Structural *k*-means (S *k*-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data

Quang-Van Doan[1], Toshiyuki Amagasa[1], Thanh-Ha Pham[2], Takuto Sato[1], Fei Chen[3], Hiroyuki Kusaka[1]

[1]Center for Computational Sciences, University of Tsukuba, Japan
[2]Hanoi University of Sciences, National University Hanoi, Vietnam
[3]Research Applications Laboratory, National Center for Atmospheric Research, USA

Correspondence to: Quang-Van Doan (doan.van.gb@u.tsukuba.ac.jp)

Abstract. Dramatic increases in climate data underlie a gradual paradigm shift in knowledge-acquisition methods from physical-based models to data-based mining approaches. *k*-Means is one of the most popular data clustering/mining techniques, and it has been used to detect hidden patterns in climate systems. *k*-Means is established based on distance metrics for pattern recognition, which is relatively ineffective when dealing with "structured" data that are dominant in climate science, that is, data in time and space domains. Here, we propose (i) a novel structural similarity recognition-based *k*-means algorithm called structural *k*-means or S *k*-means for climate data mining and (ii) a new clustering uncertainty representation / evaluation framework based on the information entropy concept. We demonstrated that the novel S *k*-means could provide higher-quality clustering outcomes in terms of general silhouette analysis, although it requires higher computational resources compared with conventional algorithms. The results are consistent with different demonstration problem settings using different types of input data, including two-dimensional weather patterns, historical climate change in terms of time series, and tropical cyclone paths. Additionally, by quantifying the uncertainty underlying the clustering outcomes we for the first time evaluated the "meaningfulness" of applying a given clustering algorithm for a given dataset. We expect that this study will constitute a new standard of *k*-means clustering with "structural" input data, as well as a new framework for uncertainty representation/evaluation of clustering algorithms for (but not limited to) climate science.

1. Introduction

In recent decades, the volume and complexity of climate data have increased dramatically owing to advancements in data acquisition methods (Overpeck et al., 2011). This increase underlies a gradual shift of climate-knowledge acquisition paradigm

from using classical "first-principle" models (i.e., based on physical laws) to models and analyses directly based on data (i.e., data mining) (Kantardzic, 2011). Hence, numerous data mining techniques have been developed to shed light on the underlying nature and structure of data. Clustering, as one of the principal data mining methods, is the technique to organize a set of data into clusters that maximize the homogeneity of the elements in a cluster and the heterogeneity among different clusters (Pérez-Ortega et al., 2019). Clustering algorithms are useful to handle large, multivariate, and highly dimensional data, which are difficult for human perception. Among numerous clustering algorithms, *k*-means is one of the most well-known and widely used in most research domains (Wu et al., 2008).

The history of *k*-means can be traced back to the 1950s – 1960s, when it was developed through independent efforts (e.g., Lloyd, 1957; Forgy, 1965; Jancey, 1966; MacQueen, 1967). The name *k*-means was coined in a paper by MacQueen (MacQueen, 1967). *k*-Means has been extensively used in climate science, thanks to its ease of implementation and interpretation. It is used to explore unknown atmospheric mechanisms and/or improve predictions. The most common application is the use of *k*-means within "detection-and-attribution" framework. In the framework, specific atmospheric conditions, or events, e.g., abnormally hot weather or heavy precipitation, are detected first. Then, the causes of these atmospheric conditions are attributed to atmospheric regimes/patterns, determined by *k*-means (Esteban et al., 2005; Houssos et al., 2008; Spekat et al., 2010; Zeng et al., 2019; Smith et al., 2020). Another application is the use of *k*-means for weather or climate predictions. In such a case, rather than being used as an independent prediction method, it is used to complement existing numerical prediction systems by suggesting the occurrence probability of certain weather conditions referring analog patterns to that derived by *k*-means from historical data (Kannan and Ghosh, 2011; Gutiérrez *et al.*, 2013; Le Roux *et al.*, 2018; Pomee and Hertig, 2022). Furthermore, *k*-means is also used for future climate prediction (also known as a statistical downscaling) or for reconstructing historical data (Camus et al., 2014) using the same analog approach.

The *k*-means algorithm is an interactive clustering method. To briefly describe, it involves four processing steps: *i*) initiation: pre-definition of *k* cluster centers (or centroids); *ii*) classification: clustering of an object with similar objects; iii) centroid update: recalculation of centroids based on the updated classification; iv) convergence (equilibrium) judgement: halts the algorithm if object migrations are not observed from one cluster to another; otherwise, returns to step ii) if such migrations are observed (Pérez-Ortega et al., 2019). The dominance of *k*-means over most research fields is partly due to its simplicity and ease of use. Also, simplicity inherits the drawbacks of the algorithm, which have inspired researchers for decades to identify improvements. Consequently, these efforts have delivered a great number of *k*-means variants alongside those from the earliest time.

Improving centroid initialization represents an important issue to be resolved. *k*-Means clustering outcomes are known to be sensitive to the initialization of centroids (Sydow, 1977; Katsavounidis et al., 1994; Bradley and Fayyad, 1998; Pelleg, 2000; Khan and Ahmad, 2004; Arthur and Vassilvitskii, 2006; Su and Dy, 2007; Eltibi and Ashour, 2011). Subsequent efforts have been made to improve the calculation procedure in the classification scheme primarily because it is the most computationally time-consuming. These efforts resulted in numerous *k*-means variants (Fahim et al., 2006; Lai and Huang, 2010; Perez et al., 2012). More recent studies have focused on the fundemental basis of the classification, that is, how to define the similarity for which an object should be classified to one cluster but not another.

The conventional *k*-means classification scheme is established based on the distance paradigm, in which the similarity is determined by distance metrics including the Euclidean distance, Manhattan distance, or their general form, the Minkoski distance (Cordeiro de Amorim and Mirkin, 2012). The advantage of distance metrices lies in their ease of implementation and popularity, thus making the judgement for using them less controversial. Nonetheless, recent studies have pointed out that distance metrics defend less against noisy and irrelevant features (or dimensions, in other words) of input objects (vectors) (de Amorim, 2016). Few studies have proposed the use of feature weights to overcome this weakness (Chan et al., 2004; Huang et al., 2005; Cordeiro de Amorim and Mirkin, 2012). However, such improvements do not intentionally consider the structural relationship between vector dimensions, especially when data are time series or spatially distributed.

Atmospheric data are characterized by their temporal and spatial "structuredness". In other words, the information value of data lies in their interrelationship or trends in time and space. For example, when looking at weather maps, one might realize that locations of high or low pressures would be the first concern. Likewise, the similarity in trend, or the phase correlation between two time series might be more important than the difference in their absolute values. Thus, the distance measures, which treat the features of the input objects equally, might underestimate the inherent "structuredness" in the objects when determining the similarity between them, consequently deteriorating the clustering outcomes. However, to replace distance metrics by something different remains big challenging, because distance metrics have deep historical roots, and they undoubtedly laid the foundation for modern data mining, including clustering algorithms. As mentioned by Wang et al., "it (distance metric) is not bad, and easy to use" and "everyone else use it" (Wang et al., 2004).

Contemplating the nature of atmospheric data, a specific question raised here is whether another *k*-means approach is available that can consider the "structural" similarity in time and space between input objects. Answering this question has great practical value, particularly, for the climate informatics field owing to the unprecedented recent increase in archived data. The demand is growing for innovative and effective tools of data mining that can handle the inherent nature of climate data.

Here we propose a novel *k*-means algorithm based on the "structural" similarity recognition, called *Structural k-means* or *S k-means*. S *k*-means follows the same procedure as the generic *k*-means algorithm. It differs from the generic algorithm in incorporating a recent innovation in signal processing science, namely, the structural similarity (S-SIM) recognition concept (Wang et al., 2004) into the classification scheme. The novel S *k*-means inherits the simplicity of the generic algorithm and meanwhile can handle temporally and spatially ordered data.

We evaluate the performance of S *k*-means clustering across three representative demonstration tests. The tests cover multiple types of input data, that is, spatial distributions (weather patterns), time series (historical change in temperature), and hybrid types (tropical cyclone tracking). Using multiple data types is a unique point of this study that can robustify the conclusions through cross comparisons. The performance of S *k*-means is evaluated against three other *k*-means algorithms using different similarity/distance metrics for the classification scheme, that is, the Pearson correlation coefficient, Euclidean, and Manhattan distances, hereafter called C, E, and M *k*-means, respectively. We implement various *k* (number of centroids) configurations and multiple initializations (randomized). Eventually, 1320 model runs are conducted. Such settings ensure the robustness of the results and conclusions. The "general" silhouette analysis/score, which is a scoring method based on general similarity/distance metrics, is used to quantify the algorithm performance.

We propose a novel framework for *clustering-uncertainty evaluation/representation* based on the information entropy concept. This framework is primarily used to quantify the variability/consensus among the clustering outcomes across the different *k*-means algorithms. At the core of the framework is the newly proposed concept "*clustering uncertainty degree*," which builds on mutual information theory. Also, relevant visualization tools including connectivity matrix, heatmap, and chord diagram are proposed to represent the clustering uncertainty.

To the best of our knowledge, this study is the first to address the uncertainty issue in the climate science. Our study is the first to propose a clustering-uncertainty evaluation framework, borrowing recent-most techniques and concepts in information theory. This framework is not only used to quantify the clustering uncertainty but also to serve a more fundamental purpose, i.e., to measure the "meaningfulness" of the application of clustering for a given problem dataset. We expect that this framework together with the S *k*-means algorithm will establish a new standard in data mining and clustering studies, primarily for (but not limited to) climate science.

The remainder of this paper is organized as follows. Section 2 describes the S *k*-means algorithm. Section 3 presents the test simulation configurations. Section 4 describes the evaluation metrics and a novel framework for clustering uncertainty. Section 5 presents and discusses the results. Section 6 provides the concluding remarks.

2. Description of the algorithms

2.1 S *k*-means algorithm

S *k*-means follows the conventional procedure of generic *k*-means clustering. To express this mathematically, let define $X = \{x_1, \ldots x_i, \ldots, x_n\}$ be a set of n objects (input vectors), where $x_i \in R^d$ $(i = 1, \ldots, n)$ and $d \geq 1$ is the number of dimensions. Let $K = \{1, \ldots, k\}$ with $k \geq 2$ denote the number of groups.

For a *k*-partition, $P = \{G(1), \ldots, G(k)\}$ of X, let c_j denote the centroid of cluster $G(j)$, for $j \in K$, with $C = \{c_1, \ldots, c_k\}$ and a set of weight vectors $W = \{w_{11}, \ldots, w_{ij}\}$. Hence, the clustering problem can be formulated as an optimization problem (Selim and Ismail, 1984), which is described by the following equation:

$$P\text{: minimize } z(W, M) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} d(x_i, c_j)$$

$$\text{subject to } \sum_{j=1}^{k} w_{ij} = 1, \text{for } i = 1, \ldots, n,$$

$$w_{ij} = 0 \text{ or } 1, \text{for } i = 1, \ldots, n, \text{and } j = 1, \ldots, k \tag{1}$$

where $w_{ij} = 1$ implies that object x_i belongs to clusters $G(j)$ and $d(x_i, \mu_j)$ denotes the distance between x_i and μ_j for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

The S *k*-means algorithm consists of four steps (**Fig. 1a**), which are similar to that of generic algorithms except for step (*ii*). The steps are described as follows:

- *(i) Initialization.* Initialize k centroid vectors. Although k-means has several options for initialization, we apply a randomized scheme to initialize the centroids.
- *(ii) Classification.* Assign an object to its most similar centroid. The S k-means algorithm uses the structural similarity (S-SIM) (Wang et al., 2004) recognition technique to determine the most similar centroids instead of using distance measures, such as that in generic algorithms.
- *(iii) Centroid calculation*. Update centroid vectors by taking the mean value of the objects belonging to these clusters.
- *(iv) Convergence determination.* The algorithm stops when equilibrium is reached, that is, when there are no object migrations from one cluster to another. Technically, the algorithm converges if the sum of the mean square errors of centroids versus those in the previous step become zero in the experiments of this study. The convergence criterion is the same for all k-means variants used. A limitation of iteration is setup to 100 to avoid the infinite loop of iterations. If equilibrium is not reached, then the process is repeated from step (*ii*).

