A Self-Data-Driven Method for Lifetime Prediction of PV Arrays Considering the Uncertainty and Volatility

Liu, Yongjie; Ding, Kun; Zhang, Jingwei; Sangwongwanich, Ariya; Wang, Huai

Published in:
IEEE Transactions on Power Electronics

DOI (link to publication from Publisher):
10.1109/TPEL.2023.3337713

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

A Self-Data-Driven Method for Lifetime Prediction of PV Arrays Considering the Uncertainty and Volatility

Yongjie Liu, *Student Member, IEEE,* Kun Ding, Jingwei Zhang, *Member, IEEE,*
Ariya Sangwongwanich, *Member, IEEE,* Huai Wang, *Senior Member, IEEE,*

***Abstract*—This paper proposes a self-data-driven method for remaining useful life prediction of PV arrays based on self-condition monitoring data considering the uncertainty and volatility. First, a health indicator (HI) reconstruction method is presented to eliminate the uncertainty and volatility of condition monitoring data. Second, a nonlinear Gamma stochastic process model is established to describe the probability distribution of the degradation trend. Then, the model parameter solution is transformed into an optimization problem, and a hybrid particle swarm and grey wolf optimization algorithm (PSO-GWO) is developed to estimate the model parameters avoiding trapping in local optimization and divergence. Finally, two case studies are demonstrated to verify the effectiveness of the proposed method based on the DKASC and NREL datasets, and the performance is further evaluated in comparisons with the empirical models, statistical models and long short-term memory (LSTM) network. Experimental results demonstrate that the proposed method has excellent RUL prediction accuracy.**

***Index Terms*—Degradation modelling, nonlinear gamma process, remaining useful lifetime (RUL) prediction, health indicator (HI) reconstruction, photovoltaic array (PV).**

I. Introduction

PHOTOVOLTAIC (PV) power generation is the main form of solar energy conversion and utilization. The PV module is one of the most important components of PV systems, which are composed of solar cells, EVA, glass, back sheets, junction box, metal frame and interconnect wires, etc [1]. The schematic diagram of the PV module package structure is shown in Fig. 1. However, PV modules are usually installed outdoors and exposed to harsh environmental conditions, e.g., high temperatures, strong sandy wind, heavy rain, etc. Thus, they are prone to failure during the practical operation.

The failure mechanisms of PV modules can be divided into two categories: (1) sudden failure due to damaged wires; and (2) cumulative damage and degradation due to the long-time operation. Many methods of fault diagnosis for PV modules have been proposed in recent years to deal with sudden failures [2]–[5], which is not the focus of this work. During long-time operation, some components of PV modules (see in Fig. 1) may suffer from degradation and irreversible damage, which will reduce the efficiency and useful lifetime of PV modules, and even cause serious safety accidents such as fire hazards [6], [7]. Therefore, predictive maintenance is essential to ensure the reliable and efficient operation of PV systems. The process of predictive maintenance is to monitor the health condition by the condition monitoring system (CMS), predict the remaining useful lifetime (RUL), and make an optimal maintenance decision based on RUL prediction results. Moreover, the results of the RUL prediction can be used as an evaluation indicator for PV power plant transactions and have a high economic value.

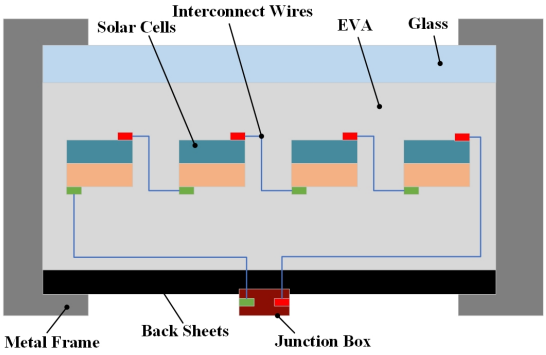


Fig. 1. Schematic diagram of the PV module package structure

Generally, the methods of RUL prediction can be divided into two main categories: physics model-based approaches and data-driven-based approaches. As a traditional approach, the physics model-based method describes the degradation process by building a mathematical model based on the analysis of the factors affecting the degradation and the material deformation of the PV module [8]–[11]. However, it is challenging to construct an accurate physical degradation model of PV modules for RUL prediction in practical applications due to the complex structure, the numerous degradation failure factors and the inclusion of complex physical and chemical alteration processes. Therefore, these factors limit the application of the physical model-based approach in real operating environments.

Compared with the physics-based model, the data-driven methods for RUL prediction do not rely on specific degradation mechanisms. From the type of data collected by the mon-

This work was supported in part by the National Natural Science Foundation of China under Grant 51777059, in part by the Fundamental Research Funds for the Central Universities under Grant B230201004, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20201163. *(Corresponding author: Kun Ding)*

Yongjie Liu, Kun Ding and Jingwei Zhang are with the College of Mechanical and Electrical Engineering, Hohai University, Changzhou, China (e-mail: yoli@energy.aau.dk; dingk@hhu.edu.cn; jwzhang@hhu.edu.cn;)

Ariya Sangwongwanich and Huai Wang are with the Department of Energy Technology, Aalborg University, 9220 Aalborg, Denmark (e-mail: ars@energy.aau.dk; hwa@energy.aau.dk;)

TABLE I
ANALYSIS OF THE FACTORS CAUSING UNCERTAINTY AND VOLATILITY IN PV ARRAYS

Causes	Description	Effect on output performance
Irradiance	Sunny days / Noontime daily	Output power increased
	Cloudy / Rainy / Every morning and afternoon	Output power reduced
Temperature	Temperature increased due to environmental temperature, e.g., hot summer days	Output power reduced
	Temperature reduced due to wind accelerated cooling or low environmental temperature	Output power increased
Partial shading	Cloud / Bird droppings / Trees / Buildings	Output power reduced
Snow accumulation	Snow accumulation on the surface of PV modules	Output power reduced
Short circuit fault	Accidental connections between PV modules or cables	Output power reduced
Open circuit fault	Cables damaged for a single string of PV arrays	Zero output power
	Cables damaged for multiple strings of PV arrays	Output power reduced
Hotspot	Current mismatch between cells / Soiling / Dust / Partial shading	Output power reduced
Dust and soiling	Accumulation of dust / Soil on the surface of PV modules	Output power reduced
Inverter abnormal shutdown	IGBT open or short circuit fault / Overheating Protection / Misoperation	Zero output power
Degradation of PV modules	Solar cells aging / EVA yellowing due to ultraviolet / Interconnect wires broken	Output power reduced
Operation and maintenance	PV modules cleaning / Faults detection and maintenance by manual	Performance restoration

itoring system, it can be further divided into three categories: survival model, similarity model and degradation model. Survival model is a statistical method to construct the relationship between the probability of failure and operation time [12]. However, establishing survival curves requires huge amounts of failure data at different operation times. In practical application, PV modules are expected to have a service life of up to 25 years under normal conditions. Thus, it is difficult to acquire enough failure data in a period of time. Similarity model is required to construct a library of data from health, degradation and failure. Then, the similarity between the degradation curves of the component in service and the historical library is measured and the most similar degradation curve is selected as a reference [13]–[15]. However, the output power of PV arrays is strongly affected by environmental conditions. The operation conditions and degradation processes of different PV arrays are distinct, making it difficult to find a similar degradation pattern. Degradation model is to predict the RUL by using self-condition monitoring data without depending on training data from failure events, which is also defined as self-data-driven RUL prediction [16]–[18]. Marios *et al*. presents a degradation model of PV arrays based on the non-linear degradation rate and the model parameters are calibrated based on historical data [16]. In [17], a comparative study of several statistical models including the simple linear regression (SLR) and autoregressive integrated moving average (ARIMA) model based on the performance degradation data is discussed.

Based on the above literature analysis, tremendous efforts have been devoted to designing prediction models for improving the accuracy and efficiency of RUL prediction models [8]–[18]. Nevertheless, based on the authors' best knowledge, the effect of uncertainty and volatility on health indicators (HI) has not been adequately investigated in the long-term outdoor operation of PV arrays. HI is the information that can directly or indirectly represent the health status and provide an accurate assessment of the operational performance for the PV array. It is essential for performance degradation analysis and the RUL prediction of the PV array. HI needs to be generic and can be applied to different environments and systems. Based on the electrical data collected by the PV system, health indicators are classified into three categories [19]: I-V curves, empirical health indicators and performance ratio (PR). The output power of PV systems is easy to measure and doesn't need to build empirical models. In addition, real-time scanning of the I-V curve will interrupt the operation and affect the power generation efficiency of the PV system. Therefore, the PR is used as a health indicator for performance assessment and RUL prediction of PV arrays during the long-term operation in this paper. In long-term outdoor operation, it is not only the degradation of the PV modules that leads to a reduction of the output power, i.e., the extracted HI is reduced, but also sudden faults such as partial shading, short-circuit faults, snow accumulation, etc. that can lead to an abnormal reduction of the output power. Then, after regular operation and maintenance (O&M), the power will be restored to normal condition again, which makes the data collected more random. Moreover, the causes and effects of the uncertainty and volatility for PV arrays are summarized in Table I. These uncertainties and volatilities will seriously affect the performance evaluation of PV arrays and the accuracy of the RUL prediction.

To obtain effective health indicators considering the effects of uncertainty and volatility, a new health indicator reconstruction method is proposed in this study. It includes the improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) is presented to separate the trend and random terms, extract the effective trend term signal features and eliminate the random term signals due to environmental fluctuations or other interference. The CEEMDAN is an improved variant of the empirical mode decomposition (EMD) and EEMD [20]. The EMD method is developed to deal with the nonlinear and non-stationary signal, which decomposes a complicated signal into several intrinsic modal functions (IMFs) and a residual component (R) with different frequencies and scales. However, EMD has the problem of modal mixing in the signal decomposition process. Then, the EEMD method is presented to solve this problem by adding noise in the decomposition process, but the residual noise after decomposition leads to a large reconstruction error and low calculation efficiency of the method. Finally, the CEEMDAN is proposed to effectively solve the problem of modal mixing and noise residual by adding adaptive Gaussian white noise. In

this paper, the augmented dickey-fuller (ADF) [21] is incorporated into the CEMMDAN method to assess the stationary and non-stationary of the IMFs, and then the stationary IMFs are combined to compose the trend term signal. In addition, the degradation pattern of PV arrays is also strongly affected by uncertainty factors and it is difficult to describe the degradation process of PV arrays through a defined non-linear degradation pattern or a linear degradation pattern. Thus, the parameters that determine the degradation pattern are modified by self-monitoring data in this paper. Furthermore, the probability distribution of the RUL is predicted by the non-linear Gamma stochastic process. The main contributions of this article can be summarized as follows:

1) A self-data-driven method for RUL prediction of PV arrays considering the uncertainty and volatility is proposed.
2) The ICEEMDAN is developed to extract trend signals and eliminate random signals. Then, the expected values of the monthly statistical distribution from the extracted trend signal are utilized as the new health indicator.
3) A hybrid optimization algorithm is developed to estimate model parameters of the non-linear Gamma stochastic process, which ensures the speed and accuracy of the model parameter solution.

