Scalable Weather Data Reduction for Solar PV Analysis Using Graph-Based Approach

Srijani Mukherjee[1,2,*], Laurent Vuillon[2], Denys Dutykh[3], and Ioannis Tsanakas[1]

[1]Univ. Grenoble Alpes, CEA, Liten, 73375 Le Bourget du Lac, France
[2]Univ. Savoie Mont Blanc, CNRS, LAMA, Chambéry, 73000, France
[3]Mathematics Department, Khalifa University, Abu Dhabi, PO Box 127788, United Arab Emirates
[*]Corresponding author: srijani.mukherjee@univ-smb.fr

Abstract

Efficient analysis of solar photovoltaic (PV) system performance demands processing large-scale environmental data while preserving critical trends for energy prediction. This study proposes Graph-Oriented Information Fusion (GOIF), a novel data reduction framework that employs graph-based community detection to identify representative days for solar PV performance analysis. GOIF constructs a graph with days as nodes and Euclidean-based similarities as edges, integrating daily average irradiance and temperature to capture their combined impact on PV energy output. Using Louvain modularity, it clusters days into communities and applies PageRank to select one representative day per community. GOIF represents annual data using a few days with a 1.5% error in energy yield approximation versus 7.31% for k-means while improving cluster stability (measured by standard deviation) and reproducibility. This approach reduces computational complexity without sacrificing accuracy, achieving a robust representation of yearly PV performance.

This study establishes GOIF as a robust and efficient data reduction tool for PV performance analysis, enhancing computational efficiency and decision making. Future work could focus on refining GOIF's ability to optimize data storage and retrieval, further improving its utility for long-term solar energy applications.

Keywords: Graph-based clustering, Data fusion, Solar PV performance, K-nearest neighbors, PageRank algorithm, K-means clustering.

1. Introduction

Temperature and solar irradiance are critical factors influencing solar PV performance [1, 2]. Irradiance directly determines the electrical energy generated by PV panels, while temperature affects panel efficiency [3], with variations driven by cloud cover, seasonality, and geographic location [4]. Analyzing these factors over a full year generates terabytes of data, posing significant computational challenges for PV system studies [5]. Processing such large datasets is resource-intensive, time-consuming [6], and often unnecessary for capturing essential weather patterns, making data reduction a vital strategy for efficient analysis [7].

Reducing year-long temperature and irradiance data to a manageable subset of representative days can accelerate computations while preserving key insights into weather-driven PV behavior [8]. A full year's data may provide a broad overview but can obscure critical details within specific temporal segments, such as peak irradiance periods, extreme weather events, or seasonal shifts [9]. By selecting representative days that capture these variations, data reduction enables focused analysis with significantly less computational overhead [10], freeing up resources for other tasks like system optimization or trend analysis. Moreover, storing terabytes of data is costly and cumbersome [11], whereas reduced datasets are more manageable, cost-effective, and accessible for long-term use, especially for PV plants handling extensive historical records [12]. Existing data reduction techniques vary in effectiveness. Typical Meteorological Year (TMY) analysis [13] creates a statistical "average" year from historical data but may miss local microclimates. Extreme Value Analysis (EVA) focuses on maximum or minimum conditions to study system stress, often overlooking frequent "normal" weather patterns [14]. Clustering, however, offers a more comprehensive approach by grouping similar days to retain both typical and extreme weather patterns [15].

The selection of representative days is also a well-established strategy to reduce computational complexity in energy system modeling, particularly for systems with high shares of intermittent renewables such as solar PV [16]. Traditional approaches often rely on clustering algorithms, with k-means and k-medoids being among the most widely used due to their simplicity and effectiveness in capturing the main patterns of time series data [17].

However, k-means' assumption of spherical clusters may not suit complex weather data, which often exhibits irregular and highly variable structures [18]. More recent studies have explored optimization-based methods [19] to improve the representativeness of selected days, balancing accuracy and computational cost [20]. Frameworks that incorporate meteorological features and hybrid techniques have also been proposed to better account for the variability and interdependencies inherent in renewable energy datasets [21] [22]. Despite these advances, existing methods primarily focus on distance-based or optimization-based selection, and few have addressed the underlying structure of temporal data using graph-based community detection [23].

In this paper, we introduce Graph-Oriented Information Fusion (GOIF), a novel data reduction method designed to minimize information loss in PV weather analysis. GOIF leverages temperature and irradiance data through a graph-based framework, representing days as nodes and weather similarities as edges. Using Louvain community detection, GOIF identifies distinct weather patterns and selects 10 representative days that preserve the dataset's diversity, ensuring both typical and extreme conditions are captured. This method groups days into communities based on shared weather characteristics, such as high summer irradiance or low winter temperatures, and uses PageRank to select the most central day per community, reflecting the core pattern of each cluster. This approach represents one year in 10 days—a 97% data reduction—while maintaining a low 1.5% error in energy yield approximation, as validated on the INES 2022 dataset (Section 4). GOIF's graph structure ensures comprehensive pattern capture, avoiding the oversimplification of TMY or EVA, and its mathematical framework guarantees robustness and reproducibility for large-scale PV data analysis, making it ideal for handling extensive weather datasets.

GOIF outperforms traditional methods like k-means in preserving weather information with minimal data. Over 100 runs without a fixed seed, GOIF consistently selects the same representative days with higher frequency, demonstrating superior stability (Table 3). It also achieves a lower intra-cluster standard deviation ($\bar{\sigma}_I$) than k-means, indicating tighter, more cohesive clusters that better represent weather patterns (Table 1). This improved clustering quality ensures that the selected days more accurately reflect the dataset's variability, from summer peaks to winter lows. GOIF also provides better temporal distribution, capturing diverse patterns across the year (Fig. 1), with a 1.5% error in energy yield approximation compared to k-means' 7.31%, random selection's 14.03%, and equipartition's 8.65% error (Table 4). This low error underscores GOIF's ability to retain critical weather information, enabling efficient computations for PV analysis tasks like performance

modeling, long-term trend studies, or system design optimization. By drastically reducing data volume, GOIF minimizes computational demands, making it a scalable solution for handling extensive PV weather datasets with high fidelity. To the best of our knowledge, the application of graph-oriented information fusion for representative day selection in solar PV performance analysis has not been previously reported. By leveraging graph communities, our approach captures complex relationships and temporal dependencies that conventional clustering and optimization methods may overlook, offering a significant advancement for accurate and interpretable PV modeling. This paper makes the following contributions to the field of solar PV weather data analysis:

1. **Novel Graph-Based Data Reduction Framework:** We introduce Graph-Oriented Information Fusion (GOIF), a method that leverages graph-based community detection and PageRank to select representative days, reducing annual weather data by over 97% while maintaining a low 1.5% error in PV energy yield approximation.

2. **Improved Clustering Stability and Reproducibility:** GOIF demonstrates superior stability over k-means, consistently selecting the same representative days across multiple runs, enhancing reliability for long-term PV performance analysis.

3. **Enhanced Weather Pattern Capture:** By modeling days as nodes and weather similarities as edges, GOIF captures diverse weather patterns (e.g., seasonal peaks and lows) more effectively than centroid-based methods like k-means, as validated by lower intra-cluster standard deviation and better temporal distribution.

4. **Scalable and Efficient Analysis:** GOIF reduces computational complexity, enabling efficient processing of large-scale PV weather datasets for applications like performance modeling and system optimization, with practical validation on the INES 2022 dataset.

The remainder of this paper is organized as follows: Section 2 provides a detailed methodology encompassing both the K-means algorithm and our proposed GOIF approach. Section 3 describes the sources of the data utilized in this study. Section 4 presents a comprehensive discussion of the results obtained from both approaches, accompanied by a qualitative analysis. Section 5 offers the conclusions drawn from our study and outlines potential directions for future research.

2. Methodology for GOIF

This section outlines a two-step methodology to identify representative weather patterns from year-long temperature (T) and irradiance (Q) data, encompassing our proposed Graph-Oriented Information Fusion (GOIF) with the k-means algorithm. Initially, we use a weather-only dataset, comprising 365 days of hourly T and Q measurements from a meteorological station, without any solar PV-specific data. In the first step, daily feature vectors are constructed by calculating statistical summaries (e.g., mean, standard deviation) of T and Q, followed by normalization and fusion into a unified representation. The second step applies clustering to select 10 representative days based on weather similarity. GOIF constructs a graph of days, uses the Louvain community detection to form clusters, and selects representatives via PageRank, while k-means employs centroid-based grouping.

2.1 Information Fusion

This subsection outlines the feature extraction and fusion process applied to a weather-only dataset of hourly temperature (T) and irradiance (Q) measurements over N days, sourced from a meteorological station, without PV energy data. For each day i, we derive statistical features from M hourly records to construct a feature vector $\mathbf{f}_i$, summarizing daily weather profiles. These features include mean (μ), standard deviation (σ), minimum (min), maximum (max), and quartiles (Q^1, Q^2, Q^3) for both T and Q.

The mean (μ) serves as a fundamental measure of the central tendency, representing the average value of all observations within a data set. For the i-th day with M observations $\{x_k\}_{k=1}^{M}$, the mean is calculated as:

$$\mu_x^i = \frac{\sum_{k=1}^{M} x_k}{M}.$$

This is calculated for both temperature and irradiance, resulting in $\{\mu_T^i, \mu_Q^i\}$ for each day.