S k-means is compared with E, M, and C k-means (k-means using the Euclidean distance, the Manhattan distance, and the Pearson correlation coefficient). E, M, and C k-means also follow the same procedure as indicated above except for classification scheme (*ii*), where the respective similarity/distance measures are used to determine the most similar centroids.

2.2 Structural similarity

The metrics for the structural similarity (S-SIM) recognition process were first introduced by Wang et al. (2004). It was developed to better predict the perceived quality of digital television and cinematic pictures. S-SIM is intended to improve the traditional peak signal-to-noise ratio or mean squared error in detecting similarities between "structural" signals, such as images. Intuitively, S-SIM is determined by considering the differences between two input signals (vectors x, y) across multiple aspects including "luminance", "contrast", "and structure" which represent the characteristics of human visual perception. Luminance masking is a phenomenon whereby image distortions tend to be less visible in bright regions, while contrast masking is a phenomenon whereby distortions become less visible where there is significant activity or "texture" in the image. Mathematically, S-SIM is determined as follows:

$$SSIM(x,y) = l(x,y)^{\alpha} \times c(x,y)^{\beta} \times s(x,y)^{\gamma} \quad (2)$$

where $l(x,y), c(x,y)$, and $s(x,y)$ measure similarities in luminance (brightness values), contrast and structure between sample vectors x, y with weight values α, β and γ. Let μ_x and μ_y be the mean values; σ_x and σ_y the standard deviations; σ_{xy}

the covariance of the two sample vectors x, y, then luminance, contrast, and structure similarities are defined as $l(x,y) = \frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1}$, $c(x,y) = \frac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2}$, and $s(x,y) = \frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}$. Note that c_1, c_2 and c_3 are parameters to stabilize the division with a weak denominator. Even if $c_1 = c_2 = c_3 = 0$, S-SIM still work quite well (Wang and Bovik, 2009). $l(x,y)$ measures the similarity in brightness, i.e., the difference regarding mean values; $c(x,y)$ quantifies the similarity in illumination variability, which is regarded to standard deviations; and $s(x,y)$ measures the correlation in spatial inter-dependencies between images that is close to the Pearson correlation coefficient. For simplification, here we set $c_1 = c_2 = c_3 = 0$ and weights $\alpha = \beta = \gamma = 1$ and reduce the original formula to the following:

$$\text{S-SIM}(x,y) = \frac{2\mu_x\mu_y\sigma_{xy}}{(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_y^2)} \tag{3}$$

S-SIM is symmetric index, i.e., $\text{S-SIM}(x,y) = \text{S-SIM}(y,x)$. It does not satisfy the triangle inequality or non-negativity, and thus is not a distance function. S -SIM ranges from -1 to 1, where -1 indicates totally dissimilar and 1 indicates totally similar. Wang and Bovik (2009) showed that S-SIM represents a powerful, easy-to-use, and easy-to-understand alternative to traditional distance metrics, such as Euclidean distance, for dealing with spatially and temporally structured data, i.e., data having strong spatial and temporal inter-dependencies. These inter-dependencies carry important information about the objects in the visual scene. S-SIM emerged as a "new-generation" similarity metric with an increasing number of applications outside the signal processing field, including hydrology and meteorology (e.g., Mo et al., 2014; Han and Szunyogh, 2018; Doan et al., 2021).

3. Demonstration tests

S *k*-means is applied to three representative clustering problems. These problems cover various types of input datasets that represent diverse issues, i.e., weather pattern (in terms of two-dimensional pressure data), historical climate change (in terms of time series), and tropical cyclone tracking data (the hybrid type of data containing both spatial and temporal information) (**Fig. 1b**). The details of these three tests are described below.

- *Weather pattern (WP) clustering*. Group winter weather patterns in Japan. The mean sea level pressure (SLP) was obtained using ERA-Interim reanalysis data (Dee et al., 2011). The data have a horizontal resolution of 0.75°on a regular grid but are re-gridded to an equal-area scalable earth-type grid at a spatial resolution of 200 × 200 km using nearest-neighbor interpolation method. This interpolation/regridding method is commonly applied to high-latitude

domains (Gibson et al., 2017). Data collected in winter months, that is, December, January, and February (DJF), for ten years (2005-2014) over the region from 20 – 50 °N and 115 – 165°E were used. The total number of samples used is 902. Each sample has a grid size of 35 pixels × 35 pixels.

- *Climate change (CC) clustering*. Group temperature-increase time series data collected over 70 years (1951 – 2020) from in situ weather stations run by the Japan Meteorological Agency. A simple data-quality check is implemented. Weather stations that missed (daily basis) observations for more than 10% of the total period of interest are excluded. Therefore, 134 valid weather sites remain (see Fig. 1b CC for the location of weather sites). The annual mean of each time series is calculated, and the climate change component is determined by subtracting the average of the first 30 years (1951 – 1980) from each value series.
- *Tropical cyclone (TC) tracking clustering*. Group the best TC tracks from 1951 to 2020, which are retrieved from the Japan Regional Specialized Meteorological Center (RSMC) (https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/besttrack.html). Note that the RMSC provides only the best TC tracks, which have a maximum wind speed of more than 17.2 $m \cdot s^{-1}$, e.g., wind force 8 of the Beaufort scale (Barua, 2019). These data contain the TC classification, maximum sustained wind speed, central pressure, and latitude and longitude of the TC centers with 6-hourly intervals. In this study, only TCs that passed the Japan region, defined as the region between 25 – 45°N and 126 – 150°E, are used for the analysis. Hence, the total number of TCs feeding the *k*-means is 863. Because *k*-means clustering requires identical lengths of input vectors, the TC tracks are reconstructed so that they had an equal length of 20 segments by the method proposed by Kim et al. (Kim et al., 2011), which has been applied in several studies (Choi et al., 2012; Kim and Seo, 2016).

As mentioned in the introduction, in addition to S *k*-means, the C, E, and M *k*-means methods that use Pearson correlation coefficients and Euclidean and Manhattan distances for the classification scheme are used for the tests. For this, we perform a total of $3 \times 4 = 12$ simulations. For each simulation, 11 k settings are implemented, that is, $k = 2, 4, 6, \ldots, 20$, and for each k, ten runs (randomized initializations) are realized. In summary, a total of $12 \times 11 \times 10 = 1320$ runs (model realizations) are performed for the analysis.

4. Evaluation measures

4.1 Similarity distributions

The similarity-distributions technique developed by Doan et al. (2021) to evaluate "global" pairwise relationship of input vectors is adopted for performance evaluation. In this study it is named the similarity distribution (S-distribution or S-D). The S-distribution is a probability density function of pairwise similarities of a vector set. Let $X = \{x_1, \ldots, x_n\}$ be the set of n objects; s_{ij} the pairwise similarity between two objects, which is defined as $s_{ij} = F(x_i \rightarrow x_j)$; and F the similarity function, $i = 1, 2, \ldots, n; j = 1,2, \ldots, n$. The normalized s_{ij} is defined as $s'_{ij} = (s_{ij} - \min\{s\})/(\max\{s\} - \min\{s\})$. By definition, s'_{ij} ranges from 0 to 1, with the maximum value of 1 indicating perfect similarity (self-similarity) and the minimum value of 0 indicating a lack of similarity (distance to the furthest object), thus s'_{ij} is data dependent. As F is a symmetric function, that is, $F(x_i \rightarrow x_j) = F(x_j \rightarrow x_i)$ for all similarity/distance indices of interest, i.e., S-SIM, COR, ED, and MD, duplicated values are removed. Also, self-similarity values, that is, s'_{ij} with $i = j$, are removed. Thus, $n(n-1)/2$ values remained in the final set S of s'_{ij}. The S-distribution, or S-D, is defined as the probability density function of the values of S. The S-Ds were then plotted together for comparison. In addition, statistical parameters, such as the mean, standard deviation, skewness, kurtosis, and Shannon entropy, were calculated to further diagnose the characteristics of the datasets of interest.

4.2 "General" silhouette analysis

As *k*-means clustering is an unsupervised machine learning method, it does not require "ground truth", or predefined cluster labels of an input dataset for classification. The absence of "ground truth" means that the algorithm can be validated only with internal validation criteria. Internal validation is to define the goodness of clustering outcome based on the result itself to define how clustering methods optimize the homogeneity within a cluster and maximize the difference among clusters (Hassani and Seidl, 2017). There are numerous indices for clustering internal validation, though most of them are built on Cartesian geometric algebra, which is not the case with non-distance metrics like S-SIM.

Thus, this study uses the *general* silhouette analysis method to validate the algorithms. The general silhouette analysis is the generalized form of the silhouette analysis (Rousseeuw, 1987) that can applicable also for non-distance metrices. This concept was firstly used for the evaluation of self-organizing maps by Doan et al. (2021). Silhouette analysis is a comprehensive analysis of the interpretation and validation of cluster methods. This technique offers a concise graphical representation of how well each object has been classified (Rousseeuw, 1987). The silhouette value is a measure of how coherent an object is

with its cluster versus how it is separated from other clusters. Mathematically, the general silhouette coefficient (GSC) for a given object is defined as follows:

$$GSC = \frac{b - a}{\max\{a, b\}} \tag{4}$$

where a and b are the mean intracluster distance and mean distance to the nearest cluster, respectively. Note that the distance here is the "general" distance and not the Euclidean distance, which is originally defined in the study by Rousseeuw (1987). The general distance is the reversed normalized similarity (i.e., $-s'_{ij}$) defined in subsection 4.1, which is why here we call it the general silhouette coefficient.

The GSC values ranged from −1 to +1. A higher value indicates the goodness of the cluster assignments, that is, the object is coherent with its cluster and well separated from neighboring clusters. The clustering configuration is appropriate if most objects have high scores. In contrast, if many objects have low or negative values, then the clustering configuration performs poorly. A GSC of zero indicates that the object is on or very close to the border of two neighboring clusters, and a negative GSC indicates that the object may have been assigned the wrong cluster label.

4.3 Clustering uncertainty evaluation

Evaluating the variability or uncertainty inherent in a clustering algorithm is challenging owing to the unique nature of the clustering outcome. It is difficult to define the statistical mean, standard deviation, or range between quantiles of a given ensemble of clustering realizations.

Herein, we propose a framework for the representation/evaluation of the uncertainty of the clustering problem, which is based on a pairwise comparison of clustering realizations using a quantified index called the clustering uncertainty degree (CUD). The CUD is built-up on the mutual information concept, in more detail, the adjusted mutual information index. In information theory, mutual information from two random variables is used to quantify the "amount of information" obtained for one random variable by observing another random variable. The concept of mutual information is intimately linked to the entropy concept of a random variable, which is a fundamental notion in information theory that quantifies the expected "amount of information" held in this variable. In this study, mutual information is applied to evaluate the agreement between two clustering realizations (label assignments of N objects). To do so, the mathematical formula for mutual information $I(U, V)$ between two clustering realizations U and V is defined as follows:

$$I(U,V) = H(U) + H(V) - H(U,V) \tag{5}$$

where $H(U)$ *and* $H(V)$ are the entropies of each realization and $H(U,V)$ is the joint entropy of the two. Entropies of clustering realizations are defined as the amount of uncertainty for partition sets of each realization.

$$H(U) = -\sum_{i=1}^{|U|} P(i)\log\left(P(i)\right) \tag{6}$$

$$H(V) = -\sum_{j=1}^{|V|} P'(j)\log\left(P'(j)\right) \tag{7}$$

where $P(i) = a_i/N$ and $a_i = |U_i|$ is the probability that an object pickup at random from U falls into class U_i. Similarly, for V, $P'(j) = b_j/N$, where $b_j = |V_i|$ is the probability of an object from V falling into class V_j.

$$H(U,V) = -\sum_{i=1}^{|U|}\sum_{\mathrm{j}=1}^{|V|} P(i,j)\log\left(P(i,j)\right) \tag{8}$$

where $P(i,j) = \left|U_i \cap V_j\right|/N$ is probability that an object pickup at random falls into both class U_i and V_j.

