

# Hourly-Similarity Based Solar Forecasting Using Multi-Model Machine Learning Blending

Cong Feng, Jie Zhang  
The University of Texas at Dallas  
Richardson, TX, 75080 USA

**Abstract**—With the increasing penetration of solar power into power systems, forecasting becomes critical in power system operations. In this paper, an hourly-similarity (HS) based method is developed for 1-hour-ahead (1HA) global horizontal irradiance (GHI) forecasting. This developed method utilizes diurnal patterns, statistical distinctions between different hours, and hourly similarities in solar data to improve the forecasting accuracy. The HS-based method is built by training multiple two-layer multi-model forecasting framework (MMFF) models independently with the same-hour subsets. The final optimal model is a combination of MMFF models with the best-performed blending algorithm at every hour. At the forecasting stage, the most suitable model is selected to perform the forecasting subtask of a certain hour. The HS-based method is validated by 1-year data with six solar features collected by the National Renewable Energy Laboratory (NREL). Results show that the HS-based method outperforms the non-HS (all-in-one) method significantly with the same MMFF architecture, wherein the optimal HS-based method outperforms the best all-in-one method by 10.94% and 7.74% based on the normalized mean absolute error and normalized root mean square error, respectively.

**Index Terms**—Hourly similarity, solar forecasting, machine learning blending, classification, sky imaging

## I. INTRODUCTION

With the intensify of energy consumption and environment deterioration, facilitating and integrating renewable energy become increasingly important. Solar energy has increased rapidly and the cost also continuously decreases in recent years [1]. However, it is still challenging to reliably and economically operate power systems under large penetrations of solar energy due to the uncertainty and variability in solar power. Improving solar forecasting accuracy could partially address the challenges of foreseeable large solar penetrations faced by independent system operators and independent power producers.

For short-term solar forecasting, statistical methods are popularly used due to their learning power and high computational efficiency. Based on the algorithm complexity, statistical methods can be categorized into traditional time series methods (e.g., autoregressive integrated moving average (ARIMA) method [2]), machine learning methods (e.g., support vector machine (SVM) [3]), and deep learning models (e.g., long short term memory neural network (LSTM-NN) [4]). A more comprehensive review of statistical methods for solar forecasting can be found in recent review papers [5]–[7].

To improve forecasting accuracy, different techniques have been proposed in the literature, such as dividing forecasting

into different subtasks. For example, Feng et al. [8] grouped forecasts by weather conditions and improved the forecasting performance by 20%. Temperature was used in [9] to determine a suitable forecasting model for the similar day to enhance the accuracy. Forecasting could be also differentiated by forecast error metrics, thereby generating superior forecasts [10].

In an attempt to more precisely forecast short-term global horizontal irradiance (GHI), an hourly-similarity (HS) based method is developed in this study, using a two-layer hybrid machine learning based multi-model forecasting framework (MMFF). Contributions of this paper are twofold: (i) to develop an advanced forecasting method based on an HS classification criterion; (ii) to take advantage of MMFF’s powerful learning ability.

The remainder of the paper is organized as follows: Section II explores the hourly similarities in solar feature time series. The developed HS-based MMFF framework is described in Section III. An experiment study is used to validate the proposed method in Section IV. Section V concludes the paper.

## II. HOURLY SIMILARITIES IN SOLAR DATA

### A. Solar Features

The forecasting performance of data-driven statistical models is highly influenced by their inputs. To obtain well-performing models, six solar features are extracted from two types of data source: i) GHI features: historical GHI ( $GHI$ ), clear sky GHI ( $GHI_{clr}$ ), and clear sky index ( $CSI$ ); ii) sky imaging features: mean ( $\mu$ ), standard deviation ( $\sigma$ ), and Rényi entropy ( $H$ ) of normalized sky image pixel red blue ratio ( $nRBR$ ) values. The feature space is defined as  $\mathbf{X} = (GHI, GHI_{clr}, CSI, \mu, \sigma, H) \in \mathbb{R}^{N \times 6}$  ( $N$  is the sample length). Data elimination is performed to exclude data in the early morning (before 7am in this study) and late night (after 7pm), since most GHIs are 0. More details of feature extraction can be referred in our previous work [8].

