IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 April 2024, accepted 19 May 2024, date of publication 4 June 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3407778

RESEARCH ARTICLE

Probabilistic Solar Generation Forecasting for Rapidly Changing Weather Conditions

CHENG LYU, (Graduate Student Member, IEEE), AND SARA EFTEKHARNEJAD, (Senior Member, IEEE)

Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA

Corresponding author: Sara Eftekharnejad (seftekha@syr.edu)

This work was supported by the National Science Foundation (NSF) under Grant 2144918.

ABSTRACT Probabilistic solar generation forecasting provides a better means of quantifying generation uncertainties for power grid operations by providing a range of potential power outputs rather than a single-point estimate. The traditional probabilistic models are unreliable under rapidly changing weather conditions due to fluctuating data correlations, necessitating dynamic modeling of spatio-temporal feature correlations under diverse weather scenarios. The correlations represent the interactions across space and time that reflect the impact of weather conditions on solar power output. This paper addresses this critical problem with a novel method by fusing copula theory and machine learning methods to dynamically quantify the spatio-temporal correlations among meteorological data under diverse weather conditions. The meteorological data and the functions employed to estimate spatio-temporal correlations change dynamically based on weather conditions. A data-driven environment-aware model has been developed to produce probabilistic forecasts from this data, effectively quantifying uncertainty in meteorological data. Case studies on real-world datasets demonstrate that the proposed dynamic method exhibits robust performance in solar irradiance and solar power forecasting. Moreover, the model outperforms state-of-the-art models by up to 60% higher accuracy under non-sunny conditions in autumn and winter.

INDEX TERMS Copula theory, data correlation, dynamic forecasting, probabilistic solar generation forecast.

NOMENCLATURE

α	Dependence parameters in copula functions.
$\boldsymbol{p}$	Correlation matrix in Gaussian copula.
η, μ	Parameters controlling penalty for T and w.
γ, σ	Quantile level, Standard deviation.
$\hat{y}$	Predicted results of XGBoost.
$\mathcal{I}, \mathcal{I}_L, \mathcal{I}_R$	All nodes, left nodes, and right nodes in a tree.
$\Omega(f_k)$	Complexity penalty of kth tree in XGBoost.
$\Phi()$	Uni-variate standard normal CDF.
CT_I	A data cluster with centriod x_{CT_I}.
F, F^{-1}	Distribution function, inverse distribution function.
f_k	kth independent regression tree with structure q and leaf weights w.
n	Data points number in the original weather dataset.
T, w	Number of leafs, and leaf weights in a tree.
U	A random variable that has a uniform distribution on the interval [0,1].

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

I. INTRODUCTION

The dramatic increase in the integration of solar-powered generation units is anticipated to lead to substantial changes in power grids, necessitating improved operational and planning procedures. An accurate solar generation forecast is a vital step in these operational enhancements and is especially critical for maintaining the real-time load and generation balance. Traditionally, the generation forecast is achieved using historical data, such as meteorological observations obtained from local weather stations or remote sensing devices [1]. The resulting forecast provides grid operators with a short-term (day-ahead or hour-ahead) estimate of solar

generation, allowing an optimized power dispatch, balancing the generated electricity and load, and preparing necessary measures to protect the grids. Solar power forecasting can be performed deterministically or probabilistically. The probabilistic forecasts generate predicted values as a probability distribution [1]. In recent years, the research community and utilities have acknowledged the need for probabilistic predictions to integrate uncertainty quantification into grid operations and planning [2]. Previous research has shown that probabilistic approaches are more reliable than their deterministic counterparts [3]. The focus of this paper is thus on probabilistic forecasting.

Probabilistic forecasting methods are classified into parametric techniques, in which the forecasted variable is expected to follow a prior distribution, and non-parametric approaches, in which no such assumptions are made [1]. Parametric techniques refer to the sum of a deterministic forecast and a predefined distribution of the forecast error [4]. In the category of the non-parametric probabilistic forecasts, Quantile Regression [5], [6] and Persistence Ensemble (PeEn) [7], [8] are two widely used benchmarks. An example of a probabilistic solar irradiance forecast from PeEn is shown in Fig. 1, from which some shortcomings of conventional probabilistic approaches for solar forecasts can be identified as follows:

1) Conventional probabilistic forecasts cannot adapt to sudden weather changes; often, forecasts for different days are nearly identical.
2) The Prediction Interval (PI) derived from traditional techniques is overly wide to be a credible reference for power system planning.

In light of the limitations of traditional probabilistic forecasting, the Copula Theory offers a refined solution. By capturing the dependency correlation between multiple variables, copula-based forecasts can adapt to dynamic weather changes and provide more reliable prediction intervals for enhanced decision-making in power system planning.

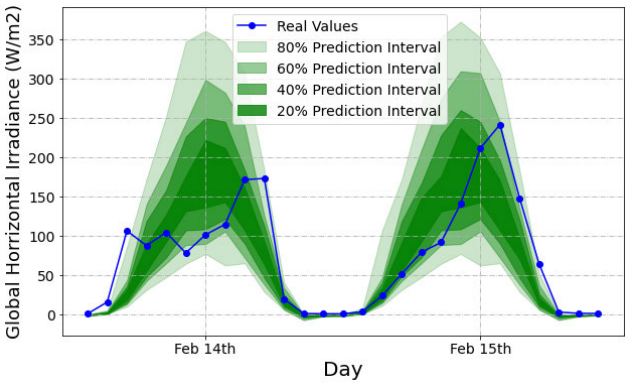

FIGURE 1. Probabilistic forecasts obtained from the PeEn method.

A. RELATED WORK ON FORECASTING BASED ON COPULA THEORY

Copulas were first introduced to probabilistic wind power forecasting in [21] and [22] to capture uncertainties in forecasting and later were applied to solar generation forecasting for the same purpose [9], [10], [11]. The application of copula theory in generation forecasting can be categorized in three ways: 1) the development of a conditional probability distribution of generation, given meteorological variables, 2) the analysis of the correlations between different variables, and 3) the generation of probabilistic forecasts using the ensemble approach.

In the first category, the copula-based joint probability distribution function (PDF) and Bayesian theory are generally utilized to form the conditional PDF of a renewable generation unit. For example, authors in [9] apply copula theory to estimate the joint distribution of Global Horizontal Irradiance (GHI) forecast and solar generation output. Conditional probabilities for solar generation output are calculated using the obtained marginal and joint distributions. Similarly, in [10] and [11], Copula-based conditional probabilistic forecast models are developed for predicting wind power and its ramp rate, respectively. In another work [12], the joint distribution between the forecasted and the real meteorological variables is modeled using copulas to capture the uncertainty inherent in the forecasted meteorological data. A conditional PDF of weather scenarios given near real-time weather predictions is thus generated [12]. The generated weather scenarios are input variables to a machine-learning model that yields probabilistic forecasts. In [13], the prior distribution of the solar power forecasts is derived first using deterministic forecasts. Copula functions, constructed through the analysis of relationships between solar power output and temperature, are used to update the prior distribution to the posterior forecast distribution, thus providing probabilistic forecasts. Furthermore, Copula-based Quantile Nonlinear Regression (CQNR) is deployed by authors in [23] for a day-ahead solar power forecast methodology. Given input variables v and predicted values x [23], the quantile equation with quantile level γ is formulated as:

$$Q_x(\gamma \mid v) = F^{-1}(C^{-1}_{x|v}(\gamma \mid v)), \tag{1}$$

where the copula function C is used to optimize the quantile equation. In a similar work [21], Quantile-Copula is applied to probabilistic wind power forecasting. In [24], a copula-based autoregressive time series forecasting model is used to forecast solar irradiance. The time series are modeled using a pairwise decomposition of conditional distributions obtained from a copula, thus providing a flexible framework to generate synthetic series for generating forecasts. In [25], the time series of wind speed, which is used for generating forecasts, is classified into multiple non-Gaussian components through the Gaussian mixture copula model. The hybrid model developed in [14] forms a joint probability distribution of solar power and weather variables using Copula theory and the Monotone Broad Learning System. The marginal

TABLE 1. Discussion of the limitations of the previous solutions and the proposed solution to address those limitations.