The rest of this paper is organized as follows. In Section II, the self-data-driven model for RUL probability distribution prediction of PV arrays considering the uncertainty and volatility is detailed. In Section III, the feasibility of the proposed method is verified based on the DKASC and NREL datasets and a comparative analysis of other methods. Finally, some significant results are concluded.

II. Methodology

A. Framework of the proposed RUL prediction method

The framework of the proposed RUL prediction method for PV arrays is shown in Fig.2 and the main procedures are summarized as follows:

Step1: Data Collection: The output power and environmental data are collected during the long-term operation of PV arrays.

Step2: HI Reconstruction: Health indicators are reconstructed to take into account the uncertainty and volatility of data collected during the long-term operation of PV arrays. Firstly, the median absolute deviation (MAD) [22] is used to remove rough errors from the measured data. The ICEEMDAN is presented to extract the effective trend term signal features and eliminate the random term signals due to environmental fluctuations or other interference factors. The statistical distribution of the extracted trend signal is obtained and the expected value is used as the health indicator to evaluate the operational state for PV arrays.

Step3: Degradation Model: The degradation model is established based on the nonlinear Gamma stochastic process.

Step4: Parameters Estimation: The estimation of model parameters for the Gamma stochastic process is converted into an optimization problem. Then, the hybrid particle swarm and grey wolf optimization (PSO-GWO) optimization algorithm is proposed to accurately and reliably identify the unknown model parameters of the Gamma stochastic process based on the performance degradation percent.

Step5: RUL Prediction: The RUL probability distribution prediction is completed by the nonlinear Gamma distribution probability density function and the 95% confidence interval is calculated. Finally, a comparative analysis was performed by comparing the proposed approach with the empirical models, statistical models (ARIMA) and long short-term memory (LSTM).

The details of each step are described in the following subsections.

B. HI Reconstruction Considering the Uncertainty and Volatility

For long-term outdoor operation, it is not only the degradation of PV modules that leads to a reduction of the output power, i.e., HI are reduced, but sudden faults that can also lead to an abnormal reduction of the output power. The PV inverter may suffer abnormal shutdowns, resulting in abnormal and measurement errors of the collected data [23]. It makes the collected data more random, and there are many data with zero output power. The measured output power is shown in Fig. 3.

The outlier data will seriously affect the accuracy of the RUL prediction. Thus, the data are filtered and the data within the stable interval of the ambient condition (irradiance between 700 W/m^2 and 1200 W/m^2) are selected to eliminate the influence of environmental fluctuation factors. Then, the MAD method is used to remove rough errors in measured data. MAD has the advantage of simple and fast computation, which can detect and remove abnormal data well [22].

For a length n of measured power data $\{x_1, x_2, \ldots, x_n\}$. the median M is calculated as follows:

$$M = \underset{i=1,2,3\ldots n}{\text{median}} (x_i) \tag{1}$$

If n is odd, the median is the middle value of the measured dataset after sorting, and when n is even, the median is the mean of the values at $n/2$ and $n+1/2$ of the measured dataset.

Then, the MAD is determined as follows:

$$MAD = b \times \underset{i=1,2,3\ldots n}{\text{median}} |x_i - M| \tag{2}$$

$b = 1/\left(\Phi^{-1}(3/4)\right) \approx 1.4826$, which allows normal data to fall in the middle of the 50% region and outliers to fall in the 50% region on both sides [22]. Finally, the determination coefficient D is calculated for each measured data and is expressed as follows:

$$D = \frac{|x_i - \text{median}_{j=1,\cdots n}(x_j)|}{MAD} \tag{3}$$

The value is defined as outlier data when the D is greater than the threshold value and the threshold value is generally set at 2.5. Finally, the outlier data is filtered out. Next, after filtering out the outlier data, the output power under different environments is converted to STC using Eq.4, which eliminates the dependence of the output characteristics on irradiance and temperature for PV arrays. Eq.4 is defined as follows [24]:

$$P_{\mathrm{C-STC}} = \frac{P_{\mathrm{DC}}}{G_{\mathrm{POA}}\left[1 + r\left(T_{\mathrm{mod}} - T_{STC}\right)\right]/G_{\mathrm{STC}}} \tag{4}$$

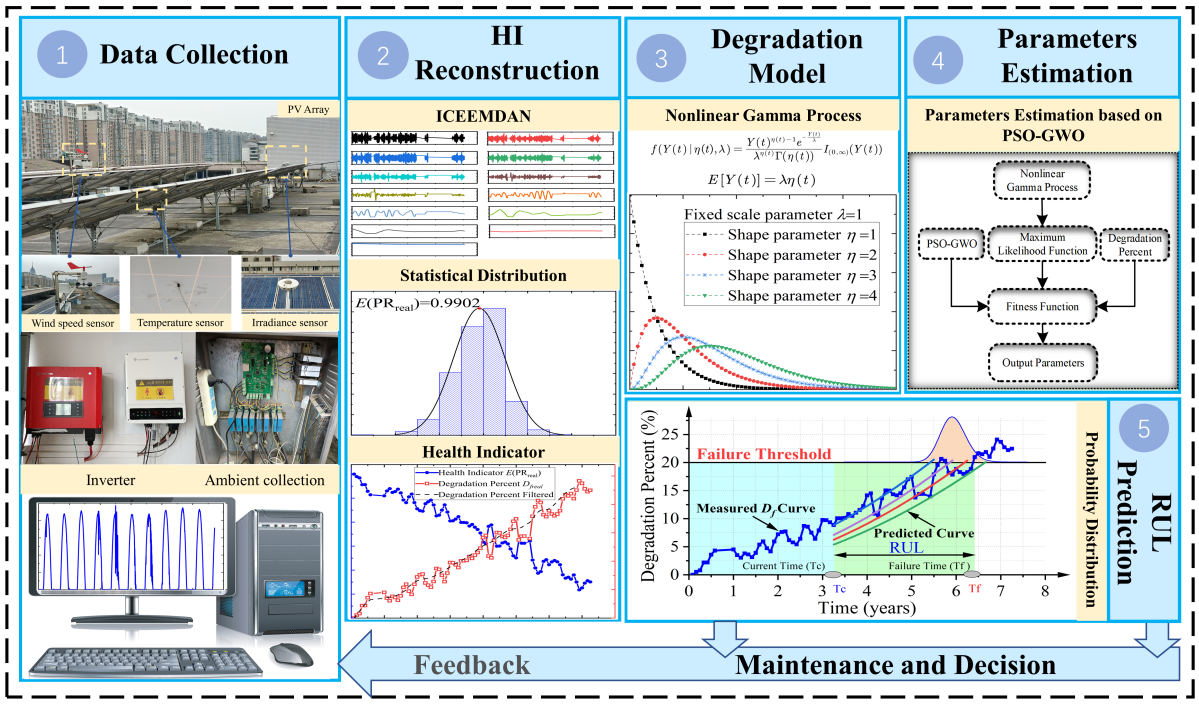


Fig. 2. Framework of the proposed RUL prediction method

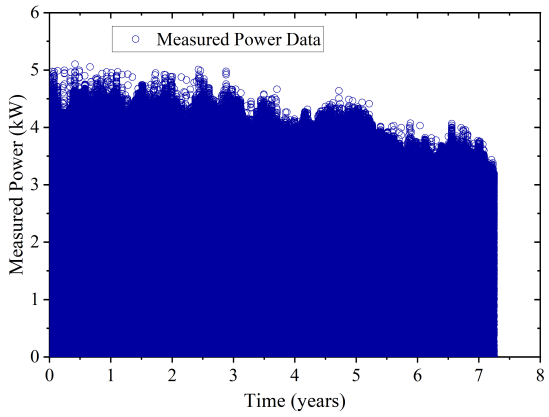


Fig. 3. Measured output power of the PV array (DKASC)

As this paper focuses on the RUL prediction of PV arrays, the DC-side output power is used to eliminate the influence of factors such as inverter degradation, etc. In addition, the method proposed can be extended to evaluate the RUL of PV systems using PR on the AC side. In Eq.4, G_{STC} and T_{STC} are irradiance and temperature under STC, i.e., 1000 W/m^2 and 25°C , respectively. P_{DC} is the maximum output power on the DC side of the PV system under actual operating conditions. r is the temperature coefficient of the PV module at STC. T_{mod} is the measured PV module temperature.

However, PV arrays are installed in complex outdoor environments and the electrical output characteristics are affected not only by irradiance but also by intermittent shadow shading and ash accumulation, etc. These cause the converted electrical output characteristics to be non-smooth and random, which will reduce the accuracy of the RUL prediction. In addition, the above data processing methods can only remove some rough error data and do not take into account the effect of non-stationary random characteristics, which cannot eliminate some minor noise and interference caused by environmental factors. Thus, ICEEMDAN is developed to separate the trend and random terms of the converted power data to STC, to extract the effective trend term signal features and eliminate the random term signals due to environmental fluctuations or other interferences. The detailed steps are as follows:

Step1: A new sequence signal $P_{\mathrm{C-STC}\,i}(t)$ is obtained by adding Gaussian white noise with a standard normal distribution to the transformed power data, which is described as follows:

$$P_{\mathrm{C-STC}i}(t) = P_{\mathrm{C-STC}}(t) + \beta_0\omega_i(t), i = 1, 2, \cdots N \tag{5}$$

where $P_{\mathrm{C-STC}\,i}(t)$ is the new signal obtained after the i-th addition of Gaussian white noise; $P_{\mathrm{C\text{-}STC}}\,(t)$ is the original signal; $w_i(t)$ is the Gaussian white noise; β_0 is the standard deviation of Gaussian white noise; N is the number of times added noise.