By definition, mutual information ranges from 0 to 1. A value of 1 indicates perfect agreement (equality) between the two clustering realizations, while values close to zero indicate that the two label assignments are largely independent. However, mutual information is weak against chance. Vinh et al. (2009) derived the expected mutual information and proposed the concept of adjusted mutual information that can defend against chance (Vinh and Epps, 2009; Vinh et al., 2010; Romano et al., 2016). Thus, random (uniform) label assignments have an adjusted mutual information score close to 0.0 for any number of clusters and objects (which is not the case for raw mutual information). Note that the adjusted mutual information is primarily developed to measure the "goodness" of clustering outcomes versus prior-known "ground truth". In this study, we diversify this primary purpose by applying the metrics to evaluate the uncertainty/consistent/convergence of clustering outcomes. Also, using the adjusted mutual information must be understood as showcase for the evaluation framework. We could also use alternative techniques, e.g., rand index, for the same purpose.

$$E[I(U,V)] \tag{9}$$
$$= \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{Nn_{ij}}{a_i b_j}\right)\frac{a_i!\,b_j!\,(N-a_i)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_j-n_{ij})!\,(N-a_i-b_j+n_{ij})!}$$

$$I'(U,V) = \frac{I(U,V) - E[I(U,V)]}{mean\{U(U), H(V)\} - E[I(U,V)]} \tag{10}$$

The core concept underlying the CUD, i.e., clustering uncertainty degree, is defined as follows:

$$CUD(U,V) = 1 - I'(U,V) \tag{11}$$

By definition, CUD is a representation of pairwise dissensus of clustering realizations. The CUD ranges from 0 to 1. A value of 1 indicates the greatest dissensus or highest uncertainty between U and V, while a value of 0 indicates perfect consensus or no uncertainty. The connectivity matrix of pairwise CUDs is defined as a $M \times M$ matrix and CUD values for a pair of clustering realizations, where M is the number of clustering realizations. The connectivity matrix naturally serves as a visualization tool to assess the general uncertainty of the clustering system. Other visualization tools are also used to visualize the CUD, including a heatmap and a chord diagram (Holten, 2006). Heatmaps work like a connectivity matrix but in a more visualized form. A chord diagram is a useful graphical method for demonstrating the interrelationships between the data in a matrix. The data are plotted radially around a circle. The relationships between data points are usually drawn as arcs that connect the data.

5. Results and discussion

5.1 S-distributions

Before analyzing the *k*-means clustering results, we diagnose the nature of the input data using S-distributions (or S-Ds). S-Ds provide "global" insights into how data vectors are related to each other in four S-SIM, COR, ED, and MD topological spaces. The results, which are shown in **Figure 2**, demonstrate an apparent difference in the shape of the S-Ds. Notably, the S-Ds for

ED and MD appear more symmetrical than those for S-SIM and COR across the three types of input data, that is, WP, CC, and TC. For S-SIM and COR, S-Ds tend to be more tailed (both sides), with skewness over the left tail. Quantitively, the standard deviation of S-Ds for S-SIM and COR exhibit higher values (approximately 0.13 – 0.20) than those for ED and MD (approximately 0.11 – 0.13) (**Table 1**), despite an exception for ED in the TC simulation. The skewness (measures the symmetry of S-Ds) exhibits negative values meaning the distributions are left-skewed. This fact is clearly confirmed in visualized results (Figure 2). Especially, S-SIM and COR exhibit higher skewed than ED and MD particularly in the CC and TC experiments. The skew-over-left of S-SIM and COR indicates that those tend to project "hierarchical affinity" of input vectors, meaning that a given vector tends to be closer to a certain group of peers and relatively far from another group located at the opposite end of similarity spectrum. In this sense, these results demonstrate that the discrimination ability of S-SIM and COR is higher than that of traditional distance metrics, such as ED or MD. In addition, kurtosis and Shannon entropy measure the flatness and "information value" (or "information gain" in the case of comparison) of distributions, respectively. Overall, kurtosis values are consistent with the visualized results in Figure 2, i.e., S-Ds of S-SIM and COR tend to spread more over two tails than those of ED and MD. Entropy, on the other hand, is likely more data dependent. It does not show obviously higher and lower trends of S-SIM, and COR than those of ED and MD.

5.2 Clustering results

As explained in Section 3, three demonstration problems, WP, CC, and TC, are conducted with different k configurations and centroid initializations, with a total number of runs of 1320. Note that this study addresses the algorithm aspects (attempting to seek general insight into the system's performance regardless of problems). We do not intend to physically interpret the specific clustering outcomes, although some phenomenal explanations are provided in the manuscript.

The clustering results are partly visualized and shown together with quantified silhouette scores in **Figures 3, 4**, and **5** for WP, CC, and TC, respectively, for the configuration $k = 4$ and the first initialization, R0 (see the Supplementary Material for more information). Here, we explain the *k*-means-detected weather patterns over the Japan region during December, January, and February (DJF) (**Fig. 3**). During DJF, the weather in Japan is dominated by a winter-type pattern. The winter type is characterized by the Siberian High (develops over the Eurasian continent) and the Aleutian Low (develops over the northern North Pacific) resulting in prevailing northwesterly winds. The wind blows cold air from Siberia to Japan and causes heavy snowfall on the western coast and sunny weather on the Pacific side of the country. This winter-type pattern is clearly captured by all *k*-means variants, that is, C2 for S, C4 for C, C3 for E, and C4 for M *k*-means (**Fig. 3**). The silhouette analysis reveals an interesting result. S *k*-means generates dominant cluster C2 over other clusters regarding its frequency (the thickness of

each cluster label in the silhouette diagram indicates the number of members in the cluster). This result is consistent with prior knowledge of the weather patterns over the region (https://www.data.jma.go.jp/gmd/cpd/longfcst/en/tourist_japan.html). Moreover, S *k*-means shows consistently highest silhouette scores than the other algorithms for all $k = 2, 4, \ldots, 20$ settings (**Fig. 6a**), followed by C *k*-means. E and M *k*-means have lower scores than S and C *k*-means.

Regarding the CC experiment, the time-series results are visualized with reference to the geographical locations of the weather stations to support interpretation (**Fig. 4**). Overall, the result shows that although it is seen over all stations, the warming trend is not geographically uniform. These regional differences are well captured by the clustering. For example, the northern part (Hokkaido) is consistently separated from other regions in terms of warming rate which is faster than the other regions. Such a result highlights the usefulness of *k*-means to detect regional differences, which is useful for building detailed appropriate climate change actions (though it is not the main concern of this study). Regarding clustering quality, the superiority of S and C *k*-means is confirmed. Like WP, S and C *k*-means exhibit relatively higher silhouette scores for the CC data compared with E and M *k*-means (**Fig. 6b**).

In addition, the TC experiment aims to determine how *k*-means works with hybrid spatiotemporal data. Like the above experiments, S and C *k*-means are likely to outperform E and M *k*-means, that is clearly reflected by their higher silhouette scores (**Fig. 5** and **Fig. 6c). Figure 5** shows the four main patterns of the TC track determined using the four clustering methods. Although there are some differences in the results among the *k*-means variants, such as the genesis and depression points, all determined patterns are characterized mainly by curved trajectories. These averaged patterns could be divided into two groups: i) not crossing and ii) crossing mainland Japan. Overall, the number of TCs in group i) was higher than that in group ii), with these tracks characterized by TCs containing both straight and re-curving TC trajectories forming to the east of 140° E (e.g., clusters 2 and 4 of S *k*-means in **Fig. 5a**). For group ii), the averaged patterns show the TC track passing through the central area of Japan (e.g., clusters 1 and 3 of S *k*-means in **Fig. 5a**)

Consistently, the higher performance of S *k*-means is observed throughout the ensemble of tests, k settings, and initializations. The performance of S *k*-means is sometimes competed by C *k*-means. The two, S and C *k*-means, outperform the distance-metrics-based E and M *k*-means. It is worth noting that these results are obtained from the silhouette analysis. Additional evaluation approaches might be needed to generalize the conclusions, although this could be challenging because most objective clustering evaluations have been developed on the Cartesian geometric algebra assumption (that could work for distance metrics but might not work for non-distance measures). Therefore, it is necessary to develop new evaluation approaches beyond the distance paradigm. Another difficulty lies in the fact that, like other clustering techniques, *k*-means is

an unsupervised machine learning technique. It works in an absence of a single "ground truth" to guide the classification. The absence of "ground truth" indicates the difficulty to define the "goodness" or meaningfulness of *k*-means clustering outcomes. In pragmatic view, clustering outcomes become meaningful if they are assigned with physical meaning or successfully used for practical purposes like a prediction. Doing so does not fall into the scope of this study (it is a huge work and must be addressed in an independent study), here we adopt another approach to gain insight into the behavior of the *k*-means variants.

By taking a careful glance at the silhouette plots shown in Figure 3, it's possible to notice a discrepancy in S k-means compared to the rest. S *k*-means is likely to generate, say, "high-ordered" clustering, *i.e.*, one dominant weather pattern (larger group size) beside several non-dominant (smaller group size). The same trend is seen with different *k* settings (not shown). This agrees well with the prior knowledge recognized by meteorological research community and local people about the winter weather patterns in Japan (explained above. The insight leads to some possible hypotheses: (i) Does S *k*-means perform better, *i.e.*, closer to the human perception, than other variants? (ii) Is achieving "highly-ordered" clustering the intrinsic property of S *k*-means?

To examine the hypotheses, we attempt to quantify the "orderliness" of clustering outcomes using the Shannon entropy. The results, illustrated in **Figure 7**, show a good agreement between the calculated entropy values versus the intuition. S *k*-means appears to have consistently lower entropy (highly ordered clustering) than the other algorithms for the WP experiment (Fig. 7a), but not for the CC, and TC experiments (Fig. 7b, c). We can dismiss the second hypothesis (ii), which posits that achieving "highly-ordered clustering" is an intrinsic property of S k-means, because it is not universally true across all experiences. Now hypothesis (i) remains. It is possible that S *k*-means can achieve clustering which fits closer to human perception. However, because we don't have prior knowledge regarding the CC and TC experiments, it is early to conclude that with the complete certainty. To diversify the clustering problems with different types of input data, or for different geographical areas, is necessary to obtain the comprehensive insight into S *k*-means.

To further the discussion from a different aspect, we examine how the similarity between objects is recognized in *k*-means variants. For intuitive comprehension, we generate "imagination" weather patterns and assess the discrimination ability of similarity/distance metrices. **Figure 8** illustrates the weather patterns including the reference (a), characterized by two extrema (Low and High) symmetrically distributed over both sides, the Gaussian noise contamination (b), the blurring (to the mean value) (c), luminance shift (d), contrast stretch (e) and the spatial shift (f). Though the Euclidean distance from these patterns (b-e) to the reference are intentionally set to be identical (=2.9), by using S-SIM, one can rank the similarities with descending order: S-SIM(d-a) = .99 > S-SIM(e-a) = .8 > S-SIM(b-a) = .67 > S-SIM(f-a) = .5 >> S-SIM(b-a) = 0. This simple demonstration

confirms the superiority of S-SIM in recognizing the difference between two-dimensional patterns, agreeing well with human intuition compared to ED. This implies that S-SIM could reduce the situation of random classification (i.e., an object is assigned to a centroid by chance) adding confidence to S *k*-means derived results. Though this result is shown for the two-dimensional data, it is believed to be true for one-dimensional structured data like time series.

