A Hybrid Data and Knowledge Driven Risk Prediction Method for Distributed Photovoltaic Systems Considering Spatio-Temporal Characteristics of Extreme Rainfalls

Yuxuan Wang, *Student Member*, *IEEE*, Bin Zhou, *Senior Member*, *IEEE*, Cong Zhang, *Member*, *IEEE*, Siu Wing Or, *Senior Member*, *IEEE*, Xiang Gao, Ziqi Da

***Abstract*—This paper proposes a hybrid knowledge-based and data-driven electrical safety risk (ESR) prediction method considering spatio-temporal characteristics of extreme rainfalls to identify distributed photovoltaic systems (DPVSs) with high risks of shutdowns induced by waterlogging. Firstly, a two-dimensional hydrodynamic partial differential model of DPVS waterlogging is formulated to deduce dynamic distributions of inundation depths under temporal-spatial heterogeneity of extreme rainfalls. A fast image segmentation driven risk partitioning algorithm is developed to extract nonuniform spatial distributions and temporal volatility of rainstorms as well as waterlogging for dividing DPVSs into multiple zones with different degrees of ESRs. Then, a knowledge-based analytical approach for leakage currents concerning inundation depths and parasitic capacitance is mathematically presented to reveal the underlying impacts of extreme rainfalls on ESRs of DPVSs. A data-driven spatio-temporal graph convolutional network is implemented to predict inundation depts of DVPSs for improving ESR prediction accuracy with limited extreme rainfall events and observation samples. Furthermore, probability density functions of spatio-temporal ESRs are formed to dynamically quantify ESR degrees triggering shutdowns of DPVSs in different partitioned zones. Finally, simulation results have validated the effectiveness of the proposed method for the spatio-temporal ESR prediction of DPVSs under extreme rainfalls.**

***Index Terms*—Distributed photovoltaics, deep learning, distribution networks, electrical safety, risk prediction.**

I. Introduction

A. Motivation

EXTREME rainfall events are becoming increasingly frequent with the dramatic global meteorological change, posing a great threat to high economic losses and widespread adverse impacts on distribution networks [1], [2]. Distributed photovoltaic system (DPVS) equipment needs to be installed outdoors for receiving solar energy and is more susceptible to extreme natural disasters [3]. According to the snapshot of global PV markets 2023 [4] published by the International Energy Agency, the installed capacity of PVs around the world has reached 1,185GW by the end of 2022. However, the operational performance and reliability of PV modules are still potential issues due to failures and electric leakages in the field. For instance, from 2002 to 2015, approximately 2500 fire incidents were recorded in nearly 550,000 PV systems in Italy. In May 2019, a severe flooding disaster completely destroyed a DPVS with a total installed capacity of 100 kW in Sanming City, China. Meanwhile, there have been occasional reports related to personal injury or casualties resulting from electric leakages caused by DPVSs. DPVSs are widely scattered across multiple locations within distribution networks, which increases the difficulty of their operations and maintenance, and thus it is becoming a pressing need to investigate the electrical safety risks (ESRs) of DPVSs under extreme climates.

In transformerless DPVSs, the common-mode circuit will result in leakage current flowing through parasitic capacitances to the ground [5], and the waterlogging caused by extreme rainfalls can lead to an increase in the parasitic capacitance of DPVSs. According to international standards IEC 62109-2:2011 [6] and DIN VDE-0126-1-1 [7], it is mandatory to ensure that the leakage currents of transformerless DPVSs do not exceed 300 mA. Once leakage currents exceed the limit value, DPVSs must be disconnected from distribution networks within 0.3 s through rapid shutdown devices (RSDs). A large number of shutdowns of DPVSs will further adversely affect the overall stability of distribution networks [8], [9]. Extreme rainfalls feature numerous centers of intense rainfalls and frequent occurrences of localized heavy downpours, exhibiting highly spatio-temporal heterogeneous distribution of precipitation [10]. These intricate spatio-temporal variations underscore the complex nature of extreme rainfalls, making the occurrence and interaction of waterlogging and ESRs of DPVSs more diverse and complex. Consequently, this paper aims to investigate an ESR prediction method for DPVSs considering spatio-temporal characteristics of extreme rainfalls to identify DPVSs with high shutdown risks, and strives to provide guidance for electrical operators for formulating waterlogging prevention strategies against potential wide-scale shutdowns of DPVSs.

B. Relevant Background

So far, extensive research works have been devoted to investigating the adverse impacts of extreme meteorological disasters on DPVSs. The impacts of typhoons on the operation status and stability of DPVSs were assessed in [11], [12], aiming to optimize the design and installation positions of DPVSs for enhancing resilience against typhoons. To assess the risks of DPVSs under lightning hazards, a high-precision equivalent circuit

This work was jointly supported by the Research Grants Council of the HKSAR Government under Grant R5020-18, the Innovation and Technology Commission of the HKSAR Government to the Hong Kong Branch of National Rail Transit Electrification and Automation Engineering Technology Research Center under Grant K-BBY1, the National Natural Science Foundation of China (52277091). (*Corresponding author: Bin Zhou*, *Siu Wing Or*)

Y. Wang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China, and also with the Hong Kong Branch of National Rail Transit Electrification and Automation Engineering Technology Research Center, Hong Kong, China (email: wang_yuxuan@hnu.edu.cn).

B. Zhou, C. Zhang, and Z. Da are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: binzhou@hnu.edu.cn; zcong@hnu.edu.cn; 756481893@qq.com).

S. W. Or is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China, and also with the Hong Kong Branch of National Rail Transit Electrification and Automation Engineering Technology Research Center, Hong Kong, China (e-mail: eeswor@polyu.edu.hk).

X. Gao is with the Industrial Training Centre, Shenzhen Polytechnic University, Shenzhen, China (e-mail: gaoxiang@szpu.edu.cn).

method considering the coupling effects and the influence of metal frames is developed in [13], which demonstrates the consideration of coupling effects in the design of DPVS lightning protection systems can mitigate the risks of lightning strikes. Nevertheless, the mechanisms of electric leakages of DPVSs are revealed in [5]. By incorporating the effects of water on parasitic capacitance, the model provides insights into how rainfall-induced waterlogging can affect the leakage current of DPVSs. Additionally, the adverse impacts of operational risks of DPVSs on distribution networks are studied in [8]-[9], [14]-[15]. Extensive shutdowns of DPVSs would result in a modification of the power flow, leading to voltage fluctuations and adversely affecting the overall stability of distribution networks. Current studies have made efforts to investigate the impacts of extreme weather conditions, such as lightning and typhoons on the operational risks of DPVSs, while investigations on waterlogging-triggered shutdown risks of DPVSs under extreme rainfalls have not been involved yet.

With the continuous development of big data technology, data-driven methods have been widely implemented in the field of risk prediction in distribution networks [16]-[18] and renewable generation systems [19]-[21], which generally demonstrate higher prediction accuracy compared with traditional approaches. However, data-driven methods lack clear physical significance, meaning that the underlying physical mechanisms or relationships may not be explicitly captured or explained, and the superior performance of deep learning relies heavily on a large collection of sample data with high quality [22]-[23]. Thus, to compensate for the above limitations and enhance models' interpretability, field knowledge is often integrated into data-driven methods, and an amount of hybrid data and knowledge methods are proposed. The hybrid mode of knowledge and data driven technology contains the following four categories: 1) Cascading mode [24]: Utilizing knowledge-based approaches to process data and using the results as the input of data-driven methods; 2) Parallel mode [25]: Knowledge-based and data-driven methods are parallelly executed, and the results are outputted through methods such as weighted averaging; 3) Feedback mode [26]: Using data-driven approaches to correct certain unknown mechanisms within knowledge-driven models; 4) Informed mode [27]: Leveraging the knowledge from physical models to reconstruct the loss function of data-driven methods. Satisfactory combinations of the complementary domain knowledge and data were proposed in [28], [29] for calendar health prognostics of Lithium-Ion batteries, and the introduction of domain knowledge can significantly improve the forecast performance compared to pure data-driven methods. In [24], a knowledge augmented training method was adopted to improve the sample efficiency and avoid the overfitting of data-driven models. The data insufficiency issue can be effectively solved and thus achieve a higher sample efficiency. A knowledge-based model is proposed in [25] to assess damage probabilities of transmission line-tower systems, and the data-driven method is developed based on historical damage data to learn correction factors, and the damage probabilities are predicted by combining outputs of both data-driven and model-driven. Moreover, a hybrid physical model-driven and data-driven framework for linearizing branch power flow was proposed in [26], in which the data-driven method is utilized to obtain the linearized errors and improve the approximation accuracy of the physical-equation-based linearization.

Taking a comprehensive view of the existing studies, they have only considered the impacts of water on the electric leakage risk of an individual DPVS, while neglecting the large-scale shutdowns of DPVSs induced by electric leakages within distribution networks under extreme rainfalls. Waterlogging-prone DPVSs and their corresponding locations cannot be visually identified, and it is hard to provide effective guidance for the operation and maintenance personnel of DPVSs to carry out differentiated strategies and implement waterlogging prevention plans during the rainy season. Moreover, DPVSs are spatially located in distribution networks, where spatio-temporal distributions of rainfalls vary greatly between different zones. The spatio-temporal heterogeneity and volatility of extreme rainfalls have been not taken into account yet in previous studies, and the spatio-temporal ESR prediction methodology for DPVSs under extreme rainfalls is still in a technological gap, lacking well-developed techniques and methods.

C. Contribution

In this paper, a hybrid knowledge-based and data-driven ESR prediction methodology for DPVSs considering spatio-temporal characteristics of extreme rainfalls is proposed to identify DPVSs with high risks of shutdowns induced by electric leakages. The pivotal contributions of this paper are threefold:

1) A risk partitioning algorithm is proposed to extract uneven spatial distributions and temporal volatility of rainstorms as well as waterlogging for improving ESR prediction accuracy of DPVSs. A two-dimensional hydrodynamic partial differential model of waterlogging is formulated to deduce dynamic distributions of inundation depths of DPVSs, and a fast fuzzy *c*-means (FFCM) based segmentation algorithm is utilized to partition DPVSs into multiple zones with different spatio-temporal characteristics of extreme rainfalls as well as inundation depths.