### B. Diurnal Patterns and Hourly Similarities

To ensure the efficiency of HS-based forecasting, diurnal patterns and hourly similarities in solar data need to be explored and identified. Table I shows the periodicity, trend, and seasonality of the six solar features, which are calculated by a time series characteristic analysis [11], [12]. The periodicity in Table I indicates the frequency an observation occurs in the time domain. All of the six solar features show *diurnal*

*patterns* (note data elimination was performed, so there are 13 samples in one day). Figure 1 visualizes the six feature vectors in time series curves, where daily patterns are observed. Also, longer-term patterns also exist in  $GHI$ ,  $GHI_{clr}$ , and  $CSI$  vectors, which are indicated by the trend and seasonality.

Table I: Time series characteristics in solar features

Feature	Periodicity	Trend	Seasonality
$GHI$	13	0.70	0.84
$GHI_{clr}$	13	0.86	0.96
$CSI$	13	0.22	0.51
$\mu$	13	0.02	0
$\sigma$	13	0.06	0
$H$	13	0.16	0.20

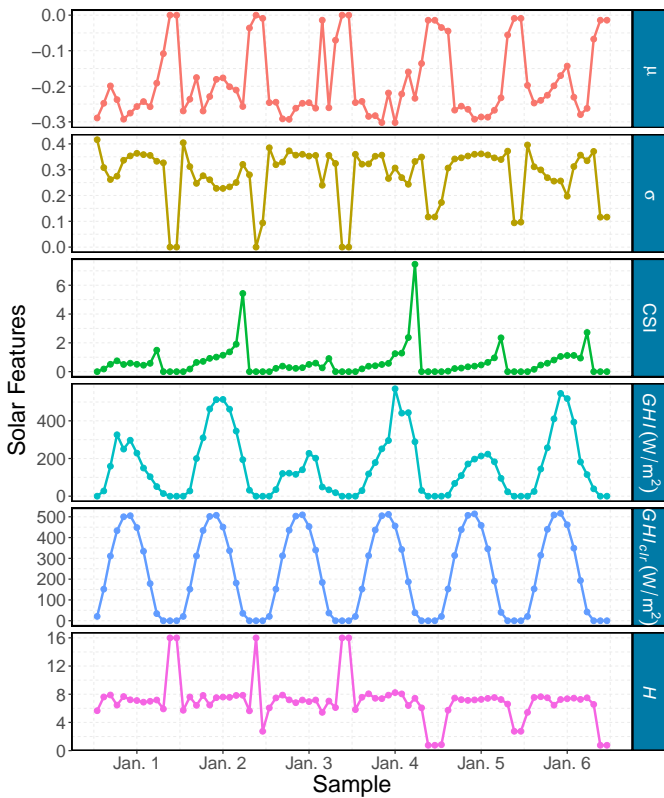


Figure 1: Time series cyclic patterns of solar features.

Figure 2 shows the distinctive statistics of solar features in different hours. For example, the median of the same feature in different hours is varying, especially for  $GHI$  and  $GHI_{clr}$ . Moreover, the value range of the same feature in different hours is also different. Thus, solar features in different hours present *statistical distinctions*, and are therefore challenging to be forecasted using a single model. This study focuses on 1HA forecasting, and the scatter plot of the current hour  $GHI$  ( $GHI_{lag}$ ) and the 1HA  $GHI$  ( $GHI$ ) indicates the most critical relationship between the output and the inputs to forecasting models. Figure 3 shows a layered pattern with scatter points

