP for all models exceeds 0.94 when the average generation was large from 8:30 to 14:00. However, the PICP is below 0.94 during other periods when the average generation is small. Additionally, from 16:30 to 18:00, the PICP for the flow-LGBM tended to increase, whereas that of the optical flow tended to decrease. This trend in the optical flow may be attributed to the narrow PINAW during that time and the high variability in the prediction accuracy.

Flow-LGBM improves the MAE and CRPS in almost all periods by learning spatial-temporal PV relationships from optical flow features. Additionally, by reducing the variability of the optical flow predictions, the PINAW is narrowed, and the PICP is stabilized. A challenge in this study was that the PINAW was narrower, and the PICP was lower during periods of low generation. Therefore, improving the PICP by adding a correction to the forecast interval width during periods of low power generation is necessary.

VII. CONCLUSION

In this study, the authors proposed a hybrid model that combined optical flow and LGBM without using meteorological data such as cloud images or weather data. The model aimed to forecast deterministic and probabilistic outputs 30 min in advance for 390 distributed PVs, relying solely on one month of PV generation and location data. The model achieved two primary objectives. First, it could use weather data to forecast the output of dispersed PVs 30 min in advance with higher accuracy than conventional models. Compared to the LGBM trained on 4.5 months of weather data, the proposed method improved the MAE of the three PVs by 18.4%. Second, it enhanced the forecast accuracy by combining machine learning and optical flow to learn the spatiotemporal relationships of PVs. Compared to the optical flow alone, the proposed method improved the MAE and CRPS of 390 PVs by 5.8% and 10.8%, respectively. Therefore, the proposed method achieved higher accuracy in forecasting the output of geographically dispersed PVs without weather data than conventional methods that use weather data or optical flow. A future challenge is to classify distributed PV systems into clusters based on multiple factors, such as geographic location and output generation, and apply specialized forecasting models to each cluster to enhance the overall accuracy of the output forecasts.

ACKNOWLEDGMENT

The OMRON Social Solutions Corporation provided power generation data from PVs at multiple locations.

REFERENCES

[1] B. Meng, R. C. G. M. Loonen, and J. L. M. Hensen, "Data-driven inference of unknown tilt and azimuth of distributed PV systems," *Sol. Energy*, vol. 211, pp. 418–432, Nov. 2020, doi: 10.1016/j.solener.2020.09.077.

[2] IEA, Paris, France. (2022). *Approximately 100 Million Households Rely on Rooftop Solar PV By 2030*. [Online]. Available: https://www.iea.org/reports/approximately-100-million-households-rely-on-rooftop-solar-pv-by-2030

[3] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, May 2020, Art. no. 109792, doi: 10.1016/j.rser.2020.109792.

[4] P. Lauret, M. David, and P. Pinson, "Verification of solar irradiance probabilistic forecasts," *Sol. Energy*, vol. 194, pp. 254–271, Dec. 2019, doi: 10.1016/j.solener.2019.10.041.

[5] M. Beykirch, T. Janke, and F. Steinke, "Bidding and scheduling in energy markets: Which probabilistic forecast do we need?" in *Proc. 17th Int. Conf. Probabilistic Methods Appl. Power Syst.*, 2022, pp. 1–6, doi: 10.1109/PMAPS53380.2022.9810632.

[6] R. R. Appino, J. Á. González Ordiano, R. Mikut, T. Faulwasser, and V. Hagenmeyer, "On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages," *Appl. Energy*, vol. 210, pp. 1207–1218, Jan. 2018, doi: 10.1016/j.apenergy.2017.08.133.

[7] D. Yang, W. Wang, C. A. Gueymard, T. Hong, J. Kleissl, J. Huang, M. J. Perez, R. Perez, J. M. Bright, X. Xia, D. van der Meer, and I. M. Peters, "A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality," *Renew. Sustain. Energy Rev.*, vol. 161, Jun. 2022, Art. no. 112348, doi: 10.1016/j.rser.2022.112348.

[8] F. Lin, Y. Zhang, and J. Wang, "Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods," *Int. J. Forecasting*, vol. 39, no. 1, pp. 244–265, Jan. 2023, doi: 10.1016/j.ijforecast.2021.11.002.