2) A knowledge-based analytical approach for leakage currents with respect to inundation depths is presented to reveal the impacts of extreme rainfalls on ESRs of DPVSs. A mathematical expression of leakage currents concerning inundation depths and parasitic capacitance is derived to calculate leakage currents of DVPSs for the subsequent ESR prediction. Then, the impacts of inundation depths on parasitic capacitance and leakage currents are analyzed under different extreme rainfall conditions.

3) A data-driven ESR prediction method is proposed to identify DPVSs with high risks of shutdowns induced by electric leakages under rainstorms. The spatio-temporal graph convolutional network (STGCN) is implemented to predict waterlogging risks within distribution networks, which can convert input data to advanced graphical representation and has strong spatio-temporal feature extraction capability to improve prediction accuracy with limited extreme rainfall events. The kernel density estimation (KDE) is utilized to quantify spatio-temporal ESR degrees triggering shutdowns of DPVSs.

II. ESR Partitioning with Spatio-Temporal Volatility of Rainfalls

A. Spatio-Temporal Characteristics of Extreme Rainfalls

Electric leakages of DPVSs are mainly caused by extreme rainfall-induced waterlogging [5]. Spatio-temporal characteristics of extreme rainfalls and waterlogging are two critical factors directly influencing ESR variations of DPVSs. As DPVSs are widely located in distribution networks, in which spatio-temporal distributions of rainfalls vary greatly between different zones, the refined characterization of spatio-temporal volatility of extreme rainfalls is crucial for ESR prediction. However, due

to the limited installed capacities of DPVSs, there are no specialized meteorological monitoring devices, and wide-area meteorological data tend to be coarser-grained and have lower latitude/longitude resolution, resulting in the lower ESR prediction accuracy of DPVSs under extreme rainfalls. Hence, the Co-Kriging [30] based spatial interpolation method is utilized to obtain high-resolution rainfall data for refined nonuniform spatial distributions and temporal volatility of rainstorms. The area scope is meshed with multiple grids, and the rainfall intensity of grid points is given by (1).

$$q_{i,j}^{t} = \sum_{m=1}^{l} \lambda_m q_m^t + \alpha(g_{i,j} - m_y + m_z) \tag{1}$$

where $q_{i,j}^{t}$ denotes the rainfall intensity of grid (x_i, x_j) at time t; x_i and x_j represent spatial coordinates on the horizontal plane; q_m^t is measured rainfall amounts of meteorological observation station m at time t; $g_{i,j}$ is the elevation data of grid (x_i, x_j), which can be derived from the digital elevation model (DEM), m_y and m_z are the global average elevation and precipitation within the distribution network area; λ_m is the weight of the meteorological observation station m; α is the weight of covariate variables. λ_m and α can be obtained by the Lagrange multiplier method. Then, the uneven spatial distributions and time-varying features of extreme rainfalls and undulation of terrains are characterized as (2).

$$\boldsymbol{Q}^n = \begin{bmatrix} q_{1,1}^t & q_{1,2}^t \cdots & q_{1,j}^t \\ q_{2,1}^t & q_{2,2}^t \cdots & q_{2,j}^t \\ \vdots & \vdots \ddots & \vdots \\ q_{i,1}^t & q_{i,2}^t \cdots & q_{i,j}^t \end{bmatrix}, \boldsymbol{\gamma}^n = \begin{bmatrix} \boldsymbol{Q}^1 \\ \boldsymbol{Q}^2 \\ \vdots \\ \boldsymbol{Q}^n \end{bmatrix}^T \tag{2}$$

where $\boldsymbol{q}_{i,j}^t = [q_{i,j}^1, q_{i,j}^2, \ldots, q_{i,j}^t]$ is the time-sequential rainfall intensity of the grid (x_i, x_j); t is the time-sequential index; $\boldsymbol{Q}^n$ denotes the spatial distribution of rainfall intensity matrix under the n-th extreme rainfall events; $\boldsymbol{\gamma}^n$ is the set of historical extreme rainfall events.

Due to historical rainfall samples are of inconsistent time-series lengths and challenging to visually express characteristics of rainfalls in different zones. Thus, statistical features are adopted to characterize the temporal-spatial heterogeneity of extreme rainfalls. Typically, the rainfall intensity follows the logarithmic normal distribution [31], as shown in (3).

$$f_{log}(q_{i,j}) = \frac{1}{q_{i,j}\sqrt{2\pi}\sigma_{i,j}} \exp[-\frac{1}{2\sigma_{i,j}^2}(\ln q_{i,j} - \mu_{i,j})] \tag{3}$$

where $q_{i,j}$ is the rainfall intensity of grid (x_i, x_j); $\sigma_{i,j}$ and $\mu_{i,j}$ denote the location parameters and scale parameters, which can be obtained by the maximum likelihood estimation (MLE). On this basis, temporal-spatial features of extreme rainfalls can be represented by $\sigma_{i,j}$ and $\mu_{i,j}$, and the rainfall feature dimensionality of each grid is reduced into two dimensions.

Consequently, the spatio-temporal volatility of extreme rainfalls characterized by (2) can be reformulated as follows,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \cdots \sigma_{1,j} \\ \sigma_{2,1} & \sigma_{2,2} \cdots \sigma_{2,j} \\ \vdots & \vdots \ddots \vdots \\ \sigma_{i,1} & \sigma_{i,2} \cdots \sigma_{i,j} \end{bmatrix}, \boldsymbol{M} = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} \cdots & \mu_{1,j} \\ \mu_{2,1} & \mu_{2,2} \cdots & \mu_{2,j} \\ \vdots & \vdots \ddots & \vdots \\ \mu_{i,1} & \mu_{i,2} & \cdots \mu_{i,j} \end{bmatrix} \tag{4}$$

where $\boldsymbol{\Sigma}$, $\boldsymbol{M}$ denote matrices composed of $\sigma_{i,j}$, $\mu_{i,j}$, respectively. Coordinating $\boldsymbol{\Sigma}$ and $\boldsymbol{M}$, the temporal-spatial heterogeneity of extreme rainfalls can be characterized.

B. DPVS Waterlogging Model Under Extreme Rainfalls

The water level monitoring system for urban waterlogging detection can record real-time inundation depths under extreme rainfalls. However, due to the limited number of waterlogging monitors, historical inundation depth samples are usually small. Additionally, inundation depths of urban regions without waterlogging monitors cannot be recorded. The two-dimensional hydrodynamic model with partial differential equations can accurately reflect spatio-temporal variations of inundation depths, which is widely used for waterlogging simulations under extreme rainfalls [32], and thus it is utilized to generate waterlogging samples under historical rainfalls for increasing the historical sample size and fill in the data gaps in the areas with missing data. Because the terrains of installation locations of DPVSs are relatively flat, this paper considers water is balanced by the forces in the vertical direction and neglects the effects of wind, Coriolis force, and it is assumed that water is only subject to the pressure of the surrounding water and the frictional resistance of the ground. The above assumptions can significantly reduce computation time while maintaining computational accuracy. The modified hydrodynamic model with two-dimensional partial differential equation for DPVS waterlogging is formulated as (5)-(7).

$$(1-\theta)\frac{\partial d_{i,j}^t}{\partial t} + \frac{\partial(\lambda u d_{i,j}^t)}{\partial x} + \frac{\partial(\lambda v d_{i,j}^t)}{\partial y} = q_{i,j}^t \tag{5}$$

$$-\frac{\partial h_{i,j}}{\partial x} = \frac{k^2 n^2 u\sqrt{u^2+v^2}}{(d_{i,j}^t)^{4/3}} + \frac{u q_{i,j}^t}{g d_{i,j}^t} \tag{6}$$

$$-\frac{\partial h_{i,j}}{\partial y} = \frac{k^2 n^2 v\sqrt{u^2+v^2}}{(d_{i,j}^t)^{4/3}} + \frac{v q_{i,j}^t}{g d_{i,j}^t} \tag{7}$$

where $d_{i,j}^t$ refers to the inundation depth of DPVSs located on grid (i,j) at time t; $h_{i,j}$ denotes the water surface elevation, and satisfies $h_{i,j}$=$g_{i,j}$+$d_{i,j}$; k is the unit conversion factor; m is the Manning coefficient; g is the gravitational acceleration; w denotes the water surface elevation; θ is the coverage of buildings within grids; λ is the connection ratio of two neighboring grids, which can be approximated as $1-\sqrt{\overline{\theta}}$. $\overline{\theta}$ is the average coverage of buildings within two grids. The finite difference method is used to solve two-dimensional partial differential equations. For convenience, it is assumed that the water can only flow between grids connected across the edges, as shown in Fig. 1.