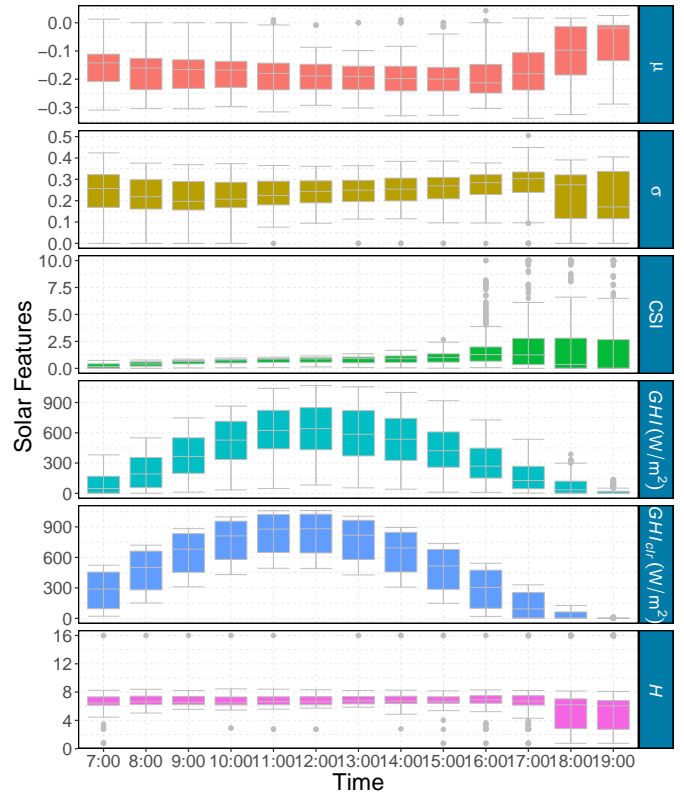


Figure 2: Hourly statistics of solar features. The line in the box is the median. The interquartile range box represents the middle 50% of the data. The upper and lower bounds are maximum and minimum values of the data, respectively, excluding the outliers. The outliers are data outside two and half times of the interquartile range.

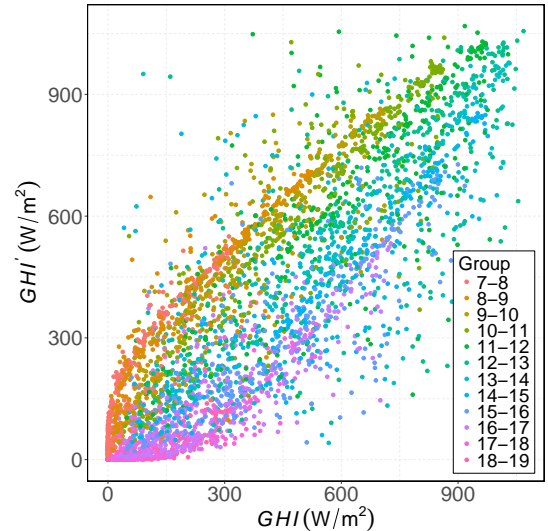


Figure 3: Scatter plot of the current hour  $GHI$  ( $GHI$ ) and the 1HA  $GHI$  ( $GHI'$ ). Different colors mean different groups. For example, group A-B has the data with current time hour A and next hour B.

in the same group dependently converged, which also verifies the *hourly similarity* in solar data. Thus, an HS-based method is expected to improve the the GHI forecasting performance.