The standard deviation (σ) quantifies the variability or dispersion of the data around the mean. A larger standard deviation indicates a wider spread within the dataset, while a smaller value suggests that the data points cluster closer to the mean. The standard deviation is calculated as the square root of the variance (σ^2), which is calculated as:

$$\left(\sigma_x^i\right)^2 = \frac{\sum_{j=1}^{M}(x_j - \mu_i)^2}{M}.$$

For temperature and irradiance, this is expressed as:

$$\sigma_T^i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(T_j^i - \mu_T^i)^2} \quad \text{and} \quad \sigma_Q^i = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(Q_j^i - \mu_Q^i)^2}.$$

The minimum (min) and maximum (max) values provide essential insight into the spread of the data and potential outliers. For the i-th day, we record $\min_T^i$, $\max_T^i$, $\min_Q^i$, and $\max_Q^i$, which represent the lowest and highest values for temperature and irradiance, respectively. These values are particularly useful for identifying extreme weather events or seasonal variations.

Quartiles divide the dataset into four equal parts, offering a robust summary of the data distribution. For the i-th day, we calculate the first, second and third quartiles for temperature and irradiance: $Q_T^{1,i}$, $Q_T^{2,i}$, $Q_T^{3,i}$ and $Q_Q^{1,i}$, $Q_Q^{2,i}$, $Q_Q^{3,i}$. The interquartile range (IQR), defined as the distance between Q^1 and Q^3, indicates the spread of the data around the median (Q^2). Data points outside this range are considered potential outliers.

Combining these features, we construct a high-dimensional feature vector $\mathbf{f}_i$ for the i-th day:

$$\mathbf{f}_i = \left[\mu_T^i, \sigma_T^i, \min_T^i, \max_T^i, Q_T^{1,i}, Q_T^{2,i}, Q_T^{3,i}, \mu_Q^i, \sigma_Q^i, \min_Q^i, \max_Q^i, Q_Q^{1,i}, Q_Q^{2,i}, Q_Q^{3,i}\right]$$

The initial dataset contained 14 meteorological features, but we reduced it to four key metrics—mean temperature (meanT), mean irradiance (meanQ), temperature standard deviation (stdT), and irradiance standard deviation (stdQ)—forming $\mathbf{F} \in \mathbb{R}^{N\times 4}$. Other features were excluded as they are derived metrics highly correlated with mean and std, introducing redundancy and multicollinearity, which could lead to overfitting and reduced clustering stability. Additionally, these derived statistics are less robust to outliers and noise in the dataset, potentially misrepresenting daily irradiation patterns under variable weather conditions. We reduce the feature vector to:

$$\mathbf{f}_i = [\mu_T^i, \sigma_T^i, \mu_Q^i, \sigma_Q^i] \in \mathbb{R}^4$$

The low-dimensional case was retained because these four features effectively capture the central tendency and variability of temperature and irradiance—critical drivers for solar PV performance—while maintaining computational efficiency and interpretability.

2.2 Feature Normalization

Following feature extraction, a feature matrix $\mathbf{F} \in \mathbb{R}^{N \times d}$ is constructed from all feature vectors $\mathbf{f}_i$, where N is the number of days and d is the number of features. To ensure that all features contribute equally to the clustering process, the feature matrix is normalized using five different normalization techniques—Standard, Min-Max, Robust, Maximum Absolute, and Quantile Transformer—tested for both GOIF and k-means, with optimal selection evaluated in Section 4. Without normalization, features with larger scales (e.g., temperature in °C) could dominate features with smaller scales (e.g., irradiance in $W\,m^{-2}$), skewing distance-based clustering. In Standard normalization, each feature j is scaled to zero mean and unit variance:

$$\mathbf{F'}_j = \frac{F_j - \mu_j}{\sigma_j}$$

where μ_j and σ_j are the mean and standard deviation of column j. Min-Max normalization maps features to [0, 1]:

$$\mathbf{F'}_j = \frac{F_j - \min(F_j)}{\max(F_j) - \min(F_j)}$$

Robust normalization uses median and interquartile range (IQR) to mitigate outliers:

$$\mathbf{F'}_j = \frac{F_j - \text{median}(F_j)}{\text{IQR}(F_j)}$$

Maximum Absolute normalization scales by the maximum absolute value, yielding [-1, 1]:

$$\mathbf{F'}_j = \frac{F_j}{\max(|F_j|)}$$

Quantile Transformer normalization enforces a uniform distribution:

$$\mathbf{F'}_j = Q_{\text{unif}}\left(Q_{F_j}(F_j)\right)$$

where Q_{F_j} is the empirical cumulative distribution function (CDF) and Q_{unif} is the uniform quantile function.

The normalized matrix $\mathbf{F'}$ ensures consistent feature influence, enabling accurate Euclidean distance calculations for GOIF's graph construction and k-means clustering in the subsequent step.

2.3 Weather Profile Clustering

Following feature normalization, this subsection outlines the clustering of days with similar weather profiles using the normalized feature matrix $\mathbf{F}' \in \mathbb{R}^{N\times 4}$. Two approaches are employed to group days based on their weather characteristics. The first utilizes the k-means algorithm, a partitioning method that assigns days to clusters by minimizing distances to iteratively updated centroids. The second implements Graph-Oriented Information Fusion (GOIF), a graph-based method that models days as nodes and pairwise Euclidean distances as edges, leveraging community detection to identify weather pattern clusters.

2.3.1 K-means Clustering

K-means clustering partitions the normalized feature matrix $\mathbf{F}' \in \mathbb{R}^{N\times 4}$ into $K = 10$ clusters to identify days with similar weather profiles [24]. This established method iteratively minimizes the within-cluster sum of squared Euclidean distances:

$$J(\mathbf{F}', S) = \sum_{k=1}^{K} \sum_{i\in S_k} \|\mathbf{f}'_i - \mu_k\|^2$$

where S_k is the k-th cluster, μ_k is its centroid, and $\|\cdot\|^2$ denotes squared Euclidean distance.

The process initializes K centroids $\{\mu_1, \mu_2, \ldots, \mu_K\}$ randomly from $\mathbf{F}'$. Each day i is assigned to cluster S_k with the nearest centroid:

$$S_k = \{i \mid k = \arg\min_{k'} \|\mathbf{f}'_i - \mu_{k'}\|^2\}$$

followed by updating each centroid as the mean of assigned points:

$$\mu_k = \frac{1}{|S_k|} \sum_{i\in S_k} \mathbf{f}'_i$$

where $|S_k|$ is the number of days in S_k. These steps repeat until convergence, defined by centroid shifts below a threshold $\epsilon = 10^{-4}$.

For each cluster S_k, a representative day $\mathbf{r}_k$ is selected as the point closest to the centroid:

$$\mathbf{r}_k = \arg\min_{i\in S_k} \|\mathbf{f}'_i - \mu_k\|^2$$

capturing the cluster's typical weather profile (mean and variability of temperature and

Algorithm 1: K-means Clustering for Weather Profiles

Input: $\mathbf{F}' \in \mathbb{R}^{N\times 4}$, number of clusters $K = 10$
Output: Centroids $\{\mu_k\}$, representatives $\{\mathbf{r}_k\}$

begin
  Randomly initialize centroids $\{\mu_k\}_{k=1}^{K}$ from $\mathbf{F}'$;
  repeat
    for *each* $\mathbf{f}'_i$ **do**
      Assign to cluster S_k where $k = \arg\min_{k'} \|\mathbf{f}'_i - \mu_{k'}\|^2$;
    for *each cluster* S_k **do**
      Update $\mu_k = \frac{1}{|S_k|} \sum_{i\in S_k} \mathbf{f}'_i$;
  until *centroid shift* $< \epsilon = 10^{-4}$;
  for *each cluster* S_k **do**
    Select $\mathbf{r}_k = \arg\min_{i\in S_k} \|\mathbf{f}'_i - \mu_k\|^2$;

irradiance). Algorithm 1 details the steps for this algorithm.

2.3.2 Graph-Oriented Information Fusion (GOIF)

Graph-Oriented Information Fusion (GOIF) is a novel clustering method designed to partition the normalized feature matrix $\mathbf{F}' \in \mathbb{R}^{N\times 4}$ into desired number of communities (present study utilizes 10 communities), each representing days with similar weather profiles based on temperature and irradiance features [25]. In contrast to k-means, which relies on centroid-based partitioning, GOIF employs a graph-based approach to model relational structures within the data. Here, days are represented as nodes in an undirected graph, and edges connect days with comparable weather characteristics, derived from $\mathbf{F}'$. Community detection then identifies clusters where intra-community connections exceed inter-community links, corresponding to distinct weather patterns such as prolonged sunny or overcast periods [26].

The GOIF process comprises three steps. First, a graph is constructed using the k-nearest neighbors (k-NN) method to establish edge connections based on Euclidean distances. Second, the Louvain algorithm maximizes modularity to detect 10 communities within the graph. Third, a representative day is selected from each community using PageRank for subsequent analysis. These steps are detailed below, with performance compared to k-means in Section 4.

Graph Construction for Community Detection

An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed to represent N days as nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, with edges $\mathcal{E}$ encoding weather profile similarity derived from $\mathbf{F}' \in \mathbb{R}^{N\times 4}$. For each node v_i, the k-nearest neighbors (k-NN) method [27] identifies the k days with the smallest Euclidean distances:

$$d_{i,j} = \|\mathbf{f}'_i - \mathbf{f}'_j\|_2$$

where $\mathbf{f}'_i$ is the feature vector of day i, and $\|\cdot\|_2$ denotes the L2 norm. Edges are assigned as:

$$\mathcal{E}_{i,j} = \begin{cases} 1 & \text{if } v_j \in \mathcal{N}_i(k) \text{ or } v_i \in \mathcal{N}_j(k) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{N}_i(k)$ is the set of k nearest neighbors of v_i. The graph is undirected, so $\mathcal{E}_{i,j} = \mathcal{E}_{j,i}$, and a node's degree may exceed k due to mutual neighbor selections. The parameter k is user-defined; in this study, we have used $k = 10$ to balance connectivity and sparsity.