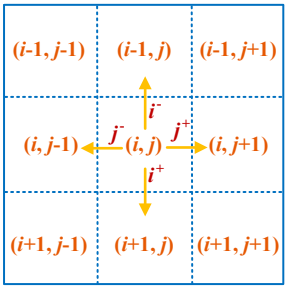


Fig. 1 Schematic diagram of the flood flowing direction

Take water flow from grid (i,j) to grid $(i,j+1)$ as an example, $\partial h_{i,j}/\partial x$ is approximated as $(h_{i,j+1} - h_{i,j})/\Delta x$, while $\partial h_{i,j}/\partial y$ is approximated as $\frac{1}{2}[(h_{i,j-1} - h_{i,j+1})/2\Delta y + (h_{i-1,j+1} - h_{i+1,j+1})/2\Delta y]$, where Δx and Δy denote the length and width of the grid, respectively. Then, water flow Q_{i,j^+} from grid

(i,j) to grid $(i,j+1)$ is formulated as (8-9), and water flow $Q_{i^+,j}$ from grid (i,j) to grid $(i+1,j)$ is derived as (10-11).

$$Q_{i,j^+} = \lambda \Delta y \Delta t d_{i,j}^t \frac{q_{i,j}^t / dg - \sqrt{{q_{i,j}^t}^2 / g^2 {d_{i,j}^t}^2 - |\nabla h_{i,j^+}|}}{2k^2 n^2 |\nabla h_{i,j^+}| {d_{i,j}^t}^{-4/3}} \frac{h_{i,j+1} - h_{i,j}}{\Delta x} \quad (8)$$

$$|\nabla h_{i,j^+}| = \sqrt{(\frac{h_{i,j+1} - h_{i,j}}{\Delta x})^2 + \frac{1}{4}(\frac{h_{i,j-1} - h_{i,j+1}}{2\Delta y} + \frac{h_{i-1,j+1} - h_{i+1,j+1}}{2\Delta y})^2} \quad (9)$$

$$Q_{i^+,j} = \lambda \Delta y \Delta t d_{i,j}^t \frac{q_{i,j}^t / dg - \sqrt{{q_{i,j}^t}^2 / g^2 {d_{i,j}^t}^2 - |\nabla h_{i^+,j}|}}{8k^2 n^2 |\nabla h_{i^+,j}| {d_{i,j}^t}^{-4/3}} \bullet \frac{(h_{i,j-1} - h_{i,j+1} + h_{i+1,j-1} - h_{i+1,j+1})}{\Delta y} \quad (10)$$

$$|\nabla h_{i^+,j}| = \sqrt{(\frac{h_{i+1,j} - h_{i,j}}{\Delta x})^2 + \frac{1}{4}(\frac{h_{i,j-1} - h_{i,j+1}}{2\Delta y} + \frac{h_{i+1,j-1} - h_{i+1,j+1}}{2\Delta y})^2} \quad (11)$$

Similarly, Q_{i,j^-} and $Q_{i^-,j}$ which respectively denote water flow from grid (i,j) to grid $(i,j\text{-}1)$ and grid $(i\text{-}1,j)$ can be derived by modifying (8-11), and the inundation depth of grid (i,j) is calculated by (12-13),

$$d_{i,j}^{t+1} = d_{i,j}^t + \frac{Q_{i,j^+} + Q_{i,j^-} + Q_{i^+,j} + Q_{i^-,j} - A_p}{\Delta x \Delta y} + q_{i,j}^t \Delta t \quad (12)$$

$$A_p = c\delta S_p \sqrt{2g d_{i,j}^t} \quad (13)$$

where A_p represents drainage capacity of grid (i,j); c denotes the number of drainage wells; δ is the drainage factor; S_p implies the cross-sectional area of drainage wells. Consequently, inundation depths of DPVSs can be deduced by the waterlogging distributions of the grids where they are located.

$$\mathcal{D}^n = \begin{bmatrix} \boldsymbol{d}_{1,1}^t & \boldsymbol{d}_{1,2}^t \cdots \boldsymbol{d}_{1,j}^t \\ \boldsymbol{d}_{2,1}^t & \boldsymbol{d}_{2,2}^t \cdots \boldsymbol{d}_{2,j}^t \\ \vdots & \vdots \ddots \vdots \\ \boldsymbol{d}_{i,1}^t & \boldsymbol{d}_{i,2}^t \cdots \boldsymbol{d}_{i,j}^t \end{bmatrix}, \boldsymbol{\beta}^n = \begin{bmatrix} \mathcal{D}^1 \\ \mathcal{D}^2 \\ \vdots \\ \mathcal{D}^n \end{bmatrix}^T \quad (14)$$

where $\boldsymbol{d}_{i,j}^t = [d_{i,j}^1, d_{i,j}^2, \ldots, d_{i,j}^t]$ represents the time-sequential inundation depths of the grid (x_i, x_j); $\mathcal{D}^n$ denotes the spatial distribution of inundation depth of the n-th extreme rainfall events; $\boldsymbol{\beta}^n$ is a set that combines actual measured waterlogging distributions and model-generated waterlogging distributions.

C. FFCM Algorithm Driven ESR Partitioning for DPVSs

Due to significant variations of rainfall intensity and waterlogging across different zones of the distribution network, DPVSs located in zones with heavier rainfalls and deeper inundation depths are more susceptible to triggering higher degrees of ESRs, thereby leading to rapid shutdowns owing to the high leakage currents. Treating all DPVSs in the distribution network as an entity to predict the ESRs of DPVSs would result in substantial errors. Thus, it is vital to partition DPVSs into multiple zones with different rainfall intensities and inundation depths to identify waterlogging-prone areas for improving ESR prediction accuracy.

On the basis of Section II-A and II-B, spatio-temporal characteristics of historical rainfalls are formulated as two matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{M}$ according to their distribution characteristics, and inundation depths of DPVSs under $n\text{-}th$ rainstorms are formulated as $\mathcal{D}^n$. As inundation depths do not follow a specific distribution, the average inundation depth $\boldsymbol{D}$ under historical rainstorm events is calculated to represent the distribution characteristics of waterlogging. Coordinating $\boldsymbol{\Sigma}$, $\boldsymbol{M}$, and $\boldsymbol{D}$, the temporal-spatial heterogeneity of extreme rainfalls and waterlogging for ESR partitioning can be characterized as follows,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \cdots \sigma_{1,j} \\ \sigma_{2,1} & \sigma_{2,2} \cdots \sigma_{2,j} \\ \vdots & \vdots \ddots \vdots \\ \sigma_{i,1} & \sigma_{i,2} \cdots \sigma_{i,j} \end{bmatrix}, \boldsymbol{M} = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} \cdots & \mu_{1,j} \\ \mu_{2,1} & \mu_{2,2} \cdots & \mu_{2,j} \\ \vdots & \vdots \ddots & \vdots \\ \mu_{i,1} & \mu_{i,2} & \cdots \mu_{i,j} \end{bmatrix},$$
$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots d_{1,j} \\ d_{2,1} & d_{2,2} \cdots & d_{2,j} \\ \vdots & \vdots & \ddots \vdots \\ d_{i,1} & d_{i,2} & \cdots d_{i,j} \end{bmatrix}, \boldsymbol{F} = \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{M} \\ \boldsymbol{D} \end{bmatrix} \quad (15)$$

where $\boldsymbol{F}$ is a feature matrix that unites $\boldsymbol{\Sigma}$, $\boldsymbol{M}$, and $\boldsymbol{D}$, and it can characterize the spatio-temporal volatility of extreme rainfalls as well as waterlogging triggering ESR variations of DPVSs.

$\boldsymbol{F}$ is used as the input for the ESR partitioning algorithm. It is worth noting that the feature matrix $\boldsymbol{F}$ contains three sub-matrices $\boldsymbol{\Sigma}$, $\boldsymbol{M}$, $\boldsymbol{D}$, and each element of the matrix can be regarded as a pixel point. Thus, $\boldsymbol{F}$ can be regarded as the pixel matrix of a color image with three channels of $\boldsymbol{\Sigma}$, $\boldsymbol{M}$, $\boldsymbol{D}$, which contains spatio-temporal characteristics of extreme rainfalls and waterlogging. Consequently, the image segmentation algorithm is implemented to subdivide the three-channel image $\boldsymbol{F}$ for partitioning distribution networks into multiple zones with different rainfalls and waterlogging depths. In the computer version field, superpixels are irregular blocks composed of neighboring pixels with similar features [33]. Superpixel algorithms are usually used to obtain adaptive neighboring information of an image for incorporating adaptive local spatial information and improving the segmentation effect. To generate superpixels, the image is projected into CIE Lab color space [34], and each pixel point is transferred to a 5-dimensional vector $V[L,a,b,x,y]$. Simple linear iterative clustering (SLIC) [35] is utilized to generate superpixels. Firstly, the initial clustering center is defined as,

$$V_i = [l_i, a_i, b_i, x_i, y_i]^T \quad (16)$$

where l_i is the brightness of the color space; a_i and b_i denote the chromaticity coordinates of clustering center i; x_i and y_i represent the spatial coordinates of clustering center i;

Then, a clustering process is performed by searching for all pixel points within a range of twice the step size around the cluster center. The distances between each pixel point and the cluster centroids are calculated as follows,

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \quad (17)$$

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \quad (18)$$

$$D = \sqrt{\left(\frac{d_c}{m}\right)^2 + \left(\frac{d_s}{S}\right)^2} \quad (19)$$

where d_c and d_s implies the chromaticity and spatial coordinates of pixel j and clustering center i, respectively; m denotes fixed factor; $S = \sqrt{p/n}$ is the length of the sides of the square grid, p is the number of pixels, n is the number of superpixels.