[9] O. Gandhi, W. Zhang, D. S. Kumar, C. D. Rodríguez-Gallegos, G. M. Yagli, D. Yang, T. Reindl, and D. Srinivasan, "The value of solar forecasts and the cost of their errors: A review," *Renew. Sustain. Energy Rev.*, vol. 189, Jan. 2024, Art. no. 113915, doi: 10.1016/j.rser.2023.113915.

[10] K. Hu, S. Cao, L. Wang, W. Li, and M. Lv, "A new ultra-short-term photovoltaic power prediction model based on ground-based cloud images," *J. Cleaner Prod.*, vol. 200, pp. 731–745, Nov. 2018, doi: 10.1016/j.jclepro.2018.07.311.

[11] L. Wei, T. Zhu, Y. Guo, C. Ni, and Q. Zheng, "CloudpredNet: An ultra-short-term movement prediction model for ground-based cloud image," *IEEE Access*, vol. 11, pp. 97177–97188, 2023, doi: 10.1109/ACCESS.2023.3310538.

[12] J. Liu, H. Zang, T. Ding, L. Cheng, Z. Wei, and G. Sun, "Sky-image-derived deep decomposition for ultra-short-term photovoltaic power forecasting," *IEEE Trans. Sustain. Energy*, vol. 15, no. 2, pp. 871–883, Apr. 2024, doi: 10.1109/TSTE.2023.3312401.

[13] C. W. Chow, S. Belongie, and J. Kleissl, "Cloud motion and stability estimation for intra-hour solar forecasting," *Sol. Energy*, vol. 115, pp. 645–655, May 2015, doi: 10.1016/j.solener.2015.03.030.

[14] L. Cheng, H. Zang, Z. Wei, T. Ding, and G. Sun, "Solar power prediction based on satellite measurements—A graphical learning method for tracking cloud motion," *IEEE Trans. Power Syst.*, vol. 37, no. 3, pp. 2335–2345, May 2022, doi: 10.1109/TPWRS.2021.3119338.

[15] D. Aicardi, P. Musé, and R. Alonso-Suárez, "A comparison of satellite cloud motion vectors techniques to forecast intra-day hourly solar global horizontal irradiation," *Sol. Energy*, vol. 233, pp. 46–60, Feb. 2022, doi: 10.1016/j.solener.2021.12.066.

[16] J. Yan, L. Hu, Z. Zhen, F. Wang, G. Qiu, Y. Li, L. Yao, M. Shafie-khah, and J. P. S. Catalao, "Frequency-domain decomposition and deep learning based solar PV power ultra-short-term forecasting model," *IEEE Trans. Ind. Appl.*, vol. 57, no. 4, pp. 3282–3295, Jul./Aug. 2021, doi: 10.1109/TIA.2021.3073652.

[17] J. Dong, M. M. Olama, T. Kuruganti, A. M. Melin, S. M. Djouadi, Y. Zhang, and Y. Xue, "Novel stochastic methods to predict short-term solar radiation and photovoltaic power," *Renew. Energy*, vol. 145, pp. 333–346, Jan. 2020, doi: 10.1016/j.renene.2019.05.073.

[18] T. Yao, J. Wang, Y. Wang, P. Zhang, H. Cao, X. Chi, and M. Shi, "Very short-term forecasting of distributed PV power using GSTANN," *CSEE J. Power Energy Syst.*, early access, Oct. 12, 2022, doi: 10.17775/CSEEJPES.2022.00110.

[19] H. Verbois, Y.-M. Saint-Drenan, A. Thiery, and P. Blanc, "Statistical learning for NWP post-processing: A benchmark for solar irradiance forecasting," *Sol. Energy*, vol. 238, pp. 132–149, May 2022, doi: 10.1016/j.solener.2022.03.017.

[20] B. Schulz, M. El Ayari, S. Lerch, and S. Baran, "Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting," *Sol. Energy*, vol. 220, pp. 1016–1031, May 2021, doi: 10.1016/j.solener.2021.03.023.

[21] R. Tawn and J. Browell, "A review of very short-term wind and solar power forecasting," *Renew. Sustain. Energy Rev.*, vol. 153, Jan. 2022, Art. no. 111758, doi: 10.1016/j.rser.2021.111758.