Computational cost is another important factor, especially in a practical sense. We measure the computational cost of each experiment and show the results in **Figure 9**. Overall, S and C *k*-means require more time to complete the same task than E and M *k*-means. Roughly, S *k*-means required 5 – 6 times more computational time than E *k*-means. C *k*-means was comparable to S *k*-means. M *k*-means required less computational time than E *k*-means. Such a tradeoff between higher performance and computational cost should be considered when selecting an algorithm. Nevertheless, the computational cost is not a big issue at least limited to the settings of this study; for example, the time to finish a run is less than a minute, which is very small compared to the numerical weather prediction or climate simulation. In addition, the computational issue can be solved with the advancement of computational ability or by using a parallel computational approach.

5.3 Uncertainty evaluation

The results for the clustering uncertainty evaluation framework (CUEF) are discussed here. The clustering uncertainty degree (CUD) is shown in **Figure 10** (for $k = 4$ and run R0; the collective results are shown in **Figure 12**). As explained in Section 4, two visualization tools, i.e., heat maps and chord diagrams, are used to visualize the clustering uncertainty. For example, **Figure 10a** (WP) show the CUD values for S relative to C, E, and M *k*-means are 0.67, 0.75, and 0.77, respectively, with the heatmap. Note that the maximum CUD (=1) indicates the absolute disagreement between two clustering assignments and the minimum CUD (=0) indicates the absolute consensus between the two. The chord diagram demonstrates the pairwise relationship in a more qualified manner. One can easily determine which algorithms (S, C, M, or E) have less consensus with another (wide arc length on the circle means less consensus), and vice versa. For example, E and M *k*-means show the high consensus with each other. S *k*-means shows less uncertainty/high consensus relative to E and M compared with C *k*-means, particularly in the CC and TC experiments. Note that we run the four *k*-means variants with the randomized centroids each time. Additional runs using the same starting centroids for the four *k*-means variants show that the uncertainty related to the clustering algorithm selection remains regardless of using the same or randomized staring centroids.

In addition to the algorithm-wise uncertainty, we evaluate the initialization-wise uncertainty. The pairwise CUDs between runs (i.e., R0 – R9 for each simulation) are shown in **Figure 11** for the WP, CC, and TC experiments with each *k*-means

variant. The results demonstrate the smaller uncertainty regarding initialization than that owing to the selection of *k*-means algorithms. Particularly, the initialization-wise CUDs are much lower than the algorithm-wise CUD for WP and TC. Meanwhile, in CC, the initialization-wise and algorithm-wise CUDs do not exhibit apparent differences except for $k < 6$ (**Fig. 12**).

The above results demonstrate the effectiveness of CUEF (with CUD as the core concept used within visualization framework including heatmaps or chord diagrams) to quantitively represent/evaluate the uncertainty inherent in clustering outcomes. Heatmaps and chord diagrams are useful in offering intuitive and general comprehension of uncertainty and consensus among the outcomes. CUEF is used to evaluate algorithm-wise (*k*-means variants in this study, but it can be used to compare clustering algorithms, *e.g.*, affinity propagation, DBSCAN, self-organizing map, *etc.*), and initialization-wise uncertainties. Note that there are several techniques for improving the cluster initialization such as *k*-means++ (Arthur and Vassilvitskii, 2007). The result from additional simulations using *k*-means++ shows that the technique could help to reduce, though not wholly remove, the uncertainty regarding initialization.

In addition, clustering uncertainty must be understood in a broader context. It can be induced by also input data. **Figure 10 and 11** shows the consistently higher CUDs for WP than those for CC and TC. It means that WP yields more random clustering outcomes regardless of the algorithm used. In the other words, input data itself can possess uncertain source for clustering. This makes sense because different data have different topologies, which can make them unsuitable or even invalid for a clustering solution. The question of whether it is valid or meaningful to apply a clustering solution to a dataset is more important than how to find the best method of clustering.

In this sense, CUEF can be used to measure the *meaningfulness of clustering application* to a given problem. As the big data era is coming, clustering analysis could play vital role in discovering unseen structures of atmospheric data that are massive and inaccessible to human perception. The last decades have witnessed a wide range of clustering applications from detecting atmospheric regimes/patterns from data (Esteban et al., 2005; Houssos et al., 2008; Spekat et al., 2010; Zeng et al., 2019; Smith et al., 2020) to using these extracted patterns for weather forecasts and climate predictions (Kannan and Ghosh, 2011; Gutiérrez et al., 2013; Le Roux et al., 2018; Pomee and Hertig, 2022) or even reconstructing historical data (Camus et al., 2014). So far, tremendous efforts have been invested in either proposing/improving clustering algorithms or inventing criteria for evaluating the goodness of the results.

A fundamental question is posed about "what is the right thing to do rather than how to do it right?" In other words, it is about how to justify the choice of clustering solution rather than looking for the way to do it right. In this sense, CUEF could help users justify the choice based directly on their data rather than relying on the experiences or literature reviews (select it because others are using it). This value of CUEF is significant in a time of unprecedented expansion of climate data and clustering algorithms, diversifying the needs in data mining. We recommend CUEF as a necessary procedure (or standard) for clustering techniques. Even though the final decision on whether to apply a clustering solution might depend on multiple factors, e.g., the purpose of further analysis, CUEF eventually can support the result explanation and help to robustify the discussion.

6. Summary and remarks

This study proposes (i) a novel *k*-means algorithm primarily for mining climate data and (ii) a clustering-uncertainty evaluation framework. The novel *k*-means algorithm, called S *k*-means, is characterized by its ability to deal with inherent spatiotemporal "structuredness" in climate data. In detail, S *k*-means incorporates the recent innovation in signal recognition regarding structural similarity into the classification scheme, which has been primarily established based on the distance metrics paradigm.

The performance of S *k*-means is evaluated against the other *k*-means variants, C, E, and M *k*-means, i.e., *k*-means using the Pearson correlation coefficient, Euclidean, and Manhattan distance (C, E, and M *k*-means, respectively). Three demonstration tests, i.e., clustering weather patterns (spatial-related data), historical climate change (time series) for long-term recorded weather station data, and best tracks of tropical cyclones (spatiotemporal hybrid), eleven k settings ($k = 2, 4, \ldots, 20$), and for each k and an ensemble of ten randomized initializations, are implemented, resulting in a total of 1320 runs produced to generate robust results.

The quantitative approaches, i.e., similarity distribution (S-D) and general silhouette analysis, are used to evaluate the performance of the algorithms. S-D diagrams were used to diagnose the topological relationship of input datasets in different distance/similarity spaces. The results show that structural similarity groups are likely to have a higher ability to discriminate the data (characteristics that might be useful for clustering) than conventional distance metrics. Regarding the clustering results, the general silhouette analysis shows consistently higher scores for S and C *k*-means compared with E and M *k*-means. The superiority of S *k*-means clustering is followed by C *k*-means clustering. Both S and C *k*-means consistently outperform E- and M-*k*-means. The trade-off between the clustering performance and computational resource requirement is revealed, as S *k*-means requires five to six times more computational time than E *k*-means.

S *k*-means could be promising as a new standard for climate-data clustering/mining, which is a rising research field within the big data context. Nevertheless, certain issues must be noted when interpreting the results of this study. First, as *k*-means clustering is an unsupervised data-mining method, it works under an assumption of no "ground-truth" labeling information. Therefore, there is no absolute reference to define the goodness of the clustering result. In this study, the goodness of the algorithm is evaluated based on an objective calculus approach using the general silhouette analysis/score. Though, this score is free from the Cartesian geometry assumption, thus allowing the algorithms to be compared with non-distance metrics, it is suggested that more evaluation and diversifying clustering problems are needed to gain deeper insight into the algorithm.

Finally, another important contribution of this study is that we built a framework for clustering uncertainty evaluation for the first time, and it is primarily applicable to climate research. The evaluation framework is built on the mutual information concept. This is the first time this concept has been adapted for clustering uncertainty evaluations in the form of the "clustering uncertainty degree" (CUD). CUD measures pairwise discrepancies among clusters, and the collective CUDs provide an overall picture of the consistency/uncertainty of the cluster algorithms. Naturally, CUD can be used to evaluate whether a given problem (input data) is preferable for clustering. In other words, if the cluster algorithm provides higher uncertainty in its outcomes, then it is not appropriate for use, and vice versa. For example, for what shown in this study, the WP problem caused more uncertainty in clustering than the CC and TC problems. Thus, it is questioned about the "meaningfulness" of the clustering application for WP compared with CC and TC. We expect this clustering-uncertainty-evaluation framework will change the conventional agenda of data clustering by adding a procedure to evaluate its application's meaningfulness/effectiveness for a given data.

Code availability

The exact version of the model used to produce the results in this study and the input data and scripts used to run the model and plot all the simulations presented in this paper have been archived on Github (https://github.com/doan-van/S-k-means) or Zenodo (https://zenodo.org/record/6976609).

Author contribution

Quang-Van Doan designed the model and developed the model code. Toshiyuki Amagasa, Thanh-Ha Pham, Takuto Sato, Fei Chen and Hiroyuki Kusaka helped to design the test experiments. Thanh-Ha Pham, and Takuto Sato helped to analyze the results. Quang-Van Doan prepared the manuscript with contributions from all co-authors.

Competing interests

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

Acknowledgements

The first author, Quang-Van Doan, is sponsored by JSPS KAKENHI Grant Nos. 20K13258, JSPS KAKENHI Grant Nos. 19H01155, and JSPS KAKENHI Grant Nos. 21K03656.

References

de Amorim, R. C.: A Survey on Feature Weighting Based K-Means Algorithms, J. Classif., 33, 210–242, https://doi.org/10.1007/s00357-016-9208-4, 2016.

Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, Stanford, 2006.

Arthur, D. and Vassilvitskii, S.: k-means++: the advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, USA, 1027–1035, 2007.

Barua, D. K.: Beaufort Wind Scale, in: Encyclopedia of Coastal Science, edited by: Finkl, C. W. and Makowski, C., Springer International Publishing, Cham, 315–317, https://doi.org/10.1007/978-3-319-93806-6_45, 2019.

Bradley, P. S. and Fayyad, U. M.: Refining Initial Points for K-Means Clustering, 91–99, 1998.

Camus, P., Menéndez, M., Méndez, F. J., Izaguirre, C., Espejo, A., Cánovas, V., Pérez, J., Rueda, A., Losada, I. J., and Medina, R.: A weather-type statistical downscaling framework for ocean wave climate, J. Geophys. Res. Oceans, 119, 7389–7405, https://doi.org/10.1002/2014JC010141, 2014.

Chan, E. Y., Ching, W. K., Ng, M. K., and Huang, J. Z.: An optimization algorithm for clustering using weighted dissimilarity measures, Pattern Recognit., 37, 943–952, https://doi.org/10.1016/j.patcog.2003.11.003, 2004.

Choi, K.-S., Cha, Y.-M., and Kim, T.-R.: Cluster analysis of tropical cyclone tracks around Korea and its climatological properties, Nat. Hazards, 64, 1–18, https://doi.org/10.1007/s11069-012-0192-7, 2012.

Cordeiro de Amorim, R. and Mirkin, B.: Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, Pattern Recognit., 45, 1061–1075, https://doi.org/10.1016/j.patcog.2011.08.012, 2012.

Doan, Q.-V., Kusaka, H., Sato, T., and Chen, F.: S-SOM v1.0: a structural self-organizing map algorithm for weather typing, Geosci. Model Dev., 14, 2097–2111, https://doi.org/10.5194/gmd-14-2097-2021, 2021.

Eltibi, M. F. and Ashour, W. M.: Initializing k-means clustering algorithm using statistical information, Int. J. Comput. Appl., Volume: 29, Number: 7, 2011.

Esteban, P., Jones, P. D., Martín-Vide, J., and Mases, M.: Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees, Int. J. Climatol., 25, 319–329, https://doi.org/10.1002/joc.1103, 2005.

