Note that this is the responses to the reviewer 1's comment. The revised manuscript is not included here, because it is not required in this stage.

Response to Reviewer's comment

Manuscript by Doan et. al. presents a S k-means clustering framework, improving on standard k-means clustering, and demonstrate their application to several climate datasets.

Manuscript presents a methods focused study, which however lacks sufficient discussion to demonstrate the benefits of the proposed algorithmic improvements to standard k-means algorithm. Section "Results and Discussions" focus more on Results and less on Discussion, which is the critical weakness of the manuscript in its current form.

We appreciate the reviewer for his/her critical, and insightful comments, which are very helpful in improving this manuscript. We have addressed all the comments point-by-point adding appropriate discussions, some of which are based on current results and some on additional tests and analyses. In summarization, additional discussions are to address:

a) how can S k-means capture the "structuredness" of input data,
b) uniqueness, and new insight that S k-means enables, quantified by the Shannon entropy,
c) the novelty of the clustering uncertainty evaluation in a broader context,
d) additional explanations regarding methods.

We hope the reviewer satisfy with the responses.

1. Manuscript is missing several key references from the reference list. Wang et. al. 2004, Wang and Bovik, 2009 Mo et al., 2014; Han and Szunyogh, 2018; Doan et al., 2021

We have added these references to the revised manuscript.

2. One of the motivation for the proposed work, as discussed in introduction, is to mine the unique "structuredness" of temporal and spatial climate data (Line 67-81). However, rest of the manuscript focused on comparison of various clustering methods based on Silhouette scores, uncertainty degree etc. Proposed S k-means consistently shows better scores than the other methods, but if and how it better captures the "structuredness" of the data need to be discussed, since that's the key contribution of the study.

We agree. We have added discussion on how S k-means captures the "structuredness" of the data into the manuscript. We focus on the ability of the algorithm to distinguish the difference between "imagination" data, which are generated for intuitive comprehension. An additional figure (Figure R1, which corresponds to Figure 8 in the revised manuscript) has also been added to the manuscript.

"To understand how S k-means clusters data, it is essential to see how the algorithm recognizes the similarity between objects. For intuitive comprehension, we generated "imagination" two-dimensional air pressure patterns and showed them in Figure 8. In the figure, the reference pattern (a) illustrates two air pressure extrema (Low and High) located symmetrically on the left and right sides. Other patterns for comparison are Gaussian noise contamination (b), blurring (identically distributed pattern) (c), luminance shift (d), contrast stretch (e), and spatial shift (f). The patterns other than the reference are intentionally generated so that those ED (Euclidean distance) to the reference are identical (=2.9). With S-SIM, the similarities are ranked in 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 example demonstrates well the ability of S-SIM in recognizing the difference between two-dimensional patterns, which ED cannot do. More interestingly, the similarities ranked by S-SIM fits well with human perception. For example, the similarity between c and a is 0 (no similarity). With using S-SIM, S k-means can avoid by-chance centroid assignments, which ED-based k-means might not. Though this example shows the two-dimensional data, it could be the same for the time series, where a temporal instead of a spatial relationship characterizes the structuredness of data."