Louvain Modularity Maximization

The Louvain Modularity Maximization algorithm [28] is a widely adopted and efficient method for community detection in large networks [29]. It detects communities in the graph $\mathcal{G}$ by maximizing modularity, partitioning N nodes into clusters reflecting distinct weather patterns. The algorithm begins by assigning each node to its own community. It then calculates the potential gain in modularity for each node if moved to a different community of its neighbors. Modularity Ω, a quantitative measure of community quality, is defined as

$$\Omega = \frac{1}{2M} \sum_{i,j} \left[A_{i,j} - \gamma \frac{k_i k_j}{2M} \right] \delta(c_i, c_j)$$

where M is the total edge count, $A_{i,j}$ is the adjacency matrix (1 if nodes i and j are connected, 0 otherwise), k_i is the degree of node i, $\gamma = 1$ is the resolution parameter, and $\delta(c_i, c_j)$ is 1 if nodes i and j share a community, 0 otherwise. The term within square brackets represents the difference between the actual number of edges between communities and the expected number based on the degree distribution. By maximizing Ω, the algorithm creates communities with high internal edge density and relatively few inter-community

edges, effectively partitioning the graph into clusters that reflect distinct weather patterns within the data. The algorithm relocates the node with the highest potential modularity gain to the community that maximizes this gain and repeats these steps iteratively until no further improvement in modularity can be achieved.

Selection of Representative Days

In our study, we used the *PageRank* algorithm to exploit the inherent structure of the constructed similarity graph $\mathcal{G}$. Within each of the 10 communities in $\mathcal{G}$, the PageRank algorithm [30] selects a representative day by analyzing the graph's connectivity. Adapted from directed graph applications, PageRank assigns each node v_i a score $PR(i)$ in the undirected similarity graph $\mathcal{G}$, identifying days central to their weather pattern community. Unlike centroids averaging features, PageRank emphasizes connectivity using random walk [31]. it ensures days linked to many others, like hubs in a network, score higher. The score $PR(i)$ is computed iteratively:

$$PR(i) = \frac{1-\alpha}{N} + \alpha \sum_{j \in \mathcal{N}_i} \frac{PR(j)}{\deg(j)}$$

where N is the total days, $\alpha = 0.85$ is the damping factor, $\mathcal{N}_i$ is v_i's neighbors, and $\deg(j)$ is v_j's degree. This splits into two parts: $(1-\alpha)/N$ gives every day a small base score (e.g., 0.15/N if $\alpha = 0.85$), ensuring all contribute slightly; the sum $\alpha \sum_j PR(j)/\deg(j)$ adds a share of each neighbor's score, scaled by their connections—busy neighbors boost v_i more. Iterating until stable, the highest $PR(i)$ per community marks the representative, capturing a key weather profile (e.g., a standout sunny day). Algorithm 2 details this process.

This method enhances selection by prioritizing relational importance over averages, reflecting dominant weather conditions like persistent heat or cloudiness. PageRank's k-NN-derived edges tie scores to weather similarity, offering interpretable centrality for analysis.

3. Dataset Description

This study employs two datasets to evaluate weather profile clustering and PV performance: a weather-only dataset and a PV-specific dataset. Both are available upon request,

Algorithm 2: Graph-Oriented Information Fusion (GOIF) Clustering

Input: $\mathbf{F}' \in \mathbb{R}^{N\times 4}$, normalized feature matrix; k, number of neighbors; $K = 10$, number of communities

Output: Representative days $\{\mathbf{r}_k\}$ for K communities

// Construct Undirected Graph;
Initialize $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$;
Initialize $\mathcal{E} = \emptyset$;
for *each* $v_i \in \mathcal{V}$ **do**
    Compute $d_{i,j} = \|\mathbf{f'}_i - \mathbf{f'}_j\|_2$ for all $v_j \in \mathcal{V}$;
    Set $\mathcal{N}_i(k)$ as k nodes with smallest $d_{i,j}$;
    for *each* $v_j \in \mathcal{N}_i(k)$ **do**
        $\mathcal{E}_{i,j} = \mathcal{E}_{j,i} = 1$;
Set $\mathcal{G} = (\mathcal{V}, \mathcal{E})$;
// Louvain Community Detection;
Assign each v_i to its own community $c_i = i$;
while *modularity* Ω *increases* **do**
    for *each* $v_i \in \mathcal{V}$ **do**
        Move v_i to neighbor's community maximizing Ω;
    if *no moves occur* **then**
        Break;
Merge into $K = 10$ communities;
// PageRank Selection;
Initialize $PR(i) = 1/N$ for all v_i;
while $\max_i |PR_{new}(i) - PR(i)| > \epsilon = 10^{-4}$ **do**
    for *each* $v_i \in \mathcal{V}$ **do**
        $PR_{\text{new}}(i) = \frac{1-\alpha}{N} + \alpha \sum_{j\in\mathcal{N}_i} \frac{PR(j)}{\deg(j)}$, where $\alpha = 0.85$;
    Update $PR(i) = PR_{\text{new}}(i)$;
for *each community* c_k, $k = 1$ *to* K **do**
    $\mathbf{r}_k = \arg\max_{i\in c_k} PR(i)$;

with the code accessible online.

Weather data is sourced from the French Ministry of Ecological Transition's RT-RE-bâtiment platform, which is compliant with 2012 thermal regulations for building energy performance. Available at RT-RE-bâtiment, it covers France's eight H-class climate zones (e.g., H1a), reflecting diverse conditions. Hourly air temperature (°C) and normal irradiance (W/m²) are extracted for 365 days, forming daily feature vectors $\mathbf{F'} \in \mathbb{R}^{N\times 4}$ as

described in Section 2.1.

PV data is collected from an 8-module string at the Institut National de l'Énergie Solaire (INES), Le Bourget-du-Lac, France (45.643958°N, 5.875885°E, 233 m elevation). Facing south with a 30° tilt, the string has a rated power of 2708.56 Wp. The dataset includes hourly average temperature (°C), irradiance (W/m²), and DC energy output (Wh) over one year, enabling validation of representative days against energy yield.

The complete codebase, implementing GOIF and k-means clustering, is publicly available at GitHub.

4. Results and Discussion

This section assesses the clustering performance of k-means and Graph-Oriented Information Fusion (GOIF) across five normalization techniques (Section 2.2), using the weather-only dataset and annual PV energy yield dataset (Section 3). Cluster quality is evaluated via the Average Intra-Cluster Standard Deviation ($\bar{\sigma}_I$), with representative day selection and stability analyzed to compare GOIF's effectiveness against k-means. A lower $\bar{\sigma}_I$ indicates tighter, more cohesive clusters. The metric is computed in three steps:

1. **Standard Deviation within Each Cluster (σ)**: For each cluster k, the standard deviation of both temperature, T_i, and irradiance, Q_i, is computed:

$$\sigma_{k,T} = \sqrt{\frac{\sum_{i=1}^{N_k}(T_{ij} - \overline{T_k})^2}{N_k - 1}}, \qquad \sigma_{k,Q} = \sqrt{\frac{\sum_{i=1}^{N_k}(Q_{ij} - \overline{Q_k})^2}{N_k - 1}}.$$

   Here, k denotes the cluster index, and N_k represents the number of days within cluster k. T_{ij} and Q_{ij} represent the temperature and irradiance values for a day i in the cluster k, respectively, while $\overline{T_k}$ and $\overline{Q_k}$ represent the mean temperature and mean irradiance within the cluster k, respectively.

2. **Intra-Cluster Standard Deviation (σ_I)**: The intra-cluster standard deviation for the cluster k is then calculated by averaging the standard deviations of temperature and irradiance:

$$\sigma_{I,k} = \frac{\sigma_{k,T} + \sigma_{k,Q}}{2}.$$

Method	Standard	Min-max	Robust	Maxabs	Quantile
K-means	0.316	0.072	0.212	0.074	0.081
GOIF	0.273	0.065	0.184	0.066	0.083

Table 1: Average intra-cluster standard deviation for different normalization techniques

3. **Average Intra-Cluster Standard Deviation ($\bar{\sigma}_I$)**: Finally, to determine the optimal normalization technique, we calculate the average intra-cluster standard deviation across all features and clusters:

$$\bar{\sigma}_I = \frac{\sum_{k=1}^{K} \sigma_{I,k}}{K},$$

where K represents the total number of clusters. A lower value of $\bar{\sigma}_I$ indicates tighter groups with smaller variations within each group, reflecting a more effective normalization technique for the community detection process.

For this study, the parameters for GOIF that we use are $K = 10$, $k = 10$ and $\gamma = 0.85$.

Impact of Normalization on σ_I

Our evaluation based on intra-cluster standard deviation in Table 1 reveals that Min-max normalization consistently achieves the lowest values in both the K-means and the GOIF approach. This finding suggests that scaling all features to a common range between 0 and 1 using Min-Max normalization may lead to more well-defined weather communities. During community detection algorithms, features with larger scales can dominate the distance calculations used to group similar days. Min-max normalization mitigates this effect by transforming both temperature and irradiance data to the same range, effectively giving them equal weight in the clustering process. This leads to a more balanced consideration of both weather aspects when forming weather communities, resulting in tighter clusters with lower values of σ_I.