Step2: The first mode component $IMF_1(t)$ and the residual component $r_i(t)$ of $P_{\mathrm{C-STC}j}(t)$ are obtained by EMD decomposition:

$$\mathrm{IMF}_1(t) = \frac{1}{N}\sum_{i=1}^{N} IMF_1^i(t) \tag{6}$$

$$r_1(t) = P_{\mathrm{C-STC}}(t) - IMF_1(t) \tag{7}$$

Step3: Using $r_1(t)$ as the original sequence, the newly constructed sequence $r_1(t) + \beta_1 E_1(\omega_i(t))$ is decomposed using EMD as the algorithm to obtain the second mode component $IMF_2(t)$ and the residual component $r_2(t)$:

$$IMF_2(t) = \frac{1}{N}\sum_{i=1}^{N} E_1\left(r_1(t) + \beta_1 E_1\left(\omega_i(t)\right)\right) \tag{8}$$

$$r_2(t) = r_1(t) - IMF_2(t) \tag{9}$$

where $E_k(\bullet)$ is the k-th order IMF generated by the EMD; β_1 is the standard deviation of Gaussian white noise.

Step4: After repeating Step 3, $(k+1)$ modal components $IMF_{k+1}(t)$ and the k-th residual component $r_k(t)$ are obtained:

$$IMF_{k+1}(t) = \frac{1}{N}\sum_{i=1}^{N} E_1\left(r_k(t) + \beta_k E_k\left(\omega_i(t)\right)\right) \tag{10}$$

$$r_k(t) = r_{k-1}(t) - IMF_k(t) \tag{11}$$

Step5: Repeat Step 4 until the residual components can not be further decomposition to obtain all IMF components of CEEMDAN. The final residuals $R(t)$ are expressed as:

$$R(t) = P_{\mathrm{C-STC}}(t) - \sum_{k=1}^{K} IMF_k(t) \tag{12}$$

K is the number of modal components obtained.

Step6: The ADF is used to access the stationary of each IMF component. If there is a unit root the component is non-stationary, if there is no unit root the component is stationary. Then, the non-stationary components are discarded and the stationary components are superimposed with the residual series to obtain the extracted trend term signal P_{trend}. It can be expressed as follows:

$$P_{\mathrm{trend}}(t) = \sum_{i=1}^{M} IMF_i(t) + R(t) \tag{13}$$

M is the number of stationary components after ADF calibration.

The number of times added Gaussian white noise N is set to 100 and the standard deviation of Gaussian white noise is 0.2. The power data converted to STC is decomposed by ICEEMDAN to obtain the IMFs and residual signals as shown in Fig.4. Each IMF represents the component of the original signal at different frequencies, as can be seen in Fig.4, IMF1 has the highest frequency and IMF2 to IMF14 are decreasing in frequency. In order to build an accurate RUL prediction model, random signals need to be filtered out, however, too much of the random signal is removed resulting in distortion of the original signal. Thus, ADF calibration is used to filter out the non-stationary IMF component signals, and the calibration results are listed in Table II. "1" indicates a stationary signal, "0" indicates a non-stationary signal, and the trend term signal is obtained by superimposing the components of the stationary signal to obtain the trend term signal after filtering out the random signal from the original signal, as shown in Fig.5. It can be seen from Fig.5 that the extracted trend term signal eliminates the random errors of the original signal, which makes the signal present a trend distribution, retaining the

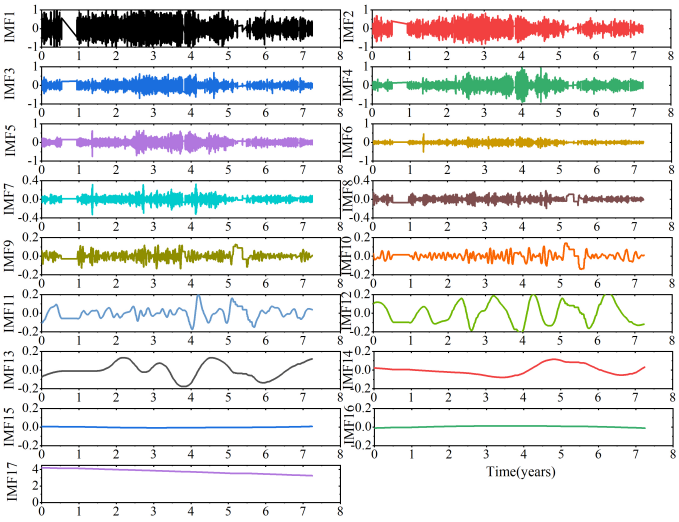


Fig. 4. Results for IMFs and the residual signal after ICEEMDAN decomposition

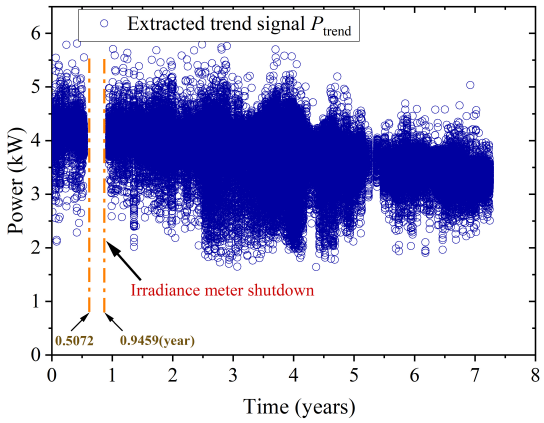


Fig. 5. The extracted trend signal P_{trend} (DKASC)

main features of the original signal and eliminating the random errors caused by noise and interference, etc.

Next, the obtained trend signal is then used to calculate the performance ratio using Eq.14, as the actual operating environment is significantly different from the STC conditions and the conversion equation is inherently inaccurate. Therefore, the converted power at the initial installation of the PV array $P_{\text{C-STC}}$ $(t = \text{initial})$ is used as a reference value to calculate the performance ratio under actual operation PR_{real} :

$$PR_{\text{real}} = \frac{P_{\mathrm{C-STC}}(t)}{P_{\mathrm{C-STC}}(t = \text{ initial })} \tag{14}$$

Then, RUL prediction of PV arrays is a research problem on long-time scales, as well as to avoid performance declines over

TABLE II
THE RESULTS OF THE ADF CALIBRATION

	F1	F2	F3	F4	F5	F6	F7	F8	F9
ADF	1	1	1	1	0	1	1	0	0
	F10	F11	F12	F13	F14	F15	F16	Residual	
ADF	0	1	0	0	0	0	0	1	

short periods of time due to human operation and intermittent failures. In this paper, the PR_{real} is divided on a monthly basis and the PR_{real} distribution for each month is statistics and its expected value is used as an indicator for the health assessment of the PV array. The statistical results of the distribution of PR_{real} for some months are shown in Fig.6. As can be seen from the figure, the expected value $E\left(PR_{\text{real}}\right)$ of PR_{real} for each month, which gradually decreases from the value 1 with a long-time operation, reflects its degradation trend. In January of the first year of PV array installation, $E\left(PR_{\text{real}}\right)$ =0.9902; January of the 3rd year, $E\left(PR_{\text{real}}\right) = 0.9004$; January of the 5th year, $E\left(PR_{\text{real}}\right) = 0.8197$; January of the 7th year, $E\left(PR_{\text{real}}\right) = 0.7643$. Finally, $E\left(PR_{\text{real}}\right)$ is used as a health

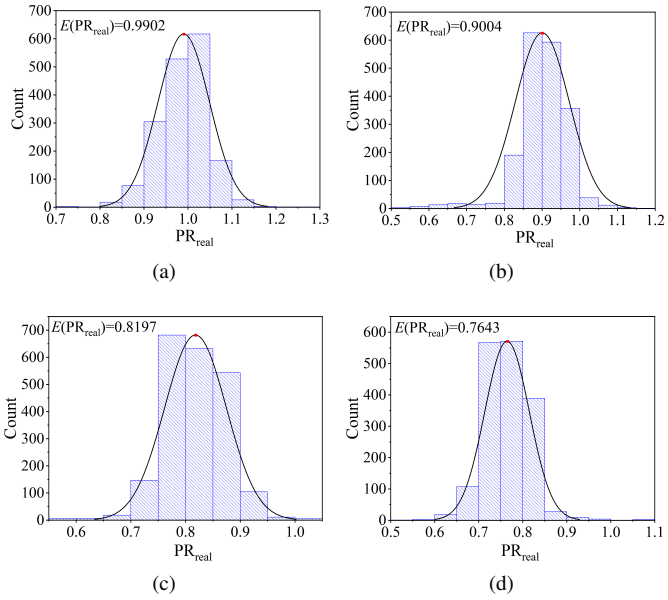


Fig. 6. The statistical results of the distribution of PR_{real}: (a) January of the 1st year; (b) January of the 3rd year; (c) January of the 5th year; (d) January of the 7th year

indicator to evaluate the operational status and degradation trend of the PV array, and the performance degradation percent D_f of the PV array is calculated as:

$$D_f = (1 - E(PR_{real}))^* \, 100\% \tag{15}$$

The curve of the health indicator and degradation percent is shown in Fig.7. It can be seen that extracted health indicators have shown a clearly deteriorating trend.

C. Degradation Model based on Nonlinear Gamma Stochastic Process

For PV arrays, the main causes of performance degradation with long-term operation are the accumulation of irreversible damage caused by physical and chemical alterations [6], so the degradation trend should be a strong monotonicity. It can be seen from Fig.7., the performance degradation percent is sliding averaged using the smooth function in MATLAB and the degradation trend shows a monotonic incremental trend. In the outdoor environment, due to the randomness of the operating environment and the differences between individual

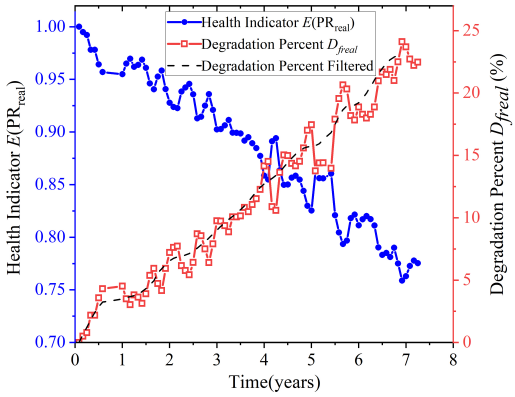


Fig. 7. The health indicator and degradation percent curve for the PV array (DKASC)

PV modules, the performance degradation process of PV arrays is also a stochastic process. The Gamma degradation process is an independent non-negative incremental stochastic degradation process, which can be used to describe the slow degradation due to wear, corrosion, fatigue, etc, or damage due to a large shock [25]. Thus, it is potential to characterize the performance degradation of PV arrays.