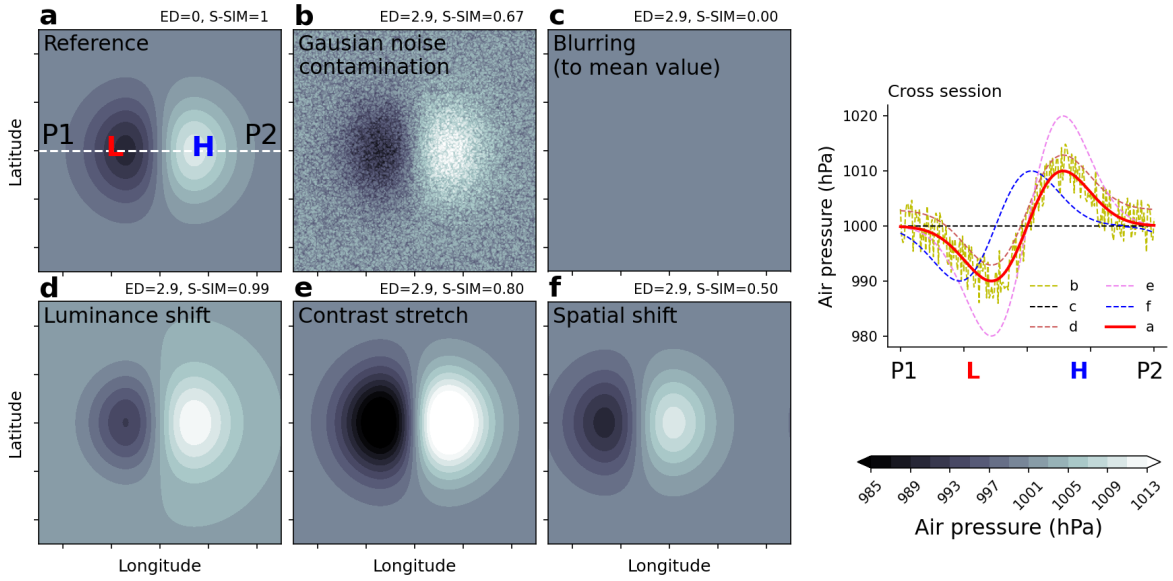

Figure R1. Imagination air pressure patterns. Subpanels are the reference (a), Gaussian noise contamination (b), blurring (to mean value) (c), luminance shift (d), contrast stretch (e), and spatial shift (f). The ED (Euclidean distance) and S-SIM (structural similarity) values shown above each panel are those calculated to reference one (a). The rightmost subpanel shows the cross-session (between two points P1 and P2 in a)) with L, H indicating the location of imagination Low and High air pressure extrema.

3. Structural similarity metric (Section 2.2) is the most important part of the study. However, several symbols/terms in equations 2, 3 and on lines 142-145 are not defined or explained. In particular the equations for luminance, contrast and structure. And the cited articles (Wang et. al. 2004, Wang and Bovik, 2009) that developed the similarity metrics are missing from the reference list. That makes it difficult to understand the similarity metric. Aside from describing

equations for S-SIM, there are disussions, in methods section or later, as to how these structural metrics capture the spatial and temporal structuredness of climate data.

We have added more the explanation for equations 2, 3 clarifying the concept of luminance, contrast and structure similarities. All symbols/terms and notations have been checked to assure that they are all appropriated defined. Regarding the second part of the question, we believe that it is appropriately addressed in the above response to the previous comment. We kindly ask the reviewer to go back to check it.

4. Discussion of clustering results in Section 5.2 is very high level. Question remains, aside from slightly higher scores what unique and new insights does the S k-means clustering enabled?

We have added new insight into S *k*-means clustering results, focusing on its unique characteristics. To support the arguments, we use the Shannon entropy concept and calculate clustering entropy which is shown in Figure R2 (corresponding to Figure 7 in the manuscript). Following are the details of the discussion, which have been also added in the revised manuscript.

> "Insight into the uniqueness of S k-means has great practical implications/instructions for ending users. As explained above, clustering patterns alone are significant only after the physical meaning is assigned or used for a practical purpose like a prediction. Doing either 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 discuss the S *k*-means performance. Looking carefully at Figure 3, one might realize an anomaly of S k-means compared with the others in Silhouette plots. S k-means is likely to generate what we call "high-ordered" clustering, i.e., one dominant weather pattern (the larger group size) besides several non-dominant ones (the smaller group size). The same trend is consistent with different k settings. This finding agrees well with prior knowledge of Japanese winter weather patterns. Recall that winter in Japan, due to its specific location, is characterized by dominated winter-type pattern (Low in the east and High in the west), which is well recognized by the meteorological research community and local people. The finding leads to hypotheses: (i) Does S k-means perform clustering closer to human perception than other algorithms? (ii) Is it an intrinsic property of S k-means that tends to generate "highly-ordered" clustering?
>
> To examine the hypotheses mentioned above, we adopt Shannon entropy to quantify the "orderliness" of clustering (see Method part for more details). The results, illustrated in Figure 7, show a good agreement with calculated clustering entropy values with human intuition. That is, lower entropy (highly ordered clustering) is consistently seen in S k-means, rather than other algorithms, for the WP experiment (Fig. 7a). But this trend is not confirmed for other experiments like CC, and TC (Fig. 7b, c). Because we did not ensure it for CC and TC, we can eliminate the second hypothesis, i.e., "highly-ordered clustering" is not an intrinsic property of S k-means. Now we have the first hypothesis remaining, i.e., S k-means is likely to generate clustering closer to human perception. Note it is still too early to conclude about the superiority of S k-means based on only

what is shown here. More investigation with a wide range of data types (with well-defined prior knowledge) is needed to gain conclusive insight into the algorithm."