Performance of K-means vs. Graph Community Detection

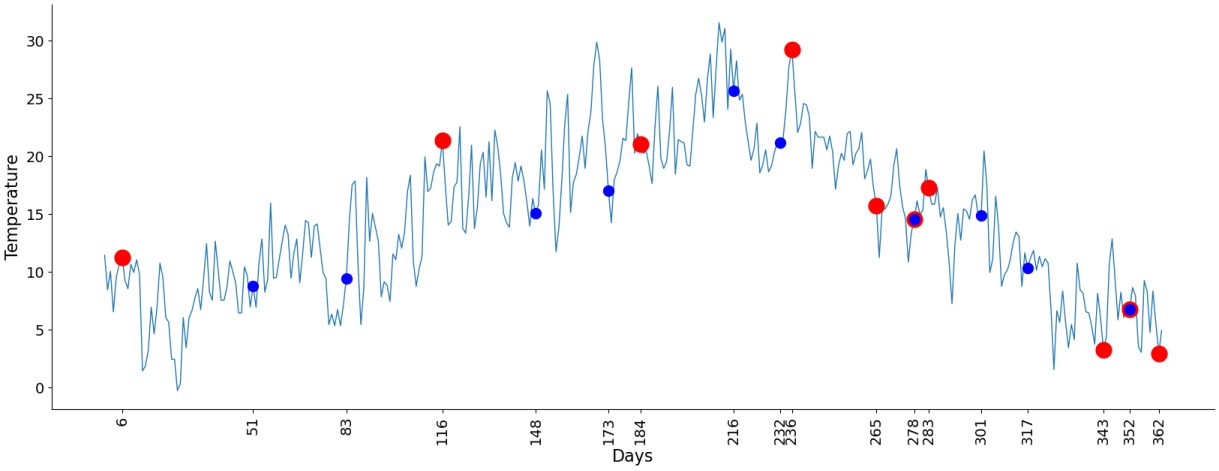
The results indicate that the graph-based community detection algorithm consistently produces lower σ_I values compared to K-means clustering, except for quantile normaliz-

ation, as shown in Table 1. This suggests that the graph-based approach, which takes advantage of the inherent structure of the data by representing days as nodes and similarities as edges, is better suited for capturing the underlying relationships between weather profiles. This can be attributed to the ability of graph-based methods to identify non-spherical clusters, which are more representative of the complex relationships present in the weather data. K-means, on the other hand, assume spherical clusters, potentially leading to suboptimal partitioning when dealing with more intricate data structures.

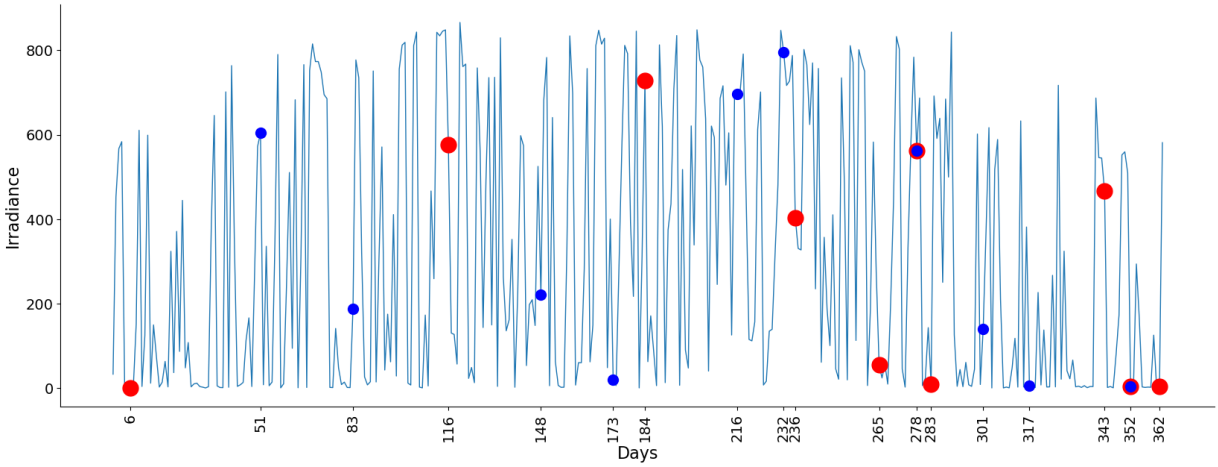
Effectiveness of GOIF for Selecting Representative Dates

Panels (a) and (b) of Figure 1 represent the distribution of temperature and irradiance for year-long weather data, respectively. The red dots represent the days chosen by K-means, followed by the centroid, and the blue dots represent the days chosen by our proposed GOIF method. The distribution of temperature and irradiance for representative days chosen by GOIF, marked with blue dots, exhibits a more distinct separation between clusters and better captures the seasonal trends compared to the centroid-based selection. Panels (a) and (b) of Figure 2 illustrate the daily irradiation curves for the 10 clusters identified by GOIF and k-means, respectively, using Min-Max normalization. Each subplot shows the average hourly irradiance (grey) for all days in a cluster, overlaid with the irradiance curve of the representative day selected by the respective method (blue for GOIF, red for k-means). Clusters are ordered by mean irradiance, from lowest (top-left) to highest (bottom-right). GOIF's representatives align more closely with cluster averages, capturing seasonal and daily variations effectively, with the GOIF irradiance pattern being particularly close to the average, especially for the low irradiance clusters. This enhanced alignment for low irradiance conditions is crucial, as it allows for more accurate representation of low performance by the solar panels, which is essential for assessing efficiency and energy yield under suboptimal weather conditions.

This can be explained by the nature of GOIF. Unlike the K-means centroid, which simply represents the average feature vector of a cluster, GOIF utilizes the concept of *PageRank* within the constructed graph to identify data points with higher connectivity and influence within their respective communities. These influential data points, chosen as representatives, have a higher likelihood of reflecting the core characteristics of their corresponding weather patterns.



(a)

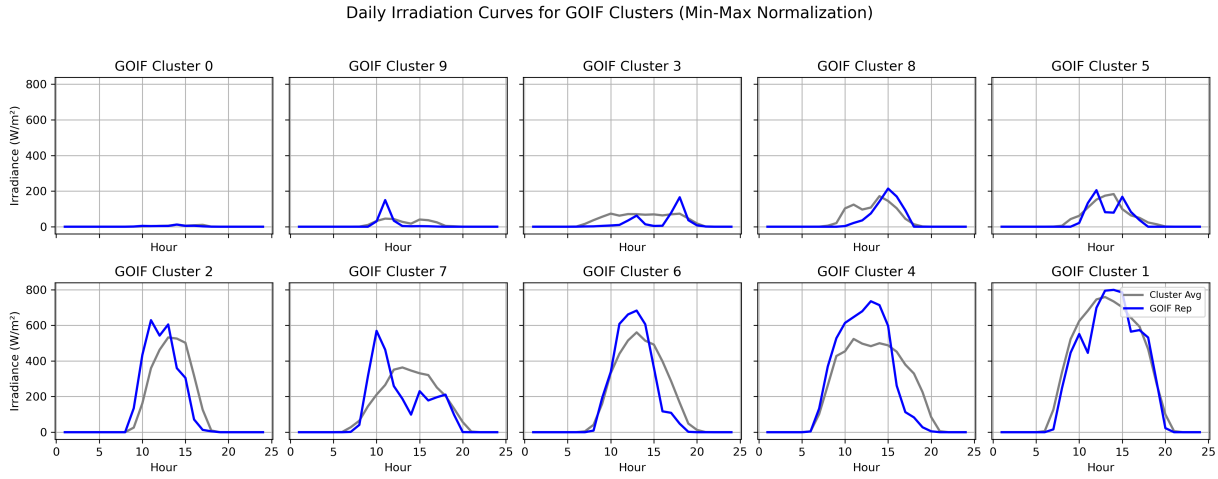


(b)

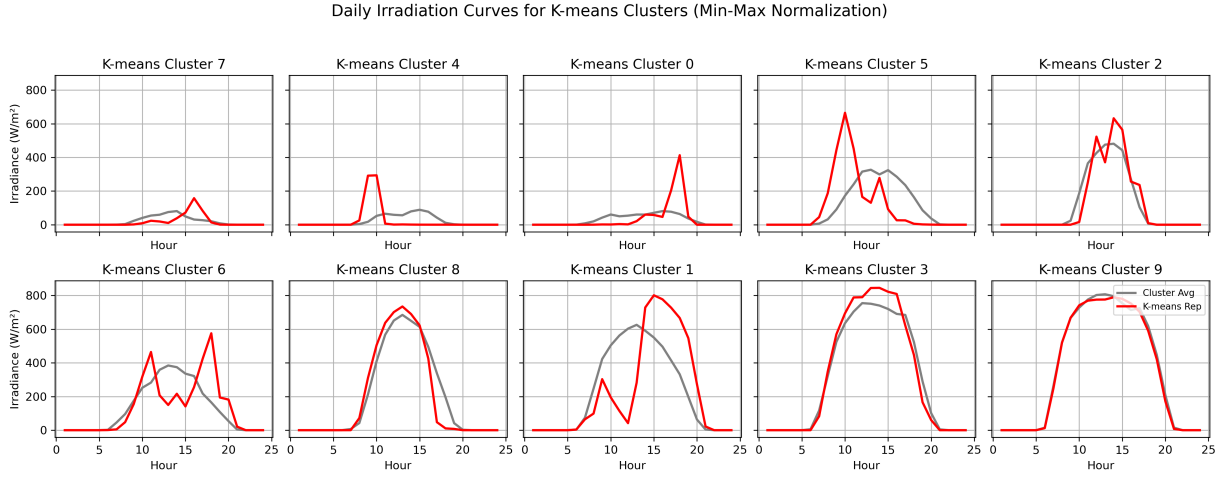
Figure 1: Distribution of Temperature (a) and Irradiance (b) for year-long data with the representative days marked by the dots. Red dots are chosen by K-means with centroid, and blue dots are chosen using our proposed GOIF method.