### III. HOURLY-SIMILARITY BASED FORECASTING FRAMEWORK USING MMFF

The flowchart of the developed HS-based forecasting framework using MMFF method is illustrated in Fig. 4. The training dataset  $\mathbf{T} := (\mathbf{X}, \mathbf{Y})$  is hourly partitioned into 12 subsets  $\mathbf{T}_k$  ( $k = 1, 2, \dots, 12$ ), each of which is used to train a MMFF model independently. The MMFF model is a two-layer machine learning based method for short-term forecasting. Multiple machine learning algorithms with several kernels generate forecasts,  $\tilde{\mathbf{Y}}$ , individually in the first layer. Then the forecasts are blended by another machine learning algorithm in the second layer, which gives the final forecasts,  $\hat{\mathbf{Y}}$ . Machine learning algorithms used in the MMFF method include artificial neural networks (ANN), SVM, gradient boosting machines (GBM), and random forests (RF). MMFF method has shown to perform better than single algorithm machine learning methods [13]. The MMFF method can be expressed as:

$$\tilde{y}_{t,ij} = f_{ij}(\mathbf{x}_t) \quad (1)$$

$$\hat{y}_t = \Phi(\tilde{\mathbf{y}}_t) \quad (2)$$

where  $t$  is the time index,  $f_{ij}(\ast)$  is the  $i$ th first-layer model using kernel  $j$ ,  $\tilde{y}_{ij}$  is the forecast provided by the model  $f_{ij}$ ,  $\mathbf{x}_t \in \mathbf{X}$  is the input vector to the first-layer models,  $\tilde{\mathbf{y}} = \{\tilde{y}_{ij}\}$  is the combination of the first-layer forecasts,  $\hat{y}_t$  is the final forecast at time  $t$ , and  $\Phi$  is the blending algorithm in the second layer. Note that several blending algorithms can be applied in the second layer, and the best MMFF model,  $\text{MMFF}_{k_i}$ , with a certain blending algorithm in each hour forecasting is selected to constitute the final forecasting framework. This training process is evaluated through a 10-folder cross-validation in the training dataset. At the forecasting stage, the MMFF model of a certain hour is selected to generate forecasts for that specific hour. The GHIs at 7am are forecasted by a 1-day-ahead (1DA) persistence of cloudiness model, since most GHIs before 7am are close to zero, which do not provide enough information to build an efficient learning model.

### IV. CASE STUDY

#### A. Data Description and Experimental Setup

The hourly GHI and sky imaging data released by the National Renewable Energy Laboratory (NREL) was adopted in the case study, which was collected at a location in Colorado (latitude = 39.74° North, longitude = 105.18° West, elevation = 1,828.8 m). The training data was randomly selected from each month, and the remaining data was used for testing. The ratio of the number of training samples to testing samples was 3:1. The experiment was carried out on a laptop with an Intel Core i7 2.6 GHz processor and a 16.0 GB RAM.

Two groups of models were setup for comparison, which were the *HS*-based group and non-similarity based (*all-in-one*) group. In both groups, several machine learning algorithms with multiple kernels were selected as the blending algorithm in the second layer for each forecasting subtask. Details of the kernel information are listed in Table II [14]. In the HS-based group,  $C_{opt,h}$  is a combination of multiple MMFF models with the best blending algorithm (i.e., 1DA persistence of cloudiness, RF, SVM<sub>1</sub>, RF, RF, SVM<sub>1</sub>, ANN<sub>1</sub>, ANN<sub>2</sub>, SVM<sub>1</sub>, ANN<sub>3</sub>, RF, GBM<sub>3</sub>, and RF) for each forecasting subtask. The 1HA persistence of cloudiness model is included in the all-in-one group.

Table II: Blending algorithm pool for the MMFF second layer

Algorithm	Model Name	Kernel, learning, or loss function
ANN	ANN <sub>1</sub>	Standard back-propagation
	ANN <sub>2</sub>	Momentum-enhanced back-propagation
	ANN <sub>3</sub>	Resilient back-propagation
SVM	SVM <sub>1</sub>	Radial basis function kernel
	SVM <sub>2</sub>	Linear kernel
GBM	GBM <sub>1</sub>	Squared loss
	GBM <sub>2</sub>	Laplace loss
	GBM <sub>3</sub>	T-distribution loss
RF	RF	CART aggregation

