Note that this is the responses to the reviewer 2's comment. The revised manuscript is not included here, because it is not able to be uploaded in this stage.

Response to Reviewer 2's comments

This study by Doan et al. presents the use of a S k-means clustering as a better alternative for climate and atmospheric science to clustering data than traditional k-means methods. This study introduces a novel framework to identify uncertainty within clustering methodologies and said framework introduces a methodology by which researchers can compare different clustering techniques with each other in a way that doesn't require a ground truth dataset to exist by which to compare results to. The study presents the methdology in an excellent manner that seems like it would be easy to replicate/apply to future studies.

S k-means is a useful technique that adapats SSIM techniques, traditionally used in image comparison analysis, to be applied to climate data. It is an improved technique, compared to the traditional distance metric comparisons, as this takes into account both spatial and temporal differences in datasets. This manuscript does a good job at summarizing the use of the aforementioned techniques with respect to three example tests for typical climate situations in which clustering is used. **However, this manuscript lacks in the discussion and summary sections. The manuscript needs to emphasize more as to the usefulness of this new uncertainty framework compared to current available methodology. The results are well explained, but there is a lack of discussion about how this brings a significant change to current techniques/how this improves current understanding and techniques.**

Thanks, the reviewer for the positive feedbacks. We agree that the discussion about the significant contribution of the proposed methodologies versus current techniques is needed to improve the quality of the manuscript. Following the advice of the reviewer, we add following discussion into the revised manuscript. The summary section is revised accordingly. We hope the reviewer satisfy with this revision.

> "Another benefit of CUEF is that it can measure the meaningfulness of clustering given data. To date, clustering algorithms including *k*-means have been used primarily to either explore unknown atmospheric patterns or support predictions. The most common approach is using clustering techniques within the framework of "detection-and-attribution", i.e., detect specific atmospheric events, e.g., abnormally hot weather or heavy precipitation, then attribute the causes to atmospheric regimes/patterns revealed by clustering analysis (Esteban et al., 2005; Houssos et al., 2008; Spekat et al., 2010; Zeng et al., 2019; Smith et al., 2020). Clustering techniques are also used for weather forecasts or climate predictions (Kannan and Ghosh, 2011; Gutiérrez et al., 2013; Le Roux et al., 2018; Pomee and Hertig, 2022) or for reconstructing historical data (Camus et al., 2014).
>
> No doubt, clustering analysis largely contributes to advancing climate sciences alongside other data analysis and numerical modeling techniques. The essence of the technique lies in its ability to extract knowledge (patterns) from data. It allows researchers to

discover unseen structures hidden in data which is massive and inaccessible to human perception. So far, tremendous efforts have been invested in either proposing/improving clustering algorithms or inventing criteria for evaluating the goodness of the results. Such efforts could be classified as "attempts to do things right." A question posed here is more fundamental in the sense of how to justify the selection, i.e., "whether it is the right thing to do (the right method to select)?". This study proposes a quantitative framework in which the users could justify the selection directly based on the data rather than relying on the literature review (select it because other researchers use it). Such kind of justification is more or less a fallacy due to the diversifying clustering problems in climate science, and the variety of clustering algorithms. Also, climate data, whose types and amounts are increasing at an unprecedented pace, is adding challenges to experience-based justification. According to the authors' knowledge, the CUEF is the first attempt to address this issue. Though there is still free room for further development, CUEF is believed to constitute a new standard for climate data clustering. We recommend CUEF as a necessary procedure before applying clustering techniques. Even though the justification might depend on multiple factors other than the data-oriented uncertainty, such as how the clustering results will be used in further analysis processes, CUEF can support the explanation and discussion of the clustering results."

1. Table 1 provides a nice summary of different metrics compared between the different k-means models used in the study. In the text, the mean and standard deviation are mentioned from the table, however, the other metrics are not mentioned at all other than in passing. The Shannon metric needs to be explained more and some presentation of the data should be given in the text to give the reader some context as to its meaning and how it is used in this study.

Thank you for the noticing. In the revised manuscript, we have added the explanation related to other metrics. We copied Figure 2 and Table 1 here together with additional explanation for reference.