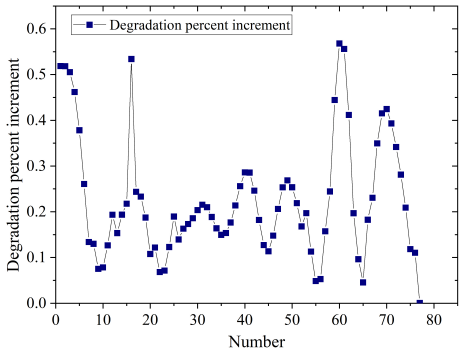


Fig. 8. Degradation percent increment

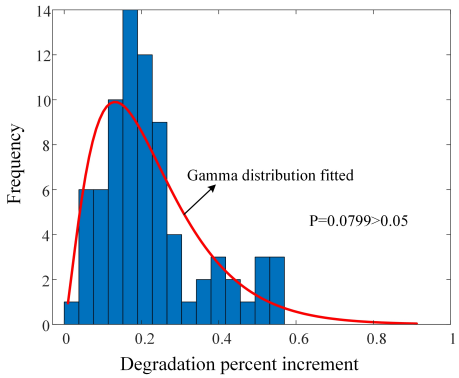


Fig. 9. Calibration histogram of the gamma distributiont

If $\{Y(t), t > 0\}$ is a Gamma stochastic process, which has the following properties: (1) $Y(t)$ is the degradation percent and $Y(0) = 0$; (2) $Y(t)$ has independent increments, at time intervals $[t_1, t_2]$, $[t_3, t_4]$ $(t_1 < t_2 < t_3 < t_4)$, and the

increments $Y(t_4)-Y(t_3)$ and $Y(t_2)-Y(t_1)$ are independent of each other; (3) The degradation increment $\Delta Y(\mathrm{t}) = Y(t+\Delta t)-Y(t)$ follows a Gamma distribution, $\Delta Y(t) \sim \mathrm{Gamma}(\Delta\eta(t),\lambda)$; $\Delta\eta(t)=\eta(t+\Delta t)^q-\eta(t)^q$ is a function of time, when $q=1$ for a linear degradation process, otherwise for a non-linear degradation process, and $\eta(0)^q=0$, and thus, $Y(t+\Delta t)-Y(t) \sim \mathrm{Gamma}\left(\eta(t+\Delta t)^q-\eta(t)^q,\lambda\right)$; η is the shape parameter of the Gamma distribution, $\eta>0$; λ is the scale parameter of the Gamma distribution, $\lambda>0$. The probability density function (PDF) of $Y(t)$ can be expressed as follows:

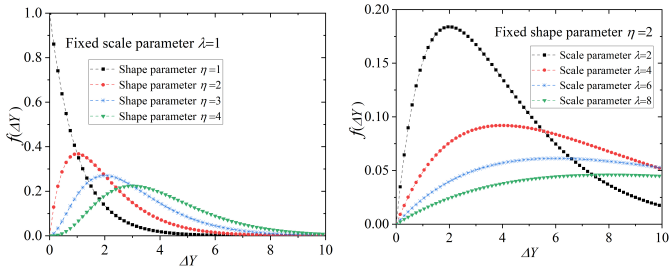
$$f(Y(t)\mid\eta(t),\lambda)=\frac{Y(t)^{\eta(t)-1}e^{-\frac{Y(t)}{\lambda}}}{\lambda^{\eta(t)}\Gamma(\eta(t))}I_{(0,\infty)}(Y(t)) \tag{16}$$

where $\Gamma(\eta)=\int_0^\infty x^{n-1}e^{-x^d}\,\mathrm{d}x$ is a Gamma function, and the $I_{(0,\infty)}(Y(t))$ is given as:

$$I_{(0,\infty)}(Y(t))=\begin{cases}1 & Y(t)\in(0,\infty)\\ 0 & Y(t)\notin(0,\infty)\end{cases} \tag{17}$$

The expectation and variance $E[Y(t)], Var[Y(t)]$ of the Gamma process $\{Y(t),t>0\}$ are denoted as follows:

$$\begin{cases}E[Y(t)]=\lambda\eta(t)\\ Var[Y(t)]=\lambda^2\eta(t)\end{cases} \tag{18}$$



(a) (b)

Fig. 10. The probability density curve of the Gamma distribution: (a) Shape parameter η change; (b) Scale parameter λ change

The performance degradation percent increment of PV arrays is shown in Fig.8 and the distribution of degradation percent increment is fitted by Gamma distribution as shown in Fig.9. Then, the goodness-of-fit is tested by Kolmogorov-Smirnov (K-S) to obtain P=0.0799>0.05 [26], which indicates that the degradation percent increment of PV arrays satisfies the gamma distribution. It is demonstrated that the performance degradation of PV arrays can be described by using the Gamma process. The model parameters that need to be solved are the shape parameter η, and the scale parameter λ. The distribution of the Gamma probability density curve for different parameters is illustrated in Fig.10. The above results further reveal the feasibility of using the Gamma process to describe the degradation process, which can be used to depict degradation data with different probability distributions by modifying the model parameters.

D. Parameter Estimation based on the Hybrid Optimization Algorithm

The model parameters to be solved for the Gamma stochastic degradation process model include the shape parameter η and the scale parameter λ. Generally, η is a parameter that varies with time, i.e., $\eta(t)=kt^q, q=1$ the Gamma stochastic degradation process is a linear degradation process, otherwise, it is a non-linear degradation process. Thus, the model parameters $[k,\lambda,q]$ will need to be solved for the non-linear degradation process. The performance degradation percent $\{D_{freal}(t),t>0\}$ for the PV array obtained after the HI reconstruction is a gamma stochastic process. The increment of the degradation percent $\Delta D_{freal(i)}=D_{freal(i)}-D_{freal(i-1)}$ follows a Gamma distribution and is expressed as follows.

$$\Delta D_{freal(i)} \sim \mathrm{Gamma}\left(k\left(t_{(i)}^q-t_{(i-1)}^q\right),\lambda\right) \tag{19}$$

The probability density function of $\Delta D_{freal(i)}$ can be expressed as follows:

$$f\left(\Delta D_{freal(i)}\mid k\Delta d_{(i)},\lambda\right)=\frac{1}{\lambda^{k\Delta d_i}\Gamma\left(k\Delta d_{(i)}\right)}\Delta D_{freal(i)}{}^{k\Delta d_{(i)}-1}e^{-\frac{\Delta D_{freal(i)}}{\lambda}} \tag{20}$$

where $\Delta d_{(i)}=t_{(i)}^q-t_{(i-1)}^q$. Then, the maximum likelihood function is expressed as follows:

$$\begin{aligned}L\left(k\Delta d_{(i)},\lambda\right)=&\sum_{i=1}^{m}\left(k\Delta d_{(i)}-1\right)\ln\Delta D_{freal(i)}-kt_{(m)}\ln\lambda\\&-\sum_{i=1}^{m}\ln\Gamma\left(k\Delta d_{(i)}\right)-\frac{D_{freal(m)}}{\lambda}\end{aligned} \tag{21}$$

The model parameters $[k,\lambda,q]$ are estimated by solving for the extreme value of the maximum likelihood function. The traditional approach is to take the logarithm of Eq.22, then take the partial derivatives of the model parameters k,λ,q and solve the system of equations by an iterative algorithm. However, the iterative method has certain limitations, which require the objective function to be derivable and continuous. In addition, it is susceptible to local convergence and divergence due to the influence of initial values. Thus, to overcome the above problems, the model parameter estimation of the gamma stochastic process is converted into an optimization problem, the target of which is to minimize the value of objective function $F\left(D_{freal},\boldsymbol{\Theta}\right)$ by searching the solution vector $\boldsymbol{\Theta}$ within the range of the parameter. The objective function is formulated as follows:

$$\begin{cases}F\left(D_{freal},\boldsymbol{\Theta}\right)=1/\left[\sum_{i=1}^{m}\left(k\Delta d_{(i)}-1\right)\ln\Delta D_{freal(i)}\right.\\ \qquad\left.-kt_{(m)}\ln\lambda-\sum_{i=1}^{m}\ln\Gamma\left(k\Delta d_{(i)}\right)-\frac{D_{freal(m)}}{\lambda}\right]\\ \boldsymbol{\Theta}=\{k,\lambda,q\}\end{cases} \tag{22}$$

where m denotes the number of measured degradation percent data.

In recent years, metaheuristic optimization algorithms have been widely used to solve complex mathematical problems, they have global search capabilities to avoid leading to local optima and are able to handle computational and non-linear

problems without gradient constraints. For example, the Grey Wolf Optimization (GWO) [27], Sine Cosine Algorithm (SCA) [28], Cultural Algorithm (CA) [29] and Particle Swarm Optimization (PSO) [30], etc., have been successfully applied to solve problems in a variety of fields. Although the above algorithms have high solution accuracy and robustness, they are still prone to local optima and slow convergence. Therefore, a hybrid particle swarm and grey wolf optimization algorithm (PSO-GWO) [31], by combining the advantages of the different algorithms to overcome the disadvantages of the individual algorithms, is developed to accurately and reliably identify the unknown model parameters of the nonlinear Gamma stochastic process and is compared with other optimization algorithms. In addition, the method can also be used to extract the model parameters for PV arrays.

In order to evaluate the performance of PSO-GWO, it is compared with seven state-of-the-art algorithms, including GWO, PSO, CA, SCA, Wale Optimization Algorithm (WOA) [32] and Dragonfly Algorithm (DA) [33]. For a fair comparison, the maximum iteration number and population size of all approaches were set to 100 and 50, respectively. Besides, all algorithms are completed using the MATLAB script and executed on the computer with Intel(R) Core(TM) i7-9750H CPU@2.60 GHz,16GB(RAM), and 64 bits operating system. The convergence curves of all compared algorithms are provided in Fig.11. As can be seen from Fig.11, PSO-GWO has a faster convergence speed than the other algorithms, and it converges with the minimum fitness value, indicating a better solution accuracy. Then, the extracted parameters are used to reconstruct the degradation percent curve using Eq.18 and the real degradation percent curve is compared with the estimated curve of the different algorithms as shown in Fig.12. Furthermore, the root mean square error (RMSE), the mean absolute error (MAE), and correlation coefficient (R^2) between the real and estimated degradation curves are used as the criterion for evaluating the effectiveness of the algorithm and accuracy of the model:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{points}} \sum_{i=1}^{N_{points}} \left(D_{freal,i} - D_{fmodel,i}\right)^2} \tag{23}$$

$$\mathrm{MAE} = \frac{1}{N_{points}} \sum_{i=1}^{N_{points}} \left|D_{freal,i} - D_{fmodel,i}\right| \tag{24}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N_{points}} \left(D_{freal,i} - D_{fmodel,i}\right)^2}{\sum_{i=1}^{N_{points}} \left(D_{fmean} - D_{fmodel,i}\right)^2} \tag{25}$$

where N_{points} is the number of points on the degradation percent curve, $D_{freeal,i}$ and $D_{fmodel,i}$ are the real and estimated value at the i-th point in the degradation percent curve, respectively.