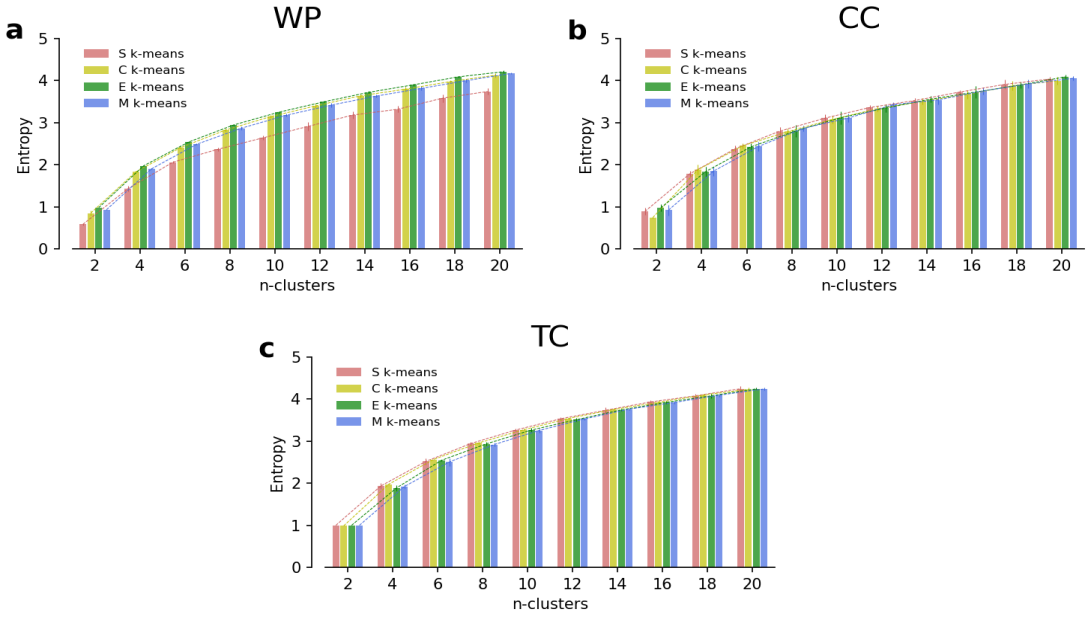

*Figure 2. Shannon entropy of clustering results. Comparison of the average silhouette score (S-score) of S, C, E, and M k-means for k = 2, 4, ..., 20 for three demonstration experiments: WP (a), CC (b), and TC (c). The uncertainty range in each line indicates the standard deviations of the scores among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

5. I am glad to see S k-means being compared with three other k-means variants. They were all run for a 11 different 'k' and with 10 random ensembles each, resulting in a toal of 1320 clustering runs. BUT were all four k-means variants run with exactly the same random starting centroids for the purpose of comparison? It's important to do that for a fair comparison. Also, was a consistent convergence criteria used for all four methods? Converge criteria was mentioned on Lines 128-129, but what criteria was used in the study never discussed.

The four k-means variants have been conducted with the randomized centroids each time. It is because we aimed to compare four algorithms as they are as integrated systems. However, we also understand well the concern of the reviewer. To address the reviewer's concern and assure that our conclusion in the original manuscript is robust, we have run additional experiments assuming the same random starting centroids. In detail, the extra 132 run (3 experiments x 11 *k* settings x 4 *k*-means variants) has been conducted based on 33 pre-defined starting centroid sets (3 experiments x 11 *k* settings).

The quick conclusion is the uncertainty (related to clustering algorithm selection) remains even though the same starting centroids are used. Compare Figure 3 (additional runs) and Figure 4 (original runs) for *k* = 4; the CUD in clustering results from the four *k*-means variants is confirmed at the same level regardless the fully randomized initialization, or identical predefined centroid assumptions. It highlights that the similarity recognition scheme dominantly causes the uncertainty associated with selecting the k-means variants.

The results demonstrated the validity of the original comparison. For this reason, we do not change the structure of the original paper. However, we added a few discussions about this.

> "Note that the four k-means variants run with the randomized centroids each time. It is because we aimed to compare four algorithms as they are as integrated systems. Additional runs using the same starting centroids for k-means variants show that the uncertainty related to clustering algorithm selection remains regardless of using the same starting centroids or randomized initialization."

