A Vine-Copula Method for Outlier Identification in Photovoltaic Arrays

Haitao Li[1,*], Weiqiong Song[1], Le Zhao[2], Shuai Guo[1], Wei Song[1], Li Huang[3]
[1]China Electric Power Research Institute,
Beijing, 100192, China
[2]State Grid Beijing Electric Power Company,
Beijing, 100031, China
[3]Southeast University,
Nanjing, Jiangsu, 211189, China
[*]zxiaoxx@163.com; swq_1984@q63.com; 2645471525@qq.com; 2949297892@qq.com;350944759@qq.com; huangli_jyf@seu.edu.cn

***Abstract*—To improve the operational efficiency and reliability of photovoltaic power stations, this paper introduces a novel approach to detect outliers in photovoltaic arrays using a Vine-Copula method. The procedure is divided into two distinct phases. Initially, it identifies deviations in the direct current (DC) component of the photovoltaic (PV) system. The following phase extends this by pinpointing irregularities in the DC voltage of the array. To model the interconnection between the PV current, irradiance, and temperature, the Vine-Copula is employed in this process. The optimisation of this function is based on the Akaike information criterion. Subsequently, a conditional probability model for the PV current is developed along with a formula to determine the quantile of this probability. This interval is then employed as the primary metric for detecting and eliminating current deviations. After refining the current data, a similar approach is taken to address voltage irregularities. The results of the simulation tests indicate that this proposed method is more effective, showing lower error rates and higher accuracy in detecting outliers, compared to other methods.**

***Index Terms*—Photovoltaic array; Anomalous data identification; Vine-Copula; Confidence interval; Interdependent structure.**

I. INTRODUCTION

In recent years, the escalating severity of global environmental pollution has propelled the imperative for the energy transition, driving research and the application of photovoltaic (PV) power generation technology, which has shown notable progress [1]. However, the operational dynamics of PV arrays is susceptible to various stochastic factors, resulting in a plethora of outliers in PV output data, which pose significant impediments to its analysis [2]–[4]. High-quality PV data serves as the basis for tasks such as monitoring the performance of PV arrays, playing a crucial role in ensuring system normal operation and grid stability. Therefore, the identification of outliers in the PV data is of paramount importance [5].

Manuscript received 21 May, 2024; accepted 17 August, 2024.
This work was supported by the Science and Technology Projects of SGCC under Grant No. 5108-202218280A-2-408-XG.

The causes of abnormal PV data are diverse, encompassing factors such as communication faults, equipment anomalies, and intentional power limitations [6]. Presently, two predominant categories of methods are employed to identify abnormal PV data: probabilistic statistical methods and machine learning methods. In the realm of probabilistic statistical methods, assumptions are typically made regarding the data following specific distributions. An approach, as described in [7], relies on the central limit theorem, assuming a normal distribution for PV power data and using the three-sigma method for outlier detection. However, the efficacy of this method is constrained by the influence of random factors such as weather on the distribution of PV data, leading to limitations in outlier identification under varying weather conditions. The authors in [8] propose a method based on the sliding standard deviation to cleanse abnormal data in the operational data of PV arrays, identifying abnormal data based on the upturn of the sliding standard deviation curve. However, when dealing with large data sets of PV arrays, the efficiency of the algorithm may become a significant concern, as the sliding standard deviation computation can consume substantial computational resources.

In terms of artificial intelligence methods, the primary approach involves quantifying sample isolation using metrics such as distance, density, and degree of isolation for anomaly detection. The authors in [9] use the local outlier factor algorithm combined with empirical clustering and successfully eliminate outliers in a wind database. However, this performance of this method may be influenced by the data distribution, resulting in a less stable recognition of different types of outliers. The authors in [10] introduce an image-based algorithm using clustering methods to map normal data and outliers of wind turbines. However, this method may be affected by illumination and image quality, and it exhibits limited adaptability to different wind turbine models and environmental conditions. The authors in [11] present the isolation forest anomaly detection model, characterised by linear time complexity and efficient perception of global sparse points. However, the model may perform inadequately when dealing with locally relative

sparse abnormal points, leading to a high error rate due to its reliance on global features without sufficient consideration of local features.

Among the references cited, artificial intelligence methods typically exhibit satisfactory performance on specific PV arrays, but lack generality. Probabilistic statistical methods, especially when faced with high-dimensional data, particularly in cases requiring consideration of multivariate relationships, can encounter computational complexity issues [12]. Copula Theory provides a more accurate method for characterising interdependencies between random variables, independent of specific assumptions about data distribution, enabling the capture of complex relationships between multivariate data [13]–[15]. The authors in [16] propose a PV abnormal data identification algorithm based on the irradiance process, utilising Copula joint distribution functions to construct the probability distribution relationship between irradiance and power. However, this study does not account for the influence of temperature. The authors in [17], using Copula, establish a conditional probability model of wind turbine power-wind speed, calculating confidence intervals given wind speed and confidence levels to identify outliers. Nevertheless, the aforementioned references do not consider the comprehensive impact of irradiance and temperature on PV output data, which exhibit a strong correlation with factors such as PV power. Additionally, the various types of Copula functions pose a challenge; the use a single-type Copula function to characterise interdependence structures among high-dimensional variables may result in poor flexibility and accuracy.

To enhance recognition accuracy, this study builds on existing methods by incorporating irradiance and temperature as features, expanding the modelling of bivariate interdependence structures to trivariate interdependence structures. Furthermore, a more flexible and accurate Vine-Copula method is selected for modelling. Additionally, this study focusses on anomaly data identification in distributed PV arrays, which, in contrast to traditional centralised PV stations that only identify power data anomalies, detects anomalies in the direct current voltage and current. On this basis, this study derives confidence interval calculation formulas for PV current and voltage and validates the effectiveness of the proposed method.

II. Analysis of Photovoltaic Output Anomaly Data and Feature Selection

A. Analysis of Photovoltaic Output Anomaly Data

Compared to normal operational states, photovoltaic (PV) arrays exhibit significant variations in maximum power point power, the voltage, and the current during abnormal operational states. Various factors contribute to the occurrence of anomalies in PV arrays, including high-potential grounding faults, short-circuit faults, open-circuit faults, partial shading, dust accumulation, and ageing. These factors correspond to five distinct abnormal operational states. Table I illustrates the temporal fault characteristics of current, voltage, and power for these five abnormal operational states of PV.

From Table I, it is evident that anomalies induced by abnormal PV operational states manifest themselves primarily as decreases in current and voltage.

TABLE I. TEMPORAL FAULT CHARACTERISTICS UNDER VARIOUS ANOMALOUS OPERATIONAL STATES.

Abnormal state		Output current	Output voltage	Output power
High voltage to ground fault	Fault transient	Momentarily becomes 0	Momentary decline	Momentary decline
	After stabilisation	Essentially unchanged	Decline	Decline
Short-circuit fault	Fault transient	Momentary decline	Essentially unchanged	Momentary decline
	After stabilisation	Essentially unchanged	Decline	Decline
Open-circuit fault	Fault transient	Momentary decline	Essentially unchanged	Momentary decline
	After stabilisation	Decline	Essentially unchanged	Decline
Partial shading obstruction	Fault transient	Momentary decline	Momentary decline	Momentary decline
	After stabilisation	Decline	Decline	Decline
Accumulated dust or ageing		Gradual decline	Gradual decline	Gradual decline

Moreover, within the same abnormal state, the anomalous features of current and voltage differ. Consequently, the PV anomaly data identification method proposed in this study considers photovoltaic output current and voltage as the objects of identification, addressing each in two distinct steps for current and voltage recognition, respectively.

In addition to the decrease in voltage and current caused by abnormal PV states, there exists another subset of anomalies that surpass normal values. Based on the distinctive features of the PV data anomalies, these anomalies can be categorised into four types. The first type of anomaly is characterised by a continuous period of values exceeding the normal range. Such faults are typically caused by malfunctions in communication devices or sensors. The second type of anomaly is identified by a sustained period of values that fall below the normal range. The primary causes of this type of anomaly include PV power limitations, abnormal states, or faults in communication or sensor devices. The third type of anomaly is distinguished by adequate irradiance, yet the PV output voltage or current is recorded as zero. The main causes of this type of anomaly include inverter malfunctions, faults in communication devices or sensors, and shutdown of the generator unit. The fourth type of anomaly is marked by outliers near the normal range. These data points are the result of noise propagated by communication devices or sensor signals, random fluctuations from external input, and inaccuracies in maximum power point tracking. Recognising the diverse nature of these types of anomalies is crucial to developing a comprehensive understanding of abnormal conditions in PV data and improving the effectiveness of anomaly detection methods.