Stability and Reproducibility

To assess the stability and reproducibility of both methods, we ran each method 100 times on the same data set without any fixed seed. For each run, we recorded the chosen representative days. Subsequently, we identified the 10 most frequently chosen representative days for each method. The results of this analysis are presented in Table 2, which illustrates the frequency of selection for the top 10 representative days identified by the K-means and GOIF methods. Figure 3 illustrates our findings, with panel (a) showing



(a)



(b)

Figure 2: Daily irradiation curves for the 10 clusters identified by (a) GOIF and (b) k-means, using Min-Max normalization. Each subplot shows the average hourly irradiance (grey) and the representative day's curve (blue for GOIF, red for k-means), ordered by mean irradiance from lowest (top-left) to highest (bottom-right).

the frequency distribution of the 10 most chosen representative days using the K-means method and panel (b) showing the same for the GOIF method. The results reveal a notable difference in the consistency of day selection between the two methods. The GOIF method demonstrates a higher frequency of selecting the same days as representatives across multiple runs compared to the K-means method. This higher consistency in the GOIF method suggests a greater likelihood of reproducing similar results across numerous iterations. Table 3 presents the mean and standard deviation of the frequency with

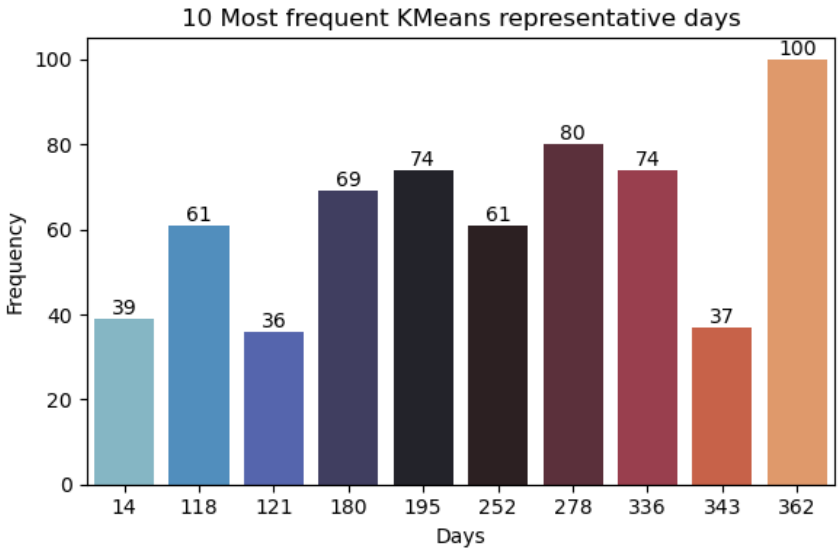
	Day	14	118	121	180	195	252	278	336	343	362
K-means	**Frequency**	39	61	36	69	74	61	80	74	37	100
	Day	63	130	165	232	247	301	310	338	347	355
GOIF	**Frequency**	97	51	100	33	72	99	74	94	86	52

Table 2: Frequency of K-means and GOIF representative days chosen over 100 runs

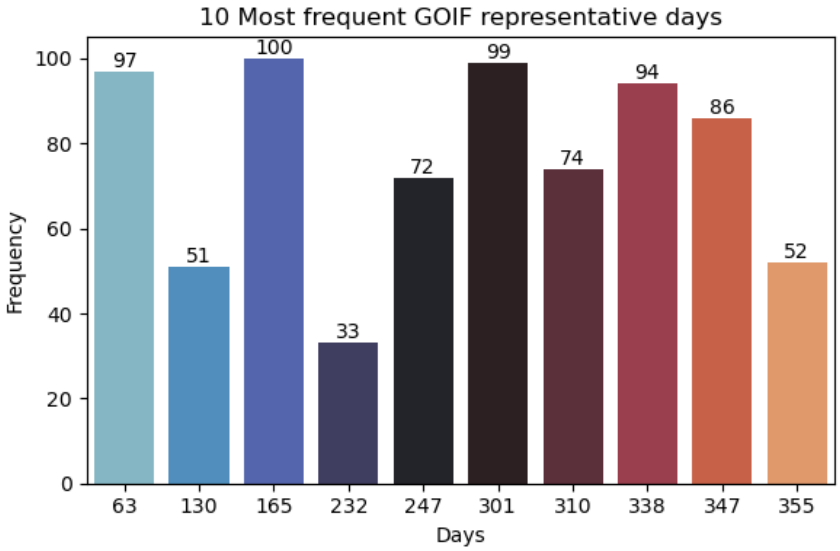
K-means	Mean	63.1
	STD	20.88
GOIF	Mean	75.8
	STD	23.66

Table 3: Mean and standard deviation (STD) of frequency of representative days

which representative days were selected by both the K-means clustering method and GOIF approach over 100 iterations. The temperature and irradiance distribution presented in Figure 1 and Figure 2 are derived from a single run of GOIF and k-means using Min-Max normalization, which yielded the lowest average intra-cluster standard deviation ($\bar{\sigma}_I$, Table 1). This run was selected to ensure optimal clustering performance. The stability of representative day selection across 100 runs without a fixed seed is separately evaluated in Tables 2 and 3. The results clearly indicate that the mean frequency of representative days chosen by the GOIF method is substantially higher compared to that of the K-means clustering algorithm. This suggests that our proposed GOIF approach is more likely to consistently identify the same representative days across different runs on the same dataset. In contrast, the K-means method exhibits greater variability in the selection of representative days, as evidenced by the lower mean frequency values. The enhanced reproducibility of the GOIF method can be attributed to its graph-based approach, which captures the inherent structure and relationships within the data more effectively than the centroid-based K-means algorithm. Using community detection algorithms along with *PageRank* measures, GOIF appears to identify more stable and representative patterns in the temporal data. This increased stability and reproducibility of the GOIF method have significant implications for long-term data analysis in solar PV applications. In contrast, the lower consistency observed in the K-means method highlights its sensitivity to initial centroid placement and the potential for converging to different local optima in different runs. While K-means remains a valuable clustering technique, its variability in selecting representative days may introduce uncertainties in long-term analyses. The superior reproducibility of the GOIF method underscores its potential as a more reliable approach for identifying



(a)



(b)

Figure 3: Frequency of ten most frequent representative days chosen by (a) K-means and (b) GOIF method.

representative days in weather data analysis for solar PV performance modeling.

Figure 4 illustrates the distribution of the graph community within the dataset of a year, employing various normalization techniques. In this representation, each node symbolizes a day of the year, with edges connecting these nodes. Nodes exhibiting similarity are placed close to each other, and the edges connecting them are correspondingly short. The