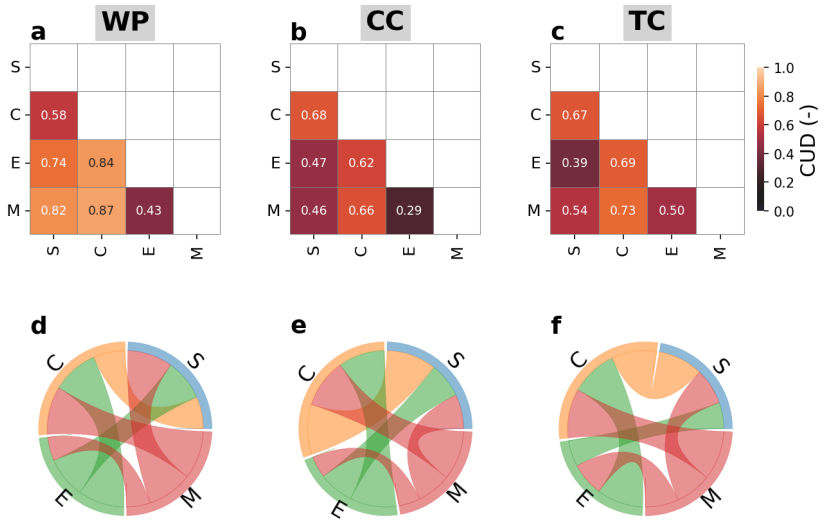

Figure 3. Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between clustering results from different k-means algorithms, i.e., S, C, E, and M k-means, for different demo experiments: WP, CC, and TC. (a, b, c) CUD in heatmaps, and (d, e, f) visualization of the interconnection using the chord diagrams. Note that the results are from the configuration with k=4 and ***the four k-means variants use the same starting centroids.***

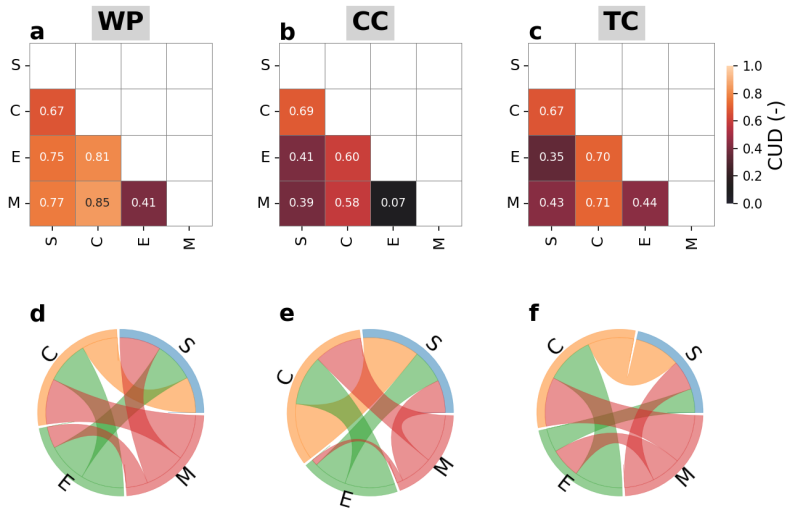

*Figure 4. (**Figure 10 in the manuscript**) Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between clustering results from different k-means algorithms, i.e., S, C, E, and M k-means, for different demo experiments: WP, CC, and TC. (a, b, c) CUD in heatmaps, and (d, e,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

Regarding the second part of the reviewer's comment, the consistent convergence criterion has been used for all four methods. We have added this information to the revised manuscript.

> "Technically, the algorithm converges if the sum of the mean square errors of centroids versus those in the previous step becomes zero. The convergence criterion is the same for all k-means variants used in this study. An iteration limitation is set up to 100 to avoid the infinite loop of iterations."

6. Lines 364-365 "As the first study to address this issue, we believe that CUEF can constitute a new standard for addressing uncertainty issues when performing data clustering in (but not limited to) climate science." -- This is an overstatement. It's well know that custering algorithms are local search methods that are sensitive to random start, however, there are number of approaches in published literature to identify good seeds and ensure that algorithms can converge to a consistent cluster set.

We partially agree with this comment. A reason is that the clustering uncertainty evaluation framework (CUEF) must be understood in **a broader context**. The clustering uncertainty is not only caused by how the algorithm is initialized. It is also caused by selecting different *k*-means algorithms, or different clustering algorithms other than *k*-means such as affinity propagation, DBSCAN, self-organizing map, etc. It is also caused by input data (for example, we confirmed that the uncertainty in clustering Japanese summer weather patterns is much higher than that we cluster winter weather patterns though not shown here).