Category	Common Characteristics, Limitations and Proposed Solutions
Development of Conditional Probability Distributions	**Associated Works:** [9]–[14]. **Characteristics:** Creating conditional PDFs through copula-based methodologies for prediction purposes. **Limitations:** Inadequate in dynamically addressing uncertainties associated with meteorological variables during rapid weather changes. **Proposed Solution:** Dynamically quantifying the spatio-temporal correlations among various variables used for forecasts, improving accuracy in rapidly changing weather conditions.
Analysis of Correlations Between Variables	**Associated Works**: [15], [16]. **Characteristics:** Analyze correlations to select relevant features for generating forecasts. **Limitations:** Fixed feature sets may not capture all relevant variations, limiting adaptability and robustness under changing conditions. **Proposed Solution:** Meteorological features for forecasts are dynamically adjusted based on prevailing weather conditions, enhancing the adaptiveness and robustness of the forecast.
Generation of Probabilistic Forecasts Using the Ensemble Approach	**Associated Works**: [17]–[20]. **Characteristics:** Employs ensemble methods coupled with copula functions to refine probabilistic forecast. **Limitations:** Forecast accuracy depends on the initial forecasts; These methods do not adequately consider the uncertainties inherent in the meteorological variables used for initial forecasts, leading to potential inaccuracies under rapidly changing weather conditions. **Proposed Solution:** Uncertainties in meteorological variables are dynamically quantified using copula functions, which generate synthetic data. Synthetic data incorporates uncertainties that are absent in historical data but may occur in the future. Training models with historical and synthetic datasets significantly improve forecast accuracy and robustness.

probability distribution of the forecast, given certain input weather variables, can be generated from the joint probability distribution.

Another application of copula is for feature selection, mainly to determine the optimal variables used for forecasting by analyzing the correlation between power generation and other variables. In [15], the relations between meteorological variables and solar power generation are analyzed by the copula theory. The most relevant variables are then fed to an LSTM model to predict mid- to long-term (monthly or yearly) solar power. Similar work is proposed in [16], where *D-vine* copulas determine suitable variables for the probabilistic solar forecast after investigating the relationship between solar power and meteorological variables. Although these works demonstrated the promise of copula for feature selection, selecting a fixed set of optimal features for all weather scenarios is not persuasive for establishing a robust predictive model, especially under rapidly changing weather conditions.

The adoption of copula theory for generating probabilistic forecasts has been demonstrated in several studies through the ensemble approach. For example, in [17], errors of deterministic forecasts are used to fit a *D-vine* copula. The generated probabilistic error from the well-trained copulas is thus added to the deterministic results to yield a probabilistic forecast. Similarly, Gaussian Copula is deployed in [18] to form the distribution of the forecast error. In a different study [19], an ultra-fast pre-selection algorithm is deployed to select the optimal features, which are utilized by Quantile Regression (QR) to yield initial forecasts. Initial forecasts from different sensors are used to train a copula function, which generates final probabilistic forecasts. In another work [20], historical meteorological data from the immediate past two weeks were first classified into groups. The data in each group is used to fit a copula function, which generates synthetic weather data for generating probabilistic forecasts.

Despite their contributions, Table 1 provides a comprehensive analysis of the limitations and proposed solutions related to previous work utilizing copula theory in renewable energy forecasting. The aforementioned studies fail to fully consider the dynamic uncertainty of meteorological variables under various weather conditions. Specifically, they do not adjust their methodologies and data to estimate spatio-temporal correlations among meteorological variables in response to rapidly changing weather conditions, which could result in suboptimal performance. This oversight can significantly impair the predictive accuracy and reliability of the models, especially during rapidly changing weather events where accurate forecasts are most crucial.

B. CONTRIBUTION

This paper thoroughly addresses the problem by integrating Copula theory and data-driven forecast methods, which enables dynamic quantification of the spatio-temporal correlations among various variables. By effectively representing the intricate relationships between diverse variables under various weather conditions, the forecasting model's accuracy and robustness significantly improve under rapidly changing weather conditions.

In summary, the contributions of this paper are:

1) The developed predictive model dynamically models spatial-temporal correlations of meteorological variables based on weather conditions. The data and functions used to quantify spatio-temporal correlations among meteorological variables are dynamically changing based on prevailing weather conditions;
2) An environment-aware model is developed to produce probabilistic forecasts that leverage the dynamically captured spatial-temporal correlations. This approach pre-trains multiple models, each designed for unique weather scenarios, with data incorporating spatio-temporal correlations relevant to those conditions. The selection of the appropriate model for forecast generation is based on the prevailing weather at the time of the forecast, thereby improving the model's adaptiveness to fluctuating meteorological conditions;

3) The developed forecast model exhibits significant robustness and precision in predicting solar irradiance and power generation under various weather conditions and geographical locations. The practical applicability of the model was substantiated through its participation in the 2022 American-Made Solar Forecasting Prize [26], where it secured a runner-up place, demonstrated exceptional performance, and highlighted its potential for solar power forecasting in real-world scenarios.

II. THE DEVELOPED MODEL FOR PROBABILISTIC SOLAR GENERATION FORECASTS

In this paper, a novel probabilistic solar power forecast model is developed by integrating the Copula theory and Extreme Gradient Boosting Tree (XGBoost). Copula theory and an XGBoost classifier are integrated to dynamically quantify spatio-temporal correlations among diverse meteorological variables under various weather conditions, producing synthetic data that capture uncertainties not present in the original dataset. The synthetic data is utilized by an environment-aware model, developed based on XGBoost regression trees, to generate probabilistic forecasts. This integration results in dependable and robust predictions, effectively addressing the uncertainties associated with solar power forecasting.

The developed probabilistic solar power forecast model uniquely incorporates data clustering, correlation analysis, synthetic data generation, and data-driven modeling to produce probabilistic forecasts, as depicted in Fig. 2.

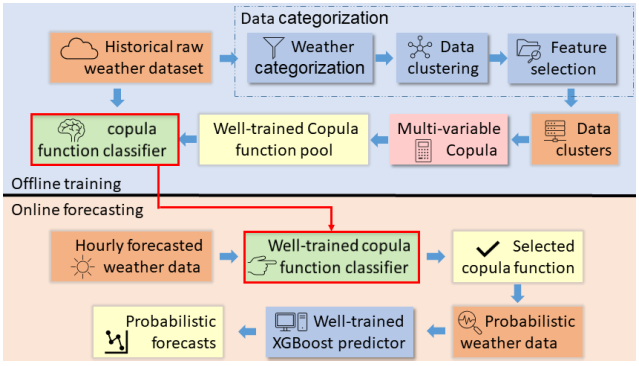

FIGURE 2. Framework of the developed model.

A. DATA CATEGORIZATION

The authors have previously demonstrated that the optimal features for accurate forecasting change dynamically depending on weather circumstances [27]. Specifically, optimal features for accurate forecasting under sunny conditions vary from those pertinent to non-sunny conditions. A decision tree model is employed to initially classify the historical weather data into three distinct categories based on the prevailing meteorological conditions, improving the efficiency of the subsequent training procedures. Specifically, the data is considered to be of the "sunny" type when the cloud cover is less than 25%; otherwise, it is classified as "cloudy" or "other" data. Precipitation and snowfall are then used to differentiate between "cloudy" and "other" data, which encompasses both "rainy" and "snowy" conditions. To eliminate the confounding effects of night-time conditions, only day-time data when the zenith angle is less than 87 degrees is retained for analysis.

The three aforementioned data groups are further classified into clusters to better quantify spatio-temporal correlation among meteorological data under diverse weather conditions. Existing literature predominantly applies clustering to the input training samples, notably meteorological variables. This method, however, is prone to noise and inaccuracies inherent in Numerical Weather Predictions (NWP), which are often employed as inputs [28]. Such vulnerabilities lead to challenges in accurately quantifying spatio-temporal correlations within these clusters, causing inaccuracies in data uncertainty estimations, particularly under varying weather conditions where NWP's reliability is compromised. Inspired by the clustering method presented in [14] and [28], clustering is based on solar power values in this study, which is extended across the entire dataset. Therefore, meteorological variables used for prediction are not clustered according to their intrinsic values but are categorized based on the target variable, the solar power output. This approach guarantees that subsequent quantification of spatio-temporal correlations among meteorological variables is tailored to specific conditions associated with a certain range of solar power output, enhancing the precision of these correlations. In this study, the optimal number of clusters, denoted as k, is determined independently for each case study through a grid-search method, resulting in $3 \times k_i$ clusters. Each cluster refers to meteorological and solar power data under a specific weather condition.

B. THE DEVELOPED COPULA-BASED FEATURE SELECTION

Utilizing the $3 \times k_i$ clusters, the next step involves selecting the most relevant features within each cluster to enhance prediction under various weather conditions. This selection is pivotal because, as demonstrated by previous research [27], the set of features that yield the most accurate forecasts can vary with changing weather conditions. By identifying these optimal features and examining the spatio-temporal correlations between them, the aim is to enhance the data correlation evaluation under varying weather conditions, ultimately leading to more precise forecasts. A bivariable copula-based method has been developed to implement feature selection.