B. Feature Selection

Environmental factors exert a significant influence on the output characteristics of a PV array, particularly the intensity of sunlight and temperature. The PV output power is nearly proportional to solar irradiance and inversely proportional to temperature. Due to the substantial correlation between

humidity and temperature, the impact of humidity can be ignored. If data cleaning methods only consider the relationship between one type of environmental factor and the output distribution, it becomes challenging to identify all different types of anomalies. Meanwhile, the influence of temperature on the output current and voltage of photovoltaics varies. As the temperature of the photovoltaic panel increases, the open-circuit voltage of the photovoltaic module decreases, while the output current of the photovoltaic module remains almost unchanged [18]. Therefore, it is necessary to perform separate data cleaning for PV voltage and current.

Table II presents the Spearman rank correlation coefficients between irradiance, temperature, and voltage-current pairs. It is evident that both temperature and irradiance exhibit a high correlation with both current and voltage. To address this, this paper proposes an outlier cleaning method based on the distribution characteristics of the output from the PV array. This method employs the Vine-Copula algorithm in two steps to remove outliers. For simplicity, the following sections collectively refer to voltage and current as target variables, and the next section will elaborate on the proposed algorithm.

TABLE II. CORRELATION COEFFICIENTS AMONG FEATURE VARIABLES.

	Current	**Voltage**	**Temperature**	**Irradiance**
Current	1	1	0.771	0.997
Voltage	1	1	0.771	0.997
Temperature	0.771	0.771	1	0.609
Irradiance	0.997	0.997	0.609	1

III. ALGORITHM CORE PRINCIPLES AND DATA CLEANSING METHODS

The Vine-Copula, capable of capturing intricate nonlinear relationships among multivariate parameters based on univariate marginal distributions, facilitates the computation of confidence intervals for target variables under conditions of temperature and irradiance. Values deviating from these confidence intervals are identified and excluded as outliers. The establishment of a Vine-Copula model generally encompasses two phases: Initially, it involves the determination of the marginal distributions for each variable; subsequently, it requires the identification of the vine structure and the optimal Copula function for each node. The following discussion delves into the algorithmic principles from these two perspectives.

A. Nonparametric Kernel Density Estimation

Establishing marginal distributions is a critical step in statistical analysis and modelling, which is typically employed to address the distribution of individual random variables. Common methodologies for constructing marginal distributions include parametric estimation, nonparametric methods, distribution fitting, and empirical distribution function approaches. In cases like photovoltaic current, voltage, irradiance, and temperature, which are often continuous variables with potentially indistinct distribution types, this study adopts the nonparametric kernel density estimation (KDE) approach. KDE establishes marginal distributions by estimating probability densities based directly on the data itself, obviating the need for pre-assuming a distribution type. The formula for nonparametric kernel density estimation is as follows:

$$\hat{F}(x)=\frac{1}{N}\sum_{n=1}^{N}G(\frac{x-X_n}{h}), \tag{1}$$

$$G(x)=\int_{-\infty}^{x}K(t)dt, \tag{2}$$

$$K(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2}), \tag{3}$$

where $\hat{F}(x)$ denotes the estimated value of the marginal distribution for the random variable x, N represents the total number of samples, X_n signifies the n^{th} sample value of the random variable x, the variable h is typically referred to as the bandwidth or smoothing parameter, and $K(x)$ refers to the Gaussian kernel function.

B. Copula Theory

1. Sklar theorem.

A Copula can be succinctly described as a function that "binds or couples a multivariate distribution function to its one-dimensional marginal distribution functions" [19]. This characterisation is primarily articulated by Sklar's theorem, which posits that for a d-dimensional distribution function F with continuous marginal distributions, there exists a unique Copula function C such that for all $\boldsymbol{x}=(x_1,\cdots,x_d)\in(\mathbb{R}\cup\{-\infty,+\infty\})^d$, the following holds

$$F\left(x\right)=C\left(F_1(x_1),F_2(x_2),\cdots F_d(x_d)\right). \tag{4}$$

The joint probability density function thereof can be articulated as such

$$f\left(\boldsymbol{x}\right)=c\left(F(x_1),\cdots,F(x_d)\right)\times f(x_1)\times f(x_2)\cdots f(x_d), \tag{5}$$

where $F(x)$ denotes the marginal distribution function of a single variable, $C(\cdot)$ signifies the Copula function, $c(\cdot)$ represents the Copula density function, and $f(x)$ stands for the univariate probability density function.

2. Types of Copula functions

The Elliptical (Ellipse-Copula) and Archimedean (Archimedean-Copula) families represent two prevalent types of Copula functions.

Elliptical Copulas, such as the normal and t-Copula functions, are a class of Copula functions that are used to describe symmetric tail dependencies. On the other hand, Archimedean Copulas, encompassing the Gumbel, Clayton, and Frank Copula functions, are employed to characterise asymmetric or asymptotically independent tail features.

The Gaussian, Clayton, and Gumbel Copulas each possess a single parameter, whereas the t-Copula encompasses two parameters, with the additional parameter controlling the strength of tail dependence in bivariate distributions. The Clayton Copula exhibits greater correlation in the lower tail than in the upper, while the Gumbel Copula, another asymmetric Copula, shows greater dependence in the upper tail than in the lower. This diversity in Copula functions underscores their suitability for different types of variables, necessitating the selection of the most appropriate Copula

function based on the specific correlation characteristics of the variables involved. Detailed formulations of these Copula functions can be found in [20].

C. Regular Vine Structure

Confronted with the complex tail dependencies of high-dimensional variables, Vine-Copula theory introduces the structure of Regular Vines, decomposing the multivariate joint probability density function into a cascading form of multiple bivariate Copula density functions. This approach effectively reduces a multidimensional variable problem to several bivariate issues, thereby addressing the inaccuracies inherent in using a singular Copula function to describe multidimensional variables. Consequently, Vine-Copula demonstrates improved efficacy in capturing various types of dependency relationships.

The Regular Vine structure is a hierarchical tree-like configuration, typically comprising multiple levels. Each level features a tree, and each branch of the tree encompasses two nodes, representing two variables. Each branch symbolises a Copula function. Hence, Regular Vines are adept at modelling intricate dependency relationships among multiple variables, with each branch affording the flexibility to choose different Copula functions tailored to varied dependency patterns. Figure 1 illustrates two commonly utilised vine structures: the C-vine and the D-vine.

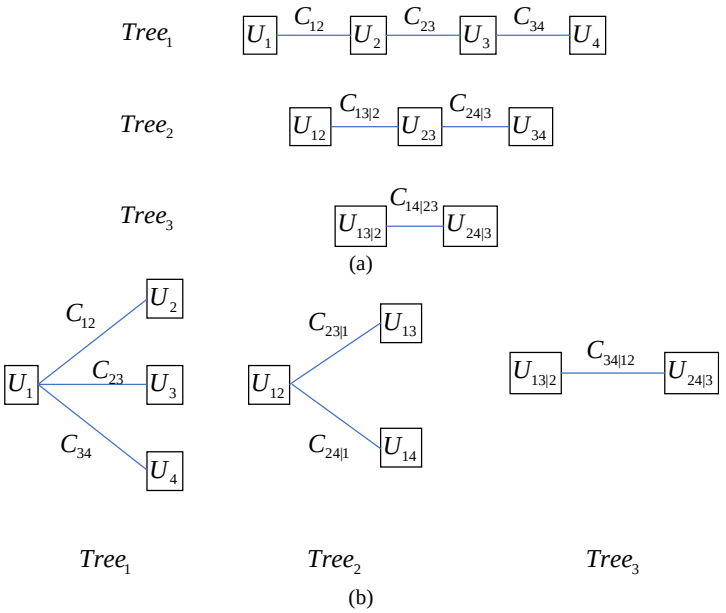


Fig. 1. Two commonly used forms of rule-based trellis structures: (a) D-vine structure; (b) C-vine structure.