Fahim, A. M., Salem, A. M., Torkey, F. A., and Ramadan, M. A.: An efficient enhanced k-means clustering algorithm, J. Zhejiang Univ.-Sci. A, 7, 1626–1633, https://doi.org/10.1631/jzus.2006.A1626, 2006.

Forgy, E. W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications, biometrics, 21, 768–769, 1965.

Gibson, P. B., Perkins-Kirkpatrick, S. E., Uotila, P., Pepler, A. S., and Alexander, L. V.: On the use of self-organizing maps for studying climate extremes, J. Geophys. Res. Atmospheres, 122, 3891–3903, https://doi.org/10.1002/2016JD026256, 2017.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing Statistical Downscaling Techniques for Their Robust Application under Climate Change Conditions, J. Clim., 26, 171–188, https://doi.org/10.1175/JCLI-D-11-00687.1, 2013.

Han, F. and Szunyogh, I.: A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability, Mon. Weather Rev., 146, 1303–1318, https://doi.org/10.1175/MWR-D-17-0040.1, 2018.

Hassani, M. and Seidl, T.: Using internal evaluation measures to validate the quality of diverse stream clustering algorithms, Vietnam J. Comput. Sci., 4, 171–183, https://doi.org/10.1007/s40595-016-0086-9, 2017.

Holten, D.: Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data, IEEE Trans. Vis. Comput. Graph., 12, 741–748, https://doi.org/10.1109/TVCG.2006.147, 2006.

Houssos, E. E., Lolis, C. J., and Bartzokas, A.: Atmospheric circulation patterns associated with extreme precipitation amounts in Greece, in: Advances in Geosciences, 9th EGU Plinius Conference on Mediterranean Storms (2007) - 9th Plinius Conference on Mediterranean Storms, Varenna, Italy, 10–13 September 2007, 5–11, https://doi.org/10.5194/adgeo-17-5-2008, 2008.

Huang, J. Z., Ng, M. K., Rong, H., and Li, Z.: Automated variable weighting in k-means type clustering, IEEE Trans. Pattern Anal. Mach. Intell., 27, 657–668, https://doi.org/10.1109/TPAMI.2005.95, 2005.

Jancey, R. C.: Multidimensional group analysis, Aust. J. Bot., 14, 127–130, 1966.

Kannan, S. and Ghosh, S.: Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output, Stoch. Environ. Res. Risk Assess., 25, 457–474, https://doi.org/10.1007/s00477-010-0415-y, 2011.

Kantardzic, M.: Data mining: concepts, models, methods, and algorithms, John Wiley & Sons, 2011.

Katsavounidis, I., Jay Kuo, C.-C., and Zhang, Z.: A new initialization technique for generalized Lloyd iteration, IEEE Signal Process. Lett., 1, 144–146, https://doi.org/10.1109/97.329844, 1994.

Khan, S. S. and Ahmad, A.: Cluster center initialization algorithm for K-means clustering, Pattern Recognit. Lett., 25, 1293–1302, https://doi.org/10.1016/j.patrec.2004.04.007, 2004.

Kim, H.-K. and Seo, K.-H.: Cluster Analysis of Tropical Cyclone Tracks over the Western North Pacific Using a Self-Organizing Map, J. Clim., 29, 3731–3751, https://doi.org/10.1175/JCLI-D-15-0380.1, 2016.

Kim, H.-S., Kim, J.-H., Ho, C.-H., and Chu, P.-S.: Pattern Classification of Typhoon Tracks Using the Fuzzy c-Means Clustering Method, J. Clim., 24, 488–508, https://doi.org/10.1175/2010JCLI3751.1, 2011.

Lai, J. Z. C. and Huang, T.-J.: Fast global k-means clustering using cluster membership and inequality, Pattern Recognit., 43, 1954–1963, https://doi.org/10.1016/j.patcog.2009.11.021, 2010.

Le Roux, R., Katurji, M., Zawar-Reza, P., Quénol, H., and Sturman, A.: Comparison of statistical and dynamical downscaling results from the WRF model, Environ. Model. Softw., 100, 67–73, https://doi.org/10.1016/j.envsoft.2017.11.002, 2018.

Lloyd, S. P.: Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM, IEEE Trans Inf. Theor19571982, 18, 11, 1957.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 281–297, 1967.

Mo, R., Ye, C., and Whitfield, P. H.: Application potential of four nontraditional similarity metrics in hydrometeorology, J. Hydrometeorol., 15, 1862–1880, 2014.

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R.: Climate Data Challenges in the 21st Century, Science, 331, 700–702, https://doi.org/10.1126/science.1197869, 2011.

Pelleg, D.: Extending K-means with efficient estimation of the number of clusters in ICML, in: Proceedings of the 17th international conference on machine learning, 277–281, 2000.

Perez, J., Mexicano, A., Santaolaya, R., Hidalgo, M., Moreno, A., and Pazos, R.: Improvement to the K-Means algorithm through a heuristics based on a bee honeycomb structure, in: 2012 Fourth World Congress on Nature and Biologically Inspired Computing (NaBIC), 2012 Fourth World Congress on Nature and Biologically Inspired Computing (NaBIC), 175–180, https://doi.org/10.1109/NaBIC.2012.6402258, 2012.

Pérez-Ortega, J., Almanza-Ortega, N. N., Vega-Villalobos, A., Pazos-Rangel, R., Zavala-Díaz, C., and Martínez-Rebollar, A.: The k-means algorithm evolution, in: Introduction to Data Science and Machine Learning, IntechOpen, 2019.

Pomee, M. S. and Hertig, E.: Precipitation projections over the Indus River Basin of Pakistan for the 21st century using a statistical downscaling framework, Int. J. Climatol., 42, 289–314, https://doi.org/10.1002/joc.7244, 2022.

Romano, S., Vinh, N. X., Bailey, J., and Verspoor, K.: Adjusting for chance clustering comparison measures, J. Mach. Learn. Res., 17, 4635–4666, 2016.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., 20, 53–65, 1987.

Selim, S. Z. and Ismail, M. A.: K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 81–87, https://doi.org/10.1109/TPAMI.1984.4767478, 1984.

Smith, E. T., Lee, C. C., Barnes, B. B., Adams, R. E., Pirhalla, D. E., Ransibrahmanakul, V., Hu, C., and Sheridan, S. C.: A Synoptic Climatological Analysis of the Atmospheric Drivers of Water Clarity Variability in the Great Lakes, J. Appl. Meteorol. Climatol., 59, 915–935, https://doi.org/10.1175/JAMC-D-19-0156.1, 2020.

Spekat, A., Kreienkamp, F., and Enke, W.: An impact-oriented classification method for atmospheric patterns, Phys. Chem. Earth Parts ABC, 35, 352–359, https://doi.org/10.1016/j.pce.2010.03.042, 2010.

Su, T. and Dy, J. G.: In search of deterministic methods for initializing K-means and Gaussian mixture clustering, Intell. Data Anal., 11, 319–338, https://doi.org/10.3233/IDA-2007-11402, 2007.

Sydow, A.: Tou, JT/Gonzalez, RC, Pattern Recognition Principles, London-Amsterdam-Dom Mills, Ontario-Sydney-Tokyo. Addison-Wesley Publishing Company. 1974. 378 S., $19, 50., Z. Angew. Math. Mech., 57, 353–354, 1977.

Vinh, N. X. and Epps, J.: A novel approach for automatic number of clusters detection in microarray data based on consensus clustering, in: 2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering, 84–91, 2009.

Vinh, N. X., Epps, J., and Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, J. Mach. Learn. Res., 11, 2837–2854, 2010.

Wang, Z. and Bovik, A. C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures, IEEE Signal Process. Mag., 26, 98–117, 2009.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process., 13, 600–612, 2004.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D.: Top 10 algorithms in data mining, Knowl. Inf. Syst., 14, 1–37, https://doi.org/10.1007/s10115-007-0114-2, 2008.

Zeng, S., Vaughan, M., Liu, Z., Trepte, C., Kar, J., Omar, A., Winker, D., Lucker, P., Hu, Y., Getzewich, B., and Avery, M.: Application of high-dimensional fuzzy *k*-means cluster analysis to CALIOP/CALIPSO version 4.1 cloud–aerosol discrimination, Atmospheric Meas. Tech., 12, 2261–2285, https://doi.org/10.5194/amt-12-2261-2019, 2019.

List of tables

Table 1. Statistical metrices of S-distributions for three demonstration input datasets, i.e., weather pattern (WP), climate change (CC), and tropical cyclone (TC). The different distance/similarity measures are structural similarity (S-SIM), the Pearson correlation coefficient (COR), Euclidean distance (ED) and Manhattan distance (MD). Statistical measures include the mean (Mean), standard deviation (STD), skewness (SKEW), kurtosis (KUR) and Shannon entropy (ENTROPY)

	WP				*CC*				*TC*			
	S-SIM	COR	ED	MD	S-SIM	COR	ED	MD	S-SIM	COR	ED	MD
Mean	0.68	0.71	0.67	0.68	0.71	0.81	0.66	0.65	0.81	0.87	0.65	0.69
STD	0.18	0.19	0.11	0.11	0.20	0.13	0.12	0.13	0.14	0.11	0.15	0.13
SKEW	-0.66	-0.81	-0.73	-0.74	-1.08	-1.25	-0.65	-0.67	-1.10	-1.67	-0.46	-0.59
KUR	-0.18	0.00	0.58	0.64	0.97	1.79	0.59	0.58	1.15	3.31	-0.32	0.03
ENTROPY	2.83	2.79	2.19	2.16	2.83	2.29	2.32	2.36	2.30	1.80	2.57	2.45

List of figures

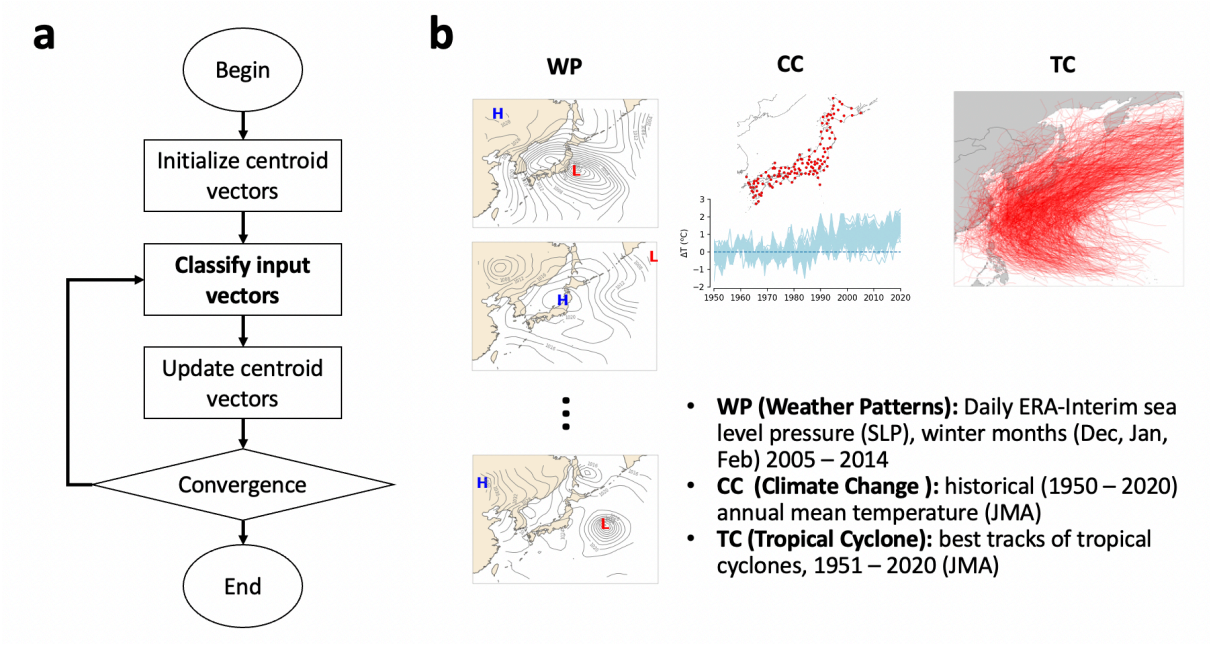


Fig. 1 Illustration of the *k*-means clustering algorithm (a) and three demonstration experiments (b). Demonstration experiments include clustering weather patterns (WPs) in terms of daily ERA-Interim sea level pressure (SLP) during winter months (December, January, and February) for ten years 2005 – 2014 over the Japan region; clustering climate change (CC) in terms of historical (1951 – 2020) annual mean temperature collected from in situ weather stations in Japan; and clustering best tracks of tropical cyclones that passed the Northwest Pacific region from 1951 – 2020. Data were obtained from the JMA