#### B. Forecasting Accuracy Assessment

To assess the forecasting accuracy, four evaluation metrics are used, which are normalized mean absolute error ( $nMAE$ ), normalized root mean square error ( $nRMSE$ ),  $nMAE$  improvement ( $Imp_A$ ), and  $nRMSE$  improvement ( $Imp_R$ ). The mathematical expressions of  $nMAE$  and  $nRMSE$  can be referred in [8]. The ( $Imp_A$ ) and ( $Imp_R$ ) metrics are, respectively, defined as:

$$Imp_A = \frac{nMAE_{M_{l,a}} - nMAE_{M_{l,h}}}{nMAE_{M_{l,a}}} \quad (3)$$

$$Imp_R = \frac{nRMSE_{M_{l,a}} - nRMSE_{M_{l,h}}}{nRMSE_{M_{l,a}}} \quad (4)$$

where  $M$  is the model name, and  $l$  is the kernel index.  $h$  and  $a$  are group indices indicating the HS-based group and the all-in-one group to which a model belongs, respectively. The focus of this study is to evaluate if the HS-based solar forecasting can improve the accuracy, compared with the all-in-one based forecasting. Therefore, the  $Imp_A$  and  $Imp_R$  were calculated

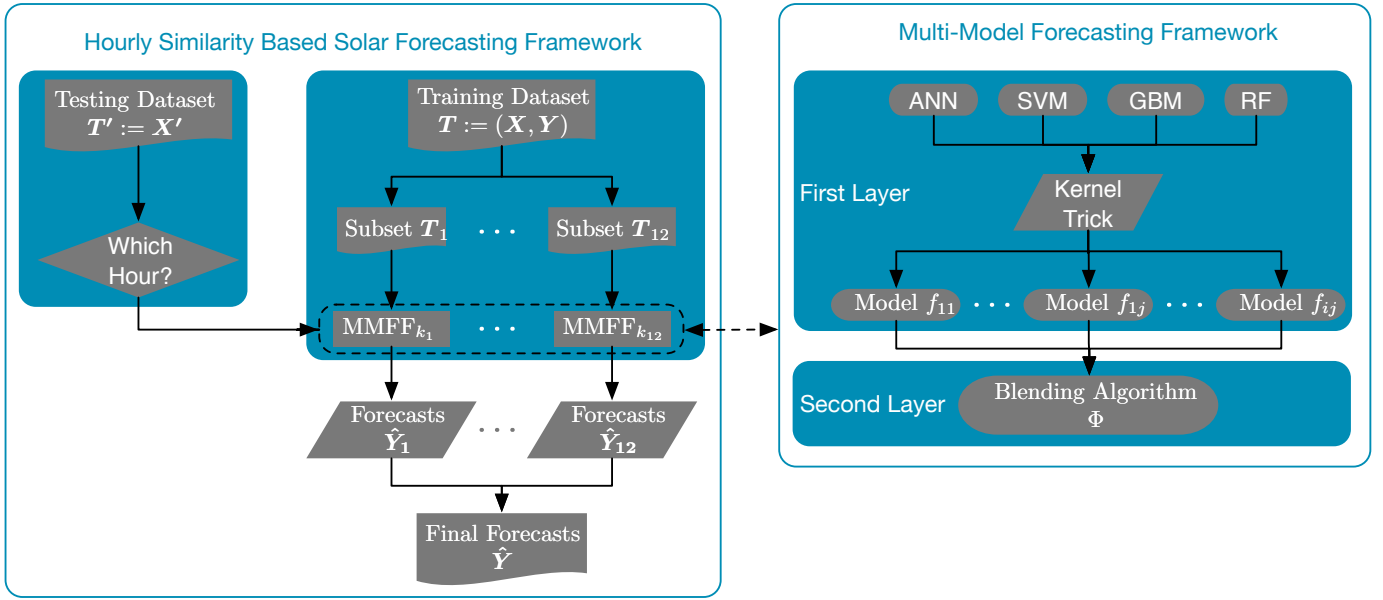


Figure 4: Flowchart of the HS-based GHI forecasting framework with MMFF.

based on the MMFF models that use the same blending algorithm in the HS-based and all-in-one groups. Note that the  $Imp_A$  and  $Imp_R$  of the final  $C_{opt,h}$  are calculated by comparing with the best model in the all-in-one group.