Nevertheless, we agree that there are number of approaches to identify good seeds to improve the convergence of *k*-means. According to knowledge of the authors, the most well-known approach is *k*-means ++ (Arthur and Vassilvitskii, 2007). The intuition behind this method is that spreading out the *k* initial cluster centroids is preferable: the first cluster centroid is chosen randomly from the input data points. Each subsequent cluster centroid is determined from the remaining input data points with probability proportional to its distance from the point's closest existing centroid. One might note that *k*-means ++ still, literarily, relies on random selection of the first "seed". For that the final "seed" set will be different from each other resulting the uncertainty in the clustering results. Also, note that random choice of seed is not bad assumption, rather it is necessary to avoid the bias of specified seeds.

To demonstrate it in more clear and evident way, we have run additional simulations, in which the ***k*-means++** scheme of initialization is used instead of fully randomized method in the original *k*-means algorithm. The results are shown in figures following demonstrate two things.

(1) Obviously, clustering uncertainty in using different *k*-means ++, i.e., S, C, E, M *k*-means++ still exists (Fig. R3). **The degree of uncertainty is the same** with that among S, C, E, M *k*-means (as shown in Fig. R4 which corresponds to Fig 9 in the original manuscript).
(2) Clustering uncertainty exists among different runs (Fig. R5). It is because *k*-means ++ is not free from random choice of seed. However, interestingly and also somehow expectedly, the clustering uncertainty caused by *k*-means++ is smaller than original *k*-means (Fig. R7 – Figure 10 in the manuscript). In TC experiment, E k-means ++ can provide a zero uncertainty. In WP experiment, the uncertainty is higher, implies that it depends much on data used.

We have added additional discussion about CUEF into the revised manuscript:

> "Our proposed clustering uncertainty evaluation framework (CUEF) must be understood in a broader context. The clustering uncertainty can be caused by selection of clustering algorithms (other than *k*-means such as affinity propagation, DBSCAN, self-organizing map, etc.), by initialization scheme, and by input data itself. Focusing on initialization scheme for k-means, one might note that that there are number of approaches to identify good seeds to improve the convergence of final outcomes. The most well-known approach is k-means ++ (Arthur and Vassilvitskii, 2007). The intuition behind this method is that spreading out the k initial cluster centroids is preferable: the first cluster centroid is chosen randomly from the input data points. Each subsequent cluster centroid is determined from the remaining input data points with probability proportional to its distance from the point's closest existing centroid. One might note that k-means ++ still, literarily, relies on random selection of the first "seed". For that the final "seed" set will be different from each other resulting the uncertainty in the clustering results. Also, note that random choice of seed is not bad assumption, rather it is necessary to avoid the bias of specified seeds.
>
> To see how uncertainty problem is solved with *k*-means ++, we have run additional simulations, in which the *k*-means++ scheme of initialization is used instead of fully

randomized method in the original *k*-means algorithm. The results (shown in the appendices) demonstrate that inter-algorithm clustering uncertainty with *k*-means ++ still exists. The degree of uncertainty is the same with that among original *k*-means variants (Fig. 9). Interestingly and somehow expectedly, the inter-run uncertainty with *k*-means++ is smaller than original *k*-means, implying that improving initialization could reduce the uncertainty of clustering results, though this depends much on the type of input data. ”