Subsequently, the FFCM clustering [36] algorithm driven image segmentation method is utilized to divide the color image constructed by multiple superpixels. Because the number of superpixels is far less than that of pixels in an image, it is faster to implement FFCM on the generated superpixels than original

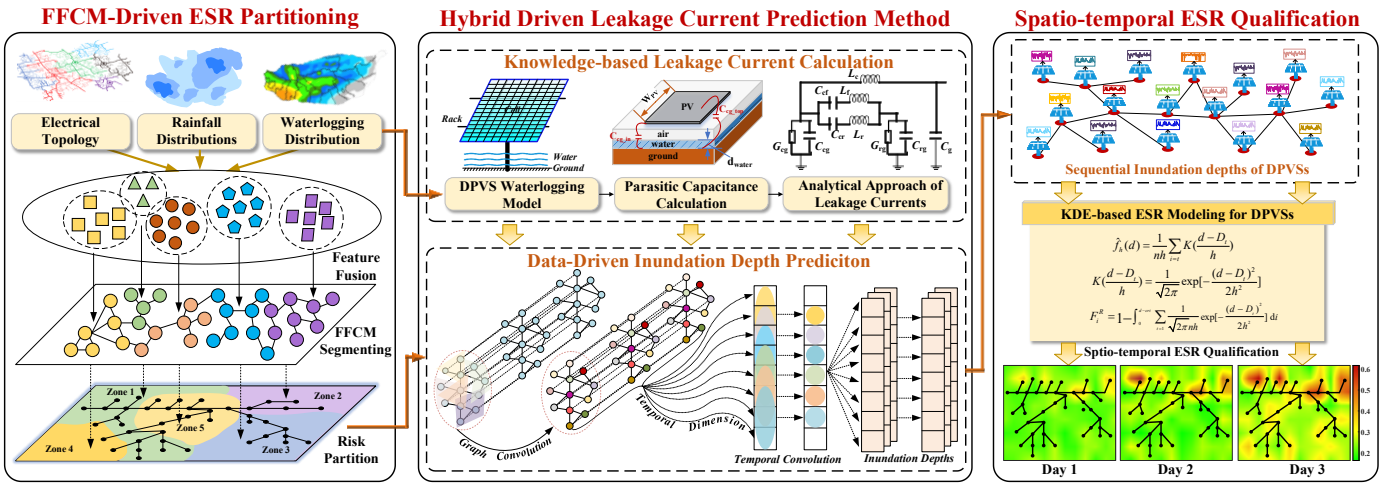


Fig. 2 The proposed hybrid data and knowledge driven ESR prediction methodology for DPVSs under extreme rainfalls

pixels for color image segmentation. The objective function of FFCM is defined as follows,

$$J_l = \min \sum_{i=1}^{n}\sum_{j=1}^{m} S_i u_{ji}^{l} \| ((1/S_i)\sum_{p\in\mathcal{R}_i} x_p) - c_j \|^2 \tag{20}$$

where n is the number of generated superpixels; i denotes the color level and subjects to $1 \le i \le n$; m is the number of clusters; l is the weighting exponent; u_{ji}^{l} represents the fuzzy membership matrix between the i-th superpixel and the j-th clustering center; S_i implies the number of pixels in the i-th superpixel R_i, and x_p is the color pixel within the i-th region of the superpixel image; c_j is the j-th clustering centroid.

By implementing the Lagrange multiplier method, the aforementioned optimization problem can be reformulated as an unconstrained optimization problem as (21),

$$L(u_{ji}, c_j, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} S_i u_{ji}^{l} \| ((1/S_i)\sum_{p\in\mathcal{R}_i} x_p) - c_j \|^2 - \lambda(\sum_{j=1}^{m} u_{ji} - 1) \tag{21}$$

where λ is the Lagrange multiplier, μ_{ji} and c_j can be derived by $\partial L(\mu_{ji}, c_j, \lambda)/\partial c_j = 0$ and $\partial L(\mu_{ji}, c_j, \lambda)/\partial \mu_{ji} = 0$, respectively. Solutions of μ_{ji} and c_j are shown in (22-23),

$$c_j = \sum_{i=1}^{n} u_{ji}^{l} \sum_{p\in\mathcal{R}_i} x_p \Big/ \sum_{i=1}^{n} S_i u_{ji}^{l} \tag{22}$$

$$u_{ji} = \| 1/S_i \sum_{p\in\mathcal{R}_i} x_p - c_j \|^{-2/(l-1)} \Big/ \sum_{t=1}^{m} \| 1/S_i \sum_{p\in\mathcal{R}_i} x_p - c_t \|^{-2/(l-1)} \tag{23}$$

A convergence coefficient η is set to assess the convergence condition of FFCM. When the membership partition matrix $\boldsymbol{U}$ composed of u_{ji} satisfies $\boldsymbol{U}^{l} - \boldsymbol{U}^{l-1} < \eta$. Then, the color image composed of pixels representing rainfalls and inundation depths is segmented into multiple sub-images. DPVSs located on the same sub-image are partitioned in a zone, and all DPVSs are partitioned into multiple zones with different rainfall and inundation depths triggering the different ESR degrees.

III. Hybrid Data and Knowledge Driven ESR Prediction Method

The proposed hybrid data and knowledge driven ESR prediction methodology for DPVSs under extreme rainfalls is shown in Fig. 2, which contains the following parts: 1) The study area for ESR prediction of DPVSs is extracted based on the electrical topology of the distribution network, other areas outside the distribution network are not taken into consideration; 2) The waterlogging model is implemented to deduce inundation depths within the distribution network, and the dataset for risk partitioning is constructed based on the fusion of actual limited measured inundation depth distributions and numerous model-generated inundation depth distributions; 3) The data-driven FFCM algorithm is utilized to divide the distribution network into multiple zones based on historical rainfalls and waterlogging; 4) Critical inundation depths inducing ESRs of DPVSs are derived from the mathematical expression of leakage currents concerning inundation depths; 5) The data-driven STGCN is implemented to predict waterlogging risks for the distribution network. 6) The KDE is utilized to fit the time-varying PDFs of inundation depths of DPVSs and dynamically quantify spatio-temporal ESR degrees triggering shutdowns of DPVSs under rainstorms. Parts I and II have been presented in Section I, and the remaining parts of the proposed methodology will be described in detail in the following Sections.

A. Knowledge-based Leakage Currents Calculation Model

Leakage currents are generated due to the parasitic capacitors between DPVSs and the ground. When a loop is formed between parasitic capacitors, the PV system, and the power grid, the common-mode voltage will create leakage currents in parasitic capacitors. The expression for leakage currents and parasitic capacitance is shown below,

$$i_{i,t}^{PV} = (C_{i,t}^{cf} + C_{i,t}^{cr} + C_{i,t}^{cg}) \frac{\mathrm{d}U_{i,t}^{PV}}{\mathrm{d}t} \tag{24}$$

where $i_{i,t}^{PV}$ imposes leakage currents of DPVS i at time t; $U_{i,t}^{PV}$ is the common-mode voltage across PV parasitic capacitors of DPVS i at time t; $C_{i,t}^{cf}$, $C_{i,t}^{cr}$, and $C_{i,t}^{cg}$ denote the parasitic capacitance between the PV cell and the frame, the PV cell and the rack, PV cell and the ground, respectively. The parasitic capacitors of a PV panel under waterlogging are shown in Fig. 3.

The common-mode voltage $U_{i,t}^{PV}$ across parasitic capacitors of the PV panel is related to the rated parameters, and less susceptible to the external environment. Conversely, parasitic capacitors are formed between the PV cell and the surrounding environment, which are vulnerable to extreme rainfall-inducing waterlogging, and the impedance in transformerless DPVSs is relatively small, making it easy to generate large leakage currents. Therefore, the parasitic capacitance of PV panels is utilized as a bridge to investigate the impacts of extreme rainfalls on the leakage currents of DPVSs. For instance, the parasitic capacitance $C_{i,t}^{cg}$ between the PV cell and the ground is calculated as (25-28).

$$C_{i,t}^{cg} = C_{i,t}^{cg-in} + C_{i,t}^{cg-top} \tag{25}$$

$$C_{i,t}^{cg-in} = \frac{\varepsilon_0 \varepsilon_{EVA} \varepsilon_{Tedlar} \varepsilon_{watter} \varepsilon_{air} W_{PV}}{[\varepsilon_{EVA}\varepsilon_{air}\varepsilon_{Tedlar} d_{i,j}^t + \varepsilon_{EVA}\varepsilon_{Tedlar}\varepsilon_{watter}(d_g \cdot \tan\beta_{PV} - d_{i,j}^t) + \varepsilon_{watter}\varepsilon_{air}\varepsilon_{Tedlar} T_{EVA}\varepsilon_{watter}\varepsilon_{air}\varepsilon_{EVA} T_{Tedlar})]} \tag{26}$$

$$C_{i,t}^{cg-top} = \frac{\varepsilon_0 \varepsilon_{\text{air//water}}}{\pi} \ln\left(1 + \frac{L_{PVef}}{T_{EVA} + T_{Tedlar} + d_g \cdot \tan\beta_{PV} + H_{PV}}\right) \tag{27}$$

$$\varepsilon_{\text{air//water}} = \frac{d_{i,j}^t \varepsilon_{water} + (T_{EVA} + T_{Tedlar} + d_g \cdot \tan\beta_{PV} + H_{PV} - d_{i,j}^t)\varepsilon_{air}}{T_{EVA} + T_{Tedlar} + d_g \cdot \tan\beta_{PV} + H_{PV}} \tag{28}$$

where $C_{i,t}^{cg-in}$ indicates the parasitic capacitance between the bottom of the PV panel and the ground; $C_{i,t}^{cg-top}$ denotes the parasitic capacitance between the top of the PV panel and the ground; W_{PV} is the width of the PV panel; H_{PV} is the width of the PV cell; d_g imposes the distance between the intersection of the PV panel and the extension of the PV panel to the ground; β_{PV} is the angle between the PV panel and the ground; L_{PVef} is the calculated length of frame capacitance; ε_0 is the absolute permittivity; ε_{EVA}, ε_{water}, ε_{air} imply relative permittivity of EVA layer, water, and air, respectively; T_{EVA} and T_{Tedlar} are the thickness of the EVA layer and Tedlar layer, respectively. Additionally, the calculation method for $C_{i,t}^{cf}$ and $C_{i,t}^{cr}$ can be referred to [5].

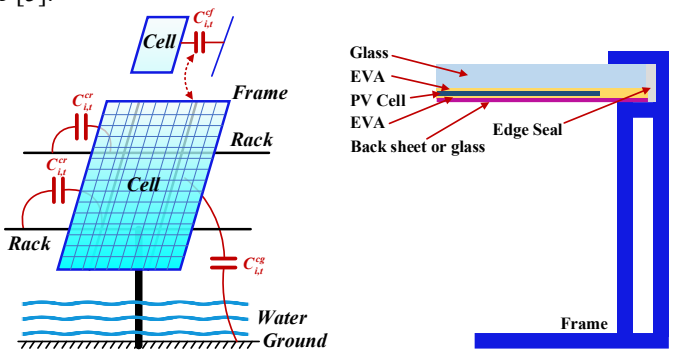


Fig. 3 Parasitic capacitors of a PV panel. (a) Schematic diagram of parasitic capacitors. (b) Two-dimensional schematic diagram of a PV panel.