"Before analyzing the *k*-means clustering results, we diagnosed the nature of the input data using S-distributions (or S-D). 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 appeared 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 tended to be more tailed (both sides), with skewness over the left tail. Quantitively, the standard deviation of S-Ds for S-SIM and COR tended to be higher (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 that measures the symmetry of S-Ds shows negative values, meaning the left-skewed distributions. Those values in Table 1 are consistent with visualization in Figure 2. Especially, S-SIM and COR tend to be higher skewed than that of ED and MD particularly in the CC and TC experiments. The consistent 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), respectively, of S-Ds. Overall, kurtosis values are consistent with visualized results in Figure 2, i.e., S-Ds of S-SIM and COR tend to spread more over two tails compared with ED and MD. Entropy, on the other hand, does not show obviously higher and lower trends of S-SIM, and COR compared with ED and MD, and it is likely more data dependent."

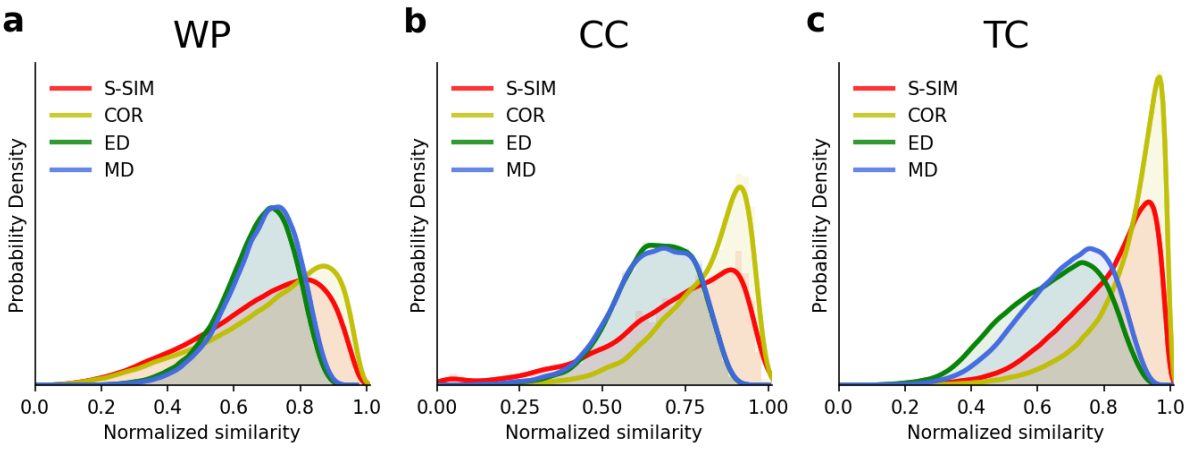

Fig. 2 (in the manuscript) Comparison of the S-distributions of normalized pairwise similarity using the structural similarity (S-SIM), the Pearson correlation coefficient (COR) the Euclidean distance (ED) and the Manhattan distance (MD) for three demonstration experiments: WP, CC, and TC. With a population 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.

Table 1 (in the manuscript). 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

2. Many references from the text are missing citations. Please check over the references in the paper to make sure all are cited, here are a few that I found that were not cited: Jancey 1966, Lloyd 1957, Wang et al. 2004, etc.

We apology for this inconvenience caused for the reviewer. We will add these references into the revised manuscript.

3. This study intends to establish both the uncertainty framework and the s k-means methodology as a new standard for data mining in the climate sciences. While the uncertainty framework definitely provides a new standard by which to test the usefulness and effectiveness of different clustering algorithms against each other, no work has been shown as to the ability of the s k-means clustering. While comparisons are shown between the s k-means to other k-means clustering measures, **we cannot objectively say from this study that the S k-means method better captured the underlying structures** within the data compared to the other k-means models. **A more comprehensive case study would be needed, rather than the short test cases, that applies the methodologies to a known problem that has a ground truth that can be compared back to.**

The reviewer is very critical on this point. “Ground truth”, if exists, is the best solution to determine the goodness of one method against another. However, there are reasons that we do not use “ground truth” in this study. First, we have no reliable “ground truth”, i.e., “real” patterns of three datasets, WP, CC, and TC. The reviewer might notice that the lack of “ground truth” is common in other atmospheric data also, not only related to our experience settings. Because it is difficult to define “true” weather pattern, even though some individuals (i.e., weather forecasters) might claim that they have. In our opinion, climate data is very, that we called, “contextual” data, i.e., a claim of a weather pattern (or typhoon pathway pattern) is exclusively data dependent, it is rather associated with broader contexts of personal experiences, knowledges. It is why we try to avoid “personal-experience involvement” in the evaluation until “universal ground truth” is available.