The D-vine structure offers greater flexibility, particularly when the dependency relationships do not follow a specific sequence. In contrast, the C-vine is better suited for scenarios with a clear dependency structure. In this paper, considering the distinct dependency relationships among the target variable, irradiance, and temperature, the C-vine structure is used.

The definition of a Regular Vine $V=\{T_1,\cdots,T_{n-1}\}$ representing the joint probability of n-dimensional variables is as follows:

– $T_1=\{N_1,E_1\}$ represents the first tree, $N_1=\{1,\cdots,n\}$ denotes the nodes of the first tree, and E_1 signifies the set of edges in the first tree;

– For $i=2,\cdots,n-1$, $N_i=E_{i-1}$ represents a node of the tree T_i;

– If there exists an edge in E_i that connects nodes a and b, then a and b must share a common node in T_{i-1} (i.e., a and b are edges in the tree T_{i-1}). This property is often referred to as the "proximity condition", as it indicates that two nodes in tree T_{i-1} are adjacent only if the corresponding edges in T_i are adjacent, meaning they share a common node;

– If the number of edges connected to each node in T_1 does not exceed 2, then the Regular-Vines is a D-vine; if for each tree $T_i(i=i,\cdots,n-1)$, there exists a unique node with $n-i$ edges connected, i.e., the root node, then the Regular-Vines structure is a C-vine;

– Specifically, in the context of modelling for photovoltaic power anomaly detection, it is crucial to ensure that, aside from the photovoltaic power itself, all other variables are treated as conditional variables. This requires that, in all generated trees, the nodes associated with photovoltaic power are linked to one single edge. This constraint is imposed to maintain the conditional independence inherent in the tree structure, thereby facilitating the effective capture of the relationships between photovoltaic power and other conditional variables.

Following adherence to the aforementioned conditions, the joint probability density function of d-dimensional variables can be delineated as follows

$$f(\boldsymbol{x})=\prod_{n=1}^{d}f_n(x_n)\times$$
$$\times\prod_{i=1}^{d-1}\prod_{e\in E_i}c_{j(e),k(e)|\,D(e)}\left(F\left(x_{j(e)}\mid\boldsymbol{x}_{D(e)}\right),F\left(x_{k(e)}\mid\boldsymbol{x}_{D(e)}\right)\right) \quad (6)$$

where E_i denotes the set of all edges in the i^{th} generated tree, $j,k\mid D$ represents the edge of the i^{th} tree determined by the

two nodes j,k, D is the set of all nodes in the i^{th} tree, and $c_{j(e),k(e)|\,D(e)}$ represents the bivariate Copula density function corresponding to the e edges. The conditional distribution function $F\left(x \mid \boldsymbol{x}\right)$ can be expressed as

$$F(x_j \mid \boldsymbol{x}_{D\cup k}) = \frac{\partial C_{j,k|D}(F(x_j|\boldsymbol{x}_D), F(x_k \mid \boldsymbol{x}_D))}{\partial F(x_k \mid \boldsymbol{x}_D)}. \quad (7)$$

For simplicity, the following function is introduced

$$h_{j,k|D}(F(x_j \mid \mathbf{x}_D), F(x_k \mid \mathbf{x}_D)) = \\ = \frac{\partial C_{j,k|D}(F(x_j|\boldsymbol{x}_D), F(x_k \mid \boldsymbol{x}_D))}{\partial F(x_k \mid \mathbf{x}_D)}. \quad (8)$$

Hence, (7) can be succinctly expressed as

$$F(x_j \mid \boldsymbol{x}_{D\cup k}) = h_{j,k|D}(F(x_j \mid \mathbf{x}_D), F(x_k \mid \mathbf{x}_D)). \quad (9)$$

D. Optimisation of the Vine-Copula Model

The flexibility of the Vine-Copula model is manifested on two fronts: first, in the availability of various vine structures to choose from and second, for each edge within the vine structure, there exists a variety of Copula functions available. Subsequently, optimisation of the Copula model is pursued, addressing both of these aforementioned dimensions.

1. Optimising Copula functions

Initially, parameter estimation for all types of Copula functions is performed based on sample fitting. Given x, y as a pair of random variables with sample sets $x=(x_1,\cdots x_i,\cdots x_N), y=(y_1,\cdots y_i,\cdots y_N)$, the maximum likelihood method is employed for parameter estimation

$$\hat{\theta} = \max_{\theta} \sum_{i=1}^{N} \ln c\left(F_x\left(x_i\right), F_y\left(y_i\right)|\theta\right), \quad (10)$$

where θ represents the parameter set of the Copula function and *N* is the total number of samples.

Subsequently, leveraging the Akaike information criterion (AIC), Copula functions are sifted. The AIC serves as a standard for evaluating the goodness of fit in statistical models, incorporating not only the likelihood function for assessing fitting quality but also considering the impact of model complexity. This aids in striking a balance between model complexity and fitting goodness, thereby mitigating the risk of overfitting. As elucidated in [21], this method is considered more suitable for the selection of the Copula function compared to alternative evaluation approaches. Based on the parameters estimation outcomes, the AIC evaluation metric for each Copula function is computed using (11)

$$f_{\mathrm{AIC}} = -2\sum_{i=1}^{N} \ln c\left(F_x\left(x_i\right), F_y\left(y_i\right)|\ \theta\right) + 2k, \quad (11)$$

where k represents the number of parameters included in the Copula function. The Copula function with the minimum AIC evaluation metric is selected as the optimisation result.

2. Optimising the vine structure

Within the vine structure, the stronger the interdependence among nodes in a regular vine, the more accurately it characterises the dependency structure of high-dimensional variables. In particular, the strength of dependence in the first generated tree exerts the most significant impact on the model accuracy. In addressing this, the present study employs the Kendall correlation coefficient to quantify the magnitude of interdependence. Furthermore, an orderly approach, specifically the sequential method, is used to optimise the vine structure. Table III elucidates the specific optimisation steps taken.

TABLE III. OPTIMISATION METHODS FOR VINE STRUCTURES.

Algorithm: Vine structure optimisation based on the sequential method.
Input: $\mathbf{U}_I, \mathbf{U}_T, \mathbf{U}_E$
Output: Optimal vine structure and its parameters
1. Compute the Kendall correlation coefficient $\tau_{j,k}$ for all possible variable pairs $\{j,k\}$, where $j,k \in \{I,T,E\}$ and $j \neq k$.
2. Consider the Kendall coefficient $\tau_{j,k}$ as weights for the edges $e_{j,k}$, solve for the maximum spanning tree: $\max \sum \lvert \tau_{j,k} \rvert$, obtaining the first generated tree $\max \sum \lvert \tau_{j,k} \rvert$.
3. Optimise the Copula function for each edge in the generated tree, calculate the conditional distribution $F(x_j \mid x_k), F(x_k \mid x_j)$ based on (7).
4. Optimise the Copula function for the edges in the second tree, calculate the conditional distribution $F(x_j \mid x_{k\cup D}), F(x_k \mid x_{j\cup D})$ based on (7).

IV. IDENTIFICATION OF ANOMALOUS DATA

A. Methodology Design

Irradiance and temperature are stochastic variables correlated with the target variable. The process involves calculating the marginal distributions of these variables and using the Vine-Copula function to articulate the correlations between irradiance, temperature, and the target variable. Given specific values of irradiance, temperature, and confidence, the conditional probability distribution of the target variable is determined, yielding the upper and lower quantile values. The interval formed by these quantile points under various irradiance and temperature conditions constitutes the confidence interval. This interval encapsulates the probabilistic distribution relationship among irradiance, temperature, and the target variable. Data points within this confidence interval are considered to conform to normal patterns under the given confidence level, reflecting more closely the real energy generation performance of photovoltaic power stations.