[22] J. Liang and W. Tang, "Ultra-short-term spatiotemporal forecasting of renewable resources: An attention temporal convolutional network-based approach," *IEEE Trans. Smart Grid*, vol. 13, no. 5, pp. 3798–3812, Sep. 2022, doi: 10.1109/TSG.2022.3175451.

[23] C. Liu, M. Li, Y. Yu, Z. Wu, H. Gong, and F. Cheng, "A review of multitemporal and multispatial scales photovoltaic forecasting methods," *IEEE Access*, vol. 10, pp. 35073–35093, 2022, doi: 10.1109/ACCESS.2022.3162206.

[24] S. Chai, Z. Xu, Y. Jia, and W. K. Wong, "A robust spatiotemporal forecasting framework for photovoltaic generation," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5370–5382, Nov. 2020, doi: 10.1109/TSG.2020.3006085.

[25] J. Simeunovic, B. Schubnel, P.-J. Alet, and R. E. Carrillo, "Spatio-temporal graph neural networks for multi-site PV power forecasting," *IEEE Trans. Sustain. Energy*, vol. 13, no. 2, pp. 1210–1220, Apr. 2022, doi: 10.1109/TSTE.2021.3125200.

[26] T. Kure, H. D. Tsuchiya, Y. Kameda, H. Yamamoto, D. Kodaira, and J. Kondoh, "Parameter evaluation in motion estimation for forecasting multiple photovoltaic power generation," *Energies*, vol. 15, no. 8, p. 2855, Apr. 2022, doi: 10.3390/en15082855.

[27] Y. Miyazaki, Y. Kameda, and J. Kondoh, "A power-forecasting method for geographically distributed PV power systems using their previous datasets," *Energies*, vol. 12, no. 24, p. 4815, Dec. 2019, doi: 10.3390/en12244815.

[28] T.-Y. L. Guolin Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and Q. Ye, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1111–1116.

[29] D. Kodaira, K. Tsukazaki, T. Kure, and J. Kondoh, "Improving forecast reliability for geographically distributed photovoltaic generations," *Energies*, vol. 14, no. 21, p. 7340, 2021, doi: 10.3390/en14217340.

[30] H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather Forecasting*, vol. 15, no. 5, pp. 559–570, Oct. 2000, doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

[31] J. R. Shewchuk, "Delaunay refinement algorithms for triangular mesh generation," *Comput. Geometry*, vol. 22, nos. 1–3, pp. 21–74, May 2002.

[32] Y. Kameda, H. Kishi, T. Ishikawa, I. Matsuda, and S. Itoh, "Multi-frame motion compensation using extrapolated frame by optical flow for lossless video coding," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2016, pp. 300–304, doi: 10.1109/ISSPIT.2016.7886053.

[33] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 69, no. 2, pp. 243–268, Apr. 2007, doi: 10.1111/j.1467-9868.2007.00587.x.

[34] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

HIROKI YAMAMOTO received the bachelor's degree from the Department of Electrical Engineering, Faculty of Science and Technology, Tokyo University of Science, in March 2022, and the master's degree from the Department of Electrical Engineering, Graduate School of Science and Technology, Tokyo University of Science, in March 2024. His research interests include time-series analysis and forecasting the output of photovoltaic (solar power) systems.

TAIKI KURE received the bachelor's degree from the Department of Electrical Engineering, Faculty of Science and Technology, Tokyo University of Science, in March 2020, and the master's degree from the Department of Electrical Engineering, Graduate School of Science and Technology, Tokyo University of Science, in March 2022. His research interests include time-series analysis and forecasting the output of photovoltaic (solar power) systems.

JUNJI KONDOH received the D.Eng. degree from Tokyo Institute of Technology, Tokyo, Japan, in 1998. He joined the Electrotechnical Laboratory (ETL), in 1998, which was reorganized to the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He then moved to Tokyo University of Science, in 2013, as an Associate Professor. His research interests include electric power systems with large amounts of renewable energy.

DAISUKE KODAIRA (Member, IEEE) received the Ph.D. degree in electrical engineering from Kyungpook National University, Daegu, South Korea, in 2020. From 2020 to 2022, he was an Assistant Professor with Tokyo University of Science. Since 2022, he has been an Assistant Professor with the Institute of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan. His research interests include reinforcement learning-based battery control, electric vehicle charging scheduling, block-chain for energy trade, forecasting photovoltaic generation, and energy demand.