1) COPULA THEORY

In statistics theory, copula theory is deployed to estimate the multivariate cumulative distribution function (CDF) of the input variables [29], thereby enabling an analysis of data correlations. More specifically, bi-variable copula functions are employed to identify the most relevant meteorological

variables for generating accurate forecasts under diverse weather conditions. Multivariate copula functions are introduced to quantify the spatio-temporal correlations among relevant meteorological variables.

According to the Sklar's theorem, the joint distribution F of variables $X = \{x_1, x_2, \ldots x_n\}$ with marginals $F_i(x_i)$ is formulated as [30]:

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)) \tag{2}$$

The CDF is in the range of zero to one. Hence, the input variables must be transformed to a standard uniform distribution, a procedure known as probability integral transform. Suppose a standard uniform distribution for variable x_n is $U_n = F_{x_n}(x_n)$, the probability integral transform is [30]:

$$\begin{aligned} F_{U_n}(u_n) &= P(U_n \le u_n) = P(F_{x_n}(x_n) \le u_n) \\ &= P(x_n \le F_{x_n}^{-1}(u_n) = F_{x_n}(F_{x_n}^{-1}(u_n)) = u_n, \end{aligned} \tag{3}$$

where F_{U_n} is the CDF of a uniformly distributed random variable, and U_n has a uniform distribution in the interval [0,1]. Thus, all the variables in X can be converted to uniform distributions [31]:

$$(u_1, u_2, \ldots u_n) = (F_1(x_1), F_2(x_2), \ldots F_n(x_n)) \tag{4}$$

The copula of the original variables $X = \{x_1, x_2, \ldots x_n\}$ is defined as the joint CDF of $(U_1, U_2, \ldots U_n)$:

$$C(u_1, u_2, \ldots u_n) = P(U_1 \le u_1, U_2 \le u_2, \ldots, U_n \le u_n) \tag{5}$$

Suppose the joint distribution of $(u_1, u_2, \ldots u_n)$ is $H(u_1, u_2, \ldots u_n)$. There exists a function $C()$ that integrates the marginal distribution and the joint distribution, which can be expressed as [30]:

$$H(u_1, u_2, \ldots u_n) = C(F_1(u_1), F_2(u_2), \ldots F_N(u_n)) \tag{6}$$

Based on the inverse transformation of a CDF of the marginal distribution, which refers to $u_i = F_i^{-1}(u_i)$, $i = 1, 2, \ldots N$, the expression of the Copula function can be obtained:

$$C(u_1, u_2, \ldots u_n) = H[F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots F_n^{-1}(u_n)] \tag{7}$$

Upon obtaining the Copula function $C()$, the joint density function of X is derived as:

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{\partial^2}{\partial x_1 \cdot \ldots \cdot \partial x_n} C(u_1, \ldots, u_n) \\ &= f_1(x_1) \cdot \ldots \cdot f_n(x_n) \cdot c(u_1, \ldots, u_n) \\ c(u_1, \ldots, u_n) &= \frac{\partial^2 C(u_1, \ldots, u_n)}{\partial u_1 \cdot \ldots \cdot \partial u_n} \end{aligned} \tag{8}$$

Therefore, Copula functions enable the independent modeling of the marginal distributions and the dependency structure for random variables with unique marginal distributions [31]. Copulas are classified into bivariate and multivariate Copulas. A bivariate Copula is a joint cumulative distribution function (CDF) of two random variables, and a multi-variate copula is used to model the joint CDF of multiple random variables.

2) SINGULARITY PROBLEM

The efficacy of copula functions is undermined when the input exhibits singularity. When the random variables X form a singular matrix, the corresponding U in (3) is also singular. The determinant of a singular matrix is zero, indicating that U_i is not invertible. Hence, the copula function C in (7) cannot be obtained. The Ledoit-Wolf shrinkage method [32] is introduced to address the problem.

Suppose U forms a singular matrix, and the corresponding covariance matrix is $\boldsymbol{S}$. To avoid singularity, $\boldsymbol{S}$ is then updated using the Ledoit-Wolf shrinkage method as,

$$\boldsymbol{S'} = (1 - \epsilon) \times \boldsymbol{S} + \epsilon \times trace(\boldsymbol{S}), \tag{9}$$

where ϵ is the shrinkage parameter and $\epsilon \approx 0$.

3) BI-VARIABLE COPULA-BASED FEATURE SELECTION

Archimedean copulas constitute an associative class of copulas, with the Clayton, Frank, and Gumbel copulas being among the most widely utilized instances of the class of Bi-variable copula. The copula function of the Clayton, Frank, and Gumbel copulas are denoted in (10)-(12) respectively [33].

$$C(u_1, u_2; \alpha) = (u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-\frac{1}{\alpha}} \quad \alpha \in (-1, \infty) \setminus \{0\} \tag{10}$$

$$C(u_1, u_2; \alpha) = -\frac{1}{\alpha} \log\left(1 + \frac{(e^{-\alpha u_1} - 1)(e^{-\alpha u_2} - 1)}{(e^{-\alpha} - 1)}\right) \quad \alpha \in (-\infty, \infty) \setminus \{0\} \tag{11}$$

$$C(u_1, u_2; \alpha) = \exp\left(-[(-\log(u_1))^{\alpha} + (-\log(u_2))^{\alpha}]^{\frac{1}{\alpha}}\right) \quad \alpha \ge 1 \tag{12}$$

Here, the Maximum log-likelihood estimation (MLE) (13) [34] is employed to estimate the dependence parameter.

$$\ell(\alpha) = \mathbf{argmax} \sum_{i=1}^{m} \log(c(F(x_1), F(x_2); \alpha)) \tag{13}$$

In (13), $F(x_1)$ is the marginal CDF of the original feature x_1. To evaluate the appropriateness of an Archimedean copula function in accurately capturing the dependence structure between two variables, the Bayesian Information Criterion (BIC) [35] is deployed:

$$BIC = -2 \times \log(c(F(x_1), F(x_2)) + 2 \times \log(N_s), \tag{14}$$

where N_s is the sample size of X.

The copula model differs from the traditional correlation analysis methods in that it does not limit the selection of marginal distributions. This flexibility allows for examining both linear and nonlinear associations between pairs of variables, as the copula function consistently estimates correlation values. Given a bi-variable copula, the Kendall's τ (15) and Spearman's rank correlation (16)

coefficients are utilized here to analyze data correlations [36].

$$\tau = 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1 \tag{15}$$

$$\rho = 12\int_0^1 C(u_1,u_2)du_1du_2 - 3 \tag{16}$$

The applicability of the coefficients is broadened by the fact that they are independent of the marginal distributions of the variables, in contrast to linear correlation coefficients. These coefficients quantify the probability of simultaneous substantial increases or decreases in the random variables, providing a valuable measure of the strength and nature of the correlation between variables.

In this work, the correlation between each of the meteorological features and the solar power is analyzed using Kendall's tau coefficient and Spearman's rank correlation coefficient in (15) and (16), respectively, based on the obtained bi-variable copulas. Instead of determining one combination of optimal features, the optimal 2,3,..., n features are obtained for each cluster through the aforementioned data correlation analysis, where n refers to the total number of features. Thus, $3 \times k_i \times (n-1)$ numbers of optimal feature sets are obtained. This approach comprehensively investigates the relationship between meteorological features and solar power, providing valuable insights into factors influencing forecasts.

C. SYNTHETIC DATA GENERATION

The generation of synthetic data aims to assess the spatio-temporal relationships among optimal feature sets, thereby quantifying the uncertainties within meteorological data that may not be apparent in original data but could emerge in future forecasting scenarios. The approach begins with developing multivariable copula functions for each optimal feature set and then selecting the best copula function to generate synthetic data related to prevailing weather conditions.

Fig. 3 demonstrates the rationale behind generating synthetic data. Initially, the process starts with using the data in optimal feature sets as the original data. Through the application of the copula function, additional synthetic samples are generated. These samples represent uncertainties not captured in the initial data, and data expansion strengthens the model's ability to handle uncertainties. Training the model with both original and synthetic data refines its forecasting robustness, making it more resilient and adaptable to unforeseen variations in weather conditions.

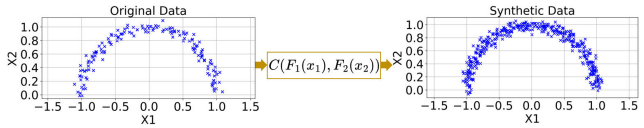

FIGURE 3. An illustration of synthetic data generation.