Since PV arrays are arranged in several strings in a DPVS, parasitic capacitance cannot be revealed by a single equivalent capacitance model. The π-shaped equivalent circuit for PV arrays is utilized to calculate leakage currents of DPVSs, as illustrated in Fig. 4. This equivalent circuit not only considers the parasitic capacitance of the DPVS but also considers the panel equivalent inductance L_f, L_r, the equivalent conductor G_{cg}, G_{rg}, and cable parameters L_c, C_c. By calculating theoretical leakage currents of DPVSs under all extreme rainfall events occurring in the distribution network area, the dataset formed by spatio-temporal rainfall data and corresponding time-sequential leakage currents of DPVSs can be obtained for the training process of the data-driven STGCN.

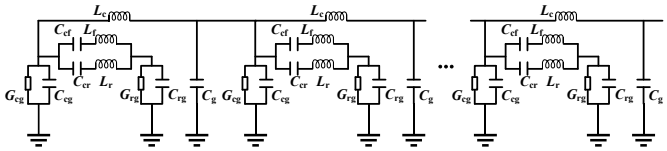


Fig. 4 Equivalent π-shaped circuit for a PV panel with parasitic capacitance

B. Data-Driven Waterlogging Risk Prediction Method

Knowledge-based inundation depth risk prediction models with clear physical mechanisms are prone to result in remarkable errors due to the uneven spatial distribution and temporal variability of extreme rainfalls. In contrast, data-driven methods can mine the relationship between extreme rainfalls and the time-sequential inundation depths from historical data to simplify the prediction process, thereby providing capabilities to handle models' uncertainties and enhancing computational efficiency. Thus, a data-driven STGCN [37] is implemented to predict waterlogging risks of distribution networks, which can convert input data to an advanced graphical structure and has the capability to learn shared representations across nodes in the graph, allowing it to transfer knowledge from observed samples to unseen ones. This transferability property enables GNNs to make accurate predictions and generalize well in application scenarios with limited training datasets [38]. Assuming meshed grids of distribution networks as nodes and correlations between nodes as edges, each partitioned region can be represented as a spatio-temporal graph as $G =< V, E >$. $V = \{v_i\}$. $E = \{e_{i,j}\}$. V is the node set consisting of meshed grids. E is the edge set, representing the correlation between node v_i and v_j. v_i and v_j can only be connected to each other if they exhibit a high correlation, and the connection relationship between nodes can be represented as an adjacency matrix $A \in \mathbb{R}^{N\times N}$, which is calculated as follows,

$$A_{ij} = \begin{cases} \exp(-dist(v_i, v_j)), & dist(v_i, v_j) > k \\ 0, & dist(v_i, v_j) \le k \end{cases} \tag{29}$$

where A_{ij} denotes the correlation coefficient between v_i and v_j; $dist(v_i, v_j)$ represents the Euclidean distance between v_i and v_j; k is the threshold, inferring that two nodes with a small correlation coefficient are considered to be unconnected for reducing computational complexity.

The STGCN presented in this paper is composed of Chebyshev graph convolution networks (CGCNs) and one-dimensional temporal convolution networks (TCNs), where CGCNs are utilized to extract spatial features of extreme rainfalls and TCNs are implemented to learn the impacts of extreme rainfalls on time-varying inundation depths of distribution networks. In CGCNs, due to the non-Euclidean nature of graphs, it is not possible to perform convolution operations in the vertex domain. The graph features are firstly converted to the spectral domain, and the convolution operation is then performed. The Laplace matrix is defined as,

$$L = D - A = I_N - D^{-1/2} A D^{1/2} \tag{30}$$

where I_N represents the n-th order identity matrix; $D \in \mathbb{R}^{N\times N}$ is the degree matrix of the previously defined graph. The eigenvalue decomposition of L is as follows,

$$L = U \Lambda U^T \tag{31}$$

where U denotes the basis composed of eigenvectors; $\Lambda \in \mathbb{R}^{N\times N}$ is the diagonal matrix consisting of eigenvalues λ_i, $i \in [1, N]$.

The Fourier transform matrix $\hat{X}_g$ of graph features $X_g \in \mathbb{R}^{N\times 1}$ can be calculated as,

$$\hat{X}_g = U^T X_g \tag{32}$$

Analogous to the traditional function convolution, the graph convolution formula is defined as,

$$(X_g * g)_G = U((U^T g) \odot (U^T X_g)) = U g_\theta U^T X_g \tag{33}$$

where symbol "⊙" represents Hadamard product; "∗" denotes graph convolution; g is the convolution kernel; g_θ is the learnable convolutional kernel, which satisfies $g_\theta = U^T g$; $\left(X_g * g\right)_G$ implies graph convolution applied in graph G.

Due to the heavy computational burden in performing graph convolution, the Chebyshev polynomial is used to fit the convolution kernel to speed up computation, which is as follows,

$$g_\theta = g_\theta(\Lambda) \approx \sum_{h=0}^{h-1} \theta_h T_h(\tilde{\Lambda}) \tag{34}$$

$$T_h(x) = 2xT_{h-1}(x) - T_{h-2}(x) \tag{35}$$

where $\tilde{\Lambda} = 2\Lambda/\lambda_{\max} - I_N$, $\lambda_{\max}$ is the maximum eigenvalue of L; $T_k(\cdot)$ denotes the k-th order Chebyshev polynomial; θ_k represents the Chebyshev coefficient.

Consequently, the graph convolution can be reformulated as,

$$g_\theta * X_g = \sum_{k=0}^{k-1} \theta_k T_k(L) X_g \tag{36}$$

where $\tilde{L} = 2L/\lambda_{\max} - I_N = U\tilde{\Lambda}U^T$; $\left(U\tilde{\Lambda}U^T\right)^k = U\tilde{\Lambda}^k U^T$.

The TCN is designed as a dilated causal convolution followed by a gated linear unit (GLU) layer, which can extract temporal features for time-sequential data of different lengths. Given a one-dimensional time series $B \in R^T$ (T is the time step) as the input and a filter $f \in R^K$ (K is the kernel size), the dilated causal convolutional form of B and f can be represented as,

$$B * f(t) = \sum_{s=0}^{K-1} f(s) B(t - d_s s) \tag{37}$$

where symbol $*$ denotes the dilated causal convolution form; d_s implies the dilated factor; s denotes the index of the kernel.

Suppose P denotes the output of dilated causal convolution, and the output of GLU can be represented as,

$$h = \sigma_1(\Theta_1 * P + b) \odot \sigma_2(\Theta_2 * P + c) \tag{38}$$

where $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ denote sigmoid and tangent hyperbolic activation function, respectively; Θ_1, Θ_2, b, c are learnable parameters of STGCN.

Waterlogging risk prediction for distribution networks can be considered to be a multivariate time series prediction problem considering spatial dependence. Meshed grids are considered to be nodes while rainfall intensity and inundation depths are regarded as the input features and labels of the nodes, respectively. Then, connecting nodes with high correlations, the graph $G_c =< V_c, E_c >$ for waterlogging risk prediction can be formed. $V_c = \{q_{i,t}, i_{i,t}\}$, where $q_{i,t}$ and $d_{i,t}$ denote the rainfall intensity and inundation depth of grid i, which are both spatio-temporal sequential data; $E_c = \{w_{i,j}\}$ denotes the correlation between nodes i and j, which can be obtained by (29). The data-driven STGCN based model of waterlogging risk prediction for distribution networks can be represented as a learning function f mapping H steps historical rainfalls to K steps future inundation depths, as follows,

$$\{q_{i,t-H+1}, q_{i,t-H+2}, ..., q_{i,t}, G_C\} \xrightarrow{f} \{d_{i,t+1}, d_{i,t+2}, ..., d_{i,t+K}\} \tag{39}$$

where $\left\{q_{i,t-H+1}, q_{i,t-H+2}, \dots, q_{i,t}\right\} \in \mathbb{R}^{N\times H\times F}$ represents the H steps historical rainfalls of grid i; $\left\{d_{i,t+1}, d_{i,t+2}, \dots, d_{i,t+A}\right\} \in \mathbb{R}^{N\times A\times 1}$ denotes the K steps future inundation depths of grid i; N is the number of nodes; F is the feature dimension of nodes.

C. ESR Modelling for DPVSs under Extreme Rainfalls

It is obligated in international standards [6]-[7] of the operation and maintenance of DPVSs with peak leakage currents exceeding 300 mA must be rapidly disconnected. Otherwise, it may occur electric leakage accidents. Critical inundation depths $d\text{-}cri$ of DPVSs triggering electric leakages can be derived from the knowledge-based leakage current calculation model (24-28), and the ESR degrees of DVPSs can be qualified by the probabilities of inundation depths exceeding their critical values. Since the nonparametric KDE does not require any assumptions about the distribution of data, it is more flexible in dealing with uncertain inundation depth distributions under uneven spatial and time-varying features of rainfalls. Thus, the KDE is implemented to fit the PDFs of predicted inundation depths and quantify spatio-temporal ESR degrees of DPVSs under extreme rainfalls [39]. The PDF of inundation depths is estimated as follows,

$$\hat{f}_h(d) = \frac{1}{nh} \sum_{i=t} K\left(\frac{d - D_t}{h}\right) \tag{40}$$

$$K\left(\frac{d - D_t}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(d - D_t)^2}{2h^2}\right] \tag{41}$$

where $D_t = \{d_{i,t}, d_{i,t+1}, .., d_{i,t+L_p}\}$ represents the inundation depths from time slot t to $t + L_p$; L_p is the ESR warning rolling cycle; h denotes the bandwidth, determining the smoothness of fitted distributions; $K(\cdot)$ implies the kernel function, and the Gaussian kernel function is adopted in this paper. The bandwidth h has a significant impact on the quality of the KDE. Thus, the principle of minimizing the asymptotic mean integrated squared error (AMISE) is used for optimizing the bandwidth of the KDE.