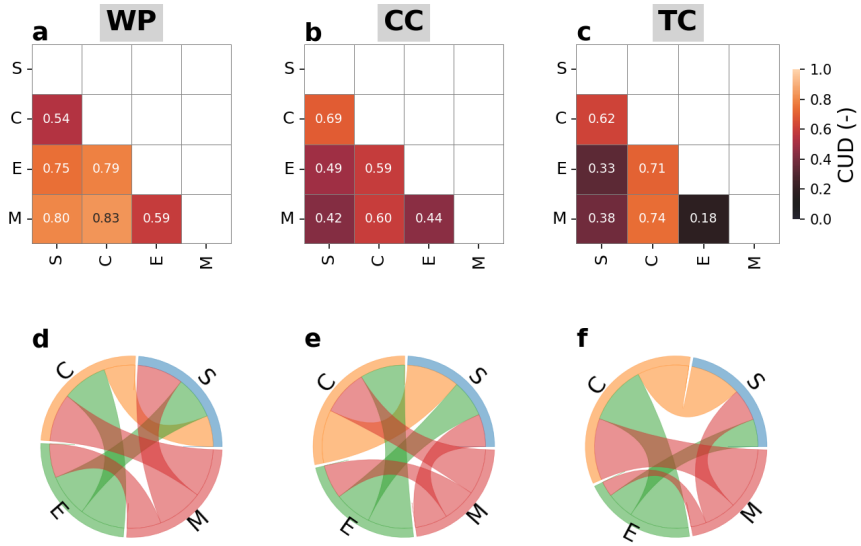

*Figure 5. Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between clustering results from different k-means algorithms, i.e., **S, C, E, and M k-means++**, for different demo experiments: WP, CC, and TC. (a, b, c) CUD in heatmaps, and (d, e, f) visualization of the interconnection using the chord diagrams. Note that the results are from the configuration with k=4 and the first initialization run.*

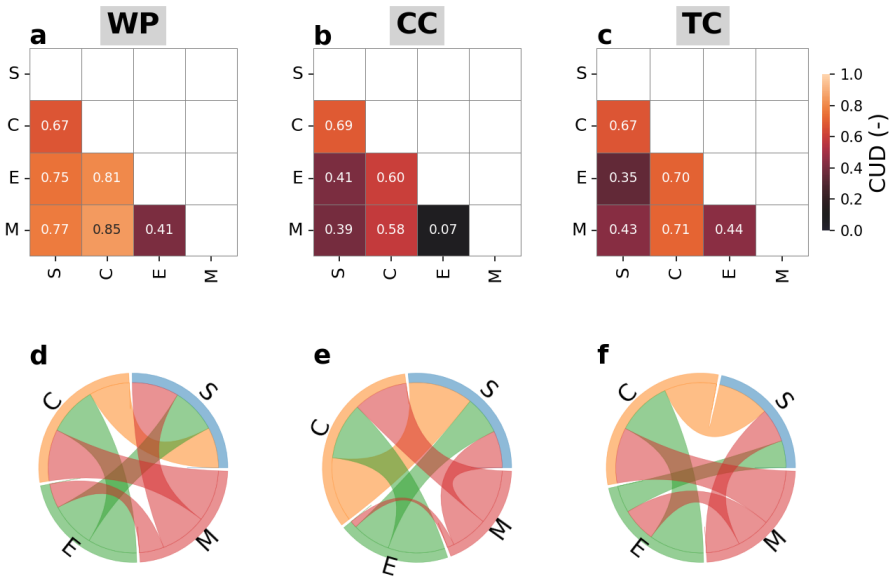

*Figure 6 (**Figure 10 in the manuscript**) Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between clustering results from different k-means algorithms, i.e., **S, C, E, and M k-means**, for different demo experiments: WP, CC, and TC. (a, b, c) CUD in heatmaps, and (d, 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.*