Specifically, the dependence structure between variables in each aforementioned $3 \times k_i \times (n-1)$ optimal feature set is estimated using multivariable copulas. Thus, joint CDFs can be obtained from multivariable copulas to generate synthetic meteorological data to incorporate the uncertainty of the weather conditions. Vine copula and Gaussian copula are introduced simultaneously in the research for such a purpose. The simultaneous integration of these distinct copula functions mitigates the limitations associated with the exclusive reliance on a single copula type, thereby enhancing model flexibility. The adaptability of the model is underscored by the selective application of either Vine or Gaussian copula, contingent upon the characteristics of the real-time weather conditions.

1) VINE COPULA

Vine copulas model dependencies among random variables by implementing a nested structure of bi-variate copulas, known as pair-copula [30]. The nested structure, and thus the relationship between the pair-copulas, determines the classification of the vine copula as either a C-vine, R-vine, or D-vine copula. The joint density function of variables is formulated as X as D-vine, C-vine, and R-vine copulas in equations (17)-(19) respectively.

$$f(x_1,\ldots,x_n;\boldsymbol{\alpha}) = [\prod_{j=1}^{n-1}\prod_{i=1}^{n-j} c_{i,(i+j)|(i+1),\ldots,(i+j-1);\alpha_j}] \cdot \prod_{h=1}^{n} f_h(x_h) \tag{17}$$

$$f(x_1,\ldots,x_n;\boldsymbol{\alpha}) = [\prod_{j=1}^{n-1}\prod_{i=1}^{n-j} c_{j,(i+j)|1,\ldots,(j-1)\alpha_j}] \cdot \prod_{h=1}^{n} f_h(x_h) \tag{18}$$

$$f(x_1,\ldots,x_n;\boldsymbol{\alpha}) = \prod_{h=1}^{n} f_h(x_h) \cdot \prod_{j=1}^{n-1}\prod_{e\in E_i} c_{j(e),r(e)|D(e);\alpha_j}(F(X_{j(e)}|X_{D(e)}), F(X_{r(e)}|X_{D(e)}) \tag{19}$$

The dependence parameter α_j of each pair-copula is estimated using (13). In (19), $e = j(e)$, and $r(e)|D(e)$ is the combination that determines each pair-copula, and E is the combination set. The conditioning sets D and conditioned sets j, r are utilized to establish the order of the arguments within the pair copula.

2) GAUSSIAN COPULA

Gaussian copulas model the variables' dependence as a Gaussian distribution. The correlation matrix of the multivariate normal distribution is utilized as the parameter to describe the variables' dependence in the Gaussian copula.

The Gaussian copula function is denoted as [33]:

$$C(u_1, \ldots u_n; \boldsymbol{p}) = \boldsymbol{\Phi_p}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)) \tag{20}$$

Nevertheless, either Vine copula or Gaussian Copula has its limitations. Gaussian copula makes it hard to capture the tail structure, and Vine copula is inefficient for high-dimensional data. This work introduces a novel solution to address the described research gap in comprehensively considering the volatile characteristics of weather conditions. The developed method applies the introduced types of multivariable copulas simultaneously, while the optimal copula is dynamically chosen based on future weather conditions at the time of the forecast. An XGBoost classifier is applied for this purpose. Thus, the spatio-temporal correlation between meteorological variables is *dynamically* quantified.

XGBoost [37], a scalable machine learning algorithm based on ensemble tree boosting, excels in classification tasks. By aggregating multiple weak decision trees, XGBoost demonstrates resilience against data outliers. Given its robustness, XGBoost is well-suited for applications within this study, notably in scenarios where meteorological data may include outliers due to rapidly changing weather conditions.

In the developed method, four introduced types of multivariable copulas are utilized simultaneously, resulting in a total of $3 \times k_i \times (n-1) \times 4$ copula functions, thereby creating a comprehensive pool of copula functions, denoted as set $\mathcal{C}$. An XGBoost-based classifier is then employed to identify and select the most suitable copula function from this pool based on the prevailing weather conditions. The classifier is trained using meteorological data $X = \{x_1, \ldots, x_n\}$ as input features, and the copula functions $\mathcal{C} = \{C_1, \ldots, C_{3\times k_i\times(n-1)\times 4}\}$ as labels. During training, the optimal copula functions under prevailing weather conditions are pre-analyzed based on final forecasting performance, enabling the precise selection of appropriate functions for accurate weather condition analysis.

For the developed multi-class copula function classification problem, the prediction as is formulated as:

$$\hat{C} = \mathbf{argmax}_j \sum_{k=1}^{k} f_{k,j}(x), \tag{21}$$

where $f_{k,j}(x)$ is the output of the k-th tree for class j (copula function j). The softmax function, defined as the loss function for each tree, is then applied in this study to convert raw predictions into probabilities.

$$L(C, \hat{C}) = -\sum_{j=1}^{k} \mathbf{1}(C = j) \log\left(\frac{\exp(\sum_{k=1}^{k} f_{k,j}(x))}{\sum_{s=1}^{k} \exp(\sum_{k=1}^{k} f_{k,s}(x))}\right) \tag{22}$$

Aggregating trees and improving the model using the greedy method, the objective of the applied XGBoost is defined as:

$$\begin{cases} \mathcal{L}^{(t)} = \sum_{i=1}^{n} L(C_i, \hat{C}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \\ \Omega(f_t) = \eta T_t + \frac{1}{2}\mu ||w_t||^2 \end{cases}, \tag{23}$$

where $f_t(x)$ is the independent regression tree with structure q and weights w at the t-th iteration. It is computationally challenging to list all tree structures in $f(\cdot)$ for optimization. To simplify the calculation, the second-order Taylor expansion and the greedy algorithm are applied to add branches of trees for optimization iteratively. The objective function (23) can be simplified as:

$$\mathcal{L}' = \frac{1}{2}\left[\frac{(\sum_{i\in\mathcal{I}_L} g_i)^2}{\sum_{i\in\mathcal{I}_L} h_i + \mu} + \frac{(\sum_{i\in\mathcal{I}_R} g_i)^2}{\sum_{i\in\mathcal{I}_R} h_i + \mu} - \frac{(\sum_{i\in\mathcal{I}} g_i)^2}{\sum_{i\in\mathcal{I}} h_i + \mu}\right] - \eta, \tag{24}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$. The objective $\mathcal{L}$ is thus readily optimized to find the leaf weights of the entire tree.

The developed approach, summarised in Algorithm 1, allows for selecting the copula function that best captures the underlying dependencies between the meteorological features and the solar power, providing a more accurate representation of the complex relationships between these variables. The well-trained copula function classifier selects different well-trained copula functions for each time step in the forecasting process. The selected copula function is then employed to generate synthetic samples $\mathcal{S}$, which are subsequently deployed for probabilistic forecasts.

D. PROBABILISTIC FORECASTING

An environment-aware model is developed to generate probabilistic forecasts using the synthetic data generated in the previous section. This model is integrated with the copula function classifier, enhancing its adaptability to fluctuating weather conditions by selecting appropriate data and models for forecast generation.

After the data-categorization step, $3 \times k_i \times (n-1)$ optimal feature sets corresponding to various weather scenarios are established. Each set is employed to train an individual XGBoost regression tree. The determination of the specific model for forecasting is guided by future weather conditions, as informed by the copula function classifier. A Huber loss function is employed for the XGBoost regression trees, which is defined as (25). Through the Huber loss function, the forecast model achieves robustness without compromising the precision essential for accurate prediction.

$$L_\delta(y, p) = \begin{cases} \frac{1}{2}(y-p)^2 & \text{for } |y-p| \le \delta, \\ \delta(|y-p| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{25}$$

The Huber loss function merges aspects of both mean squared error and mean absolute error, allowing for outlier impact mitigation and maintaining the model's predictive

Algorithm 1 Adaptive Selection of Copula Functions Based on Weather Conditions Leveraging XGBoost

Input: Meteorological data $\mathcal{X} = \{x_1, \ldots, x_n\}$, $3 \times k_i \times (n-1)$ number of optimal feature sets
Output: Trained XGBoost model, selected copula function $\hat{C}$, synthetic samples $\mathcal{S}$
Initialize pool of copula functions $\mathcal{C}$
for each optimal feature set **do**
    Estimate joint CDFs using Vine and Gaussian copulas
    Add all four well-trained copulas to $\mathcal{C}$
end for
$\mathcal{C}$ contains $3 \times k_i \times (n-1) \times 4$ copula functions
for each meteorological data $x_i \in \mathcal{X}$ **do**
    for each copula function $C_j \in \mathcal{C}$ **do**
        Evaluate final forecasting accuracy using synthetic data generated from C_j
    end for
    Find the best copula function for x_i
end for
Initialize XGBoost model with trees $f_{k,j}$ for each class j
for $t \leftarrow 1$ to T **do**
    for each instance $x_i \in \mathcal{X}$ **do**
        Compute gradient g_i and Hessian h_i for x_i
        Update tree structures and leaf weights
    end for
    Aggregate trees, update model using greedy method
    Optimize $\mathcal{L}^{(t)}$ with regularization term $\Omega(f_t)$
end for
$\hat{C} \leftarrow \arg\max_j \sum_{k=1}^{K} f_{k,j}(x)$
Return trained XGBoost model

accuracy. The δ parameter establishes a threshold for applying quadratic loss, enhancing sensitivity to minor discrepancies.