Table III: Overall forecasting performance evaluation

Dataset Group	Model	$nMAE$	$nRMSE$	$Imp_A$	$Imp_R$
HS-based	$C_{opt,h}$	<b>6.55</b>	<b>10.07</b>	10.94	7.74
	ANN <sub>1,h</sub>	7.08	10.49	<b>11.79</b>	<b>16.13</b>
	ANN <sub>2,h</sub>	7.20	10.66	<b>10.76</b>	<b>14.67</b>
	ANN <sub>3,h</sub>	7.10	10.51	<b>11.54</b>	<b>15.89</b>
	SVM <sub>1,h</sub>	8.02	12.43	<b>-1.32</b>	0.95
	SVM <sub>2,h</sub>	7.54	10.92	4.35	11.72
	GBM <sub>1,h</sub>	7.46	10.87	5.47	12.14
	GBM <sub>2,h</sub>	7.23	10.72	5.40	13.57
	GBM <sub>3,h</sub>	7.45	11.10	4.94	10.24
	RF <sub>h</sub>	<b>6.73</b>	<b>10.26</b>	<b>12.60</b>	<b>17.04</b>
All-in-one	ANN <sub>1,a</sub>	8.05	12.51	N/A	N/A
	ANN <sub>2,a</sub>	8.07	12.50	N/A	N/A
	ANN <sub>3,a</sub>	8.02	12.50	N/A	N/A
	SVM <sub>1,a</sub>	8.09	12.55	N/A	N/A
	SVM <sub>2,a</sub>	7.84	12.46	N/A	N/A
	GBM <sub>1,a</sub>	7.89	12.37	N/A	N/A
	GBM <sub>2,a</sub>	<b>7.64</b>	<b>12.36</b>	N/A	N/A
	GBM <sub>3,a</sub>	7.84	12.37	N/A	N/A
	RF <sub>a</sub>	7.70	12.37	N/A	N/A
	P <sub>a</sub>	<b>7.35</b>	<b>10.91</b>	N/A	N/A

Note: The units of all the evaluation metrics are %. The models with the same model name and numerical subscripts (if any) use the same algorithm. The letter subscript  $h$  or  $a$  of the models indicates the group it belongs to.

Table III lists the forecasting evaluation metrics of all the HS-based models and all-in-one models. The  $C_{opt,h}$  model has the smallest forecasting  $nMAE$  and  $nRMSE$  among all the models in both groups, as shown in bold green. The P<sub>a</sub> model

is the IHA persistence of cloudiness model [8] that is the best model in the all-in-one group, as shown in bold black in Table III. The best HS-based method (i.e.,  $C_{opt,h}$ ) outperforms the best all-in-one method (i.e., P<sub>a</sub>) by 10.94% and 7.74% based on  $nMAE$  and  $nRMSE$ , respectively. If using a unique blending algorithm in the MMFF models (exclude the  $C_{opt,h}$  model) to forecast every hour GHI (HS-based group), RF<sub>h</sub> outperforms other blending algorithm models in the HS-based group. In the all-in-one group, GBM<sub>2,a</sub> is the best machine learning model, as shown in italic and bold black in Table III. By comparing the  $Imp_A$  and  $Imp_R$ , the RF<sub>h</sub> has the most significant improvement, followed by the ANN models (ANN<sub>1,h</sub>, ANN<sub>2,h</sub>, ANN<sub>3,h</sub>). Compared with all-in-one models, the HS-based models present better performance except for the SVM<sub>1,h</sub>, based on  $nMAE$  (as shown in bold red). This SVM model used the radial basis function kernel, which normally needs a large dataset to train. Dividing the training dataset into 12 subsets reduced the sample amount, which possibly caused underfitting during the SVM training. **Overall it is found that the HS-based GHI forecasting improved the accuracy significantly.**