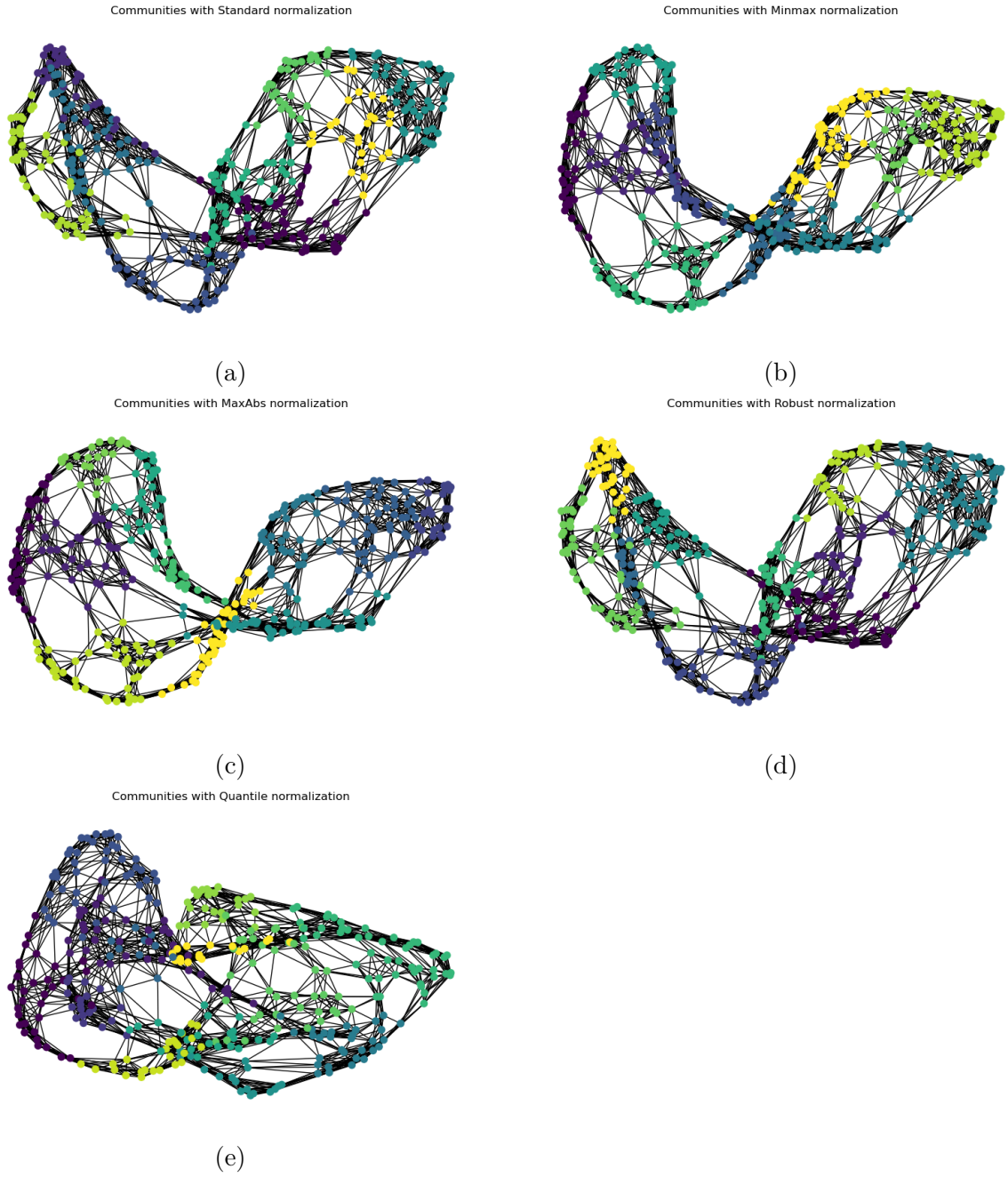


Figure 4: Community structure of annual weather data using different normalization techniques. Each node represents a day, with edges connecting similar days based on Euclidean distances. Colors denote distinct communities identified by the Louvain algorithm. Subfigures show: (a) Standard Normalization, (b) Min-Max Normalization, (c) Maximum Absolute Normalization, (d) Robust Normalization, (e) Quantile Transformer Normalization.

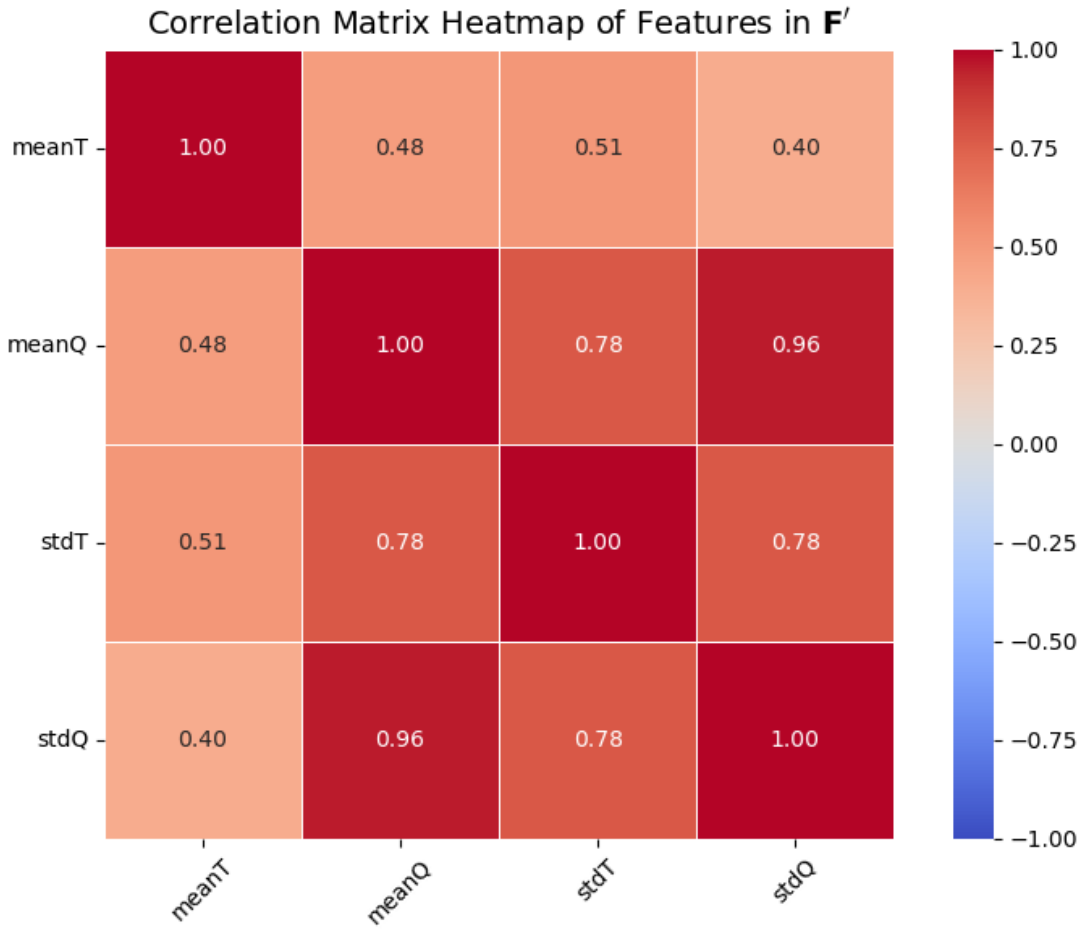


Figure 5: Pearson Correlation matrix heatmap of the features in $\mathbf{F}$

community structure within the data set is visually delimited by color coding. Nodes belonging to the same community are assigned identical colors, facilitating the identification of distinct groups or clusters of days that share similar characteristics. This color-based differentiation allows for immediate visual recognition of community boundaries and the overall community structure within the annual data. This graphical representation serves as a powerful tool for identifying temporal patterns, seasonal effects, and other cyclical phenomena within the year-long dataset. It provides a clear, intuitive visualization of the complex relationships and groupings present in the data, as well as valuable information on the underlying structure of the annual measurements.

Comparison of yearly energy yield

Method	E_{est} (kWh)	E_{act} (kWh)	Error (%)
GOIF	4108.90	4171.65	1.5
k-means	3867.06	4171.65	7.31
Random (avg.)	3586.35	4171.65	14.03
Equipartition	3810.92	4171.65	8.65

Table 4: Energy Yield Estimation Error (INES 2022)

Graph-Oriented Information Fusion (GOIF) is evaluated for solar PV yield estimation using 2022 data from an 8-module PV string (2708.56 Wp) at INES, Le Bourget-du-Lac, France (Section 3). The dataset provides hourly temperature, irradiance, and DC energy output (Wh), yielding an actual annual energy, $E_{\text{act}} = 4171.65$ kWh, after excluding faulty and missing data ($\approx 7 days$). Figure 5 presents the correlation matrix heatmap for the four features in $\mathbf{F}$ (mean temperature, mean irradiance, temperature standard deviation, and irradiance standard deviation), derived from the dataset. This heatmap shows the positive correlation between temperature and irradiance as a whole, highlighting their interdependent nature. This underscores the need to study these features simultaneously to accurately capture the meteorological dynamics influencing daily energy yield. GOIF ($K = 10$, $k = 10$) selects 10 representative days (indices: 18, 53, 103, 134, 195, 244, 246, 284, 313, 334) weighted by community sizes (w_i). K-means ($K = 10$) selects days (198, 53, 94, 176, 207, 216, 225, 280, 324, 347) with cluster sizes, while random selection (10 runs) and equipartition (every 36 days) use $w_i = 35.8$ (358/10). We utilized Min-Max normalization to produce optimized clusters as indicated in our previous study [32].

The estimated annual energy yield E_{est} is calculated as:

$$E_{\text{est}} = \sum_{i=1}^{10} w_i e_i$$

where e_i is the measured daily yield (Wh) of representative i, and $\sum w_i \approx 358$. For GOIF and k-means, w_i reflects the number of days in each community or cluster, varying based on weather similarity. Random selection picks 10 days per run (e.g., indices 5, 42, 89, etc.), repeating 10 times to average E_{est}, with each day's yield weighted by $w_i = 35.8$ to approximate the 358-day total. Equipartition selects days at fixed 36-day intervals (e.g., 0, 36, 72, ..., 324), assigning $w_i = 35.8$ to evenly distribute the year's span. The error quantifies accuracy:

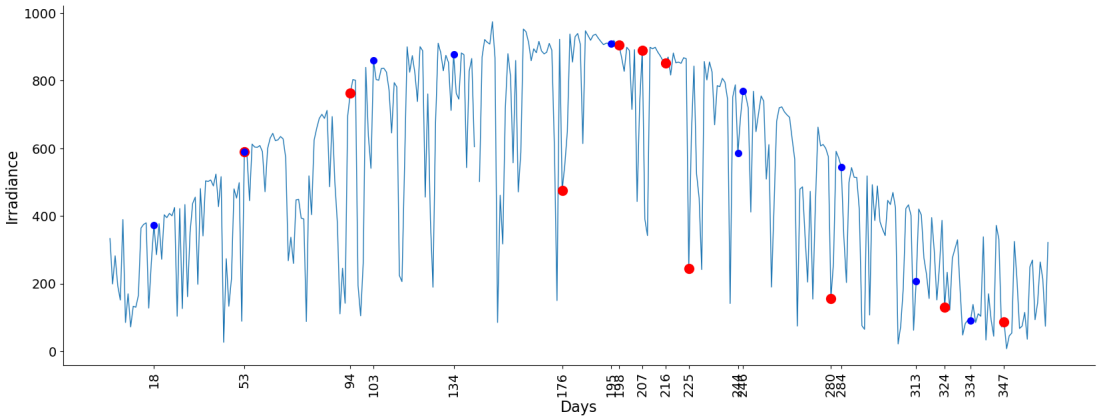


Figure 6: Irradiance distribution at 12 noon (INES 2022), with representative days chosen by GOIF (blue) and k-means (red).