The specific steps are as follows:

1. Selecting the raw data. Encompassing irradiance, ambient temperature, photovoltaic array output current, and voltage;
2. Data preprocessing. Aligning the four aforementioned data sets based on time sequences, imputing missing values, and excluding data points where voltage and current are zero to mitigate the impact of cumulative effects arising from zero data accumulation;
3. Removing current data. Initially, the marginal distributions of current, temperature, and irradiance are computed. Based on these distributions, an appropriate vine structure is selected. Subsequently, the optimal Copula function is chosen for each edge. This is followed by derivation of the conditional probability distribution

function for the current. Finally, an appropriate confidence level is set to ascertain the confidence interval for the current, and data points falling outside this interval are excluded;

4. Removing the voltage data. Following the exclusion of data in the previous step, the marginal distributions of voltage, temperature, and irradiance are recalculated. This process is repeated to determine the confidence interval for the voltage, after which the data points falling outside this voltage confidence interval are removed. The overall workflow is depicted in Fig. 2.

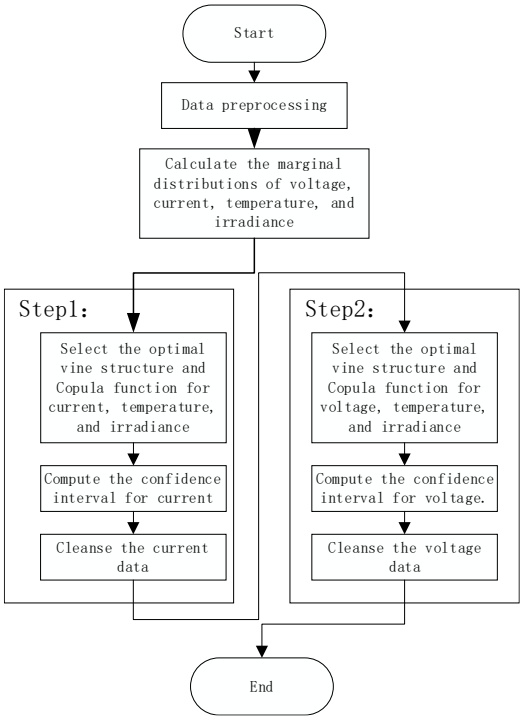


Fig. 2. The overall schematic diagram of the Two-Step Vine-Copula Method.

B. Detailed Implementation Steps

1. Calculation of marginal distribution of variables

As depicted in (12), the sample set S is a matrix of size $N\times 4$, where N represents the sample capacity, and 4 is the dimension of the feature vector. Each column of the matrix corresponds, in order, to the voltage, current, temperature, and irradiance sample sets, denoted as X_V, X_I, X_T, X_E

$$\mathbf{S}=\begin{bmatrix}\mathbf{X_1}\\ \vdots\\ \mathbf{X_n}\\ \vdots\\ \mathbf{X_N}\end{bmatrix}=\begin{bmatrix}x_{\mathrm{V},1} & x_{I,1} & x_{\mathrm{T},1} & x_{E,1}\\ \vdots & \vdots & \vdots & \vdots\\ x_{\mathrm{V},n} & x_{I,n} & x_{\mathrm{T},n} & x_{E,n}\\ \vdots & \vdots & \vdots & \vdots\\ X_{\mathrm{V},N} & X_{I,N} & x_{\mathrm{T},N} & x_{E,N}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_V\\ \mathbf{X}_I\\ \mathbf{X}_T\\ \mathbf{X}_E\end{bmatrix}^{\mathrm{T}}. \quad (12)$$

Applying the nonparametric kernel density estimation method to transform the sample set X_V, X_I, X_T, X_E into a form of marginal distribution yields a new set $\mathbf{U}_V, \mathbf{U}_I, \mathbf{U}_T, \mathbf{U}_E$. This transformed set is then utilised as the α input for the model. The distribution histograms of each random variable before and after the transformation are provided in Appendix A.

2. Current cleansing

– Identification of vine structure and copula functions

For the variable $\mathbf{U}_I, \mathbf{U}_T, \mathbf{U}_E$, optimise its vine structure following the steps outlined in Table III, and subsequently select the Copula function for each edge based on (11).

– Establishment of conditional probability model for current

Taking $\mathbf{U}_E$ as the root node, for example, the formula for the conditional probability distribution function $F(x_I \mid x_T x_E)$ of the photovoltaic current *I* with respect to temperature *T* and irradiance *E* is given by:

$$\begin{cases} F(x_I \mid x_E) = h_{I,E}(F(x_I), F(x_E)), \\ F(x_T \mid x_E) = h_{T,E}(F(x_T), F(x_E)), \end{cases} \quad (13)$$

$$F(x_I \mid x_T x_E) = \\ = h_{I,T|E}(F(x_I \mid x_E), F(x_T \mid x_E)), \quad (14)$$

where x_I, x_T, x_E represents the photovoltaic current, irradiance, and temperature, respectively; function $h_{j,k|D}(\cdot)$ corresponds to (7); $F(x_I \mid x_T x_E)$ denotes the conditional probability distribution function of the photovoltaic current.

– Solving the confidence interval for photovoltaic current

Determining the confidence interval involves calculating the upper and lower quantile points that define the interval boundaries. This calculation process is inversely related to the computation of the conditional probability distribution, as described in (13) and (14).

Figure 3 presents a schematic illustration of the quantile point calculation process. Solve following the steps shown in (15) to (18):

$$F(x_T \mid\ x_E) = h_{T,E}(F(x_T), F(x_E)), \quad (15)$$

$$F(x_I \mid\ x_E) = h^{-1}_{I,T|E}(\alpha, F(x_T \mid\ x_E)), \quad (16)$$

$$F(x_I) = h^{-1}_{I|E}(F(x_I \mid\ x_E), F(x_E)), \quad (17)$$

$$x_I = F^{-1}(F(x_I)), \quad (18)$$

where α represents the conditional probability distribution values corresponding to the quantiles, i.e., $F(x_I \mid x_T x_E) = \alpha$; function $h^{-1}_{j,k|D}(\cdot)$ denotes the inverse function of (7), and its specific form is provided in [22].

Setting the confidence probability to α implies that α percent of the data lies within the probability interval. Let

$$\beta = 1-\alpha. \quad (19)$$

Due to the uneven distribution of abnormal values in photovoltaic current data, type 2 anomalies are typically more prevalent. Therefore, setting an asymmetry coefficient κ for the confidence interval, with quantile probabilities β_1, β_2 for the upper and lower bounds, respectively, is defined as follows, expressing the probability of data points exceeding the upper bound as β_1 and falling below the lower bound as β_2:

$$\beta_1 = (1-\kappa)\beta, \quad (20)$$

$$\beta_2 = \kappa\beta. \tag{21}$$

When $\kappa = 0.5$, the confidence probability interval is symmetrical, and when $\kappa > 0.5$, the confidence probability interval is shifted upward.

For the sample set S as shown in (12), substitute $x_I = x_{I,n}$, $x_E = x_{E,n}$, and $\alpha = \beta_1$ into (15) to (18) to obtain the upper bound $x_{I,n,\text{up}}$ of the current confidence interval for the n^{th} sample. Similarly, substitute $x_I = x_{I,n}$, $x_E = x_{E,n}$, and $\alpha = \beta_2$ into (15) to (18) to obtain the lower bound $x_{I,n,\text{low}}$ of the current confidence interval for the n^{th} sample.

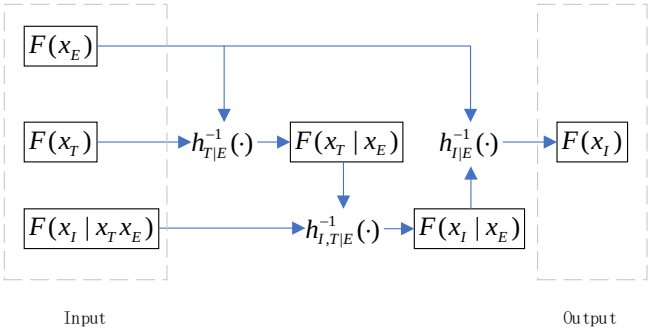


Fig. 3. Quantile calculation procedure.

– Cleanse the current data

When the current of the n^{th} sample is $x_{I,n} \in [x_{I,n,\text{low}}, x_{I,n,\text{up}}]$, the sample point is labelled as a normal data point. Otherwise, the sample point is labelled as an outlier, and the sample points labelled as outliers are removed.