A further analysis is carried out by comparing the model performance in each hour and each month. Four models are selected, which are the two best models in each group ( $C_{opt,h}$  and P<sub>a</sub>), and two best MMFF models with the unique blending algorithm in the two groups (RF<sub>h</sub> and GBM<sub>2,a</sub>). Figure 5 shows the forecasting errors in every hour. Compared to the methods in the all-in-one group, the HS-based methods have superior forecasting performance, especially when the GHI value is small in the early morning and late afternoon. The stable performance of HS-based forecasts is due to the more stable and similar pattern in the same-hour data. Considering the monthly behaviour, the HS-based methods also perform

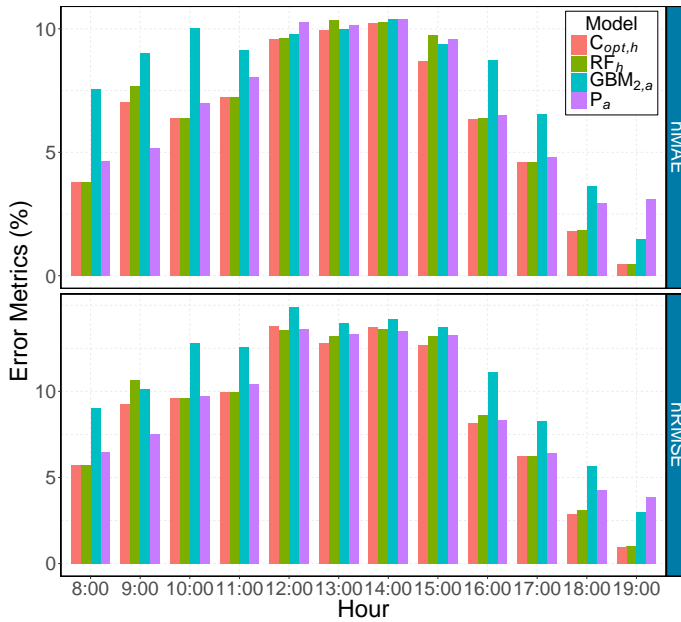


Figure 5: Forecasting errors by hour.

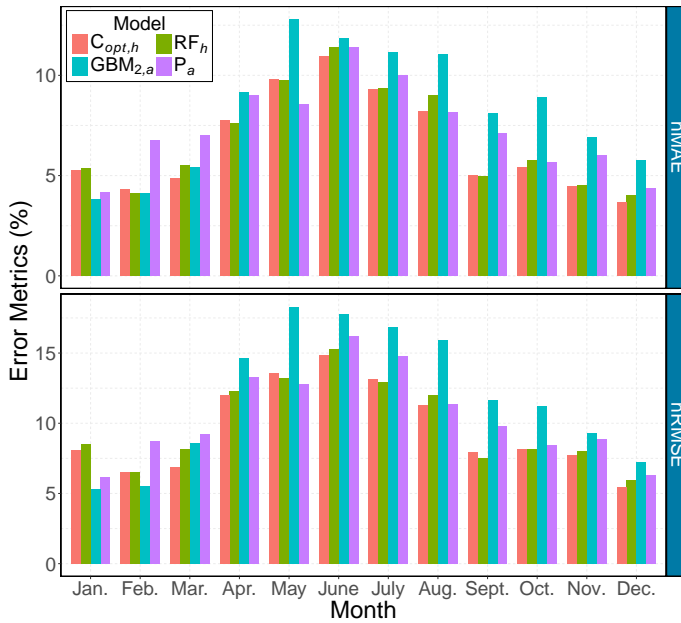


Figure 6: Forecasting errors by month.