Then, the cumulative distribution functions (CDFs) of inundation depths exceeding critical depths are formed to dynamically quantify spatio-temporal ESR degrees of DPVSs i in different partitioned zones. As inundation depths are time-varying with the evolution of extreme rainfalls, F_i^R is variable at different warning rolling cycles R.

$$F_i^R = 1 - \int_0^{d-cri} \sum_{t=1} \frac{1}{\sqrt{2\pi}nh} \exp\left[-\frac{(d - D_t)^2}{2h^2}\right] \mathrm{d}d \tag{42}$$

IV. Case Studies

A. Network Data

A typical distribution network, located in a hilly area, is introduced to verify the performance of the proposed hybrid data and knowledge driven ESR prediction method for DPVSs under extreme rainfalls. There are a number of DPVSs installed within the distribution network. As rooftop PVs are not susceptible to waterlogging, we only consider ground mounted DPVSs. Relevant parameters of DPVSs are obtained from the local power company. The spatial distribution of DPVSs within the distribution network area is illustrated in Fig. 5. Moreover, the specific parameters of a typical ground mounted DPVS are listed in Table I, and all data requirements and sources are summarized in Table II. The hourly historical rainfall data is collected from the Open-Meteo platform [40]. With its historical weather API, we have access to over 80 years of hourly rainfall data, covering any location on earth, all at a 10-kilometer resolution. It also provides forecasted rainfalls for up to 16 days. Terrain elevations of the distribution network area are obtained from shuttle radar topography mission version 3 (SRTM3) [41], the spatial resolution of which is 90 meters. Other geographic information is obtained from OpenStreetMap [42]. According to the classification criteria of rainfalls by the China Meteorological Administration (CMA), an amount of precipitation exceeding 30 mm within a 12-hour period or exceeding 50 mm within a 24-hour period is

considered a rainstorm event. Since 2004, a total of 73 rainstorm events have been observed in the studied area.

TABLE I
SPECIFIC PARAMETERS OF A TYPICAL GROUND MOUNTED DPVS

Parameters	Symbol	Value
Width of PV panels	W_{PV}	920mm
Thickness of EVA layer	T_{EVA}	0.5mm
Thickness of Tedlar layer	T_{Tedlar}	0.2mm
Thickness of PV cells	H_{PV}	0.2mm
Effective length of PV panels	L_{PVef}	30mm
Distance between PV panel and ground	d_{PV}	50mm
Installation angle	β_{PV}	40°
Absolute permittivity	ε_0	$8.85*10^{-12}$
Relative permittivity of water	ε_{water}	78.5
Relative permittivity of Tedlar layer	ε_{Tedlar}	2
Relative permittivity of EVA layer	ε_{EVA}	3
Relative permittivity of air	ε_{air}	1.0006

TABLE II
DATA REQUIREMENTS AND SOURCES FOR THE PROPOSED METHOD

Data Requirements	Data Sources
Hourly historical and forecasted rainfall data	Open-Meteo
Terrain elevations	SRTM
Geographic information	OpenStreetMap
Inundation depths of DPVSs	Waterlogging model
Specific parameters of DPVSs	Local power company

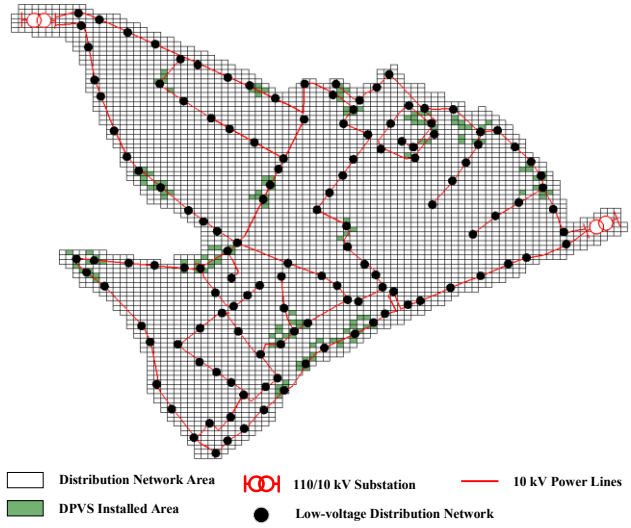
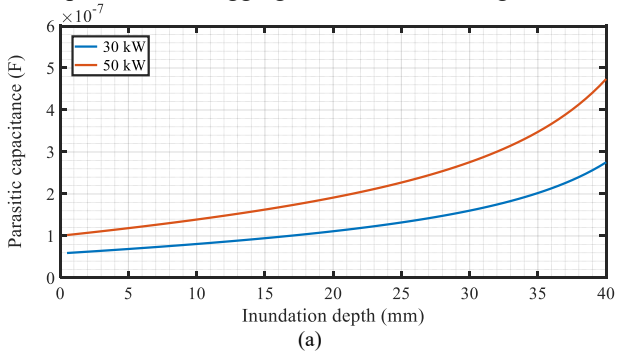


Fig. 5 Spatial distribution of DPVSs within the distribution network area

B. Analysis of Leakage Currents of DPVSs

According to the basis of Section III, leakage currents of DPVSs are significantly impacted by the parasitic capacitance. Under extreme rainfall conditions, the parasitic capacitance of DPVSs is influenced by the inundation depth. Taking DPVSs with capacities of 30 kW and 50 kW as examples, the variation patterns of parasitic capacitance and leakage currents with inundation depths of waterlogging are illustrated in Fig. 6.



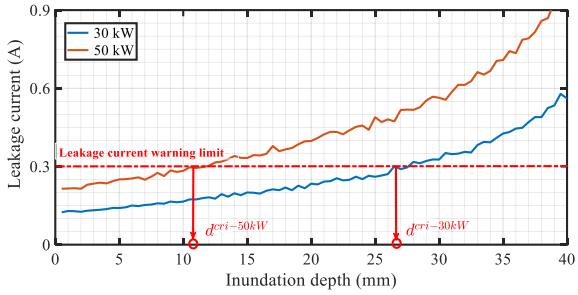


Fig. 6 Impacts of extreme rainfalls on electric leakages of DPVSs. (a) Relationship between parasitic capacitance and inundation depths. (b) Relationship between leakage currents and inundation depths

It can be observed from Fig. 6(a) that the parasitic capacitance of DPVSs tends to increase significantly as the inundation depth rises, and the parasitic capacitance of the DPVS with a capacity of 50 kW is bigger than that of the DPVS with a capacity of 30 kW. Due to leakage currents of transformerless DPVSs should be limited to less than 300 mA, 300 mA can be regarded as the threshold to determine whether electric leakages have occurred. As shown in Fig. 6(b), the red line is the leakage current warning limit. Inundation depths corresponding to the intersection point of the red line and the leakage current curve represent the critical inundation depths of DPVSs. $d^{cri-30kw}$ and $d^{cri-50kw}$ are critical inundation depths inducing electric leakages of DPVSs with capacities of 30 kW and 50 kW, respectively. Once inundation depths exceed the critical value, it will cause electric leakages.

C. Meteorological Interpolation and ESR Partitioning

Based on the historical rainfall data from four meteorological stations within the distribution network, the Co-Kriging method is implemented to interpolate the rainfall intensity across different zones of the distribution network. The results of the Co-Kriging method based meteorological interpolation for the precipitation of an extreme rainfall event within the distribution network area are shown in Fig. 7. It can be found from Fig. 7 that the Co-Kriging based meteorological interpolation method can fully reflect nonuniform spatial distributions and temporal volatility of the extreme rainfall disaster. Compared with the rainfall data from nearby meteorological observation stations, meteorological interpolation results with high spatial and temporal resolution can ensure the subsequent spatio-temporal ESR prediction accuracy of DPVSs under extreme rainfalls.

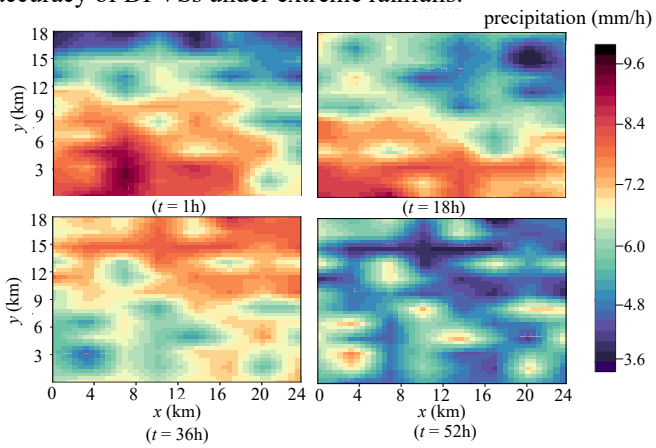


Fig. 7 Meteorological interpolation results for an extreme rainfall event

Based on meteorological interpolation results of historical rainfall intensity incorporating the distribution of calculated historical inundation depths, the FFCM driven partitioning algorithm is utilized to divide DPVSs into multiple zones with

different levels of waterlogging risks. Fig. 8 presents the ESR partitioning result of DPVSs within the distribution network. Different partitioned zones have different rainfall distributions and waterlogging risks that trigger variations of waterlogging degrees. For instance, in partitioned Zone 1, the average historical rainfall intensity is 5.6 mm/h, and the average inundation depth is 0.21 m. While in partitioned Zone 3, the average rainfall intensity is 2.1 mm/h, and the average inundation depth is 0.08 m. This indicates that DPVSs located in Zone 1 are more prone to waterlogging and electrical leakages, evidently exhibiting higher ESR degrees compared to DPVSs located in Zone 2.