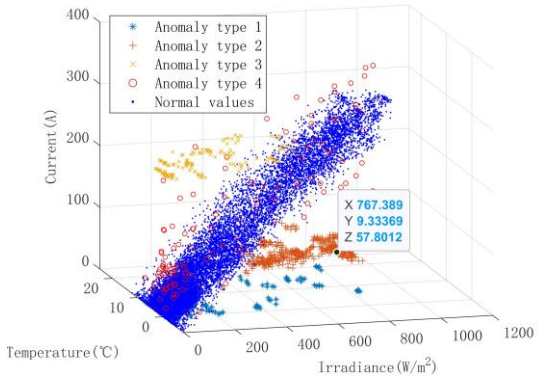
3. Voltage cleansing

After cleansing the anomalous current data, perform the operations described in Section II on the photovoltaic voltage, temperature, and irradiance data sets. Calculate the confidence interval for the photovoltaic voltage and label the anomalous data.

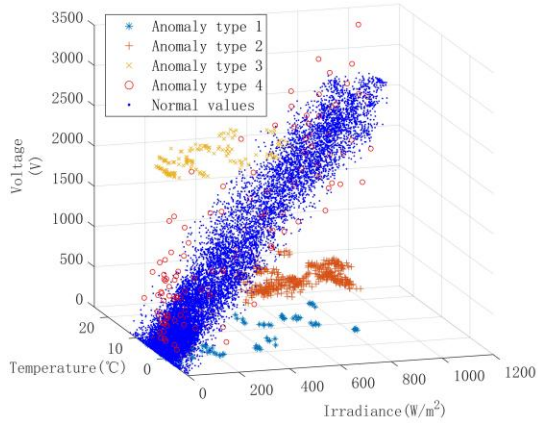
V. Case Study Analysis

A. Experimental Data

The experimental setup consists of a high-performance workstation equipped with an Intel Core i7-9700K CPU, 32 GB RAM, and an NVIDIA GTX 1080 Ti GPU, running MATLAB R2021a. The experimental data are derived from actual data collected by Tongzhou Power Supply Company from March 2022 to September 2022, with a data collection interval of 15 minutes. Each individual sample is composed of photovoltaic voltage, photovoltaic current, temperature, and irradiance. As there are no available data on faults or malfunctions in the photovoltaic array, this study artificially synthesised anomalous data by manually introducing anomalies into the original data set. The scatter plots for photovoltaic voltage and current after the introduction of artificially synthesised anomalies are illustrated in Fig. 4.



(a)



(b)

Fig. 4. Scatter plot following artificial anomaly synthesis: (a) Current; (b) Voltage.

The time series curves for photovoltaic power data before and after the introduction of artificially synthesised anomalies can be found in Appendix A. The proportions of each type of anomaly in the overall data set are as follows: Type 1 is 1 %, Type 2 is 2 %, Type 3 is 1 %, and Type 4 is 1 %.

B. Optimisation Results of Vine-Copula

The optimisation results are presented in Table IV. When modelling the dependency relationship for current, the root node of the first generated tree is U_E. Similarly, when modelling the dependency relationship for voltage, the root node of the first generated tree is also U_E.

TABLE IV. RESULTS OF VINE OPTIMISATION.

	Tree	Node	Edge	$\max\sum\lvert\tau_{j,k\mid D}\rvert$
Current	Tree 1	U_I, U_T, U_E	C_{IE}, C_{TE}	1.428
	Tree 2	$U_{I\mid E}, U_{T\mid E}$	$C_{IT\mid E}$	0.045
Voltage	Tree 1	U_V, U_T, U_E	C_{VE}, C_{TE}	1.0090
	Tree 2	$U_{V\mid E}, U_{T\mid E}$	$C_{VT\mid E}$	0.0483

Figure 5 illustrates the vine structure of the Copula model in the given example. Table V presents the specific parameters of the Vine-Copula model, including the optimal Copula function corresponding to each edge and their parameter estimation results.

Fig. 5. Vine-Copula structure: (a) I-T-E; (b) V-T-E.

TABLE V. RESULTS OF VINE-COPULA PARAMETER OPTIMISATION.

	Edge	AIC value					Optimal Copula	Parameter 1	Parameter 2
		Clayton	Gumbel	Frank	Gaussian	t			
Current	C_{IE}	-24179.08	-21887.22	-24686.31	-16315.37	-33197.44	t	0.99631	1
	C_{TE}	-883.61	-2136.37	-1764.27	-1988.70	-1996.2	Gumbel	1.4946	-
	$C_{IT\|E}$	-12.72	-43.28	-58.06	-42.94	-41.81	Frank	0.55923	-
Voltage	C_{VE}	-38050.51	-32303.46	-37130.71	-34345.05	-37334.67	Clayton	56.2069	-
	C_{TE}	-602.01	-1556.36	-1209.08	-1403.96	-1408.61	Gumbel	1.4145	-
	$C_{VT\|E}$	-11.14	-1.13	-17.46	-21.11	-19.11	Gaussian	0.060297	-

C. *Results of Anomaly Identification*

The calculated upper and lower thresholds are illustrated in Fig. 6. Figure 7(a) shows a three-dimensional scatter plot of current against temperature and irradiance. For each pair of specific environmental variables (temperature and irradiance), an upper and lower quantile point for the current is determined, together forming an upper and lower threshold surface. Points within this surface are classified as normal, whereas those outside are identified as anomalies. Figure 7(c) shows a time-series graph of the current, where the upper and lower quantile points of the current of each moment are determined by the current environmental variables, ultimately creating a threshold curve. Points within this curve are recognised as normal, whereas those outside are marked as anomalies.

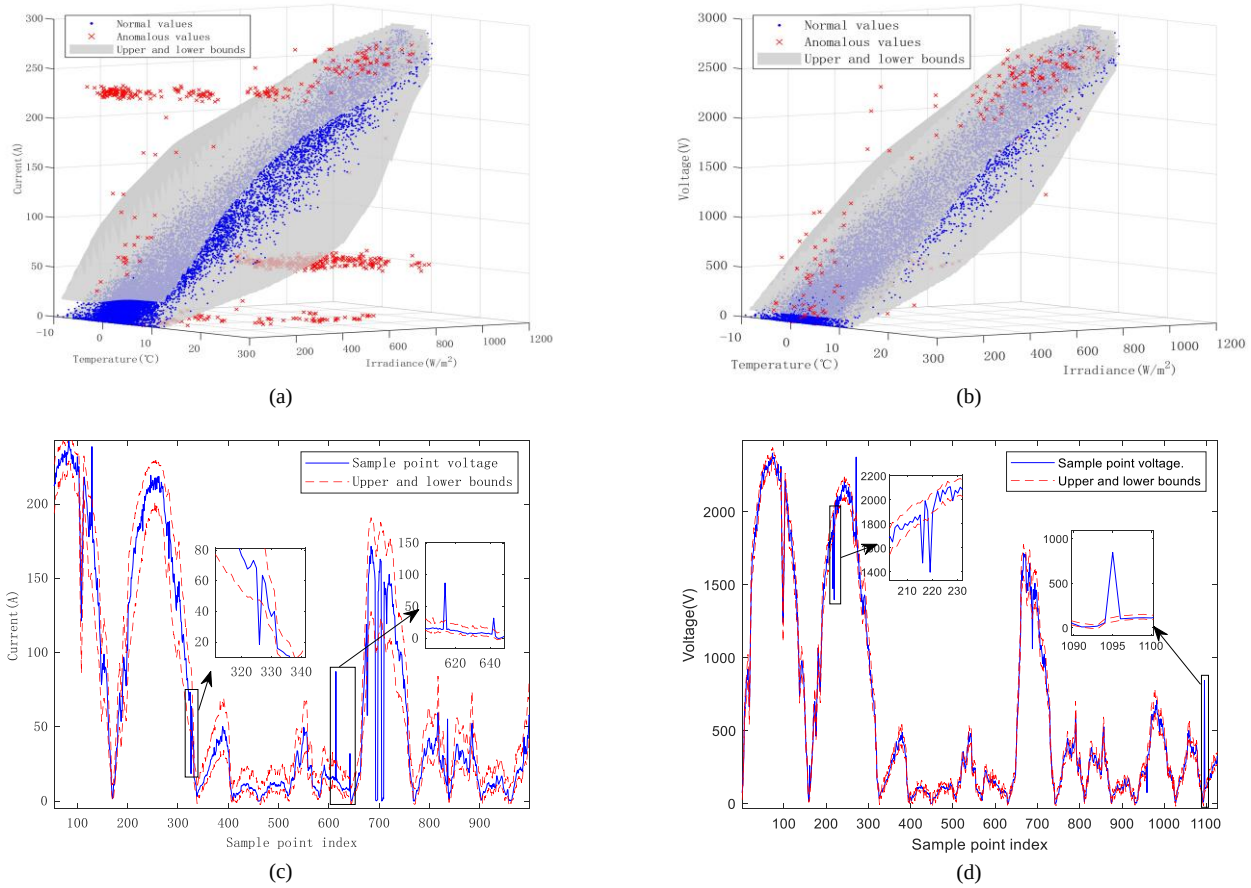


Fig. 6. Outlier identification outcome: (a) Current identification results; (b) Voltage identification results; (c) Current confidence interval; (d) Voltage confidence interval.