The statistical results, including the RMSE, MAE, R^2, time and RUL Prediction Error for all comparison algorithms are provided in TableIII. It is clear that the proposed PSO-GWO can provide better performance compared to other algorithms in model parameter extraction. Finally, the radar chart of statistical results is presented in Fig.13, which further demonstrates the effectiveness of the proposed method for solving the model parameters. For the RUL prediction error, it can be seen that PSO, CA and DA algorithms have larger RMSE and MAE, which indicates that the estimated degradation curve based on the condition monitoring data is gradually deviating from the real degradation curve with larger RUL prediction errors. GWO, SCA, WOA and PSO-GWO algorithms have similar RMSE and MAE, which also provide small RUL prediction errors. However, there is a certain uncertainty of RUL prediction error due to the real degradation curve fluctuating. In order to ensure the reliability of model parameter extraction based on condition monitoring data, the PSO-GWO algorithm with better solution accuracy is applied to identify model parameters of the Gamma process.

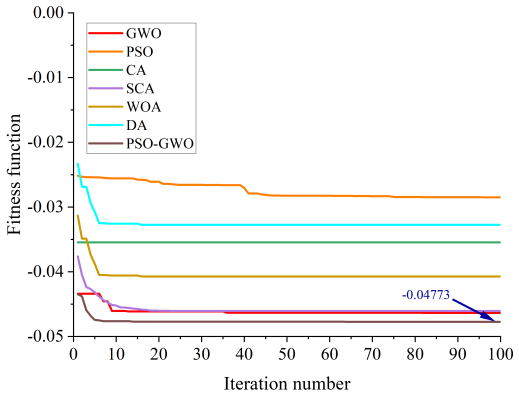


Fig. 11. Convergence curves of compared algorithms for the degradation model

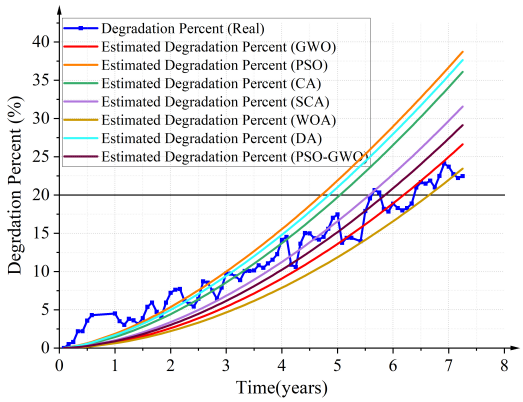


Fig. 12. Real and estimated degradation curves of different algorithms

E. Remaining useful lifetime prediction

The Remaining Useful Life (RUL) is the time duration between the Current Time (Tc) and the Failure Time (Tf). The expression for the calculation of RUL is as follows:

$$RUL = Tf - Tc \tag{26}$$

Nowadays, manufacturers of PV modules usually offer a guarantee period of 20 to 25 years, which guarantees that the power generation performance will not decrease more than

TABLE III
PARAMETER ESTIMATION PERFORMANCE COMPARISON FOR DIFFERENT ALGORITHMS

Method	RMSE	MAE	R^2	Time(s)	Error of RUL Prediction(years)
GWO	3.0522	2.624	0.8792	3.9621	0.1441
PSO	6.7745	4.9774	0.6988	**3.1824**	1.6187
CA	5.2896	3.7843	0.7751	4.2280	1.2877
SCA	3.3961	2.7174	0.8721	3.6220	0.7590
WOA	3.5225	3.1291	0.7908	3.5947	0.3155
DA	6.1493	4.4654	0.7302	4.1754	1.4873
PSO-GWO	**2.8446**	**2.5151**	**0.8953**	3.3950	0.3744

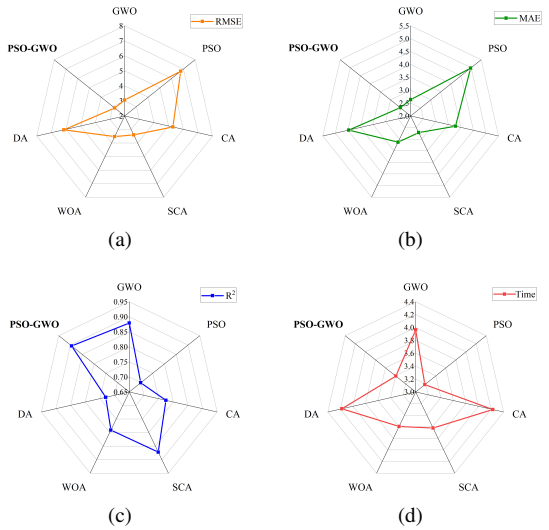


Fig. 13. Radar chart of statistical results for different algorithms: (a) RMSE; (b) MAE; (c) R^2; (d) Time

20% under STC, i.e., PR_{DC} decline of 20% during the period of operation [19]. Therefore, PR_{DC} degradation of 20% is defined as the failure threshold (end of the useful lifetime) of the PV array. It means that PV arrays will fail once the D_f reaches 20% and the degradation failure schematic diagram is shown in Fig. 14. At the current time T_{c}, the model parameters are estimated by the above-proposed method. Then, the established model is utilized to extrapolate the degradation trend curve and obtain the time-varying probability distribution of the degradation percent. Finally, the RUL is calculated until the predicted degradation percent reaches the failure threshold by using Eq.(27).

III. Experimental Verification and Comparative Analysis

A. Case Study 1: based on the DKASC Dataset

1) *Data Acquisition and Description:* In this paper, the historical data are collected from the Desert Knowledge Australia Solar Center (DKASC) located in Alice Springs in Australia. The dataset is publicly available and can be obtained from [34]. Alice Springs is a solar-rich city and the climate is a dry desert climate. The average temperature exceeds 30°C for six months of the year and the lowest average temperature in winter is 5.5°C . In addition, the daily temperature difference of up to 28 °C , and repeated changes in temperature and

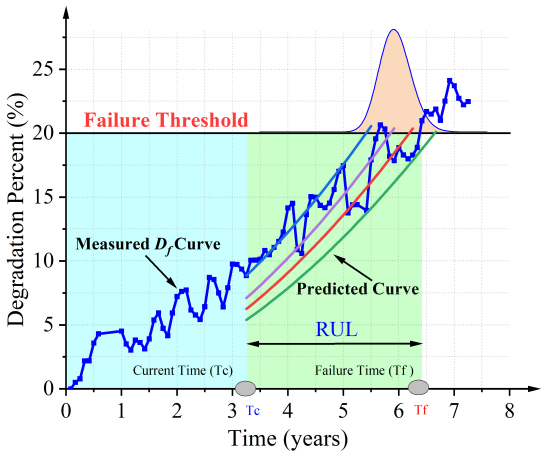


Fig. 14. RUL prediction for PV arrays

TABLE IV
PARAMETRIC INFORMATION OF THE PV SYSTEM TDG IN DKASC

Parameters	Value
Array Rating	5kW
Panel Rating	250W
Number of Panels	20
Panel Type	TDG T250M606/mono-Si
Array Area	33.5m^2
Inverter Size/Type	APS Micro-inverter YC500
Installation Completed	Wed, 7 Aug 2013
Array Tilt/Azimuth	Tilt=20, Azi=0 (Solar North)

humidity stress can cause fatigue damage to the PV modules, accelerating failure and seriously affecting the useful lifetime of PV arrays. The map of the system is shown in Fig.15 and the data used in this study is from TDG. The detailed parametric information for the TDG PV system is listed in Table IV. For long-term monitoring data, the operation time is calculated as follows:

$$T_{\mathrm{c}} = Y + (M/12) + (D/365) + (N/(365 * 24 * 60)) \quad (27)$$

$$T_{\mathrm{o}} = T_{\mathrm{c}} - T_{\mathrm{c0}} \quad (28)$$

where T_{c} is the condition monitoring time of PV arrays; T_{o} is the operation time of PV arrays; T_{c0} is the initial time of the installation for PV arrays. Y, M, D and N represent the year, month, day and minute of the condition monitoring time, respectively. The historical data was collected from 17 March 2014 to 16 June 2021, a total run of 7.25 years, with a sampling frequency of one point every 5 minutes, which makes a total of 756546 data sample points.

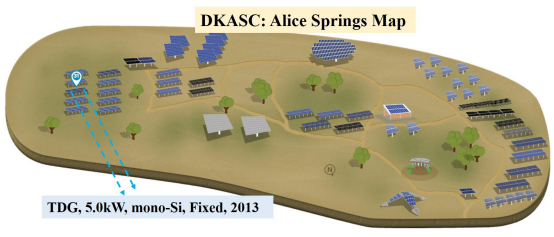


Fig. 15. The location of PV arrays on DKASC in Australia [34]

2) *Prediction Performance of the Proposed Method:* The collected data of the DKASC is shown in Fig.3, which contains a large amount of outlier data and the output power varied with the environmental conditions. Thus, an effective health indicator is obtained by the health indicator reconstruction method and the degradation percent is used to establish the degradation model based on the nonlinear Gamma process. The solution of the model parameters is transformed into an optimization problem and the model parameter set $\boldsymbol{\Theta}$ is estimated by a hybrid PSO-GWO optimization algorithm. The details of the health indicator reconstruction, degradation model establishment and model parameter set $\boldsymbol{\Theta}$ estimation are described in Section II.

In order to evaluate the effectiveness of the proposed method and the prediction accuracy of the RUL. The top 50% of the DKASC dataset (run to 4^{th} year with a degradation percent of 14.1501%) used to estimate the degradation model parameter set $\boldsymbol{\Theta}$. The model parameters $\boldsymbol{\Theta}$ are estimated as $k = 7.2117, q = 1.2595, b = 0.3192$ and the time-varying probability density function (PDF) of the Gamma distribution is calculated. Fig.16 presents the PDF, the real and estimated degradation percentages and the 95% confidence interval (CI) for the estimated value. It can be seen that the estimated degradation percent (expectation of PDF) is quite close to the real degradation percent and all the real degradation percents are well within the 95% CI of the estimated degradation percents. Table V shows the details of the degradation trend predictions. RMSE, MAE and R^2 were then applied as indicators to evaluate the predictive accuracy of the degradation trend. The RMSE, MAE and R^2 were 1.8283, 1.4498 and 0.9206, respectively. In addition, as can be seen in Fig.16, the width of the distribution for PDF is gradually larger over time, which indicates that the uncertainty of prediction is increasing. Finally, the predicted time to failure for the PV array to reach 20% degradation is 5.5652 years, and then the RUL is calculated to be 1.5652 years using Eq.26. The actual time to failure of the PV array is 6.333 years and the actual RUL is 2.333 years, so the prediction error of the RUL is 0.7678 years.