During training, $3 \times k_i \times (n-1) \times 4$ copula functions are derived from $3 \times k_i \times (n-1)$ optimal feature sets, resulting in the same number of synthetic sample feature sets. Each sample set, along with its corresponding original data, is used to train an XGBoost regression tree. A substantial number, specifically $3 \times k_i \times (n-1) \times 4$, of XGBoost regression trees are pre-trained. This pre-training process ensures that each model is finely tuned to accurately reflect the uncertainty captured by its feature set, enhancing the overall predictive accuracy in varying weather conditions. Through this comprehensive approach, the research leverages the dynamic nature of copula functions to provide a robust framework of weather-adaptive predictive modeling.

To generate forecasts using Numerical Weather Prediction (NWP) data, the trained XGBoost-based copula function classifier, and the trained XGBoost regression trees, the developed process is outlined as follows:

1) The trained copula function classifier uses the NWP data to select the best-fit copula function for future weather conditions;
2) The copula function determined in step 1 is used to generate synthetic samples, denoted as $\mathcal{S}$. The XGBoost regression tree that has been trained with feature types corresponding to those in $\mathcal{S}$ is selected;
3) Finally, $\mathcal{S}$ are used by the selected trained XGBoost regression tree to produce solar power forecasts. These forecasts are refined into probabilistic forecasts using the Gaussian Kernel Density Estimation [38] in (26).

$$\hat{f}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{(x - X_i)^2}{2h^2}\right) \tag{26}$$

III. CASE STUDIES

The forecasting performance of the developed model is initially examined through solar irradiance predictions, where it is compared to benchmark models in both literature [39] and industry. Subsequently, to further validate the model's effectiveness, the solar power data used in [14] are utilized for comparative analysis with models developed in [14] and [39], representing contributions in recent literature.

A. DATA CHARACTERISTICS AND EVALUATION CRITERIA

The data used for solar irradiance forecasting were sourced from the Open Weather database [40] and were gathered in Seattle, Washington, between January 1, 2019, and December 31, 2021. Meteorological data such as hourly temperature, zenith angle, dew points, feel-like temperature, air pressure, relative humidity, average wind speed, wind degree, cloud cover, and visibility are employed to forecast solar irradiance. The training dataset spans from January 1, 2019, to December 31, 2020, whereas the validation data is the last 20 % of training data. The test dataset spans from January 1, 2021, to December 31, 2021.

Following a comparison of the proposed model with benchmarks utilizing the aforementioned solar irradiance data, the proposed model's performance has been analyzed against benchmarks in the literature for solar power forecasting. The data used for solar power forecasting is identical to the dataset used in [14], spanning from 2017 to 2018 in Yulara, Australia. The meteorological data for solar power prediction comprises solar irradiance, temperature, wind direction, and wind speed, measured at 5-minute intervals. To ensure a fair comparison between the proposed model and the model in [14], the training data is consistent with the data used to develop the model in [14]. Additionally, both models were evaluated using the same testing data from 2018.

To evaluate the probabilistic forecast, two statistical metrics, i.e., the normalized Continuous Ranked Probability Score (nCRPS) (27) and the Pinball loss (28), are deployed.

$$\text{nCRPS}(F, y) = \frac{1}{y_{max}} \int_{-\infty}^{\infty} (F(\hat{y}) - 1_{y \le \hat{y}})^2 d\hat{y} \times 100\% \tag{27}$$

The nCRPS is similar to the normalized mean absolute error (nMAE) for a deterministic forecast [41]. Pinball Loss is

expressed as:

$$Pinball\ \ Loss_{\gamma}(y,\hat{y}) = \begin{cases} (y-\hat{y})\gamma, & if\ y \geq \hat{y} \\ (\hat{y}-y)(1-\gamma), & if\ \hat{y} > y \end{cases} \tag{28}$$

Since γ refers to the quantile level, the Pinball loss is used to evaluate the accuracy of quantile forecasts. A lower CRPS or Pinball loss implies higher forecasting accuracy, and a zero value is a perfect result. The CRPS and Pinball Loss are essential for evaluating probabilistic forecasts, as CRPS evaluates the accuracy and calibration of entire distributions, while Pinball Loss specifically measures the precision of quantile forecasts used in forming prediction intervals. Together, these metrics evaluate the reliability and accuracy of models.

B. BENCHMARK

The study utilizes benchmarks including PeEn, QR, and models introduced in prior research [14], [39] for a comprehensive comparative analysis. Specifically, the proposed model is compared with PeEn, QR, and the model in [39] for hourly solar irradiance forecasting. For solar power forecasting on a 5-minute interval basis, the proposed model is compared against the model in [14] and [39]. Next, these benchmarks are explained briefly.

1) A HYBRID FORECASTING MODEL COMBINING QUANTILE REGRESSION-BASED MONOTONE BROAD LEARNING SYSTEM (QRMBLS) WITH COPULA THEORY

In recent literature, authors in [14] developed a comprehensive model that integrates QR, MBLS, and Copula theory, effectively merging probabilistic forecasts with the spatial-temporal correlations of meteorological variables through Gaussian Copulas. This model provides probabilistic solar power forecasts by generating joint probability distributions that incorporate both solar power and meteorological data, thus obtaining predictions based on forecasted meteorological data.

2) GENERALIZED LAPLACE-BASED LONG-SHORT TERM MEMORY NETWORK (GL-LSTM)

The model developed in [39] is also used as the benchmark from recent literature for comparison. Authors in [39] developed a modified LSTM network whose output adheres to a generalized Laplace distribution, incorporating an innovative loss function derived from the CRPS. The hyperparameters of the Generalized Laplace LSTM (GL-LSTM) model—such as learning rate, layer count, neurons per layer, and dropout rate—were rigorously optimized through a comprehensive cross-validation process, evaluating numerous parameter configurations to enhance model performance.

3) PERSISTENCE ENSEMBLE

The persistence ensemble (PeEn) method [42] is a widely used benchmark in the industry.

4) QUANTILE REGRESSION

Quantile Regression (QR) [43] is also a benchmark in the industry. The applied QR model is built with the same training data and optimized with the cross-validation approach.

C. DATA CATEGORIZATION

Historical weather data is first divided into three categories: sunny, cloudy, and other. Each data group is then subdivided into smaller clusters using the described clustering approach, enabling a more granular analysis of the relationships between variables within each group. To identify the optimal number of clusters, the grid-search method is applied based on the Within-Cluster Sum of Squares (WSS) (29), a metric used to evaluate the clustering performance. A lower WSS value indicates that the data points are closer to their respective centroids, suggesting a better clustering solution. The clustering performance is illustrated in Fig. 4.

$$WSS = \sum_{i=1}^{k} \sum_{x_i \in CT_I} \left\| x_i - x_{CT_I} \right\|^2 \tag{29}$$

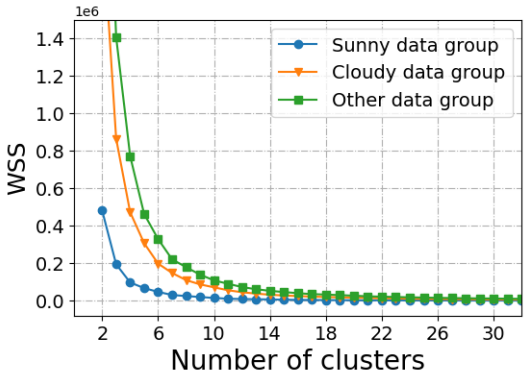

FIGURE 4. The variations of WSS with the increase in clusters.

Upon determining the optimal clusters using the Elbow method, the bi-variable copula-based approach is formulated to determine the optimal features in each cluster. For instance, consider one cluster of rainy data where the pair correlation of marginal distributions of each variable and GHI is illustrated in Fig. 5. Here, the cloud cover and humidity show a strong negative correlation, while temperature has a strong positive correlation with GHI. Kendall's τ and Spearman's rank ρ values corresponding to each variable are presented in Table 2, from which the same observation can be obtained. Consequently, the optimal features in each data cluster are chosen based on Kendall's τ and Spearman's rank ρ values. It is imperative to acknowledge that the relationship between GHI and meteorological variables is not uniform but varies significantly across distinct clusters, each representing a unique weather condition.