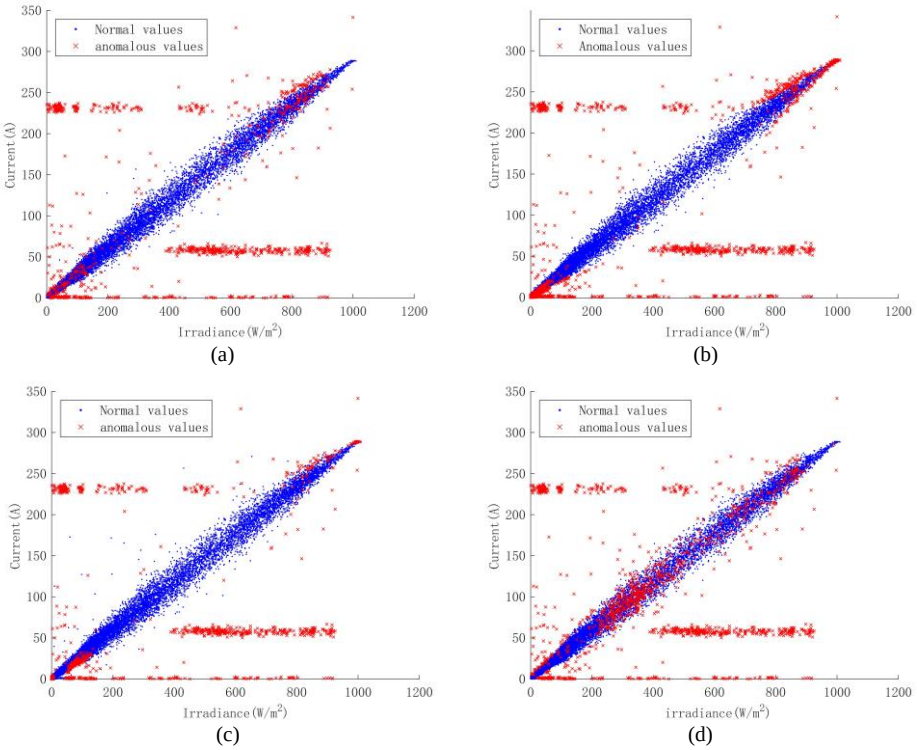


Fig. 7. Comparison of outlier identification results: (a) Our method; (b) Bivariate Copula; (c) Vine-Copula-PTE; (d) Quantile method.

D. Comparison of Different Anomaly Detection Methods

To compare the efficacy of different algorithms, the experimental data are subjected to anomaly detection using the method described in this paper, the one-step Vine-Copula method, the two-step Copula method, and the quantile method. In the one-step Vine-Copula approach, temperature and irradiance are treated as conditional variables for identifying anomalies in photovoltaic power, hereafter referred to as Vine-Copula-PTE. Both the two-step Copula method and the quantile method initially identify anomalies in the current using irradiance as the conditional variable, followed by the identification of anomalies in the voltage using temperature as the conditional variable.

To facilitate comparison, two metrics are defined: the accuracy rate of anomaly detection and the false positive rate for normal data. The definition of the accuracy rate for anomaly detection is as follows:

$$TR_i = \hat{N}_i / N_i, \tag{22}$$

$$TR = \sum_i \hat{N}_i / \sum_i N_i, \tag{23}$$

where TR_i represents the correct recognition rate for the anomaly type i, $\hat{N}_i$ denotes the number of data points correctly identified as the anomaly type i, N_i represents the total number of data points for the anomaly type i, and TR overall signifies the overall correct anomaly recognition rate.

The definition for the false positive rate of anomaly detection is as follows

$$FR = N_{\text{error}} / N_{\text{total}}, \tag{24}$$

where FR represents the false recognition rate, N_{error} is the number of normal values incorrectly identified as anomalies, and N_{total} represents the total number of sample points.

The results of anomaly detection using different methods are presented in Table VI.

TABLE VI. OUTLIER IDENTIFICATION RESULTS.

Method	TR_1	TR_2	TR_3	TR_4	TR	FR	Runing time (s)
Our method	0.965	0.876	1	0.54	0.878	0.05	632
Bivariate Copula	0.977	0.802	1	0.494	0.813	0.07	376
Vine-Copula-PTE	0.955	0.865	1	0.483	0.831	0.04	493
Quantile method	0.733	0.861	0.982	0.473	0.806	0.06	309

To facilitate observation and comparison, the final results of anomaly detection are presented using a two-dimensional scatter plot of irradiance versus current. Identification results are illustrated in Fig. 7.

Among the four methods, the two-step Vine-Copula approach exhibits the highest precision in detecting anomalies. Compared to the two-step Copula method, the method discussed in this paper demonstrates superior identification rates and lower false positive rates. This improvement is attributed to the inclusion of additional characteristic variables, which further narrows the confidence interval, thus enhancing the sensitivity of the confidence interval in detecting anomalies. The Vine-Copula-PTE method, which only identifies anomalies in power, overlooks the details in voltage and current. On the contrary, our method considers the varying impacts of temperature on current and

voltage. The quantile method segments the target variable data set based on the magnitude of the conditional variables, employing the quantile approach within each group. This method ignores the dependency relationship between the target and conditional variables, resulting in a lower identification rate.

Although our two-step Vine-Copula method is slightly more computationally intensive due to its high-dimensional modelling, it optimises the use of system memory and processor time by employing a more streamlined data handling and processing approach. Despite its complexity, the method does not excessively burden computational resources, making it suitable for scenarios with large data volumes where computational efficiency is critical. By minimising redundant calculations required by traditional methods, our approach reduces overall computational costs. This makes our method not only faster on a per-data set basis but also more scalable across larger data sets. These enhancements ensure that our method is particularly beneficial for organisations seeking to implement robust anomaly detection capabilities with limited computational resources.

In addition to this, to provide a clearer understanding of the practical implications of the various anomaly detection methods used in this study, we conducted a comparative analysis focussing on the execution times. Our findings reveal that the proposed two-step Vine-Copula method, while slightly more computationally intensive due to its high-dimensional modelling, offers a favourable balance between execution speed and detection accuracy compared to other methods. In particular, traditional methods, while faster, often sacrifice accuracy and may not adequately capture complex dependencies in the data. The execution time for the two-step Vine-Copula method averaged approximately 632 seconds per data set, which is competitive considering the enhanced detection capabilities it provides.

E. Photovoltaic Current Prediction Based on Anomaly Recognition

To further compare the precision of different anomaly detection methods, after removing anomalies using various approaches, the anomalous data are reconstructed using the method detailed in [23]. This reconstruction results in different data sets. Subsequently, each data set is trained using a bidirectional long short-term memory network (Bi-LSTM). The same set of data is then predicted using the different trained models. The reconstructed data are depicted in Fig. 8, and the prediction results are shown in Fig. 9.

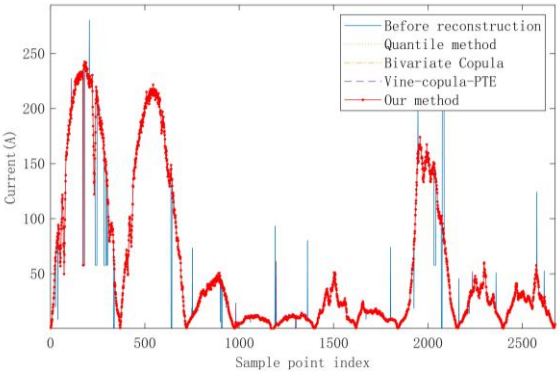


Fig. 8. Results of data reconstruction.