Nevertheless, we add discussion about the ability of S *k*-means in capturing the “structuredness” of the data with additional analysis and plotting (show below) to address different aspect of the reviewer’s comment. Here we focus on the ability of the algorithm to distinguish the difference between objects. With comparing imagination weather patterns (generated for intuitive comprehension) we demonstrate that using S-SIM could provide better (closer to human intuitive) similarity recognition than distance metrices. The following

discussion and the additional figure (Figure 8 in the revised manuscript) are 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."

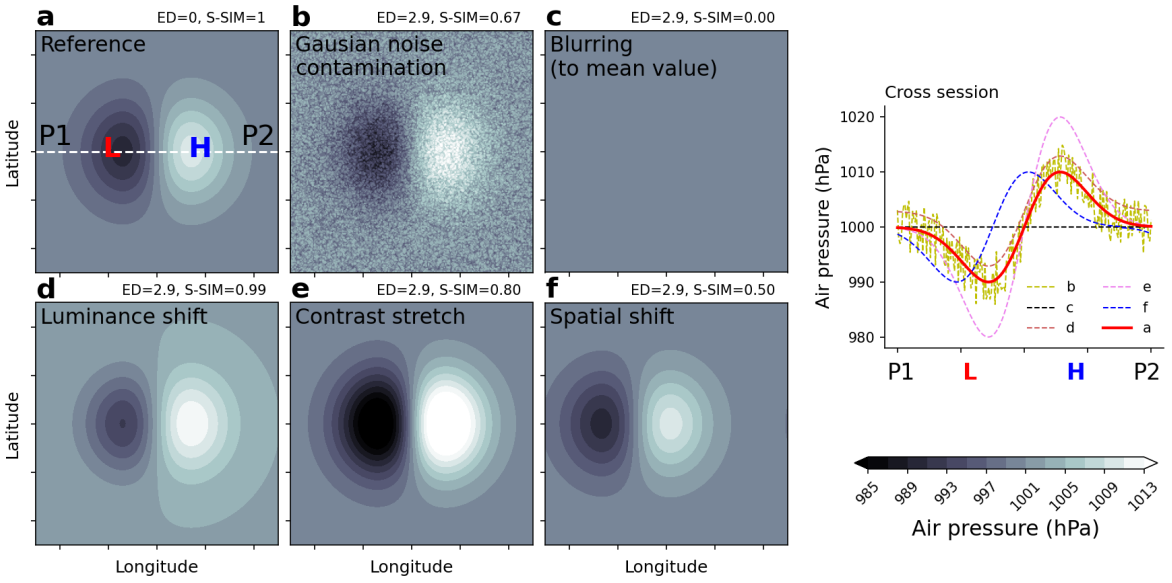

Figure 8 (in the revised manuscript). Imagination air pressure patterns. Subpanels are the reference (a), Gaussian noise contamination (b), blurring (to mean value) (c), luminance shift (d), contrast stretch (e), and spatial shift (f). The ED (Euclidean distance) and S-SIM (structural similarity) values 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.

4. The use of 3 different test case scenarios to test the uncertainty framework was a great idea and well presented. It gives good insight into how this methodology can be used in the wide-array of applications in climate science.

Thanks the reviewer for this compliment.

5. Lines 370-374. This question of applying the framework to see whether data is suitable for clustering is a much more novel approach and useful to the science than comparing the initializations. There are many other methodologies and ways to get suitable initializations for clustering and help datasets to converge on useful clustering.

Thank you. Indeed, our study emphasize the effectiveness of the CUED for when comparing algorithms and datasets. Though initializations could cause the uncertainty but some improvement such as k-means++ could help to reduce uncertainty and preserve the consistency in clustering results. The discussion regarding this comment can be found the above answer (to the general comment).

6. Lines 370-374. It is tough to say with respect to WPs that clustering my be ineffective. WPs present a lot of uncertainty compared to other types of climate data, so without care as to what is being analyzed/searched for in the data, uncertainty analysis may present false positives for datasets that would not be suitable for clustering. This isn't a problem with the methodology, the authors do note that these are inherently a data issue, which this methodology does not take into account. The authors could do to make note of similar situations in the manuscript for those who would use this method in the future.

The reviewer is correct. We add some clarification to avoid the potential misinterpretation to the revised manuscript.