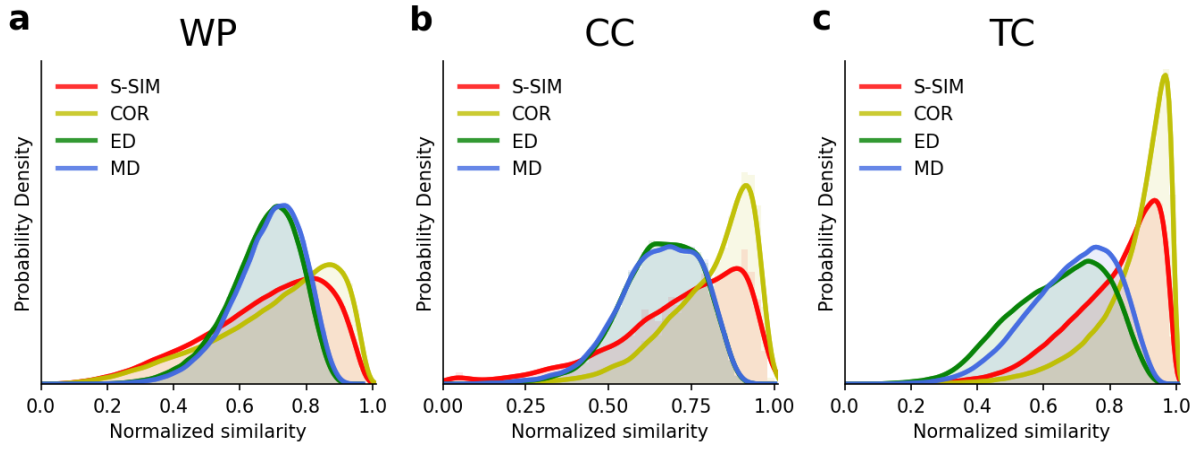


Fig. 2 Comparison of the S-distributions of normalized pairwise similarity using the structural similarity (S-SIM), the Pearson correlation coefficient (COR) the Euclidean distance (ED) and the Manhattan distance (MD) for three demonstration experiments: WP, CC, and TC. With a population size of N, $\frac{N(N-1)}{2}$ values of pairwise similarity are observed because S-SIM, COR, ED and MD are symmetric measures and self-similarity is excluded. Values are normalized from 0 to 1. The maximum similarity is 1, which corresponds to completely similar, and the minimum similarity is 0, which corresponds to the lowest pairwise similarity.

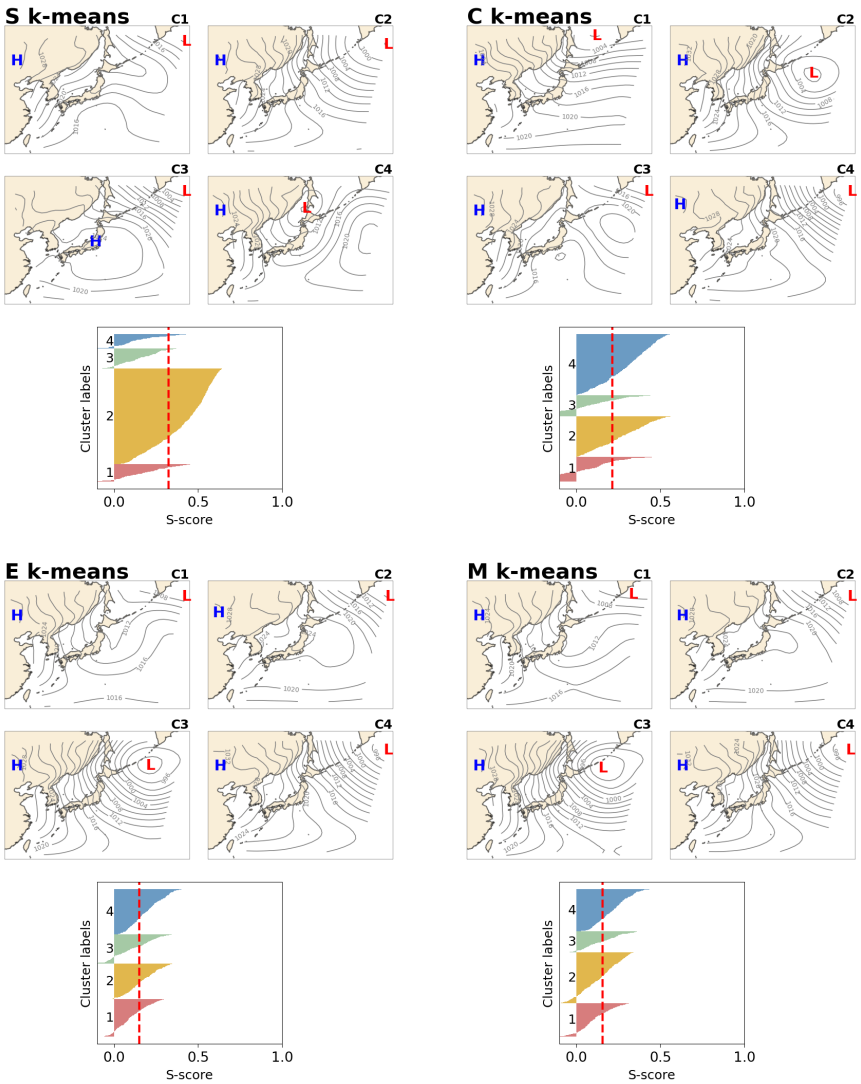


Fig. 3 Result for the WP experiment. The winter SLP pattern revealed by S, C, E, and M *k*-means with $k = 4$. "H" indicates the location of the high, and "L" indicates the location of the low. General silhouette analysis results are shown below the maps, where the x-axis indicates the score and the y-axis presents the labels of clusters numbered 1 – 4. Input data are ERA-Interim SLP data, which were re-gridded to Cartesian coordinates with a resolution of 200 x 200 km and grid size of 35 x 35. Daily data for December, January, and February collected over ten year 2005 – 2014 were used.

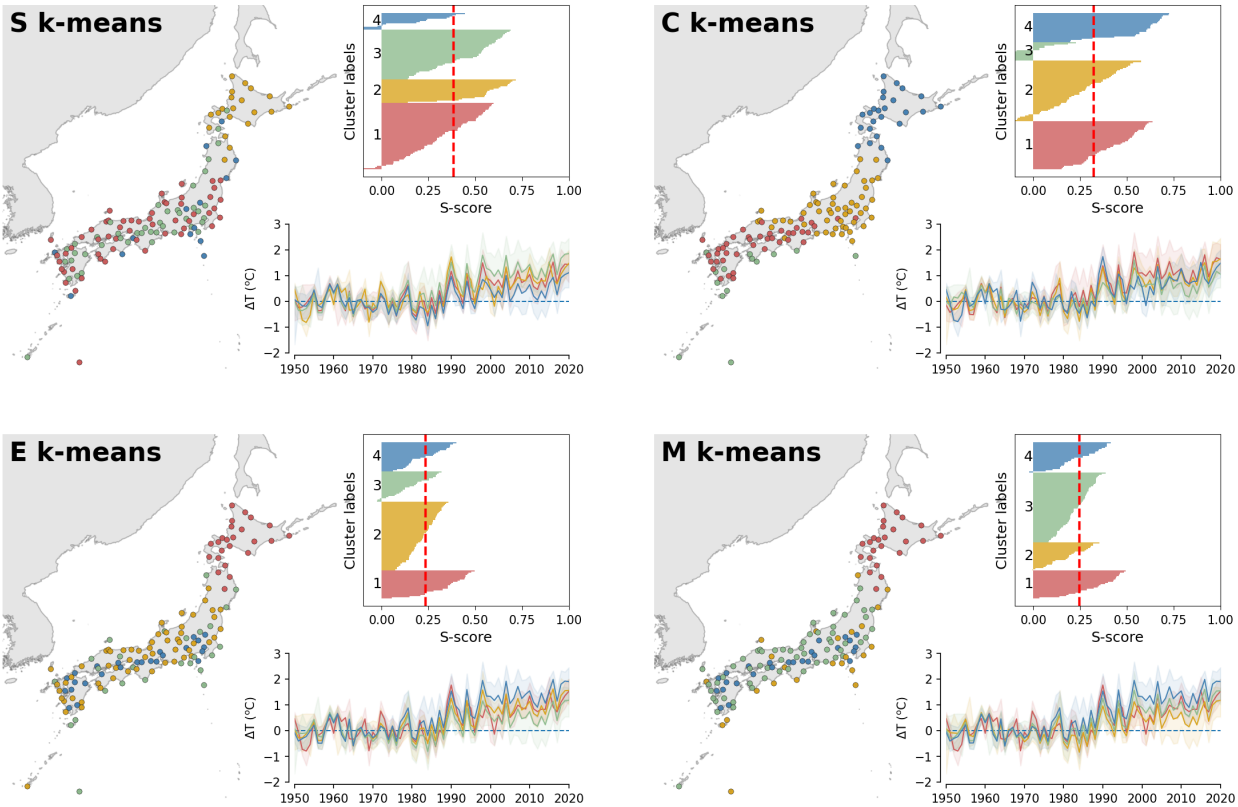


Fig. 4 Result for the CC experiment for clustering of climate change (temperature increase) time series over 134 weather stations over the entirety of Japan. Patterns were revealed by S, C, E, and M *k*-means, with $k = 4$. Input data correspond to annual mean data collected over 70 years from 1951 – 2020 (subtracted by the mean of the first 30 years) and observed temperature achieved at in situ weather stations (dots in map) operated by the JMA. Time series of centroids and input vectors are shown in below panels together with general silhouette analysis results, where the x-axis indicates the score (S-score) and the y-axis presents the labels of clusters numbered 1 – 4.