Upon identifying various optimal feature sets for each cluster, a pool of copula functions can be established using each of the optimal feature sets. This pool is utilized for training an XGBoost-based copula function classifier for future forecasting purposes, as illustrated in Fig. 2.

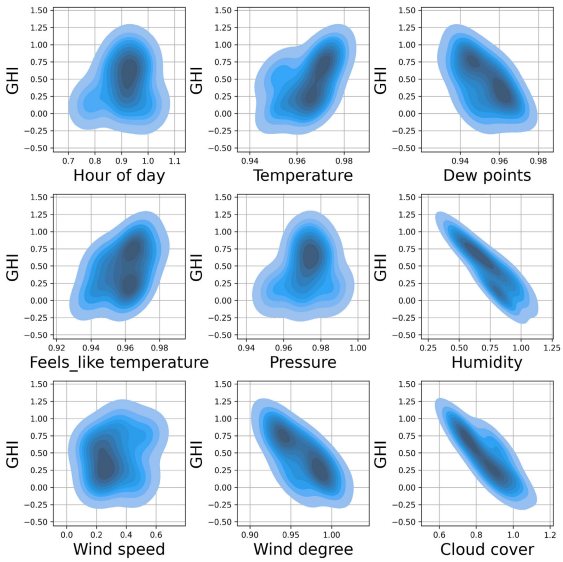

FIGURE 5. Pair correlation of marginal distributions when modeling copula functions of each meteorological variable and GHI in an example cluster.

TABLE 2. Kendall's τ and Spearman's rank ρ values between GHI and different meteorological variables.

	Hour	Temperature	Dew point	Feel-like temp	Pressure
τ	0	**0.43**	-0.28	0.42	0.15
ρ	-0.05	**0.62**	-0.43	0.61	0.33
	Humidity	**Wind speed**	**Wind degree**	**Clouds cover**	
τ	**-0.62**	0	-0.32	**-0.64**	
ρ	**-0.78**	0.073	-0.51	**-0.81**	

D. THE COPULA FUNCTION CLASSIFIER

The hyperparameters of the XGBoost classifier utilized in this study are refined through cross-validation, setting the learning rate at 0.001, the maximum tree depth at 6, and employing L2 regularization to prevent over-fitting by penalizing model complexity. An illustrative example of the forecasts and the dynamically selected optimal copula function for each time step is presented in Fig. 6, where night-time data has been excluded, as previously mentioned.

Fig. 6 shows solar irradiance forecasts for a sunny day, including an hour of cloudy conditions and a following rainy day. It illustrates how the trained copula function classifier selects an appropriate copula function for each time step based on hourly meteorological data. The labels "VC-3" and "GC-7" represent the Vine and Gaussian copula functions, estimated using three and seven meteorological features, respectively. The trained XGBoost regression trees use synthetic samples derived from these copula functions to produce probabilistic forecasts. The specific XGBoost regression tree applied varies with each time step, tailored to the copula function selected for diverse weather conditions.

This example highlights the adaptability of the forecasting model to different weather conditions by dynamically adjusting the data and functions used to estimate the spatio-temporal correlation between meteorological variables. It also demonstrates that an accurate forecast requires

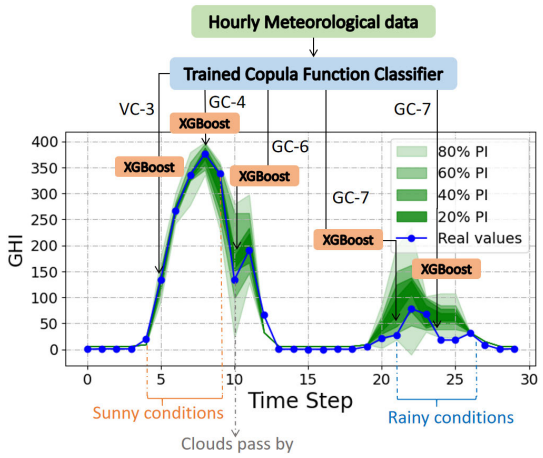

FIGURE 6. An example of forecasts for a sunny day (with one-hour cloudy condition) and a following rainy day. Night-time data is filtered out.

fewer features under stable weather conditions, e.g., sunny weather, and more under unstable conditions, e.g., cloudy or rain. The forecasts show the preference for Vine copula functions when fewer features are involved, aligning with their effectiveness in lower-dimensional data scenarios, whereas Gaussian copula functions are chosen as the data dimensions increase.

E. COMPARISON WITH BENCHMARK MODELS FOR DAY-AHEAD HOURLY SOLAR IRRADIANCE PREDICTION

The forecasting performance of the proposed model is compared to the benchmark methods. Various models generate day-ahead hourly solar irradiance forecasts, and their performance is evaluated based on nCRPS and Pinball loss metrics. Table 3 compares the developed model with benchmark models, including "no-CC" (the proposed model without copula classifier), "no-VineC" (the proposed model without Vine copulas, only using Gaussian copulas), PeEn, QR, and the GL-LSTM developed in [39]. The developed model consistently outperforms other models in terms of nCRPS and Pinball loss, achieving the lowest forecasting error across all periods except for July-August, where the GL-LSTM model [39] shows a slightly better performance in nCRPS. Nonetheless, whereas the proposed model's forecasts are slightly improved compared to the GL-LSTM model, its superiority becomes more pronounced under non-sunny or rapidly changing weather conditions. This advantage is further elucidated in the subsequent section, highlighting the robustness and effectiveness of the developed model across a broader range of meteorological scenarios.

The improved accuracy of the developed model can be attributed to the effective combination of the copula classifier and multiple types of copulas. This combination allows the model to better account for the complex relationships between variables under diverse weather conditions. The comparison of the proposed model with its modified versions, i.e., "no-CC" and "no-VineC" cases, emphasizes the significance of both the copula classifier and the application

of multiple types of copulas within the overall forecasting framework. By selecting the optimal copula function from among Vine copulas and Gaussian copulas, instead of solely relying on one type of copula, the model attains the flexibility to represent a wide range of variable dependencies. Furthermore, such a structure enhances the robustness of the model to variations in meteorological data and improves the forecast accuracy.

Fig. 7 presents the hourly variability and central tendencies of GHI across seasonal intervals, corresponding to the test dataset employed for error computation in Table 3. This visualization underscores the pronounced impact of seasonal shifts on solar irradiance and corroborates the robustness of the proposed model under diverse meteorological scenarios.

In Table 4, the peak memory usage during the training process, time for offline training, and online forecast for various models are documented. The computational experiments were executed on a system equipped with a 12th Gen Intel(R) Core(TM) i7-12700H CPU, 2300 MHz, and 14 cores and complemented by 16 GB of RAM. It is observed that offline training for the developed model is more extensive than other models; however, this process is a one-time requirement.

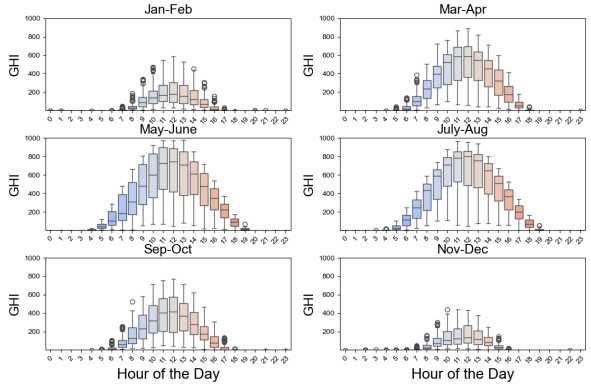

FIGURE 7. Seasonal GHI variations by hour.

TABLE 3. The comparison of the developed model with benchmark models in terms of prediction performance.

nCRPS	Jan-Feb	Mar-Apr	May-Jun	Jul-Aug	Sep-Oct	Nov-Dec
proposed	**5.2**	**6.8**	**6.7**	6.7	**5.9**	**6.3**
no CC*	6.2	7.4	7.1	7.4	6.3	7.4
no VineC*	5.7	7.6	7.3	7.2	6.6	6.7
PeEn	8.8	10.5	9.5	9.2	9.6	9.1
QR	7.2	9.8	10.2	9.8	8.3	7.4
GL-LSTM[39]	5.6	7.1	7.3	**6.6**	7.7	6.4

Pinball loss	Jan-Feb	Mar-Apr	May-Jun	Jul-Aug	Sep-Oct	Nov-Dec
proposed	**21.5**	**45.1**	**52.6**	**46.8**	**34.1**	17.3
no CC*	22.2	47	53.2	49.6	34.9	18
no VineC*	22.4	45.3	54.3	48.2	34.2	17.9
PeEn	50.4	87.2	85.7	80.1	56.9	35.8
QR	35.8	78.2	92.1	83.1	54.4	26.4
GL-LSTM[39]	25.2	49.2	56.5	48.2	37.7	**17**

*no -CC refers to the case of removing the copula classifier step in the proposed model. A fixed optimal feature set and a fixed type of copula function is applied to all weather conditions.
*no -VineC refers to removing the application of Vine-copula in the proposed model, only using Gaussian copulas.