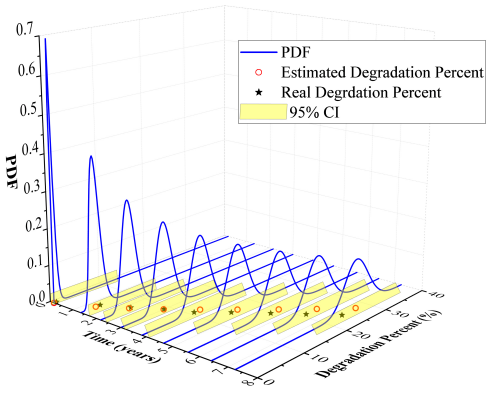


Fig. 16. Degradation rate prediction by using the reconstruction health indicator (DKASC)

3) *Comparison with Other Methods :* In order to further validate the effectiveness of the proposed method, three different types of methods are used for comparative analysis, including the empirical models (Polynomial model, Power exponential model, Exponential model, Logistic model), statistical model (ARIMA) and deep neural network (LSTM). Moreover, to evaluate the performance of the RUL prediction methods under different training samples, 40% and 70% of the data are used as training samples to train the model. Correspondingly, the remaining data samples are used as a testing set to examine the accuracy of the prediction model. The three metrics including RMSE, MAE and RUL error are selected to evaluate the accuracy of different methods for degradation trend prediction. The degradation percent extracted by the HI reconstruction method proposed in this paper is used as the training and testing datasets in all comparison algorithms to ensure comparability of the experimental.

TABLE V
COMPARISON OF ESTIMATED AND REAL DEGRADATION PERCENT AND 95% CI (DKASC)

CM Time (year)	Real DP (%)	Estimated DP (%)	95% CI
0.1667	0.5100	0.0	[-13.47, 13.47]
1.333	3.9046	3.0	[-10.47, 16.47]
2.167	6.1656	5.8	[-7.67, 19.27]
3.000	8.8667	8.8	[-4.67, 22.27]
3.833	10.8893	12.2	[-1.27, 25.67]
4.667	14.3948	15.8	[2.33, 29.27]
5.583	17.9036	19.8	[6.33, 33.27]
6.4167	20.9748	23.4	[10.13, 37.07]
7.25	24.1144	27.6	[14.13, 41.07]

The empirical models (Polynomial model, Power exponential model, Exponential model, Logistic model) are built by using least squares to solve for the model parameters based on the training samples. Least squares minimize the sum of squares of the errors between the training data and the fitted function. The LSTM is trained based on training samples and completes long-term sequence ahead prediction. The long-term sequence ahead prediction structure of the LSTM is shown in Fig.17. The long-term sequence ahead prediction is completed by the iterative approach, which is based on the single-step ahead prediction process. The historical data of $\{x_0,\ x_1, \cdots, x_k\}$ are used to estimate the degradation rate of the next step x_{k+1}. At the next step, the predicted degradation rate x_{k+1} would be added to the historical sequence data $\{x_1, x_2, \cdots, x_{k+1}\}$ and the sequence is used to estimate the degradation rate of the next step x_{k+2}. By repeating the above prediction process, the prediction results for the long-term degradation rate series $\{x_{k+1},\ x_{k+2}, \cdots, x_{k+2}\}$ are obtained and n denotes the total number of predicted data points. In LSTM, the size of the hidden layer is set to 500, and the learning rate and epoch are set as 0.01 and 1000, respectively. Moreover, the Adam optimizer is applied to train the model. Similarly, the proposed method and ARIMA are trained based on 40% and 70% of the data respectively and to predict the remaining data points. For comparability, the expected value of the PDF for the proposed method in this paper is used as the estimated degradation rate curve.

The prediction results of different algorithms with 40% (3.3333th year with a degradation percent of 10.06%) and 70% training samples (5.1667^{th} year with a degradation percent

TABLE VI
COMPARISON OF DIFFERENT ALGORITHMS WITH 40% AND 70% TRAINING SAMPLES (DKASC)

	40% Training				70% Training			
	RMSE	MAE	Predicted RUL	RUL error	RMSE	MAE	Predicted RUL	RUL error
Polynomial model	2.64	2.20	7.382	1.049	1.74	1.44	6.613	0.28
Power exponential model	2.72	2.28	8.0445	1.7115	1.74	1.44	6.6140	0.281
Exponential model	18.50	13.69	4.7240	1.609	7.27	6.36	5.5445	0.7885
Logistic model	8.50	7.68	349.9142	343.5812	7.12	6.71	31.1348	24.8054
ARIMA	2.4808	1.9741	7.0833	0.7503	**1.7927**	**1.3416**	6.25	0.083
LSTM	4.2058	3.3589	N/A	N/A	3.5441	3.0200	7.25	0.917
Proposed model	**2.3877**	**1.9473**	5.6431	**0.6899**	1.8133	1.3654	6.3782	**0.0453**

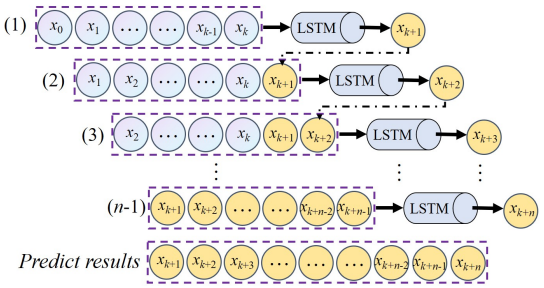


Fig. 17. The long-term sequence ahead prediction structure of the LSTM

TABLE VII
SPECIFICATION OF PV MODULE SM55

Parameters	Value
Maximum power ($P_{mpp,stc}$)	55W
Voltage at maximum power point ($V_{mpp,stc}$)	17.4 V
Current at maximum power point ($I_{mpp,stc}$)	3.15 A
Open-circuit voltage ($V_{OC,stc}$)	21.7 V
Short-circuit current ($I_{SC,stc}$)	3.45 A
Temperature coefficient of current (K_i)	1.2mA/°C
Temperature coefficient of voltage (K_v)	-0.77 V/°C
Limited Warranty	25 Years

of 14.39%) are shown in Fig.18. Fig.18 also shows the real degradation percent, the estimated degradation percent and the corresponding 95% confidence interval. It can be seen that the degradation trends predicted by the polynomial and power exponential models in the empirical models are closer to the real degradation percents. It also demonstrates the effectiveness and significance of the HI reconstruction proposed in this paper, which can achieve better results by using simple empirical models with the effective feature indicators constructed. If effective feature indicators are not extracted, the large amount of uncertainty and volatility in the raw data can lead to higher model fitting errors and make it difficult to obtain a useful model for degradation trend prediction. Moreover, the ARIMA and the nonlinear gamma process approach proposed also have better prediction accuracy and the real degradation percents are well within the 95% CI of the estimated degradation percents by the proposed method. Exponential and Logistic models gradually deviate from the real degradation percents over time.

For better observation of comparative results, the RMSE, MAE, RUL prediction results and the RUL prediction error of different algorithms with 40% and 70% training samples are listed in Table VI. As can be seen, the RMSE and MAE of the proposed method are smaller than other algorithms with 40% training samples indicating better prediction performance. The ARIMA algorithm has the better prediction performance using 70% training samples, with RMSE and MAE of 1.7927 and 1.3416 respectively, while the RMSE and MAE of the proposed method are 1.8133 and 1.3654, with no large difference between them. Besides, RUL prediction errors of the proposed method are 0.6899 and 0.0453 years with 40% and 70% training samples respectively, which is smaller than other algorithms. The degradation percent prediction error of the LSTM is larger in comparison with the Exponential model, Logistic model, ARIMA and the proposed method. The possible reasons for that are summarized as follows: (1) Due to the small training samples, no efficient model can be obtained; (2) Long-time series prediction is inherently challenging for the LSTM model due to its recursive loop structure and forgetting long term historical information, which suffers from the accumulation of prediction errors. The prediction accuracy of all algorithms also increases with training samples from 40% to 70%.

B. Case Study 2: based on the NREL Dataset

1) *Data Acquisition and Description:* The studied PV system is located at the site of the PV system on the west side of the Solar Energy Research Building at the National Renewable Energy Laboratory (NREL). The PV system is composed of 2 arrays, each of which contains 5 sub-strings with 14 SM55 solar modules produced by Siemens, with a total capacity of approximately 7.42kW. One of the PV arrays is selected for the study in this paper. The historical data was collected from 13 May 1993 to 31 December 2014, a total run of approximately 20.19 years, with a sampling frequency every 15 minutes, and it collected a total of 470960 data samples. The filed power data and environmental data (module temperature and irradiance) including metadata, can be obtained from the publicly accessible NREL PV Data Acquisition database [35]. The specification parameters of the SM55 are given in Table VII.