Method	Cluster Sizes	MAE	RMSE
GOIF	[35, 25, 40, 42, 33, 40, 25, 46, 53, 19]	11.03	13.27
k-means	[16, 33, 72, 29, 60, 32, 50, 48, 9, 9]	14.35	18.70

Table 5: Error in representing whole year from representatives (INES 2022)

$$\text{Error} = \frac{|E_{\text{est}} - E_{\text{act}}|}{E_{\text{act}}} \times 100\%$$

Table 4 shows GOIF's $E_{\text{est}} = 4108.90$ kWh (1.5% error), outperforming k-means (3867.06 kWh, 7.31%), random (3586.35 kWh avg., 14.03%), and equipartition (3810.92 kWh, 8.65%). Fig. 6 plots the irradiance during solar noon (12-noon) for the whole year, with GOIF (blue) and k-means (red) representatives marked. GOIF's days better capture seasonal peaks (e.g., days near summer highs) and lows (e.g., days in winter), reflecting diverse weather patterns, while k-means misses key variations (some days with different irradiance dip). To strengthen the evaluation we calculated the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between the full year's energy yield from the INES 2022 dataset and the reconstructed year using representative days selected by GOIF and k-means. The reconstruction process assigned each day's energy yield to the yield of its nearest representative day, with cluster sizes varying due to the inherent partitioning of the methods: GOIF clusters had sizes of [35, 25, 40, 42, 33, 40, 25, 46, 53, 19] days, while k-means clusters exhibited sizes of [16, 33, 72, 29, 60, 32, 50, 48, 9, 9] days, reflecting their differing approaches to grouping. Table 5 shows for GOIF, the MAE was 11.03 kWh and RMSE was 13.27 kWh, indicating a close approximation to the full dataset. In contrast,

k-means showed a higher MAE of 14.35 kWh and RMSE of 18.70 kWh, suggesting greater deviation. These time-resolved error metrics complement the aggregated PV output errors, highlighting GOIF's superior ability to capture daily yield variations across the year. The lower errors for GOIF, particularly in its more balanced cluster sizes, underscore its effectiveness in representing the full dataset's temporal dynamics, especially under varying irradiance conditions, including low-irradiance scenarios where accurate panel performance assessment is critical. This demonstrates GOIF's efficiency as a data reduction tool. By representing the whole year in 10 days, GOIF reduces data volume over 97% while maintaining a low 1.5% yield error highlights the precision. For computational tasks like energy modeling, using GOIF's 10 days instead of a full year minimizes error while significantly scaling down data.

5. Conclusions and Future Research

This study advances weather pattern analysis for solar PV applications by developing Graph-Oriented Information Fusion (GOIF) as a data reduction method, evaluated on weather-only and PV-specific datasets (Section 3). Min-Max normalization proves optimal, scaling temperature and irradiance to [0, 1] to achieve the lowest intra-cluster standard deviation ($\bar{\sigma}_I$, Table 1). This ensures balanced feature representation, which is critical for effective clustering in PV data analysis where weather variability drives computational complexity.

A central finding is the enhanced capability of our proposed approach, the Graph-Oriented Information Fusion (GOIF) method, over traditional k-means clustering. GOIF significantly enhances data reduction efficiency compared to baseline models (such as k-means clustering). By grouping one-year data to 10 weather-similar communities and selecting high-connectivity representatives, GOIF reduces data volume by over 97% while maintaining a low 1.5% error in energy yield approximation (Table 4). It captures diverse weather patterns, including seasonal peaks and lows (Fig. 6), outperforming k-means (7.31% error), random selection (14.03%), and equipartition (8.65%). GOIF's stability over 100 runs (Table 3) further demonstrates its reliability in consistently selecting representative days, unlike k-means's centroid-driven variability. This efficiency enables scalable PV computations—e.g., using 10 days instead of a full year for energy modeling—without significant loss in accuracy, as validated by the INES 2022 dataset (Section 4).

These findings position GOIF as a robust framework for weather data reduction in

solar PV analysis. By compressing large datasets while preserving weather-driven energy insights, GOIF supports efficient computational workflows, reducing processing demands for tasks like performance modeling or system design. This data reduction approach enhances the scalability of PV data analysis, which is crucial for sustainable energy systems handling extensive weather and energy datasets.

Future Research Directions

Future work can extend GOIF's data reduction capabilities for broader PV applications. Applying it to multi-year, multi-variate data or climate zones could test its robustness across temporal and regional weather variations, ensuring consistent reduction accuracy. Incorporating additional features—cloud cover, humidity, or PV-specific metrics (e.g., efficiency)—might improve clustering granularity, further minimizing data while retaining critical patterns.

Integrating GOIF with machine learning (e.g., neural networks) could optimize representative selection, enhancing compression ratios for specific PV tasks like long-term trend analysis [33, 34]. Exploring alternative graph algorithms (e.g., Graph Neural Networks) or tuning parameters (k, γ) may refine reduction efficiency for varied datasets, such as coastal vs. inland PV sites. Embedding meteorological rules (e.g., seasonal thresholds) into GOIF could ensure that representatives align with physical phenomena, improving the interpretability of PV data workflows.

These advancements could establish GOIF as a versatile data reduction tool, enabling efficient analysis of large-scale PV datasets. By optimizing weather data compression, this research supports scalable, data-driven insights for solar energy systems, contributing to sustainable energy goals through enhanced computational efficiency.

6. Appendix: Mathematical Descriptions

This appendix provides detailed mathematical formulations of statistical measures and normalization techniques used in the analysis of PV weather data. These methods are foundational for clustering and data reduction, ensuring that temperature and irradiance features are appropriately scaled and interpreted.

Mean

The mean (μ) is a measure of central tendency, representing the average value of a dataset. For a dataset with M observations $\{x_k\}_{k=1}^{M}$, such as daily irradiance values, the mean is calculated as:

$$\mu = \frac{1}{M} \sum_{k=1}^{M} x_k.$$

In PV analysis, the mean of irradiance or temperature over a year provides a baseline for understanding typical weather conditions, aiding in the identification of representative days.

Standard Deviation

The standard deviation (σ) quantifies the variability of data around the mean, which is crucial for assessing weather pattern consistency in PV studies. It is the square root of the variance (σ^2):

$$\sigma = \sqrt{\frac{1}{M} \sum_{k=1}^{M} (x_k - \mu)^2},$$

where μ is the mean. A lower standard deviation, as achieved by GOIF (Table 1), indicates tighter clusters, ensuring that selected days closely represent their weather patterns, which is vital for accurate data reduction.

Quartiles

Quartiles divide a dataset into four equal parts, providing insights into the distribution of weather data. The first quartile (Q_1) is the 25th percentile, the second quartile (Q_2) is the median (50th percentile), and the third quartile (Q_3) is the 75th percentile. The interquartile range (IQR) is:

$$\text{IQR} = Q_3 - Q_1.$$

In PV weather analysis, quartiles help identify typical and extreme conditions (e.g., high irradiance days in Q_3), supporting robust normalization methods that mitigate outlier effects.

Minimum and Maximum

The minimum (min) and maximum (max) values are the smallest and largest observations in a dataset, respectively. For irradiance data, they indicate the range of solar exposure, highlighting potential outliers (e.g., cloudy vs. sunny days). These values are essential for normalization techniques like min-max, which scale data based on this range.

Standard Normalization

Standard normalization (z-score normalization) transforms features to have a mean of 0 and a standard deviation of 1, which is useful for clustering methods sensitive to scale. For a feature matrix $\mathbf{F} \in \mathbb{R}^{N \times d}$ (e.g., N days, d features like temperature and irradiance), the normalized matrix $\mathbf{F}'$ is:

$$\mathbf{F}' = (\mathbf{F} - \mu) \odot \text{diag}(\sigma^{-1}),$$

where μ is the mean vector:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mathbf{F}_i,$$

σ is the standard deviation vector:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{F}_i - \mu)^2},$$

and $\odot$ denotes element-wise multiplication.

Benefits: It ensures features are comparable, improves distance-based clustering (e.g., k-means), and reduces outlier impact.

Limitations: Assumes a normal distribution, which may not hold for irradiance data, and remains sensitive to extreme values.

Min-Max Normalization

Min-max normalization scales features to a fixed range, typically $[0, 1]$, preserving their original distribution. The normalized matrix $\mathbf{F}'$ is:

$$\mathbf{F}' = (\mathbf{F} - \min(\mathbf{F})) \odot \text{diag}((\max(\mathbf{F}) - \min(\mathbf{F}))^{-1}),$$

where $\min(\mathbf{F})$ and $\max(\mathbf{F})$ are the minimum and maximum vectors across features.

Benefits: It ensures uniform scaling, is ideal for algorithms like neural networks, and retains weather data distributions (e.g., irradiance peaks).

Limitations: Highly sensitive to outliers (e.g., an unusually cloudy day), which can skew the range, and it discards the original scale, potentially affecting interpretability in PV analysis.