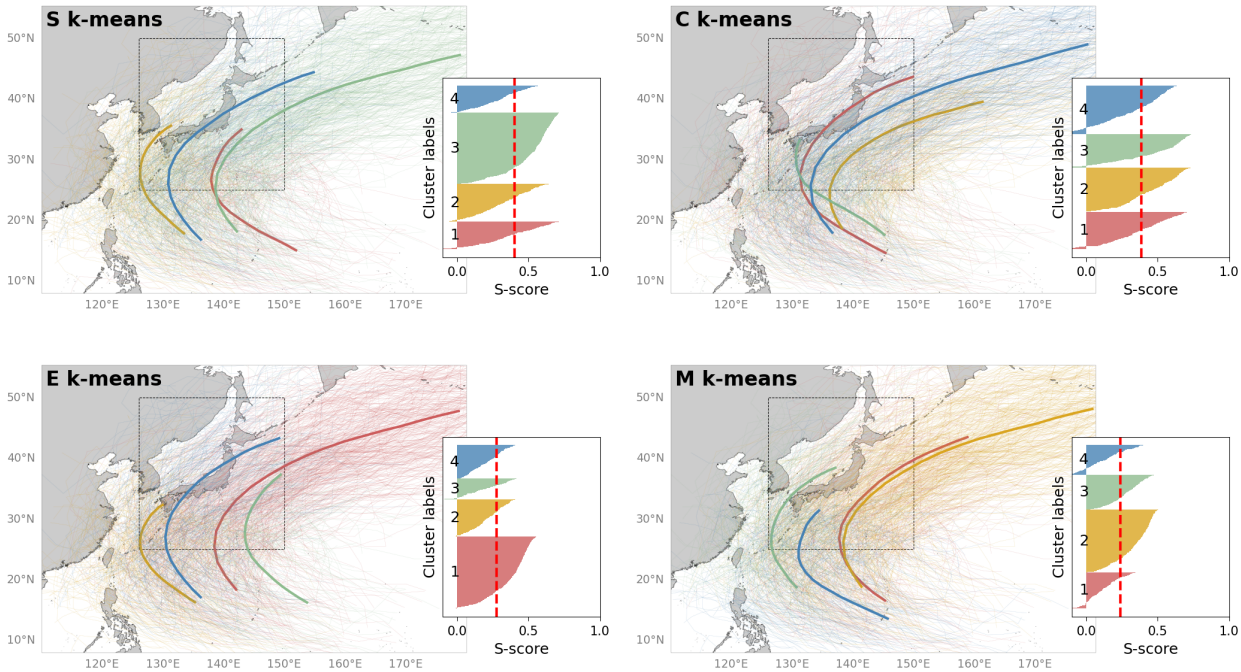


Fig. 5 Results of the TC experiment for clustering tropical cyclone paths. The pattern was revealed by S, C, E, and M *k*-means, with $k = 4$. Input data are the best TC tracks obtained by the JMA from 1951 – 2020. Only TCs that passed the dashed box in the map are used to feed the *k*-means. Thus, a total of 863 TC tracking data points are used. The left side of each panel show the general silhouette analysis results, where the x-axis indicates the score (S-score) and y-axis presents the labels of clusters numbered 1 – 4. The centroid TC path is illustrated by the bold line, and the color is consistent with that in the silhouette diagram.

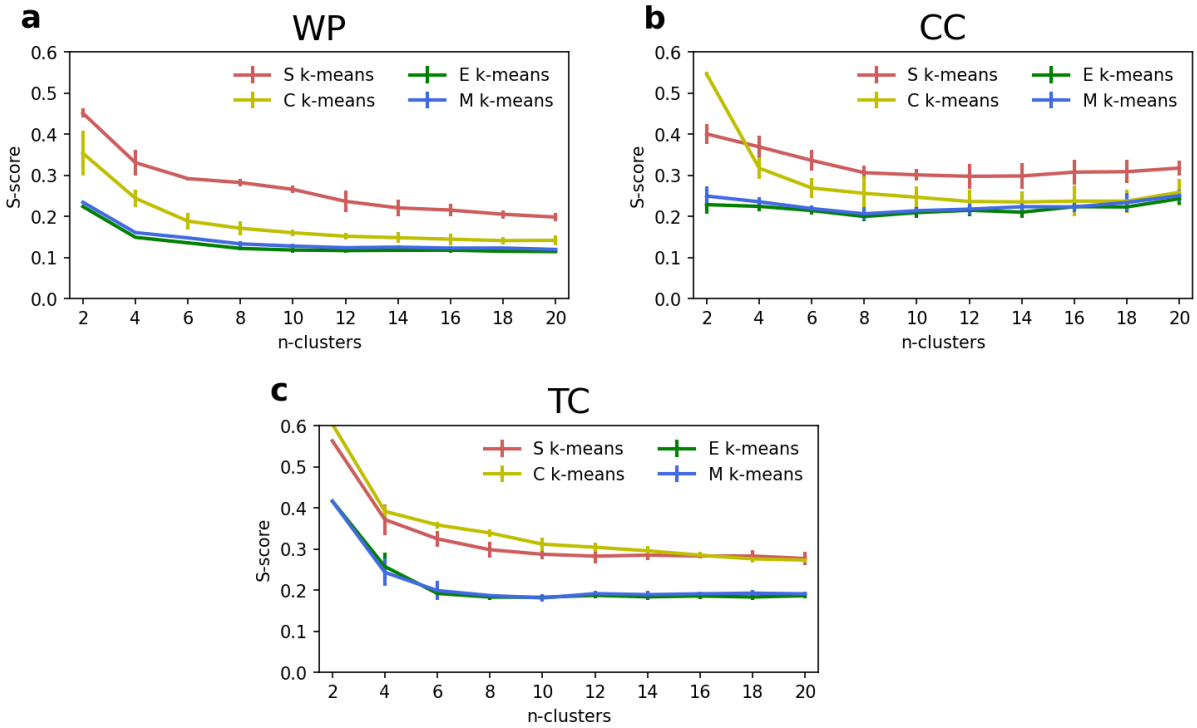


Fig. 6 Comparison of the average silhouette score (S-score) of S, C, E, and M *k*-means for k = 2, 4, …, 20 for three demonstration experiments: WP (a), CC (b) and TC (c). The uncertainty range in each line indicates the standard deviations of the scores among 10 runs with randomized initializations.

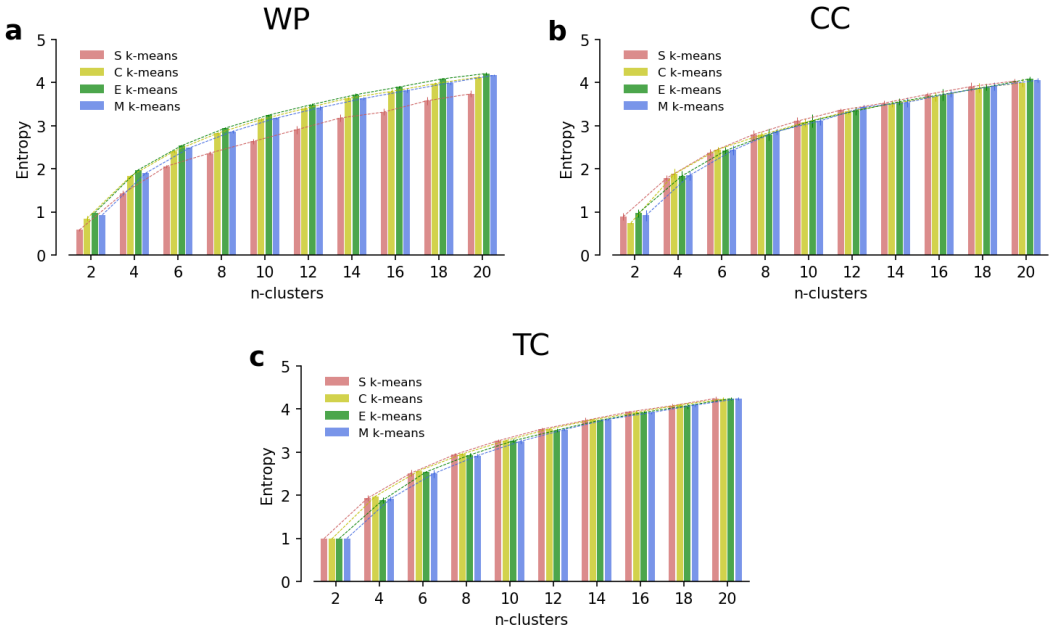


Fig. 7 Shannon entropy of clustering results. Comparison of the average silhouette score (S-score) of S, C, E, and M k-means for k = 2, 4, …, 20 for three demonstration experiments: WP (a), CC (b) and TC (c). The uncertainty range in each line indicates the standard deviations of the scores among 10 runs with randomized initializations.

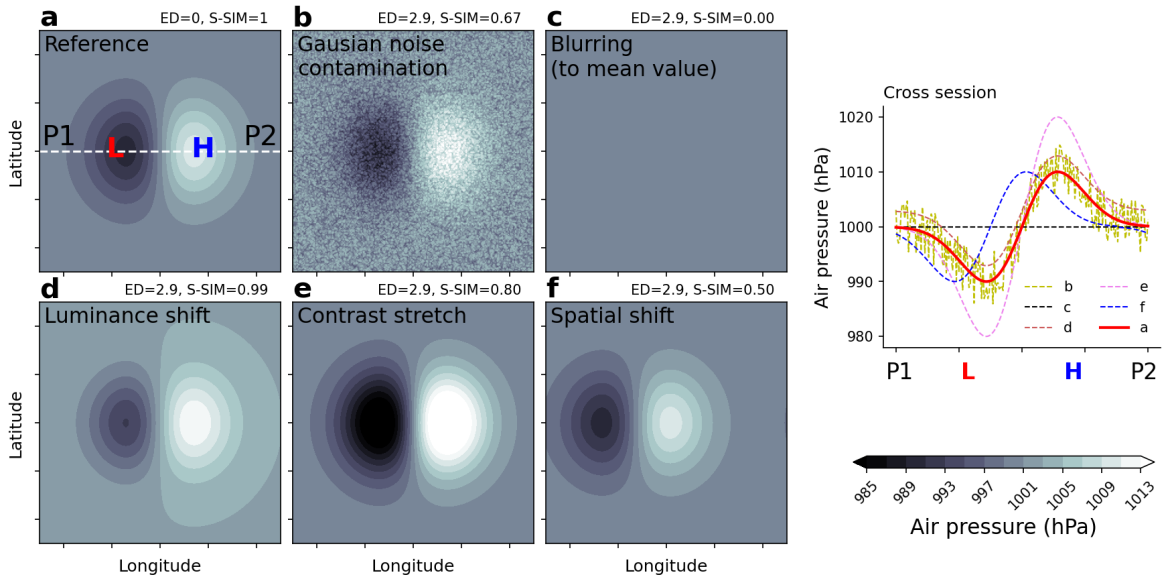


Fig. 8 Imagination air pressure patterns. Subpanels are the reference (a), Gaussian noise contamination (b), blurring (to mean value) (c), luminance shift (d), contrast stretch (e), and spatial shift (f). The ED (Euclidean distance) and S-SIM (structural similarity) values shown above each panel are those calculate to the reference one (a). The rightmost subpanel shows the cross session (between two points P1 and P2 in a)) with L, H indicates the location of imagination Low and High air pressure extrema.