> "Note that CUEF provide a method to quantify the uncertainty/consistency of clustering solutions from the data science aspect. However, the decision whether to adopt clustering techniques could depend on another factor, such as how the results will be used and interpreted. In such a case, CUEF could be used to support explanation regarding the robustness or the clustering results."

7. Some figures need revision, specificaly figures 3, 4, and 5. In Figure 3, the silhouette score charts are very small compared to the WP plots. Make them a similar size and make the text size more legible. Figures 4 and 5 have the silhouette score charts inside of the other figures. There is far too much going on inside these figures as it is, and adding the silhouette plots inside here makes it more cluttered and confusing to understand. Move them outside the plots and enlargen them.

We have replotted the Figures 3, 4, and 5 exactly following the suggestions of the reviewer. The replotted figures are attached below for reference.

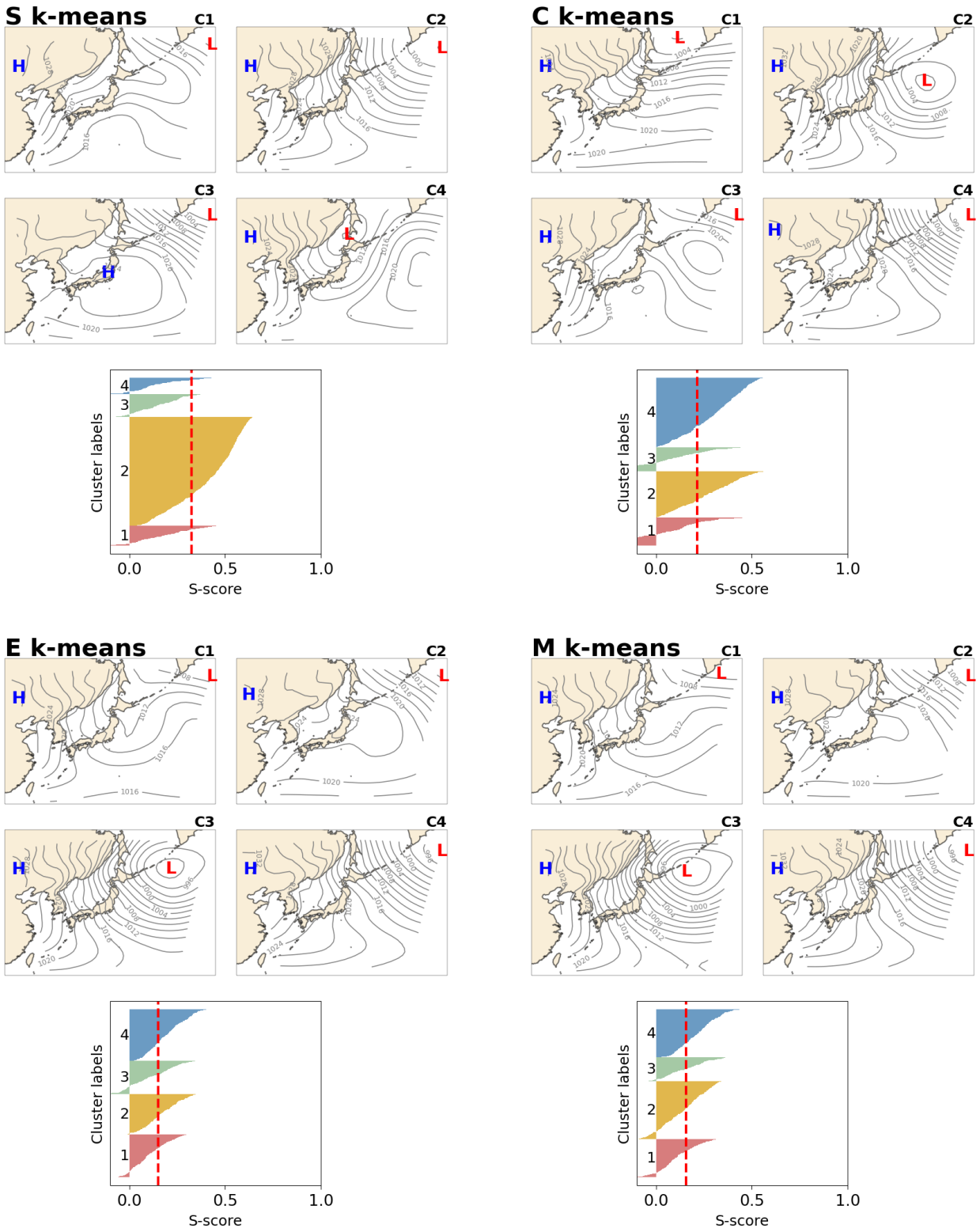

Fig. 3 (in manuscript). The silhouette score charts become bigger and text size more legible.