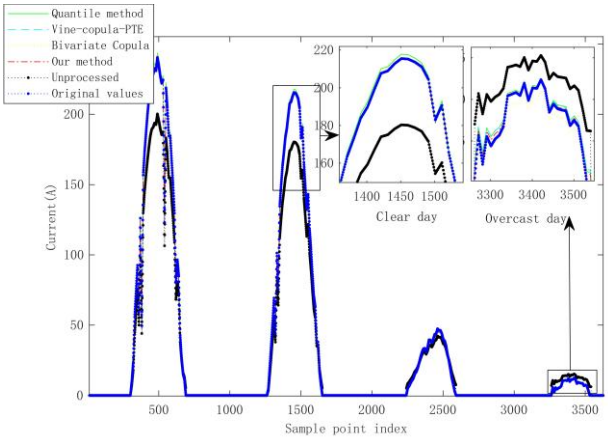


Fig. 9. Comparison of the prediction results.

The mean absolute error (MAE) and root mean square error (RMSE) for the prediction results of different data sets are shown in Table VII. It is evident that the photovoltaic data anomaly detection method proposed in this paper effectively reduces prediction errors. By identifying anomalies or faulty data during the data collection process, the prediction model can better adapt to real-world conditions, improving the prediction accuracy. This, in turn, helps power systems to plan and manage energy supply more effectively.

TABLE .VII. COMPARATIVE ANALYSIS OF PREDICTION ERRORS.

Method	MAE/MW	RMSE/MW
Our method	1.06	1.03
Vine-Copula-PTE	2.43	1.56
Bivariate Copula	3.61	1.90
Quantile method	2.01	1.42
Unprocessed	133.4	11.55

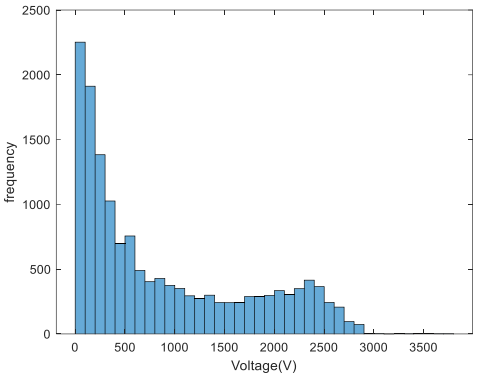
VI. CONCLUSIONS

This paper addresses the limitations of existing anomaly detection methods in photovoltaic systems and introduces a novel two-step Vine-Copula method for high-dimensional dependency structure modelling. The key findings from our experimental simulations and analysis are as follows.

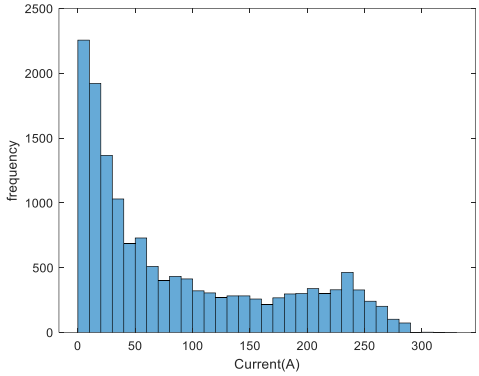
1. The proposed method analyses anomalous output data under abnormal operational conditions of photovoltaic arrays. By focussing on both photovoltaic current and voltage anomalies rather than just power, our approach uncovers hidden discrepancies more effectively, thereby increasing the accuracy of anomaly detection.
2. Using the Vine-Copula theory, we have established robust models for photovoltaic current and voltage as functions of temperature and irradiance. These models are finely tuned to optimise the structure and parameters, which facilitated the development of precise formulas for computing confidence intervals for current and voltage.
3. Our method significantly enhances the prediction accuracy of photovoltaic current outputs. By processing data through our advanced anomaly detection and reconstruction protocol, we demonstrate that our approach surpasses other existing methods in predicting photovoltaic behaviours accurately.
4. The method is designed for practical efficiency, requiring just a single modelling step to establish confidence intervals at various levels. Once configured, the model only needs input of current environmental conditions (temperature and irradiance) to determine real-

time thresholds for the target variables, making it highly applicable to real-world operational environments.

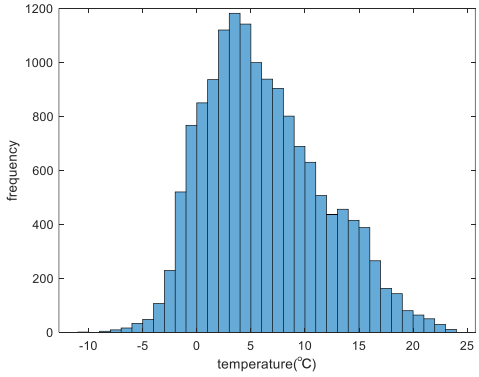
APPENDIX A



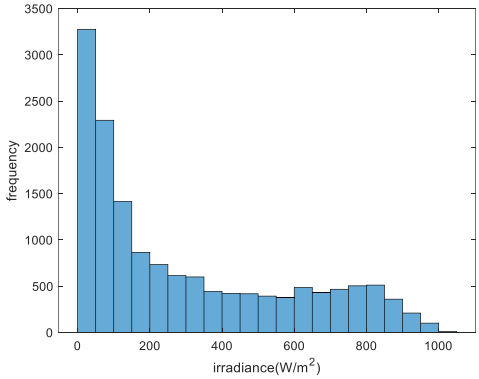
(a)



(b)

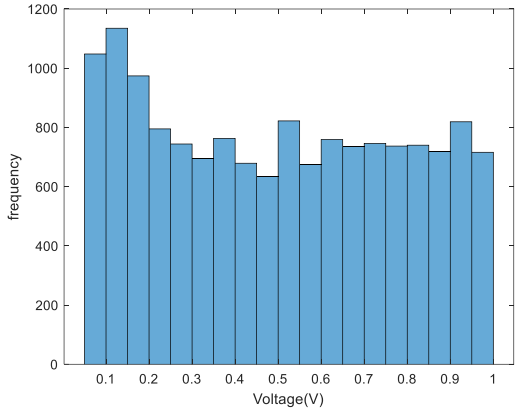


(c)

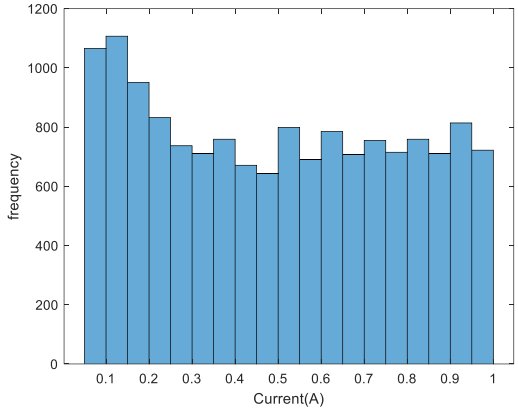


(d)

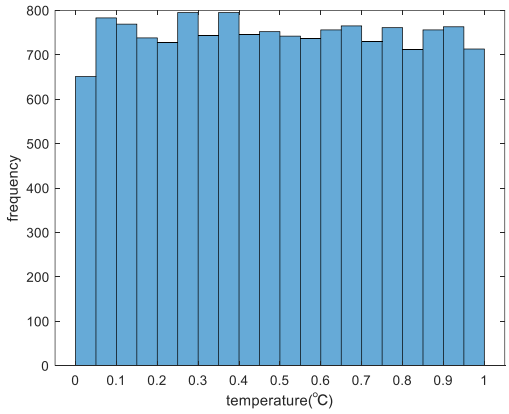
Fig. A-1. Probability distribution histograms for each variable: (a) Voltage; (b) Current; (c) Temperature; (d) Irradiance.