2) *Experimental Results and Analysis:* The collected data of the NREL is shown in Fig.20 (a), and the measured power data also contains a large amount of uncertainty and volatility. Then, the extracted trend term signals, health indicator and degradation percent curve are shown in Fig.20 (b) and Fig.21, respectively. It further validates the effectiveness of the HI indicator reconstruction method proposed in this paper, and the extracted degradation percent curve can effectively describe

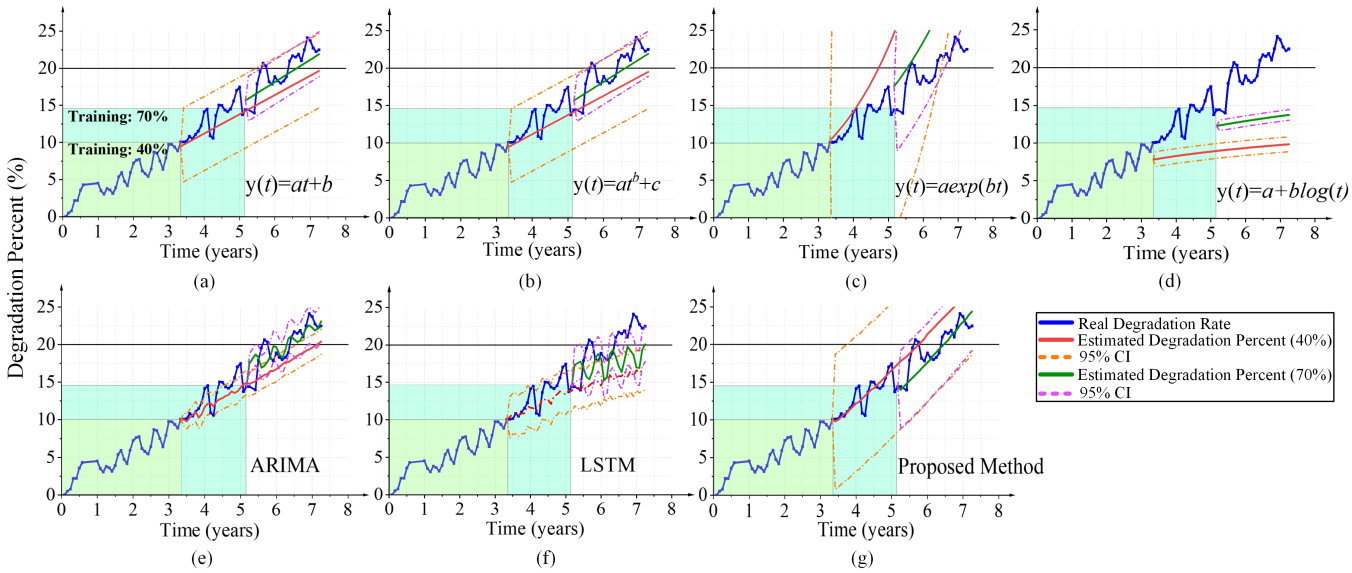


Fig. 18. Degradation percent prediction results based on different algorithms using 40% and 70% training samples (DKASC): (a) Polynomial model; (b) Power exponential model; (c) Exponential model; (d) Logistic model; (e) ARIMA; (f) LSTM; (g) Proposed model.

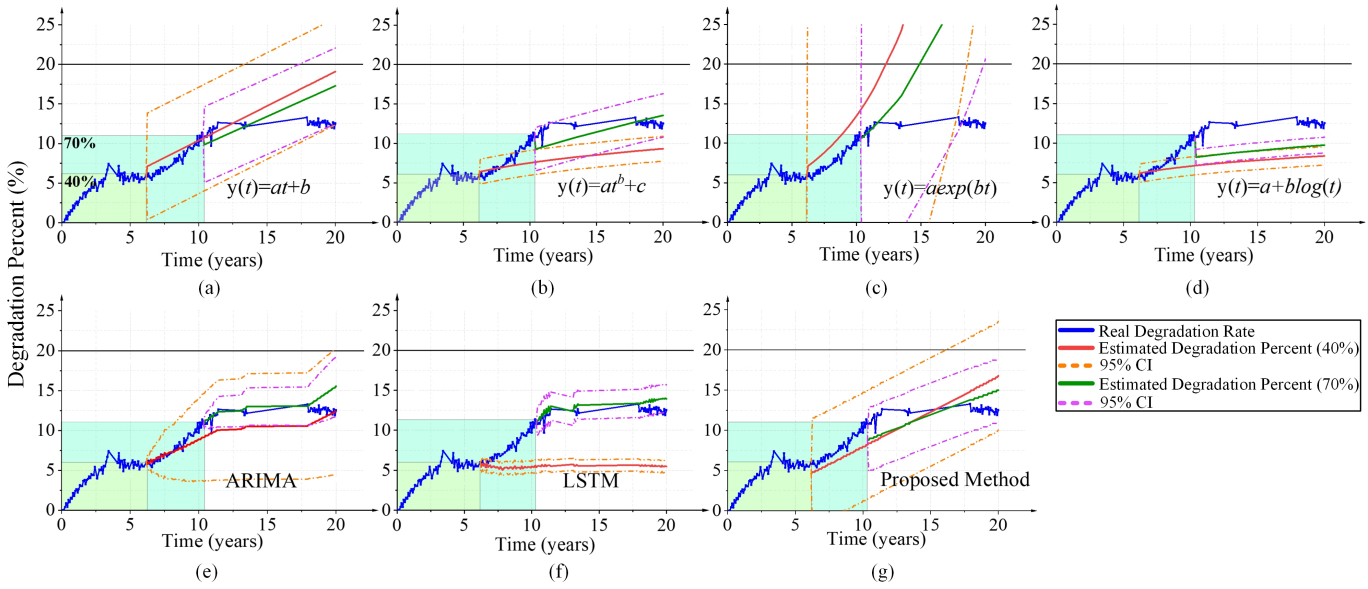


Fig. 19. Degradation percent prediction results based on different algorithms using 40% and 70% training samples (NREL): (a) Polynomial model; (b) Power exponential model; (c) Exponential model; (d) Logistic model; (e) ARIMA; (f) LSTM; (g) Proposed model.

TABLE VIII
COMPARISON OF ESTIMATED AND REAL DEGRADATION PERCENT AND 95% CI (NREL)

CM Time (year)	Real DP (%)	Estimated DP (%)	95% CI
0.1667	0.06	1.2	[-5.17, 7.56]
1.5833	3.6873	4.6	[-1.77, 10.96]
3.1667	5.4840	6.6	[0.24, 12.96]
5.3333	6.0484	8.6	[2.24, 14.96]
6.8333	6.1589	9.6	[3.24, 15.96]
8.5	7.7932	10.6	[4.24, 16.96]
10	9.7193	11.6	[5.24, 17.96]
11.4167	12.6411	12.2	[5.84, 18.56]
18.75	13.1438	15.4	[9.04, 21.76]

the performance degradation trend of the PV array. Due to the variation in the performance of PV modules and operating environments, the degradation percent is different between different PV systems. For the NREL dataset, the PV array has been in operation for 20 years without reaching the failure threshold (20% degradation percent). Similarly, 50% of the NREL dataset (run to 7.5th year with a degradation percent of 7.0371%) is used to estimate the degradation model parameter set $\boldsymbol{\Theta}$. The model parameters $\boldsymbol{\Theta}$ are estimated as $k = 5.1826, q = 0.437, b = 0.8764$, and the PDF of the gamma distribution is calculated, as shown in Fig.22. It can also be seen that the estimated degradation percent (expectation of PDF) is close to the real degradation percent

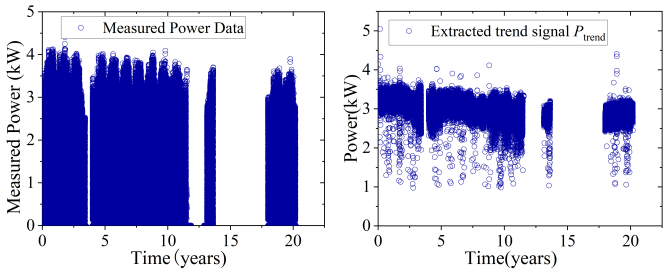


Fig. 20. The measured output power and extracted trend signal of the PV array (NREL): (a) Measured output power; (b) Extracted trend signal P_{trend}

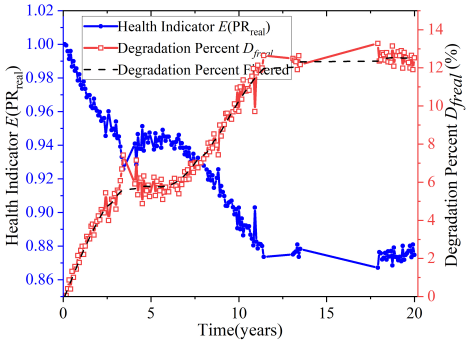


Fig. 21. The health indicator and degradation percent curve for the PV array (NREL)

and all the real degradation percents are well within the 95% CI of the estimated degradation percents. The details of the degradation trend predictions are listed in Table VIII. The RMSE, MAE and R^2 were 2.3284, 2.0648 and 0.7171, respectively. The time to failure at a degradation percent of 20% is calculated as 29.7290 years via Eq.18 and the RUL is 22.229 years.

The comparative analysis is conducted similar to that in Case Study 1. The prediction results of different algorithms with 40% (6.1667th year with a degradation percent of 5.98%) and 70% (10.3333th year with a degradation percent of 10.78%) training samples are shown in Fig.19. As can be seen, the ARIMA and the proposed method are closer to the real degradation percent. More comparative results of RMSE, MAE, RUL prediction results and the RUL prediction error of different algorithms with 40% and 70% training samples are listed in Table IX. The RMSE and MAE of the proposed method are 2.14 and 2.16, which are also smaller than other algorithms with 40% training samples. Similarly, the ARIMA model has better prediction accuracy at 70% training samples with RMSE and MAE of 1.5489 and 1.1792 respectively, while the RMSE and MAE of the proposed method are 1.6506 and 1.3635, without much difference between them. In empirical models, the polynomial has higher accuracy by using 40% training samples, and the power exponential model has higher accuracy with 70% training samples. Finally, the failure times were predicted to be 23.1139 and 34.1484 years respectively by the method proposed with 40% and 70% training samples. The RUL predicted are 16.9472 and 23.8151

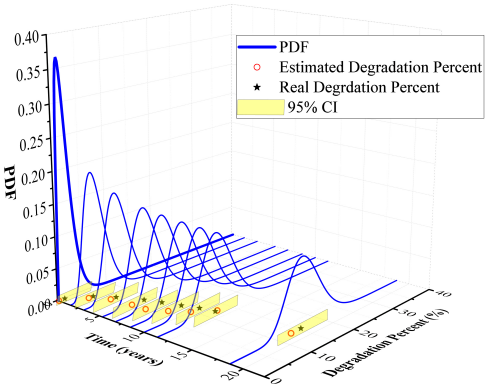


Fig. 22. Degradation rate prediction by using the reconstruction health indicator (NREL)

years, and the RUL error cannot be calculated as the PV array has not reached the failure threshold. In addition, the prediction accuracy of all algorithms also is improved with increasing training samples. However, the prediction accuracy of all algorithms is lower compared to Case Study 1 due to the more complex nonlinear degenerate form of the NREL dataset compared to the DKASC dataset.