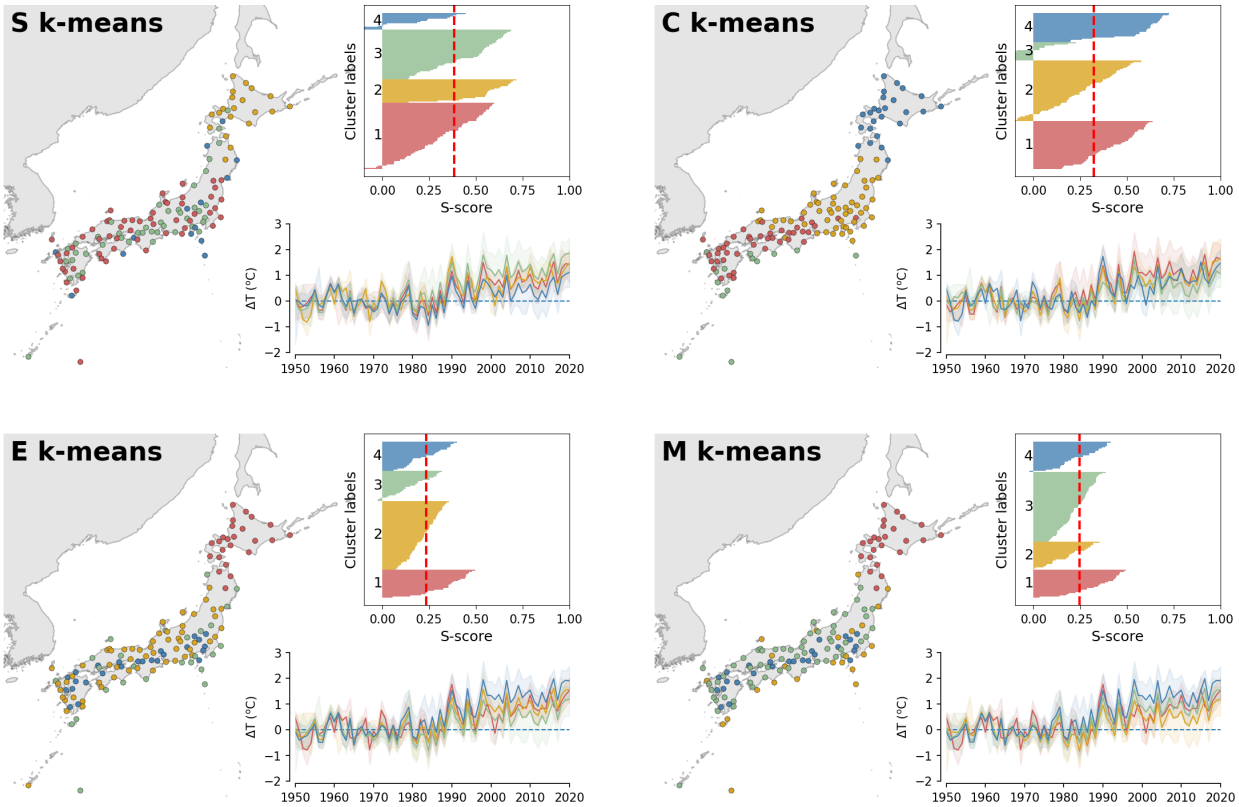

Fig 4 (in the revised manuscript). Moved the silhouette score charts outside of maps and enlarged the charts and made the text size legible.