Despite the high memory usage, the developed model can complete online forecasting tasks in about eight seconds, which is efficient enough for near real-time applications. Therefore, despite its substantial memory consumption and complex architecture during training, the model effectively supports real-world applications with satisfactory online forecasting speed and advanced functionalities.

TABLE 4. Memory usage, running time for offline training and online forecasting of different models.

	Developed Method	GL-LSTM	QR	PeEn
Offline training (s)	11980	3780	20	1
Online forecasting (s)	8.2	1.2	0.02	0.0006
Peak memory usage (MB)	670	462	401	32

F. COMPARISON WITH THE BENCHMARK MODEL FOR DAY-AHEAD MINUTELY SOLAR GENERATION PREDICTION

To further validate the proposed model's robustness under diverse weather conditions, solar power data from [14] are employed for a comparative study with models presented in [14] and [39], representing advanced models in recent literature. To ensure an equitable comparative analysis, the model in [14] is replicated with meticulous adherence to its specifications described in the paper, utilizing identical data and hyperparameters as delineated in [14]. Further details on these parameters are accessible in [44].

The model developed in [14] considers spatio-temporal correlations of meteorological variables for solar power forecasting, integrating Copula theory and machine learning methodologies. This methodology distinguishes between sunny and non-sunny days to independently quantify spatio-temporal correlations for each category. However, this distinction overlooks the variability within each category, as sunny days may experience non-sunny intervals and vice versa. Thus, the model does not dynamically analyze data correlations under fluctuating weather conditions, compromising forecast accuracy during rapidly changing weather conditions. The comparative analysis highlighted in this section demonstrates the superior performance of the developed model, which is based on dynamic spatio-temporal correlation analysis, thereby confirming its increased accuracy and robustness.

The dataset utilized in this case study is the same as the data employed in [14], which can be found in [44]. For a consistent comparison, both the training and test datasets for the developed model are the same as those used in [14].

Table 5 presents a performance comparison between the proposed model and the model in [14] and GL-LSTM [39] for different seasons and varying weather conditions. The results indicate that the proposed model generally outperforms the model in [14] and GL-LSTM [39], with much more accurate forecasts during non-sunny days. For sunny-day forecasts, the proposed model demonstrates comparable accuracy to the model [14] and GL-LSTM in spring and summer while outperforming these models during autumn and winter, when weather conditions tend to be more unstable. In the case of

TABLE 5. The comparison of the developed model with benchmarks.

nCRPS (100%)	Spring		Summer		Autumn		Winter	
	Sunny*	Other*	Sunny	Other	Sunny	Other	Sunny	Other
proposed	**1.6**	**3.56**	1.6	**2.7**	**1.1**	**1.7**	**1.6**	**3.2**
model in [14]	1.7	7.3	2.0	4.7	2.5	5.4	2.4	7.3
model in [39]	1.67	5.25	**1.1**	4.5	1.6	3.3	1.8	5.8

Pinball loss	Spring		Summer		Autumn		Winter	
	Sunny*	Other*	Sunny	Other	Sunny	Other	Sunny	Other
proposed	3.45	**5.62**	3.45	**4.73**	**1.63**	**3.36**	3.29	**5.2**
model in [14]	3.5	9.2	2.3	6.3	2.4	6.7	2.8	9.0
model in [39]	**2.35**	8.7	**1.6**	7.6	2.0	5.1	**2.6**	8.68

*Sunny refers to sunny days, which may include non-sunny conditions.
*Other refers to Non-sunny days, which may include sunny conditions.

non-sunny day forecasts, the proposed model significantly outperforms the model [14], with the proposed model demonstrating up to 60% greater accuracy during autumn. These findings further highlight the ability of the proposed model to enhance the robustness and reliability of solar generation forecasting in various weather conditions. The superior performance of the proposed model is enabled by the ability to dynamically quantify the correlation between different meteorological variables under various weather conditions.

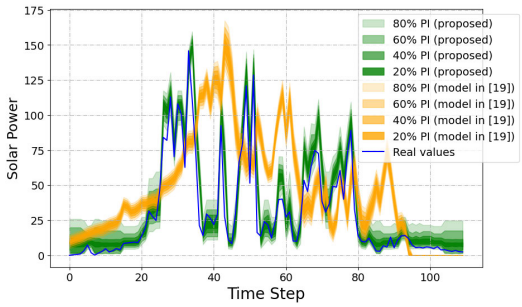

FIGURE 8. Forecast comparison of the developed model and the model in [14] under rapidly changing weather events.

Fig. 8 compares the proposed model, and the model described in [14], focusing on performance under rapidly changing weather conditions to underscore the effectiveness of our approach. Specifically, the figure illustrates forecasts for a day of heavy rainfall, generally categorized as a rare event based on its infrequency. Accurate forecasting during such rapidly changing weather conditions is crucial due to their significant potential impact on the stability and normal operation of power systems. While advanced models cited in existing literature often struggle to provide robust and precise forecasts under such challenging conditions, the proposed model demonstrates exceptional capability in predicting solar generation. This is especially noteworthy when precise forecasts are essential for ensuring the resilience and stability of power systems.

IV. CONCLUSION

In this work, a non-parametric model based on Copula theory and XGBoost is developed for probabilistic forecasting. Unlike traditional approaches that focus on identifying a single type of optimal copula function, the developed model concurrently examines multiple copula functions and selects the most suitable one according to forecasted weather conditions. This flexibility enables the model to capture the varying dependencies of meteorological variables better. By dynamically analyzing the spatio-temporal correlations of meteorological variables under diverse weather conditions, the proposed model enhances forecasting accuracy and robustness, particularly in rapidly changing weather situations when accurate forecasts are challenging.

Case studies employing real-world data from various locations and time intervals demonstrate that the developed model can substantially improve prediction accuracy when compared to benchmark models such as PeEn, QR, the GL-LSTM [39], and the model in [14], especially for non-sunny days. Furthermore, the model is adaptable to rapidly changing weather conditions, providing grid operators with valuable insights for reliable grid operations. The feasibility of the developed model has been tested in a real-world application during the 2022 American-Made Solar Forecasting Prize [26], where it secured the runner-up place and showcased its practical applicability for enhancing solar power forecasting.

Future research will aim to devise a more robust model by addressing the sensitivity of the developed model to location variations. Presently, models developed with data from a specific location may not perform adequately in other locations with completely different weather conditions. To overcome this, it is essential to incorporate spatio-temporal data analysis from multiple regions with different weather scenarios. This approach will facilitate the creation of a universally applicable model trained on a diverse dataset. Additionally, incorporating more relevant meteorological features, such as visibility, Global Normal Irradiance, and Direct Normal Irradiance (DNI), into the model could enhance forecast accuracy. This comprehensive model development strategy will ensure broader applicability and improved reliability of future forecasting models.

REFERENCES

[1] B. Li and J. Zhang, "A review on the integration of probabilistic solar forecasting in power systems," *Sol. Energy*, vol. 210, pp. 68–86, Nov. 2020.
[2] K. Doubleday, V. Van Scyoc Hernandez, and B.-M. Hodge, "Benchmark probabilistic solar forecasts: Characteristics and recommendations," *Sol. Energy*, vol. 206, pp. 52–67, Aug. 2020.
[3] 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.
[4] I. K. Bazionis and P. S. Georgilakis, "Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research," *Electricity*, vol. 2, no. 1, pp. 13–47, Jan. 2021.
[5] C. Wan, J. Lin, J. Wang, Y. Song, and Z. Y. Dong, "Direct quantile regression for nonparametric probabilistic forecasting of wind power generation," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2767–2778, Jul. 2017.
[6] Y. Yu, X. Han, M. Yang, and J. Yang, "Probabilistic prediction of regional wind power based on spatiotemporal quantile regression," in *Proc. IEEE Ind. App. Soc. Annu. Meeting*, 2019, pp. 1–16.

[7] X. Zhang, Y. Li, S. Lu, H. F. Hamann, B.-M. Hodge, and B. Lehman, "A solar time based analog ensemble method for regional solar power forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 268–279, Jan. 2019.

[8] A. Bracale, G. Carpinelli, and P. De Falco, "A probabilistic competitive ensemble method for short-term photovoltaic power forecasting," *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 551–560, Apr. 2017.