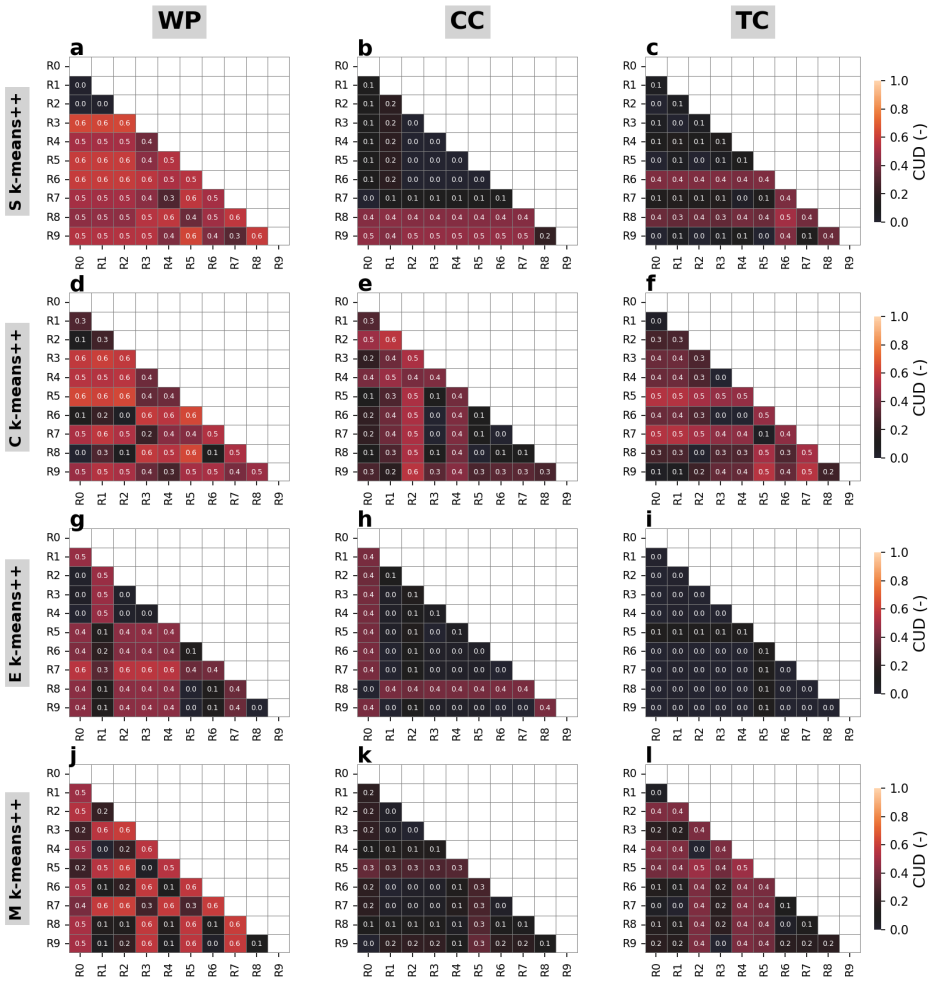

Figure 7 Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between the clustering results from different runs (10 runs indicated by R0, R1, …, R9) of different ***k-means++ algorithms, i.e., S, C, E, and M k-means++*** *(rows), for different demo experiments: WP, CC, and TC (columns). Note that the results are from the configuration with k=4 and the first initialization run.*

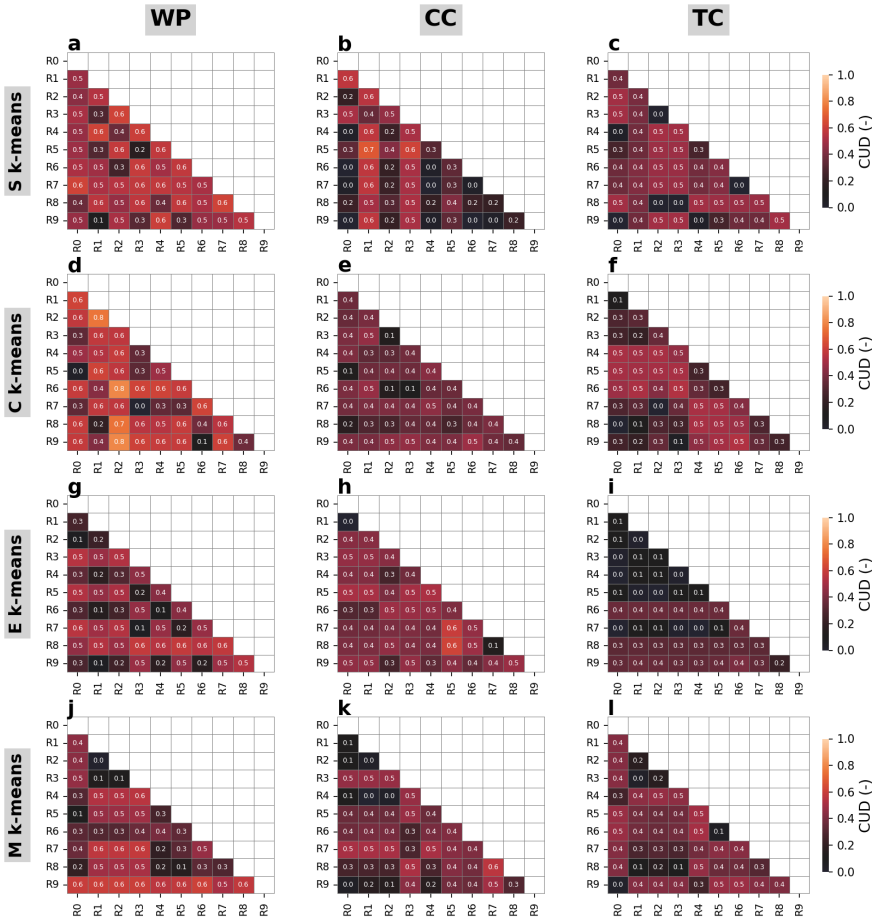

*Figure 8 (**Figure 11 in the manuscript**) Clustering uncertainty degree (CUD) based on adjusted mutual information (AMI) between the clustering results from different runs (10 runs indicated by R0, R1, ..., R9) of different **k-means algorithms, i.e., S, C, E, and M k-means** (rows), for different demo experiments: WP, CC, and TC (columns). Note that the results are from the configuration with k=4 and the first initialization run.*