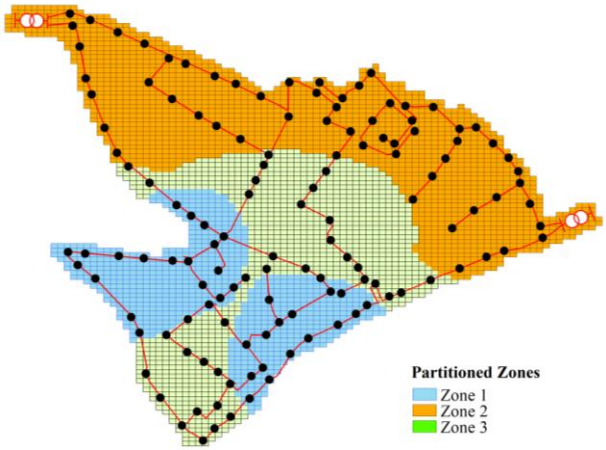


Fig. 8 ESR partitioning result of DPVSs within the distribution network

D. Spatio-Temporal ESR Prediction Results and Analysis

The dataset for the STGCN is composed of 73 rainstorm events within the distribution network area, and the sampling frequency is 1 hour. H is set to 6 and A is set to 1, which indicates that the last 6 steps of historical rainfall are used to predict the next 1 step inundation depths. The dataset is divided into the training set, validation set and test set in a ratio of 7: 2: 1, and normalized by the mean value and standard deviation of the training set, as follows,

$$\hat{x} = \frac{x - \mathrm{mean}(x_{train})}{\mathrm{std}(x_{train})} \tag{43}$$

where x is the original data of the dataset; $\hat{x}$ is the normalized data of the dataset; $\mathrm{mean}(x_{train})$ denotes calculating the mean value of the training set, $\mathrm{std}(x_{train})$ implies calculating the standard deviation of the training set. The training set is used to train the model while the validation set is used for parameter fine-tuning. The simulation is performed on Python 3.10.9 with PyTorch 2.1.0 and runs on a computing platform with an RTX 3070 GPU and 32 GB RAM.

To verify the superior capability of data-driven STGCN in extracting spatio-temporal features, several comparative baselines including random forest (RF), support vector machine regression (SVR), eXtreme gradient boosting (XGBoost), gated recurrent unit (GRU), long short-term memory (LSTM), and graph neural network (GNN) are introduced. Comparative studies are implemented in the case of a continuous extreme rainfall disaster lasting for three days in 2017. Fig. 9 illustrates predicted inundation depths of a certain DPVS which was rapidly disconnected due to high leakage currents during the heavy rain. The critical inundation depth $d\text{-}cri$ of the DPVS triggering electric leakages is 0.289 m. It can be found from Fig. 9 that inundation depths predicted by the STGCN are the closest to the real value. RF, SVR, and XGBoost are all traditional data-driven machine learning methods, which are incapable of learning the spatio-temporal correlations of inundation depths of different grids within the distribution network under the uneven spatial distributions and time-varying characteristics of extreme rainfalls. Compared with RF, SVR, and XGBoost, performances of GRU, LSTM, and GNN are feasible, owing to the capabilities of GRU and LSTM to cope with sequential data and GNN can learn the spatial feature of extreme rainfalls. However, neither GRU nor GNN can simultaneously learn the impacts of spatial and temporal characteristics of extreme rainfalls on inundation depths. Therefore, the STGCN surpasses GRU and GNN in terms of prediction accuracy.

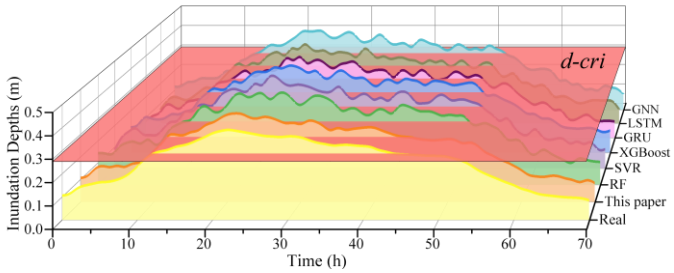


Fig. 9 Predicted inundation depths of a DPVS with different methods

Based on the predicted inundation depths of the DPVS, spatio-temporal ESR degrees can be quantified by KDE, and the warning rolling cycle is set to one day. Fig. 10 illustrates PDFs and CDFs of ESR of the DPVS under the three-day continuous rain. It can be observed from Fig. 10 that PDFs and CDFs change with the spatio-temporal variability in the evolution of extreme rainfalls as well as inundation depths of waterlogging over time.

To compare the accuracy of the optimal bandwidth with randomly selected bandwidths, the fitted PDFs under different bandwidths are evaluated using RMSE of the integral of the fitted probability density and the discrete probability of inundation depths of the DPVS within the i-th interval. The comparative results are listed in Table III. It can be seen from Table III that the RMSE of the optimal bandwidth is the smallest, which demonstrates the validity of the optimal bandwidth selection method. Furthermore, it can be deduced from CDF curves that the ESR degrees of the DPVS on Day 1 – Day 3 are 0.41, 0.79, and 0.08, respectively.

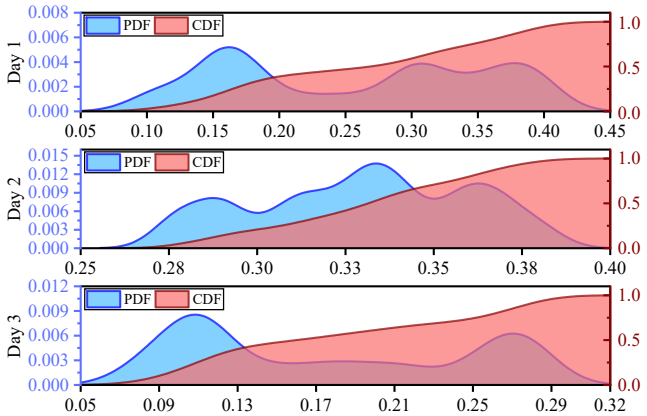


Fig. 10 PDFs and CDFs of time-varying ESRs of the DVPS

TABLE III
COMPARATIVE RESULTS OF ERRORS UNDER DIFFERENT BANDWIDTHS

Days	Bandwidths			
	Optimal	0.23	0.54	0.79
1	0.022	0.029	0.033	0.043
2	0.017	0.018	0.023	0.018
3	0.015	0.041	0.022	0.018

Moreover, spatio-temporal distribution of waterlogging risks

within the distribution network under the three-day continuous extreme rainfall event is illustrated in Fig. 11. Distribution network areas are categorized into three levels of waterlogging risks: 1) High-risk areas indicate that the maximum inundation depth on the day exceeds 0.3 m. Most DPVSs installed in these areas are prone to shutdowns due to waterlogging; 2) Medium-risk areas denote the maximum inundation depth on the day is greater than 0.1 m but less than 0.3 m. Some lower mounted DPVSs in these areas are prone to shutdowns due to waterlogging; 3) Low-risk areas represent the maximum inundation depth on the day does not exceed 0.1 m, and most DPVSs in these areas are in normal operational status.

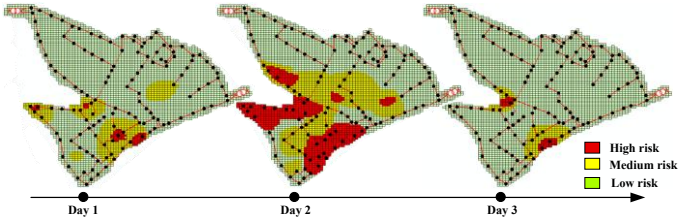


Fig. 11 Spatio-temporal waterlogging risks within the distribution network

According to the distribution of waterlogging risks, ESRs for all DPVSs can be quantified by the KDE. DPVSs with ESR probabilities exceeding 50% are considered to have high shutdown risks due to waterlogging. Four comparative methods are performed to further validate the superiority of the proposed risk prediction method for DVPSs. 1) Method I is the proposed method; 2) Method II neglects spatio-temporal characteristics of extreme rainfalls, assuming that the rainfall intensity is the same for each DPVS; 3) Method III neglects the ESR partitioning, where all DPVSs are regarded as an entity to perform spatio-temporal ESR prediction; 4) Method IV neglects both spatio-temporal characteristics of extreme rainfalls and ESR partitioning. DPVSs which are identified as high risks are compared with real shutdown events of DPVSs induced by their high leakage currents due to waterlogging. The comparative results are listed in Table IV. It can be found from Table IV that method I exhibits superior performance in all evaluation metrics, and both method II and method III are superior to method IV. It demonstrates the consideration of spatio-temporal characteristics of extreme rainfalls and risk partitioning is conducive to improving the ESR prediction accuracy of DPVSs under extreme rainfalls. After ESR partitioning, the data-driven model can better learn the characteristics of the partitioned area and make differentiated predictions. Compared with method III, method II is slightly better. This is because the precise spatio-temporal distribution of extreme rainfalls is the pivotal influence that directly affects leakage currents of DPVSs.