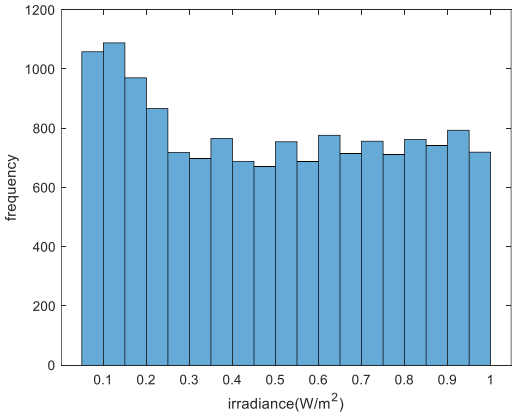
(a)



(b)

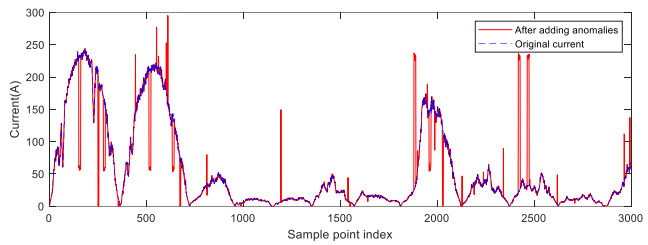


(c)

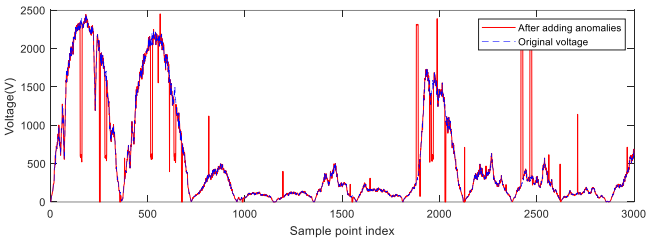


(d)

Fig. A-2. Probability distribution histograms for each variable: (a) Voltage; (b) Current; (c) Temperature; (d) Irradiance.



(a)



(b)

Fig. A-3. Comparative analysis of current (a) and voltage (b) data before and after the incorporation of anomaly data.

ACKNOWLEDGMENT

We cordially acknowledge Prof. Yu Huang from Nanjing University of Posts and Telecommunications for his valuable feedback and suggestions on this work. Additionally, we are grateful to State Grid Beijing Electric Power Company for granting us access to their photovoltaic power measurement database.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] M. Ding, W. Wang, X. Wang, Y. Song, D. Chen, and M. Sun, "A review on the effect of large-scale PV generation on power systems", *Proceedings of the Chinese Society of Electrical Engineering*, vol. 34, no. 1, pp. 1–14, 2014. DOI: 10.13334/j.0258-8013.pcsee.2014.01.001.

[2] Y. Wang, Z. Li, C. Wu, D. Zhou, and L. Fu, "A survey of online fault diagnosis for PV module based on BP neural network", *Power System Technology*, vol. 37, no. 8, pp. 2094–2100, 2013. DOI: 10.13335/j.1000-3673.pst.2013.08.024.

[3] W. Han, H. Wang, C. Wang, L. Chen, J. Zhang, and R. Sun, "Parameter identification based fault diagnosis model of photovoltaic modules", *Power System Technology*, vol. 39, no. 5, pp. 1198–1204, 2015. DOI: 10.13335/j.1000-3673.pst.2015.05.005.

[4] M. Shi, R. Yin, A. Hu, and J. Wu, "A novel photovoltaic array outlier cleaning algorithm based on moving standard deviation", *Power System Protection and Control*, vol. 48, no. 6, pp. 108–114, 2020. DOI: 10.19783/j.cnki.pspc.190484.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest", in *Proc. of 2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[6] Y. Wang, D. G. Infield, B. Stephen, and S. J. Galloway, "Copula-based model for wind turbine power curve outlier rejection", *Wind Energy*, vol. 17, no. 11, pp. 1677–1688, 2014. DOI: 10.1002/we.1661.

[7] H. Zhu and Z. Liu, "PV system output analysis of environmental factors affect", *North China Electric Power*, vol. 2014, no. 8, pp. 50–55, 2014. DOI: 10.16308/j.cnki.issn1003-9171.2014.08.015.

[8] Y. Gong, Z. Lu, Y. Qiao, and Q. Wang, "An overview of photovoltaic energy system output forecasting technology", *Automation of Electric Power Systems*, vol. 40, no. 4, pp. 140–151, 2016. DOI: 10.7500/AEPS20150711003.

[9] Y. Gong, Z. Lu, Y. Qiao, Q. Wang, and X. Cao, "Copula theory based machine identification algorithm of high proportion of outliers in photovoltaic power data", *Automation of Electric Power Systems*, vol. 40, no. 9, pp. 16–22, 2016. DOI: 10.7500/AEPS20151008006.

[10] S. Zhao, T. Zhang, Z. Li, D. Li, X. Xu, and J. Liu, "Distribution model of day-ahead photovoltaic power forecasting error based on numerical characteristic clustering", *Automation of Electric Power Systems*, vol. 43, no. 13, pp. 36–45, 2019. DOI: 10.7500/AEPS20180405002.

[11] Y. Cao *et al.*, "A comprehensive review of energy Internet: Basic concept, operation and planning methods, and research prospects", *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 3, pp. 399–411, 2018. DOI: 10.1007/s40565-017-0350-8.

[12] M. Yang and X. Huang, "Abnormal data identification algorithm for photovoltaic power based on characteristics analysis of illumination process", *Automation of Electric Power Systems*, vol. 43, no. 6, pp. 64–69, 2019. DOI: 10.7500/AEPS20180626003.

[13] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data", *The Canadian Journal of Statistics*, vol. 40, no. 1, pp. 68–85, 2012. DOI: 10.1002/cjs.10141.

[14] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence", *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 182–198, 2009. DOI: 10.1016/j.insmatheco.2007.02.001.

[15] X. L. Zhang, Q. H. Liu, B. Li, and H. M. Ma, "Analysis of output characteristics of photovoltaic system", *Advanced Materials Research*, vol. 512–515, pp. 17–22, 2012. DOI: 10.4028/www.scientific.net/AMR.512-515.17.

[16] G. Li *et al.*, "Outlier data mining method considering the output distribution characteristics for photo-voltaic arrays and its application", *Energy Reports*, vol. 6, pp. 2345–2357, 2020. DOI: 10.1016/j.egyr.2020.08.034.

[17] Y. Huang, Q. Xu, B. Xu, and Y. Lyu, "Graph cut method for dynamic zonal reserve allocation in power grid with wind power integration", *Proceedings of the CSEE*, vol. 40, no. 12, pp. 3765–3775, 2020. DOI: 10.13334/j.0258-8013.pcsee.190426.

[18] B. Xu, Q. Xu, Y. Huang, J. Song, Y. Ji, and Y. Ding, "Day-ahead probabilistic forecasting of photovoltaic power based on Vine Copula quantile regression", *Power System Technology*, vol. 45, no. 11, pp. 4426–4435, 2021. DOI: 10.13335/j.1000-3673.pst.2020.1923.

[19] P. K. Trivedi and D. M. Zimmer, "Copula modeling: An introduction for practitioners", *Foundations and Trends® in Econometrics*, vol. 1, no. 1, pp. 1–111, 2007. DOI: 10.1561/0800000005.

[20] J. Yan, "Multivariate modeling with copulas and engineering applications", in *Springer Handbook of Engineering Statistics*. Springer Handbooks, Springer, London, 2023. DOI: 10.1007/978-1-4471-7503-2_46.

[21] Y. Zhao and S. Dong, "Multivariate probability analysis of wind-wave actions on offshore wind turbine via copula-based analysis", *Ocean Engineering*, vol. 288, part 1, art. 116071, 2023. DOI: 10.1016/j.oceaneng.2023.116071.

[22] Md T. Amin, Y. Yao, J. Yu, and S. Adumene, "Probabilistic monitoring of nuclear plants using R-vine copula", *Annals of Nuclear Energy*, vol. 190, art. 109867, 2023. DOI: 10.1016/j.anucene.2023.109867.

[23] A. B. Krishna and A. R. Abhyankar, "Time-coupled day-ahead wind power scenario generation: A combined regular vine copula and variance reduction method", *Energy*, vol. 265, art. 126173, 2023. DOI: 10.1016/j.energy.2022.126173.