Robust Normalization

Robust normalization uses the median and IQR to mitigate outlier effects, suitable for skewed PV weather data. The normalized matrix $\mathbf{F}'$ is:

$$\mathbf{F}' = (\mathbf{F} - \text{median}(\mathbf{F})) \odot \text{diag}(\text{IQR}(\mathbf{F})^{-1}),$$

where $\text{median}(\mathbf{F})$ is the median vector, and $\text{IQR}(\mathbf{F}) = Q_3(\mathbf{F}) - Q_1(\mathbf{F})$.

Benefits: It is less affected by outliers (e.g., extreme temperature spikes), preserves central tendencies, and handles skewed distributions.

Limitations: It does not scale to a specific range, which some algorithms require, and is computationally heavier due to quartile calculations.

Maximum Absolute Normalization

Maximum absolute normalization scales features to $[-1, 1]$ by dividing by the maximum absolute value:

$$\mathbf{F}' = \mathbf{F} \odot \text{diag}(\max(|\mathbf{F}|))^{-1},$$

where $\max(|\mathbf{F}|) = \max_i |\mathbf{F}_i|$.

Benefits: Preserves data signs (e.g., negative temperature anomalies), scales symmetrically, and is less outlier-sensitive than min-max.

Limitations: Less effective than robust normalization for extreme outliers and unsuitable for algorithms requiring non-negative inputs.

Quantile Transformer Normalization

Quantile transformer normalization maps features to a uniform distribution, reducing skew-

ness in weather data. The normalized matrix $\mathbf{F}'$ is:

$$\mathbf{F'}_i = Q_{\text{unif}}\left(Q_F\left(\mathbf{F}_i\right)\right),$$

where $Q_F(x) = P(\mathbf{F}_i \leqslant x)$ is the empirical CDF, and $Q_{\text{unif}}(p) = p$ for $p \in [0, 1]$.

Benefits: It mitigates outliers and skewness (e.g., in irradiance distributions), improves distance-based clustering, and supports non-linear transformations.

Limitations: Computationally intensive for large datasets and may alter feature relationships, affecting PV pattern analysis.

7. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. Ethics declaration

Ethics declaration is not applicable.

9. Acknowledgments

This work has been supported by the French National Research Agency, through the Investments for Future Program (ref. ANR−18−EURE−0016 — Solar Academy). Part of this work was supported by the French National Program "Programme d'Investissements d'Avenir - INES.2S" under Grant Agreement ANR−10−IEED−00140014−01. In addition, Srijani Mukherjee would like to acknowledge the INES.2S French Institute for the Energy Transition, the PhD contract USMB, and the CSMB — Conseil Savoie Mont Blanc for their support. This publication is based upon work supported by the Khalifa University of Science and Technology under Award No. $FSU - 2023 - 014$.

References

[1] S. Dubey, J. N. Sarvaiya, and B. Seshadri, "Temperature dependent photovoltaic (pv) efficiency and its effect on pv production in the world – a review," *Energy*

Procedia, vol. 33, pp. 311–321, 2013, pV Asia Pacific Conference 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610213000829

[2] C. Cornaro and A. Andreotti, "Influence of average photon energy index on solar irradiance characteristics and outdoor performance of pv modules," *Progress in Photovoltaics: Research and Applications*, vol. 21, 04 2012.

[3] M. Bhavani, K. Vijaybhaskar Reddy, K. Mahesh, and S. Saravanan, "Impact of variation of solar irradiance and temperature on the inverter output for grid connected photo voltaic (pv) system at different climate conditions," *Materials Today: Proceedings*, vol. 80, pp. 2101–2108, 2023, sI:5 NANO 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785321044758

[4] D. Sampath Kumar, G. Yagli, M. Kashyap, and D. Srinivasan, "Solar irradiance resource and forecasting: A comprehensive review," *IET Renewable Power Generation*, vol. 14, 07 2020.

[5] G. de Freitas Viscondi and S. N. Alves-Souza, "A systematic literature review on big data for solar photovoltaic electricity generation forecasting," *Sustainable Energy Technologies and Assessments*, vol. 31, pp. 54–63, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213138818301036

[6] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263–286, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829631630488X

[7] J. Tena-García, L. García-Alcala, D. C. Lopez Diaz, and L. Fuentes-Cortes, "Implementing data reduction strategies for the optimal design of renewable energy systems," *Process Integration and Optimization for Sustainability*, vol. 6, 03 2022.

[8] A. Amin and M. Mourshed, "Weather and climate data for energy applications," *Renewable and Sustainable Energy Reviews*, vol. 192, p. 114247, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136403212301105X

[9] O. Bamisile, C. Acen, D. Cai, Q. Huang, and I. Staffell, "The environmental factors affecting solar photovoltaic output," *Renewable and Sustainable Energy Reviews*, vol. 208, p. 115073, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032124007998

[10] G. Guest, K. M. MacQueen, and E. E. Namey, "Introduction to applied thematic analysis," *Applied thematic analysis*, vol. 3, no. 20, pp. 1–21, 2012.

[11] M. Strohbach, J. Daubert, H. Ravkin, and M. Lischka, *Big Data Storage*. Cham: Springer International Publishing, 2016, pp. 119–141. [Online]. Available: https://doi.org/10.1007/978-3-319-21569-3_7

[12] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein, and G. E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 2, pp. 143–158, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3349

[13] T. Cebecauer and M. Suri, "Typical meteorological year data: Solargis approach," *Energy Procedia*, vol. 69, pp. 1958–1969, 2015, international Conference on Concentrating Solar Power and Chemical Energy Systems, SolarPACES 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610215005019

[14] D. Clarkson, E. Eastoe, and A. Leeson, "The importance of context in extreme value analysis with application to extreme temperatures in the U.S. and Greenland," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 72, no. 4, pp. 829–843, 02 2023. [Online]. Available: https://doi.org/10.1093/jrsssc/qlad020

[15] A. Arroyo, V. Tricio, E. Corchado, and A. Herrero, *A Comparison of Clustering Techniques for Meteorological Analysis*, 05 2015, pp. 117–130.

[16] M. A. Hassan, A. Khalil, and M. Abubakr, "Selection methodology of representative meteorological days for assessment of renewable energy systems," *Renewable Energy*, vol. 177, pp. 34–51, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148121008077

[17] T. Schütz, M. H. Schraven, M. Fuchs, P. Remmen, and D. Müller, "Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis," *Renewable Energy*, vol. 129, pp. 570–582, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148118306591

[18] J. Vera and J. Angulo, "An mds-based unifying approach to time series k-means clustering: application in the dynamic time warping framework," *Stochastic Environmental Research and Risk Assessment*, vol. 37, pp. 1–12, 05 2023.

[19] K. Poncelet, H. Höschle, E. Delarue, A. Virag, and W. D'haeseleer, "Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 1936–1948, 2017.

[20] A. Almaimouni, A. Ademola-Idowu, J. Nathan Kutz, A. Negash, and D. Kirschen, "Selecting and evaluating representative days for generation expansion planning," in *2018 Power Systems Computation Conference (PSCC)*, 2018, pp. 1–7.

[21] L. M. Aguiar, B. Pereira, P. Lauret, F. Díaz, and M. David, "Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting," *Renewable Energy*, vol. 97, pp. 599–610, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148116305390

[22] X. Wang, P. Guo, and X. Huang, "A review of wind power forecasting models," *Energy Procedia*, vol. 12, pp. 770–778, 2011, the Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610211019291

[23] A. Vyas, S. Abimannan, P.-C. Lin, and R.-H. Hwang, *Spatiotemporal Renewable Energy Techniques and Applications*, 04 2024, pp. 193–212.

[24] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020.

[25] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157309002841

[26] B. Colange, L. Vuillon, S. Lespinats, and D. Dutykh, "Ming: an interpretative support method for visual exploration of multidimensional data," *International Journal of Pattern Recognition and Artificial Intelligence*, 07 2020.

[27] O. Kramer, *K-Nearest Neighbors*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23. [Online]. Available: https://doi.org/10.1007/978-3-642-38652-7_2

[28] S. Ghosh, M. Halappanavar, A. Tumeo, A. Kalyanaraman, H. Lu, D. Chavarrià-Miranda, A. Khan, and A. Gebremedhin, "Distributed louvain algorithm for graph community detection," pp. 885–895, 2018.

[29] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, oct 2008. [Online]. Available: https://dx.doi.org/10.1088/1742-5468/2008/10/P10008

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 11 1998.

[31] L. László, "Random walks on graphs: A survey, combinatorics, paul erdos is eighty," *Bolyai Soc. Math. Stud.*, vol. 2, 01 1993.

[32] S. Mukherjee, L. Vuillon, D. Dutykh, and I. Tsanakas, "Dimensionality Reduction of Environmental Data for Long-Term PV Performance Analysis Using Graph Based Methods," in *EUPVSEC*, ser. EUPVSEC, Vienna, Austria, Sep. 2024, pp. 020 411–001 – 020 411–004. [Online]. Available: https://cea.hal.science/cea-04806089

[33] J. Kim, J. Obregon, H. Park, and J.-Y. Jung, "Multi-step photovoltaic power forecasting using transformer and recurrent neural networks," *Renewable and Sustainable Energy Reviews*, vol. 200, p. 114479, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032124002028

[34] S. Almaghrabi, M. Rana, M. Hamilton, and M. Saiedur Rahaman, "Multivariate solar power time series forecasting using multilevel data fusion and deep neural networks," *Information Fusion*, vol. 104, p. 102180, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523004967