better than the all-in-one methods, as shown in Fig. 6. It is seen from Figs. 5 and 6 that the 1HA persistence of cloudiness model ( $P_a$ ) also shows satisfying performance, considering its simplicity. Thus,  $P_a$  method can be used to forecast GHI at some hours, such as 9am, to further improve the forecasting accuracy. The forecasting errors have shown evident diurnal and seasonal patterns, as indicated in Figs. 5 and 6, respectively. This information could be used to build seasonal-based models [15] or to perform post-processing such as bias correction [16] in solar forecasting.

## V. CONCLUSION

In this paper, an hourly-similarity (HS) based forecasting method was developed to improve the short-term 1HA GHI forecasting. The developed HS-based method forecasted GHI of each hour separately with a two-layer machine learning based MMFF method. Several blending algorithms were used in the second-layer of the MMFF models, the optimal combination of which composes the final optimal HS-based model. A numerical experiment was performed to validate the efficiency by using the 1-year solar data with six features. The developed HS-based 1HA GHI forecasting method has shown superior performance from various perspectives, compared with all-in-one methods.

## REFERENCES

- [1] N. S. Lewis, "Research opportunities to advance solar energy utilization," *Science*, vol. 351, no. 6271, p. aad1920, 2016.
- [2] M. David, F. Ramahatana, P.-J. Trombe, and P. Lauret, "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models," *Solar Energy*, vol. 133, pp. 55–72, 2016.
- [3] R. C. Deo, X. Wen, and F. Qi, "A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset," *Applied Energy*, vol. 168, pp. 568–593, 2016.
- [4] A. Gensler, J. Henze, B. Sick, and N. Raabe, "Deep learning for solar power forecasting—An approach using AutoEncoder and LSTM neural networks," in *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 002 858–002 865.
- [5] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.
- [6] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.
- [7] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in PV output power forecast," *Solar Energy*, vol. 136, pp. 125–144, 2016.
- [8] C. Feng, M. Cui, M. Lee, J. Zhang, B.-M. Hodge, S. Lu, and H. F. Hamann, "Short-term global horizontal irradiance forecasting based on sky imaging and pattern recognition," in *IEEE PES General Meeting*. IEEE, 2017.
- [9] M. Ding, L. Wang, and R. Bi, "An ANN-based approach for forecasting the power output of photovoltaic system," *Procedia Environmental Sciences*, vol. 11, pp. 1308–1315, 2011.
- [10] J. Wu and C. K. Chan, "Prediction of hourly solar radiation with multi-model framework," *Energy conversion and management*, vol. 76, pp. 347–355, 2013.
- [11] C. Feng, E. K. Chartan, B.-M. Hodge, and J. Zhang, "Characterizing time series data diversity for wind forecasting," in *Big Data Computing Applications and Technologies (BDCAT), 2017 IEEE/ACM 4th International Conference on*. IEEE, 2017.
- [12] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data mining and knowledge Discovery*, vol. 13, no. 3, pp. 335–364, 2006.
- [13] C. Feng, M. Cui, B.-M. Hodge, and J. Zhang, "A data-driven multi-model methodology with deep feature selection for short-term wind forecasting," *Applied Energy*, vol. 190, pp. 1245–1257, 2017.
- [14] C. Feng, Z. Wang, J. Zhang, L. Castillo, M. Fajardo, and G. C. Irving, "An artificial intelligence based pseudosensor for arterial blood pressure measurement," *Artificial Intelligence in Medicine*, submitted.
- [15] S. I. Vagropoulos, G. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," in *Energy Conference (ENERGYCON), 2016 IEEE International*. IEEE, 2016, pp. 1–6.
- [16] P. Lauret, M. Diagne, and M. David, "A neural network post-processing approach to improving NWP solar radiation forecasts," *Energy Procedia*, vol. 57, pp. 1044–1052, 2014.