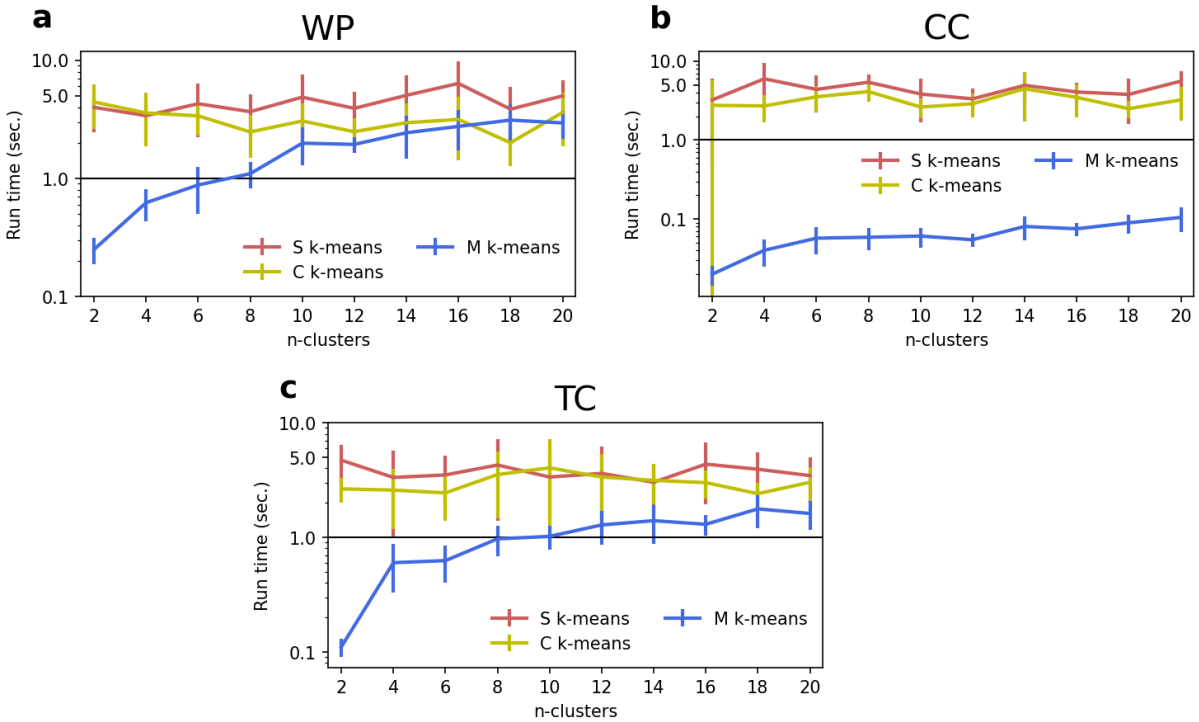
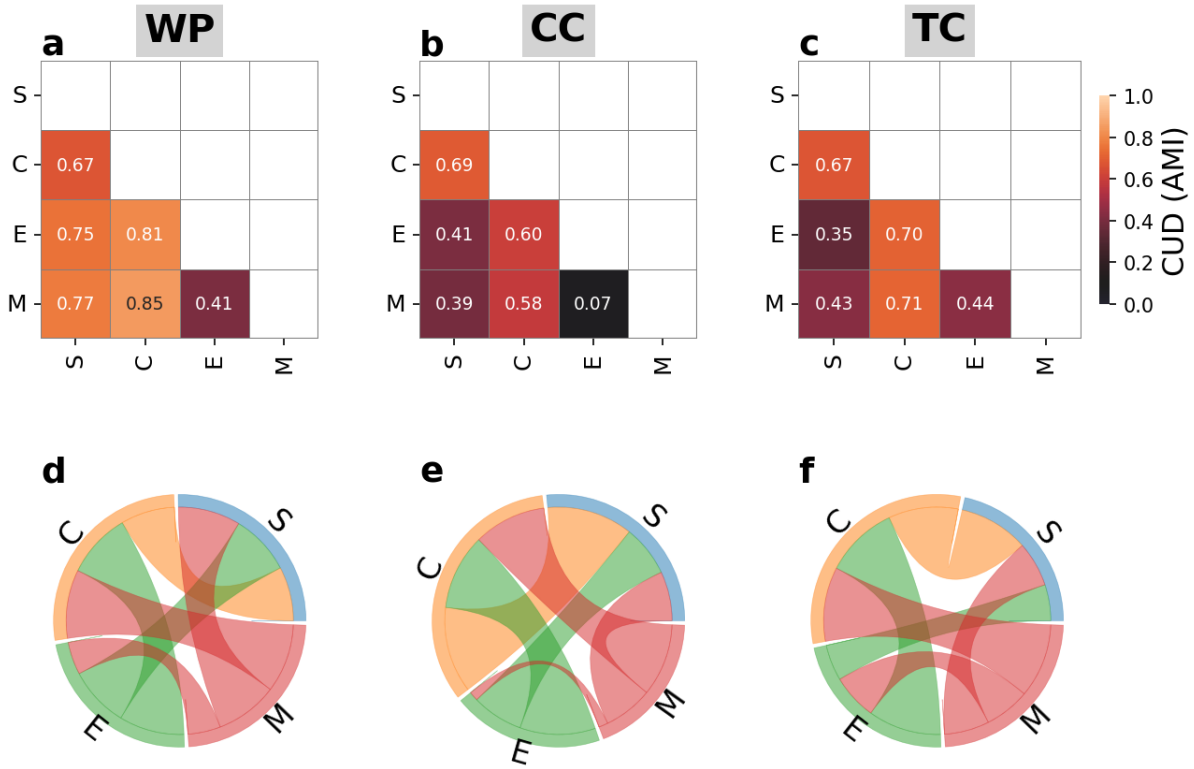


Fig. 9 Comparison of the run time (in sec) of S, C, E, and M *k*-means for k = 2, 4, …, 20 for three demonstration experiments: WP (a), CC (b) and TC (c). The uncertainty range in each line indicates the standard deviation of the scores among 10 runs with randomized initializations. Note that the y axis is logarithmically rescaled.



650 **Fig. 10 Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between clustering results from different *k*-means algorithms, i.e., S, C, E, and M *k*-means, for different demo experiments: WP, CC, and TC. (a, b, c) CUD in heatmaps, and (d, e, f) visualization of the interconnection using the chord diagrams. Note that the results are from the configuration with $k = 4$ and the first initialization run.**

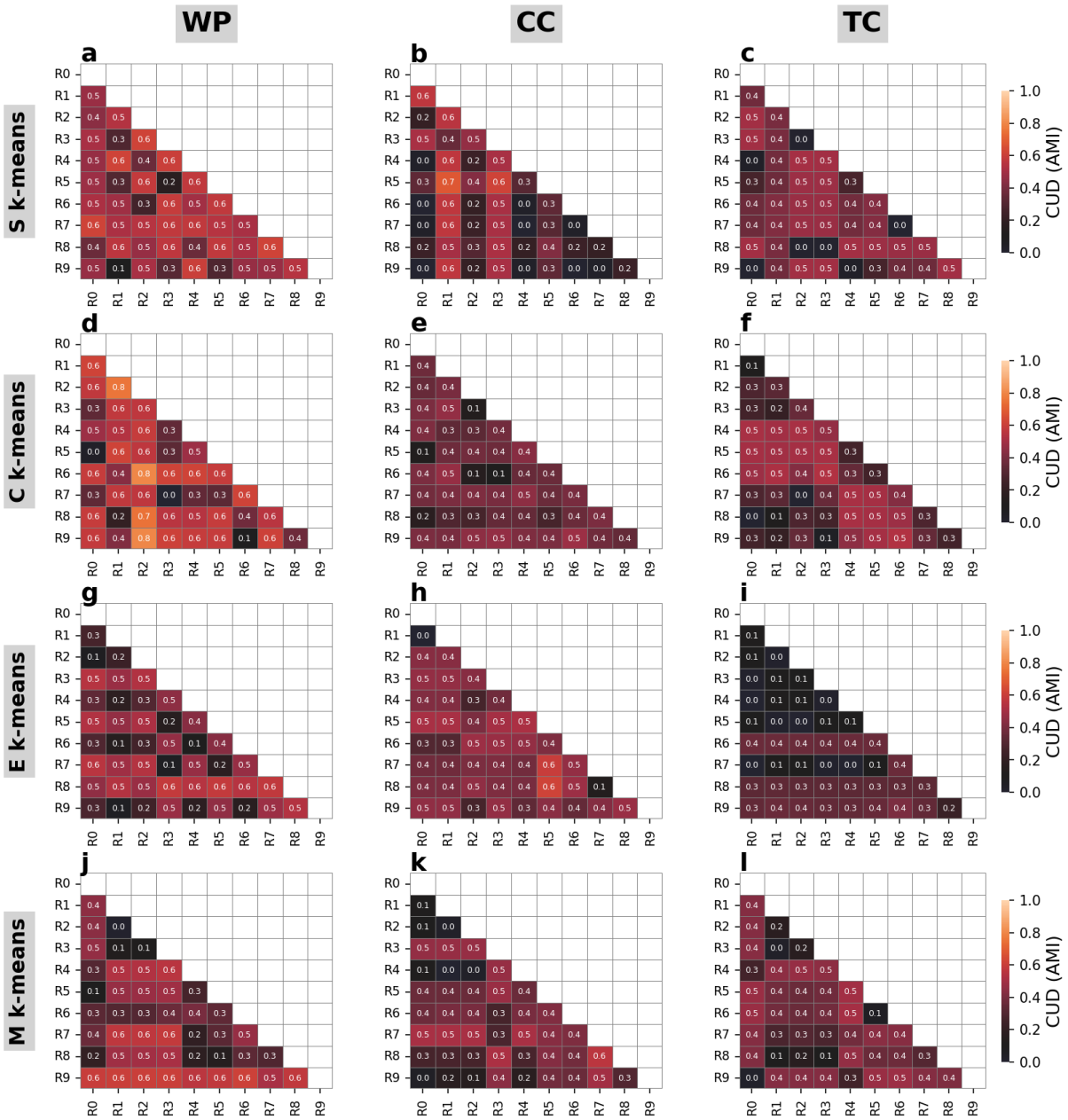

WP
CC
TC
S k-means
C k-means
E k-means
M k-means
a
b
c
d
e
f
g
h
i
j
k
l
R0
R1
R2
R3
R4
R5
R6
R7
R8
R9
CUD (AMI)
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 11 Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between the clustering results from different runs (10 runs indicated by R0, R1, …, R9) of different *k*-means algorithms, i.e., S, C, E, and M *k*-means (rows), for different demo experiments: WP, CC, and TC (columns). Note that the results are from the configuration with $k = 4$ and the first initialization run.

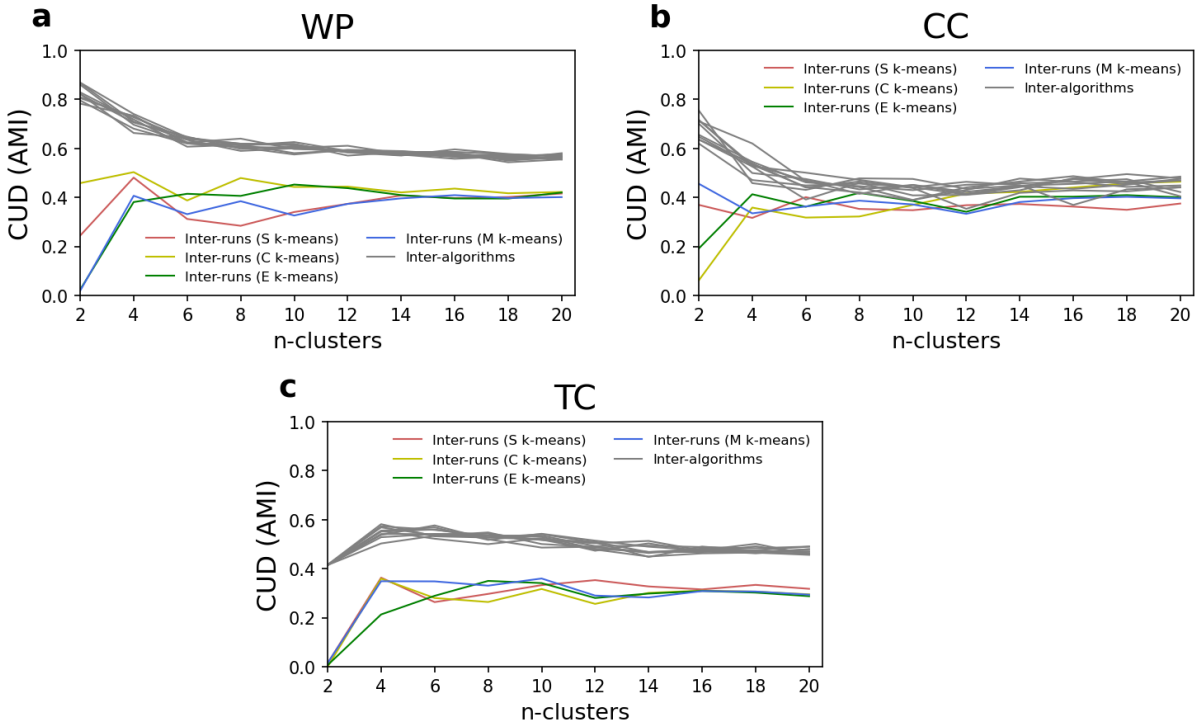


Fig. 12 Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between the clustering results from different runs (10 runs indicated by R0, R1, …, R9) of different *k*-means algorithms, i.e., S, C, E, and M *k*-means (rows), for different demo experiments: WP, CC, and TC (columns). Note that the results are from the configuration $k = 2, 4, \ldots, 20$.