TABLE IV
COMPARATIVE ANALYSIS OF PREDICTION ACCURACY OF DIFFERENT METHODS

	Method	Accuracy (%)	Precise (%)	Recall (%)	F1-score (%)
Day 1	I	**98.14**	**92.31**	**85.71**	**88.89**
	II	96.89	84.62	78.57	81.48
	III	96.27	78.57	78.57	78.57
	IV	95.65	76.92	71.43	74.07
Day 2	I	**94.41**	**86.49**	**88.89**	**87.67**
	II	92.55	83.33	83.33	83.33
	III	91.93	82.86	80.56	81.69
	IV	90.68	78.38	80.56	79.45
Day 3	I	**98.76**	**87.50**	**87.50**	**87.50**
	II	98.14	85.71	75.00	80.00
	III	98.14	85.71	75.00	80.00
	IV	96.89	71.43	62.50	66.67

V. CONCLUSION

In this paper, a hybrid knowledge-based and data-driven ESR prediction method is proposed to identify high-risk areas induced by electric leakages of DPVSs in facing the uneven spatial distributions and time-varying characteristics of extreme rainfalls. The following are the key findings of this study: 1) With the consideration of the impacts of inundation depths on the parasitic capacitance calculation model, the knowledge-based leakage current calculation model can theoretically analyze variation patterns of leakage currents of DPVSs under waterlogging conditions; 2) The STGCN can effectively capture the nonlinear relationship between the dynamic evolution of extreme rainfalls and the inundation depths within distribution networks; 3) Spatio-temporal characteristics of extreme rainfalls and risk partitioning is conducive to improving the ESR prediction accuracy of DPVSs under extreme rainfalls. Compared with the ESR prediction methods without the consideration of spatio-temporal characteristics of extreme rainfalls and ESR partitioning, the proposed can increase accuracy, precision, recall, and F1-score of ESR prediction up to 3.73%, 16.07%, 25.00%, and 20.83%; The case study results have demonstrated the superior effectiveness and applicability of the proposed hybrid data and knowledge driven method for spatio-temporal ESR prediction of DPVSs under extreme rainfalls.

REFERENCES

[1] Y. Cao, B. Zhou, and C. Y. Chung, *et al.*, "Dynamic modelling and mutual coordination of electricity and watershed networks for spatio-temporal operational flexibility enhancement under rainy climates," *IEEE Trans. Smart Grid*, vol. 14, no. 5, pp. 3450-3464, Sep. 2023.

[2] Y. K. Wu, Y. C. Chen, and H. L. Chang, *et al.*, "The effect of decision analysis on power system resilience and economic value during a severe weather event," *IEEE Trans. Ind. Appl.*, vol. 58, no. 2, pp. 1685-1695, March-April 2022.

[3] E. Galvan, P. Mandal, and Y. Sang, "Networked microgrids with roof-top solar PV and battery energy storage to improve distribution grids resilience to natural disasters", *Int. J. Electr. Power Energy Syst.*, vol. 123, p. 106239, Dec. 2020.

[4] "Snapshot 2023", IEA-PVPS. Accessed: Apr. 02, 2024. [Online]. Available: https://iea-pvps.org/snapshot-reports/snapshot-2023/

[5] W. Chen, X. Yang, and W. Zhang, *et al.*, "Leakage current calculation for PV inverter system based on a parasitic capacitor model," *IEEE Trans. Power Electron*., vol. 31, no. 12, pp. 8205-8217, Dec. 2016.

[6] *Safety of Power Converters for Use in Photovoltaic Power Systems - Part 2: Particular Requirements for Inverters*, IEC 62109-2, 2011.

[7] *Automatic Disconnection Device between a Generator and the Public Low-Voltage Grid*, DIN VDE V 0126-1-1, 2006.

[8] B. R. Prusty and D. Jena, "An over-limit risk assessment of PV integrated power system using probabilistic load flow based on multi-time instant uncertainty modeling", Renew. Energy, vol. 116, pp. 367–383, Feb. 2018.

[9] X. Fu, X. Wu, and C. Zhang, *et al*., "Planning of distributed renewable energy systems under uncertainty based on statistical machine learning," *Prot. Control. Mod. Power Syst*., vol. 7, no. 41, pp. 1-27, Oct. 2022.

[10] S. T. Hsu and T. C. Wu, "Simulated wind action on photovoltaic module by non-uniform dynamic mechanical load and mean extended wind load", *Energy Procedia*, vol. 130, pp. 94–101, Sep. 2017.

[11] C. A. J. Pantua, J. K. Calautit, and Y. Wu, "Sustainability and structural resilience of building integrated photovoltaics subjected to typhoon strength winds", *Appl. Energy*, vol. 301, p. 117437, Nov. 2021.

[12] P. Deng, M. Zhang, and Q. Hu, *et al.*, "Pattern of spatio-temporal variability of extreme precipitation and flood-waterlogging process in Hanjiang River basin", *Atmospheric Res.*, vol. 276, p. 106258, Oct. 2022.

[13] I. Hetita, A. S. Zalhaf, and D. E. A. Mansour, *et al.*, "Accurate modeling of photovoltaic systems for studying the transient effects of lightning strikes", *Energy Rep*., vol. 8, pp. 429–438, Apr. 2022.

[14] V. Sharma, S. M. Aziz, and M. H. Haque, *et al.*, "Effects of high solar photovoltaic penetration on distribution feeders and the economic impact", Renew. Sustain. Energy Rev., vol. 131, p. 110021, Oct. 2020.

[15] J. Li, Y. Zhang, and H. Fang, *et al.*, "Risk evaluation of photovoltaic power systems: An improved failure mode and effect analysis under uncertainty", J. Clean. Prod., vol. 414, p. 137620, Aug. 2023.

[16] X. Shi, R. Qiu, and T. Mi, *et al.*, "Adversarial feature learning of online monitoring data for operational risk assessment in distribution networks", *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 975–985, Mar. 2020.

[17] G. Cao, G. Wei, and P. Li, *et al*., "Operational risk evaluation of active distribution networks considering cyber contingencies", *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 3849–3861, Jun. 2020.

[18] S. Poudel, A. Dubey, and A. Bose, "Risk-based probabilistic quantification of power distribution system operational resilience", *IEEE Syst. J.*, vol. 14, no. 3, pp. 3506–3517, Sep. 2020.

[19] M. R. Ramali, N. A. Ong, and M. S. Said, *et al.*, "A review on safety practices for firefighters during photovoltaic (PV) fire", *Fire Technol.*, vol. 59, no. 1, pp. 247–270, Jan. 2023.

[20] J. C. Hernández, P. G. Vidal, and A. Medina, "Characterization of the insulation and leakage currents of PV generators: Relevance for human safety", *Renew. Energy*, vol. 35, no. 3, pp. 593–601, Mar. 2010.

[21] J. Herp, M. H. Ramezani, and M. Bach-Andersen, *et al*., "Bayesian state prediction of wind turbine bearing failure", *Renew. Energy*, vol. 116, pp. 164–172, Feb. 2018.

[22] J. Wang, F. Gao, and Y. Zhou, *et al*., "Data sharing in energy systems", *Adv. Appl. Energy*, vol. 10, p. 100132, Jun. 2023.

[23] S. Guo, B. Zhou, and Y. Yang, *et al.*, "Multi-source ensemble learning with acoustic spectrum analysis for fault perception of direct-buried transformer substations," *IEEE Trans. Ind. Appl.*, vol. 59, no. 2, pp. 2340-2351, March-April 2023.

[24] G. Ruan, J. Wang, and H. Zhong, *et al.*, "Improving sample efficiency of deep learning models in electricity market," *IEEE Trans. Power Syst.*, vol. 38, no. 5, pp. 4761-4773, Sept. 2023.

[25] H. Hou, H. Geng, and Y. Huang, *et al*. "Damage Probability Assessment of Transmission Line-Tower System Under Typhoon Disaster, Based on Model-Driven and Data-Driven Views", *Energies*, vol. 12, no. 8, Jan. 2019.

[26] Y. Tan, Y. Chen, and Y. Li, *et al.*, "Linearizing power flow model: a hybrid physical model-driven and data-driven approach", *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2475–2478, May 2020.

[27] E. Zhang, M. Dao, G. E. Karniadakis, and S. Suresh, "Analyses of internal structures and defects in materials using physics-informed neural networks," Science Advances, vol. 8, no. 7, Feb. 2022.

[28] T. Hu, H. Ma, K. Liu, and H. Sun, "Lithium-Ion Battery Calendar Health Prognostics Based on Knowledge-Data-Driven Attention", *IEEE Trans. Ind. Electron.*, vol. 70, no. 1, pp. 407–417, Jan. 2023.

[29] Q. Peng, W. Li, and M. Fowler, *et al.*, "Battery calendar degradation trajectory prediction: Data-driven implementation and knowledge inspiration", *Energy*, vol. 294, p. 130849, May 2024.

[30] S. K. Adhikary, N. Muttil, and A. G. Yilmaz, "Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments", *Hydrol. Process*., vol. 31, no. 12, pp. 2143–2161, 2017.

[31] I. Manola, B. Hurk, and H. Moel, *et al.*, "Future extreme precipitation intensities based on a historic event", *Hydrol. Earth Syst. Sci.*, pp. 3777–3788, Jan. 2018.

[32] T. Cheng, Z. Xu, and S. Hong, *et al.*, "Flood risk zoning by using 2D hydrodynamic modeling: a case study in Jinan city", *Math. Probl. Eng.*, vol. 2017, p. e5659197, Oct. 2017.

[33] X. Xiao, Y. Zhou and Y. J. Gong, "Content-adaptive superpixel segmentation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2883-2896, June 2018.

[34] B. Ly, E. Dyer, and J. Feig, *et al.*, "Research techniques made simple: cutaneous colorimetry: a reliable technique for objective skin color measurement", *J. Invest. Dermatol.*, vol. 140, pp. 3-12, Jan. 2020.

[35] R. Achanta, A. Shaji, and K. Smith, *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[36] T. Lei, X. Jia, and Y. Zhang, *et al.*, "Superpixel-based fast fuzzy c-means clustering for color image segmentation", *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753–1766, Sep. 2019.

[37] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting", *Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.

[38] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, "Transferability Properties of Graph Neural Networks". *arXiv*, Aug. 07, 2023.

[39] B. W. Silverman. "Density estimation for statistics and data analysis." Chapman and Hall, London, UK, 1986.

[40] "Free Open-Source Weather API | Open-Meteo.com". Available: https://open-meteo.com/

[41] M. AG, "OpenStreetMap in Chinese", OpenMapTiles. Available: https://openmaptiles.org/languages/zh/

[42] N. Earth Science Data Systems, "SRTM | Earthdata". Available: https://www.earthdata.nasa.gov/sensors/srtm