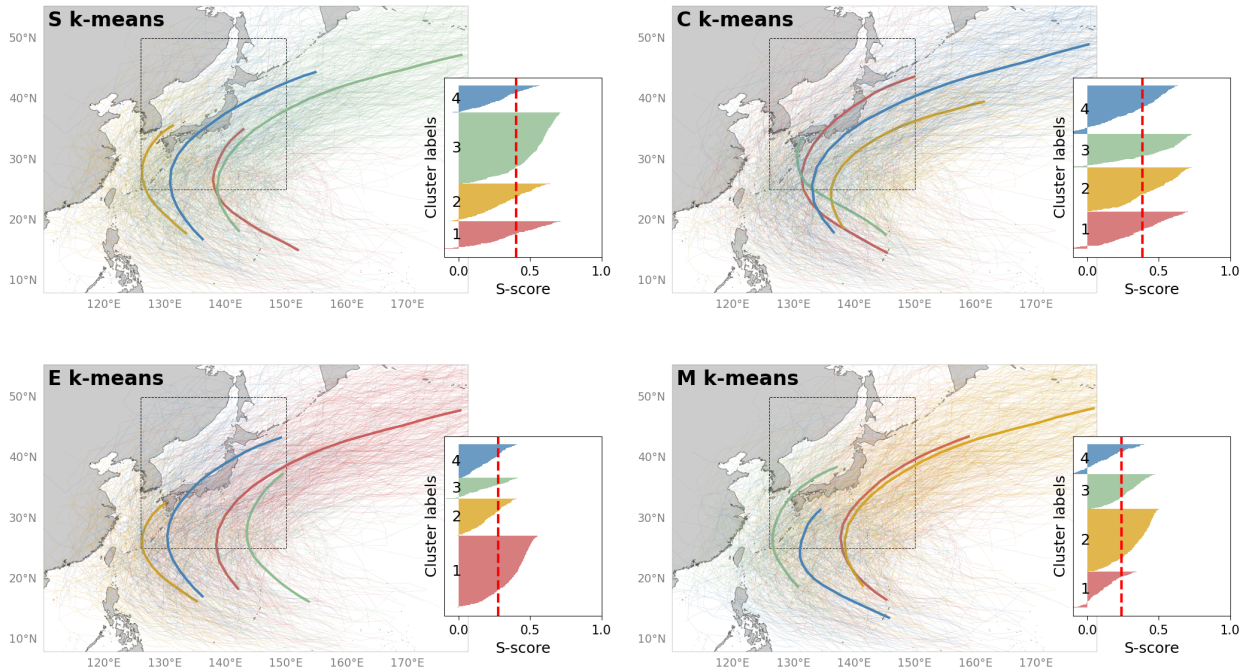

Fig 5 (in the revised manuscript). Moved the silhouette score charts outside of maps and enlarged the charts and made the text size legible.

Minor notes:

Lines 72-74: Rephrase the wording, it is confusing in this state.

We have rephrased the sentence from

> "For these reasons, k-means under the distance paradigm treats the features of the input data equally, thus mask the similarity recognition between data, consequently deteriorating the clustering outcomes."

to

> "Thus, the distance measures, which treat the features of the input objects equally, might ignore inherent "structuredness" in the objects when recognize the similarity between them. This characteristic could deteriorate the clustering outcomes."

Line 75: Remove ".It is" and use because to join the two sentences into one for better flow.

Have removed ".It is" and joined two sentences into one.

Line 125: Should cite the SSIM technique (Wang et al. 2004)

Cited Wang et al., 2004.

Lines 158-160: What's the interpolation method used?

We used nearest-neighbor interpolation for regridding the data. We add this information into the revised manuscript.

> "The data had a horizontal resolution of 0.75°on a regular grid but were re-gridded to an equal-area scalable earth-type grid at a spatial resolution of 200 × 200 km using nearest-neighbor interpolation method."

Line 238: Could explain cluster realization better/earlier. Explaining it in this sentence while also introducing a new concept could cause confusion to the reader.

Have rephrased the text for clarification.

> "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:"

Line 239-240: What do you mean by partition set? Is this the same thing as the cluster realization?

Yes, partition set is detailed form of cluster realization. We have revised the sentence for clarification.

> “Entropies of clustering realizations are defined as the amount of uncertainty for partition sets of each realization.”

Line 246: What do you mean by weakness? Is it related to the randomess you discuss in the next few lines?

Yes, mutual information is weak with random clustering (or chance). We have revised the text for easier understanding.

> “However, mutual information is weak against chance.”

Line 297: Change tense of "were" to "are".

We have revised the text accordingly:

> “These regional differences are well captured by *k*-means clustering. For example, the northern part (Hokkaido) is consistently separated from other regions in terms of temperature warming.”

Line 316: What does "completed by C K-means" mean? Is it a typo?

It is “competed” not “completed”. We have revised the text to:

> “The performance of S *k*-means is sometimes competed by C *k*-means.”