IV. Conclusions

In this paper, a self-data-driven model for RUL probability distribution prediction of PV arrays is proposed considering the uncertainty and volatility of condition monitoring data. First, a HI reconstruction method is presented to eliminate the randomness and volatility of condition monitoring data during the long-term operation of PV arrays. The reconstructed health indicators can effectively evaluate the health state of the PV array and describe the corresponding degradation trends. Second, a nonlinear Gamma stochastic process is applied to build the degradation model for predicting the probability distribution of the degradation trend. The model parameters solution is transformed into an optimization problem, and a hybrid optimization algorithm PSO-GWO is developed to estimate the model parameters, which ensures the speed and accuracy of the model parameter solution. Moreover, the proposed method for model parameter estimation is compared with iterative algorithms and various optimization algorithms, which demonstrates the superior performance of the proposed method.

In the experiment, two cases are studied to verify the effectiveness of the proposed method for RUL prediction based on the DKASC and NREL datasets, and the performance is further compared with the empirical models, ARIMA and the LSTM models with different training samples. Experimental results demonstrate that the proposed method has an excellent RUL prediction accuracy, the RUL prediction errors are 0.6899 and 0.0453 years with 40% and 70% training samples in the DKASC dataset. RUL prediction errors cannot be evaluated for the NREL dataset, which has not yet reached the failure threshold. In addition, the polynomial and power exponential models also perform well, which demonstrates the effectiveness and significance of the proposed HI reconstruction

TABLE IX
COMPARISON OF DIFFERENT ALGORITHMS WITH 40% AND 70% TRAINING SAMPLES (NREL)

	40% Training				70% Training			
	RMSE	MAE	Predicted RUL	RUL error	RMSE	MAE	Predicted RUL	RUL error
Polynomial model	3.12	2.25	21.0508	N/A	3.10	2.66	23.5157	N/A
Power exponential model	3.15	2.36	472.2661	N/A	1.39	1.17	39.0953	N/A
Exponential model	26.65	16.42	12.3369	N/A	15.61	12.12	15.1765	N/A
Logistic model	3.31	2.82	10537	N/A	3.10	3.05	15242	N/A
ARIMA	**1.2866**	**1.0416**	N/A	N/A	**1.5489**	**1.1792**	N/A	N/A
LSTM	5.1840	4.8935	N/A	N/A	2.6153	2.0712	N/A	N/A
Proposed model	2.4046	2.1683	23.1139	N/A	1.6506	1.3635	34.1484	N/A

method. Effective health indicators can enable to obtain better prediction results even by simple empirical model fitting.

REFERENCES

[1] M. Meribout, "Sensor systems for pv systems monitoring," *IEEE Trans. Instrum. Meas.*, pp. 1–1, 2022.

[2] H. Momeni, N. Sadoogi, M. Farrokhifar, and H. F. Gharibeh, "Fault diagnosis in photovoltaic arrays using gbssl method and proposing a fault correction system," *IEEE Trans. Ind. Inform.*, vol. 16, no. 8, pp. 5300–5308, 2020.

[3] Y. Liu, K. Ding, J. Zhang, Y. Li, Z. Yang, W. Zheng, and X. Chen, "Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with i-v curves," *Energy Convers. Manage.*, vol. 245, p. 114603, 2021.

[4] S. Leva, M. Mussetta, and E. Ogliari, "Pv module fault diagnosis based on microconverters and day-ahead forecast," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3928–3937, 2019.

[5] P. Jain, J. Poon, J. P. Singh, C. Spanos, S. R. Sanders, and S. K. Panda, "A digital twin approach for fault diagnosis in distributed photovoltaic systems," *IEEE Trans. Power Electron.*, vol. 35, no. 1, pp. 940–956, 2020.

[6] W. J. Jamil, H. A. Rahman, S. Shaari, and Z. Salam, "Performance degradation of photovoltaic power system: Review on mitigation methods," *Renew. Sustain. Energy Rev.*, vol. 67, no. jan., pp. 876–891, 2017.

[7] F. Carigiet, C. J. Brabec, and F. P. Baumgartner, "Long-term power degradation analysis of crystalline silicon pv modules using indoor and outdoor measurement techniques," *Renew. Sustain. Energy Rev.*, vol. 144, no. 12, p. 111005, 2021.

[8] L. Abenante, F. D. Lia, R. Schioppo, and S. Castello, "Non-linear continuous analytical model for performance degradation of photovoltaic module arrays as a function of exposure time," *Appl. Energy*, vol. 275, p. 115363, 2020.

[9] A. Dadaniya and N. V. Datla, "Degradation prediction of encapsulant-glass adhesion in the photovoltaic module under outdoor and accelerated exposures - sciencedirect," *Sol. Energy*, vol. 208, pp. 419–429, 2020.

[10] I. Kaaya, M. Koehl, A. P. Mehilli, D. Sidrach, and K. A. Weiss, "Modeling outdoor service lifetime prediction of pv modules: Effects of combined climatic stressors on pv module power degradation," *IEEE J. Photovoltaics*, vol. PP, no. 4, pp. 1–8, 2019.

[11] M. Dhimish, N. Schofield, and A. Attya, "Insights on the degradation and performance of 3000 photovoltaic installations of various technologies across the united kingdom," *IEEE Trans. Ind. Inform.*, vol. 17, no. 9, pp. 5919–5926, 2021.

[12] X. Si, T. Li, Q. Zhang, and C. Hu, "Prognostics for linear stochastic degrading systems with survival measurements," *IEEE Trans. Ind. Electron*, vol. 67, no. 4, pp. 3202–3215, 2020.

[13] W. Yu, I. Y. Kim, and C. Mechefske, "Remaining useful life estimation using a bidirectional recurrent neural network based autoencoder scheme," *Mech. Syst. Signal Process*, vol. 129, pp. 764–780, 2019.

[14] M.-H. Chang, M. Kang, and M. Pecht, "Prognostics-based led qualification using similarity-based statistical measure with rvm regression model," *IEEE Trans. Ind. Electron*, vol. 64, no. 7, pp. 5667–5677, 2017.

[15] G. Lyu, H. Zhang, and Q. Miao, "Rul prediction of lithium-ion battery in early-cycle stage based on similar sample fusion under lebesgue sampling framework," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[16] M. Theristis, A. Livera, C. B. Jones, G. Makrides, and J. S. Stein, "Nonlinear photovoltaic degradation rates: Modeling and comparison against conventional methods," *IEEE J. Photovoltaics*, vol. 10, no. 4, pp. 1112–1118, 2020.

[17] I. Romero-Fiances, A. Livera, M. Theristis, G. Makrides, J. S. Stein, G. Nofuentes, J. Casa, and G. E. Georghiou, "Impact of duration and missing data on the long-term photovoltaic degradation rate estimation," *Renew. Energy*, vol. 181, pp. 738–748, 2022.

[18] I. Kaaya, S. Lindig, K.-A. Weiss, A. Virtuani, M. Sidrach de Cardona Ortin, and D. Moser, "Photovoltaic lifetime forecast model based on degradation patterns," *Prog. Photovoltaics Res. Appl.*, vol. 28, no. 10, pp. 979–992, 2020.

[19] S. Lindig, I. Kaaya, K. Weiß, D. Moser, and M. Topic, "Review of statistical and analytical degradation models for photovoltaic modules and systems as well as related improvements," *IEEE J. Photovoltaics*, vol. 8, no. 6, pp. 1773–1786, 2018.

[20] Y. Ren, P. N. Suganthan, and N. Srikanth, "A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 236–244, 2015.

[21] Y. Hu and C. Zhao, "Fault diagnosis with dual cointegration analysis of common and specific nonstationary fault variations," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 237–247, 2020.

[22] M. Zhao, L. Ma, X. Jia, D.-M. Yan, and T. Huang, "Graphreg: Dynamical point cloud registration with geometry-aware graph signal processing," *IEEE Trans. Image Process.*, vol. 31, pp. 7449–7464, 2022.

[23] H. Wang, J. Wu, M. Ma, and Q. Chen, "A model-based power switch fault diagnosis strategy for cascaded h-bridge converter," *Microelectronics Reliability*, p. 115095, 2023.

[24] G. Belluardo, P. Ingenhoven, W. Sparber, J. Wagner, P. Weihs, and D. Moser, "Novel method for the improvement in the evaluation of outdoor performance loss rate in different pv technologies and comparison with two other methods," *Sol. Energy*, vol. 117, pp. 139–152, 2015.

[25] S. Zhao, Y. Peng, F. Yang, E. Ugur, and H. Wang, "Health state estimation and remaining useful life prediction of power devices subject to noisy and aperiodic condition monitoring," *IEEE Trans. Instrum. Meas.*, vol. PP, no. 99, pp. 1–1, 2021.

[26] Z. Wang, M. Begovic, and J. Wang, "Analysis of conservation voltage reduction effects based on multistage svr and stochastic process," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 431–439, 2014.

[27] S. Mohanty, B. Subudhi, and P. K. Ray, "A new mppt design using grey wolf optimization technique for photovoltaic system under partial shading conditions," *IEEE Trans. Sustain. Energy*, vol. 7, no. 1, pp. 1–8, 2015.

[28] Y. Wan, A. Ma, L. Zhang, and Y. Zhong, "Multiobjective sine cosine algorithm for remote sensing image spatial-spectral clustering," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–15, 2021.

[29] M. Z. Ali, P. N. Suganthan, R. G. Reynolds, and A. F. Al-Badarneh, "Leveraged neighborhood restructuring in cultural algorithms for solving real-world numerical optimization problems," *IEEE Trans. Evol. Comput.*, vol. 20, no. 2, pp. 218–231, 2016.

[30] Y. Peng, S. Zhao, and H. Wang, "A digital twin based estimation method for health indicators of dc–dc converters," *IEEE Trans. Power Electron.*, vol. 36, no. 2, pp. 2105–2118, 2021.

[31] X. Zhang, Q. Lin, W. Mao, S. Liu, and G. Liu, "Hybrid particle swarm and grey wolf optimizer and its application to clustering optimization," *Appl. Soft Comput.*, vol. 101, no. 9, p. 107061, 2020.

[32] D. Oliv, A. E. A. Mohamed, and A. E. Hassanien, "Parameter estimation of photovoltaic cells using an improved chaotic whale optimization algorithm," *Appl. Energy*, vol. 200, no. aug.15, pp. 141–154, 2017.

[33] M. K. Singla, P. Nijhawan, and A. S. Oberoi, "Parameter estimation of three diode solar pv cell using chaotic dragonfly algorithm," *Soft Comput.*, vol. 26, no. 21, pp. 11 567–11 598, 2022.

[34] "Dkasc, Alice Springs, TDG, 5.0kW, 2013. Available: https://dkasolarcentre.com.au/source/alice-springs/dka-m18-c-phase."

[35] "PV Data Acquisition: NREL. Available: https://developer.nrel.gov/docs/."