Reference:

- Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

7. Lines 370-374: "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. Although this issue is fundamentally important, to the authors' best knowledge, no studies have addressed this question or proposed a solution, at least among the climate sciences." -- this again is broad and biased inference based on the demonstrated applications and results.

We agree that the statement could sound overestimation, though we have a reason to say that. In this study we are trying to raise the awareness of clustering application research community (at least limited to climate science, where we are accustomed) that "**do right things**" is better than "**do things right**". Before solving a clustering problem (whose purposes could be to gain knowledge or to do prediction), a researcher needs to ask the first question whether it is meaningful to apply clustering approach, i.e., is this right thing to do? If the intrinsic uncertainty in the problem is too large one have to give this method up because the results even obtained are not robust enough, and discussion based on these will be misleading.

In the revised manuscript, we have revised the text in Lines 370 – 374 to, we reduce the "Although this issue is fundamentally important, to the authors' best knowledge, no studies have addressed this question or proposed a solution, at least among the climate sciences."

To

> "In the other words, this study attempts to raise awareness of clustering-application community that "before trying to do thing right, one must know it is right thing to do"."

8. Authors have termed their clustering framework to be novel, including in the title of the manuscript, which in my opinion is overstated and not justified. There are three key methodology elements in the paper + application to three select climate datasets.

Application component of study is weak and limited in scope. But author's acknowledge that application/interpretation was not the focus of their study, Lines 277-278 "We do not intend to physically interpret the specific clustering outcomes, although some phenomenal explanations are provided in the manuscript." So novelty is not in the three applications.

Three elements of methodology are adopted from published literature:

1. Structural similarity based k-means -- adopted from Wang et. al. 2004, Wang and Bovik, 2009

2. Evaluation of clustering algorithms using Similarity distributions (adopted from Doan et. al. 2021), Silhouette scores (adopted from Hassani and Seidl, 2017).

3. Clustering uncertainty degree and information theory (Vinh et al. (2009))

Building upon published literature is normal discourse of scientific research. But I suggest reconsidering the use of term "novel".

The reviewer is correct about basic structure of this study. We sincerely accept the request of the reviewer to reconsider the use of term "novel" in the title. We have revised the title to

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

We also understand that the concept of “novelty” is usually subjective depending on standing point of viewer. Apart from discussion whether it is proper to have “novelty” in the title, let us remind the reviewer about some “new values” that we added to current literature. First, S k-means algorithm is the first variant, according to knowledge of the authors, adopts **structural similarity paradigm** to cluster things. There have been a lot of variants of k-means regarding how to determine similarity/distance between objects, but most are based on distance paradigm.

The second new value is the clustering uncertainty evaluation framework. The reason we call it a **framework** because it is more than an application of a technique like mutual information. We propose the way to evaluate the meaningfulness of application of clustering solution for a given problem. We use mutual information as a showcase, though we can use different criteria such as “rand index”, which has been developed for the same purpose as mutual information. Also, remember that the adjusted mutual information (Vinh et al., 2009) is primarily developed to measure the “goodness” of clustering algorithm based on assumption of existing “ground truth”. Here we **diversify** the primary purpose by using it to evaluate the uncertainty/consistent/convergence of a clustering solution. So, using mutual information have to be understood as showcase of CUED, but not CUED itself.

We have added to revised manuscript.

> “Note that the CUEF proposed in this study it is more than an application of a technique like mutual information. The idea is to propose the way to evaluate the meaningfulness of applying a clustering solution for a given problem. Here we used mutual information as a showcase. Recall that the adjusted mutual information (Vinh et al., 2009) is primarily developed to measure the “goodness” of clustering algorithm versus prior known “ground truth”. Here we diversify the primary purpose by using it to evaluate the uncertainty/consistent/convergence of a clustering solution. Exactly saying, here we can use different techniques to do so, for rand index. So, using mutual information have to be understood as showcase of CUED, but not CUED itself.”