[9] F. von Loeper, P. Schaumann, M. de Langlard, R. Hess, R. Bässmann, and V. Schmidt, "Probabilistic prediction of solar power supply to distribution networks, using forecasts of global horizontal irradiation," *Sol. Energy*, vol. 203, pp. 145–156, Jun. 2020.

[10] M. Cui, V. Krishnan, B.-M. Hodge, and J. Zhang, "A copula-based conditional probabilistic forecast model for wind power ramps," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3870–3882, Jul. 2019.

[11] W. Dong, H. Sun, J. Tan, Z. Li, J. Zhang, and H. Yang, "Regional wind power probabilistic forecasting based on an improved kernel density estimation, regular vine copulas, and ensemble learning," *Energy*, vol. 238, Jan. 2022, Art. no. 122045.

[12] M. Sun, C. Feng, and J. Zhang, "Probabilistic solar power forecasting based on weather scenario generation," *Appl. Energy*, vol. 266, May 2020, Art. no. 114823.

[13] H. Panamtash, Q. Zhou, T. Hong, Z. Qu, and K. O. Davis, "A copula-based Bayesian method for probabilistic solar power forecasting," *Sol. Energy*, vol. 196, pp. 336–345, Jan. 2020.

[14] N. Zhou, X. Xu, Z. Yan, and M. Shahidehpour, "Spatio-temporal probabilistic forecasting of photovoltaic power based on monotone broad learning system and copula theory," *IEEE Trans. Sustain. Energy*, vol. 13, no. 4, pp. 1874–1885, Oct. 2022.

[15] S. Han, Y.-H. Qiao, J. Yan, Y.-Q. Liu, L. Li, and Z. Wang, "Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network," *Appl. Energy*, vol. 239, pp. 181–191, Apr. 2019.

[16] F. von Loeper, T. Kirstein, B. Idlbi, H. Ruf, G. Heilscher, and V. Schmidt, "Probabilistic analysis of solar power supply using D-vine copulas based on meteorological variables," in *Mathematical Modeling, Simulation and Optimization for Power Engineering and Management*, vol. 34, S. Göttlich, M. Herty, and A. Milde, Eds. Springer, 2021, pp. 51–68.

[17] A. Schinke-Nendza, F. von Loeper, P. Osinski, P. Schaumann, V. Schmidt, and C. Weber, "Probabilistic forecasting of photovoltaic power supply—A hybrid approach using D-vine copulas to model spatial dependencies," *Appl. Energy*, vol. 304, Dec. 2021, Art. no. 117599.

[18] D. L. Woodruff, J. Deride, A. Staid, J.-P. Watson, G. Slevogt, and C. Silva-Monroy, "Constructing probabilistic scenarios for wide-area solar power generation," *Sol. Energy*, vol. 160, pp. 153–167, Jan. 2018.

[19] D. van der Meer, D. Yang, J. Widén, and J. Munkhammar, "Clear-sky index space-time trajectories from probabilistic solar forecasts: Comparing promising copulas," *J. Renew. Sustain. Energy*, vol. 12, no. 2, Mar. 2020, Art. no. 026102.

[20] S. Rajabalizadeh and S. M. M. Tafreshi, "A practicable copula-based approach for power forecasting of small-scale photovoltaic systems," *IEEE Syst. J.*, vol. 14, no. 4, pp. 4911–4918, Dec. 2020.

[21] R. J. Bessa, V. Miranda, A. Botterud, Z. Zhou, and J. Wang, "Time-adaptive quantile-copula for wind power probabilistic forecasting," *Renew. Energy*, vol. 40, no. 1, pp. 29–39, Apr. 2012.

[22] S. Gill, B. Stephen, and S. Galloway, "Wind turbine condition assessment through power curve copula modeling," *IEEE Trans. Sustain. Energy*, vol. 3, no. 1, pp. 94–101, Jan. 2012.

[23] Y. Liu, Y. Zhou, Y. Chen, D. Wang, Y. Wang, and Y. Zhu, "Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: A case study in China," *Renew. Energy*, vol. 146, pp. 1101–1112, Feb. 2020.

[24] A. F. Ramírez, C. F. Valencia, S. Cabrales, and C. G. Ramírez, "Simulation of photo-voltaic power generation using copula autoregressive models for solar irradiance and air temperature time series," *Renew. Energy*, vol. 175, pp. 44–67, Sep. 2021.

[25] J. Yu, K. Chen, J. Mori, and M. M. Rashid, "A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction," *Energy*, vol. 61, pp. 673–686, Nov. 2013.

[26] (2022). *The American-Made Solar Forecasting Prize*. [Online]. Available: https://www.herox.com/SolarForecasting

[27] C. Lyu, S. Eftekharnejad, S. Basumallik, and C. Xu, "Dynamic feature selection for solar irradiance forecasting based on deep reinforcement learning," *IEEE Trans. Ind. Appl.*, vol. 59, no. 1, pp. 533–543, Jan. 2023.

[28] T. Konstantinou and N. Hatziargyriou, "Day-ahead parametric probabilistic forecasting of wind and solar power generation using bounded probability distributions and hybrid neural networks," *IEEE Trans. Sustain. Energy*, vol. 14, no. 4, pp. 2109–2120, 2023.

[29] F. Durante and C. Sempi, "Copula theory: An introduction," in *Proc. Workshop Warsaw Copula Theory Appl.* Berlin, Germany: Springer, Sep. 2009, pp. 3–31.

[30] P. Jaworski, F. Durante, W. K. Hardle, and T. Rychlik, *Copula Theory and Its Applications*, vol. 198. Berlin, Germany: Springer, 2010.

[31] W. Wu, K. Wang, B. Han, G. Li, X. Jiang, and M. L. Crow, "A versatile probability model of photovoltaic generation using pair copula construction," *IEEE Trans. Sustain. Energy*, vol. 6, no. 4, pp. 1337–1345, Oct. 2015.

[32] Y. Chen, A. Wiesel, and A. O. Hero, "Shrinkage estimation of high dimensional covariance matrices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 2937–2940.

[33] W. P. J. Philippe, S. Zhang, S. Eftekharnejad, P. K. Ghosh, and P. K. Varshney, "Mixed copula-based uncertainty modeling of hourly wind farm production for power system operational planning studies," *IEEE Access*, vol. 8, pp. 138569–138583, 2020.

[34] C. Savu and M. Trede, "Hierarchies of Archimedean copulas," *Quant. Finance*, vol. 10, no. 3, pp. 295–304, Mar. 2010.

[35] K.-H. Pho, S. Ly, S. Ly, and T. M. Lukusa, "Comparison among Akaike information criterion, Bayesian information criterion and Vuong's test in model selection: A case study of violated speed regulation in Taiwan," *J. Adv. Eng. Comput.*, vol. 3, no. 1, p. 293, Mar. 2019.

[36] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[38] S. Węglarczyk, "Kernel density estimation and its application," in *Proc. ITM Web Conf.*, vol. 23. Les Ulis, France: EDP Sciences, 2018, p. 00037.

[39] F. Lin, Y. Zhang, K. Wang, J. Wang, and M. Zhu, "Parametric probabilistic forecasting of solar power with fat-tailed distributions and deep neural networks," *IEEE Trans. Sustain. Energy*, vol. 13, no. 4, pp. 2133–2147, Oct. 2022.

[40] *Open Weather*. Accessed: Apr. 2022. [Online]. Available: https://openweathermap.org/

[41] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Appl. Energy*, vol. 157, pp. 95–110, Nov. 2015.

[42] H. Feddersen and U. Andersen, "A method for statistical downscaling of seasonal ensemble predictions," *Tellus A, Dyn. Meteorol. Oceanogr.*, vol. 57, no. 3, p. 398, Jan. 2005.

[43] R. Koenker and K. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, no. 4, pp. 143–156, 2001.

[44] N. Zhou. (2022). *Spatio Temporal Forecasting Based on Mbls*. [Online]. Available: https://github.com/ZhouNan2015/SpatioTemporalForecastingBasedOnMBLS

CHENG LYU (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Shanghai University of Electrical Power, Shanghai, China, in 2017, and the M.S. degree in electrical engineering from Syracuse University, Syracuse, NY, USA, in 2020, where he is currently pursuing the Ph.D. degree. His research project explores renewable energy forecasting and its application in power systems.

SARA EFTEKHARNEJAD (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Iran, in 2006, the M.Sc. degree from West Virginia University, Morgantown, WV, USA, in 2008, and the Ph.D. degree in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2012. She is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, USA. Her research interests include integrating renewable energy resources, uncertainty quantification in power grids, and power system stability with high penetration of renewables.