The manuscript entitled “A method for characterizing the spatial organization of deep convective cores in deep convective systems’ cloud shield” presents a comprehensive methodology for analyzing the spatial organization of deep convective cores (DCCs) within Mesoscale Convective Systems (MCSs). Recognizing the limitations of existing organization indices when used in isolation, the authors propose a multidimensional approach based on four variables: (1) the convective fraction, (2) the total area of deep convective cores, (3) a characteristic length scale of aggregation among the cores, and (4) a metric quantifying deviation from a uniform spatial distribution.

A notable innovation of the study lies in the representation of DCCs as filled squares within the deep convective region, which enables the characterization of the spatial organization of the most elemental deep convective structures, although this identification relies on strong assumptions. Applying this framework to both observational and model datasets, the authors identify four distinct modes of DCC organization. They conclude by emphasizing the method’s effectiveness, robustness, and adaptability, and suggest several promising avenues for future application.

Overall, this manuscript is of good scientific quality and aligns well with the scope of Atmospheric Measurement Techniques (AMT). The scientific content is clearly presented and well written. However, I have a few minor comments that should be addressed prior to publication. My main concern pertains to the robustness of the proposed method, as the sensitivity analyses provided are, in my view, somewhat insufficient (see detailed comments below).

We thank the referees for their thoughtful and insightful comments which have helped us improve this manuscript. In light of several converging feedbacks on the importance of the decomposition process, we propose to change the title as follows: “A method for characterizing the spatial organization of deep convective cores in deep convective systems’ cloud shield using idealized elementary convective structure decomposition”, putting emphasis on the assumption of elementary convective segmentation with circular structures of varying radii.

Another major change is that the convective area S has been replaced by the letter A to align more closely with the term 'area' rather than 'surface'. In our response, we have kept S consistent with the comments made, but all occurrences have been modified within the manuscript.

We have also updated some of the figures as requested, along with adjustments in the text. Two additional figures have been added, the first one in the end of the Results section as the new Figure 15, and the last one in the Annex B section as Figure B3. In addition, we reinforce the sensitivity study in Annex B by conducting a Monte Carlo-style experiment on propagated error within the clustering process, in order to evaluate the stability of the method with respect to each of the four key parameters. Text has been amended in tracked changes.

And in particular we have emphasized the analysis of the robustness of the method. In addition, we reinforce the sensitivity study in Annex B by conducting a Monte Carlo-style experiment on propagated error within the clustering process, in order to evaluate the stability of the method with respect to each of the four key parameters. Our point-by-point responses are detailed below, with the original comments included in black and answers in blue.

Detailed comments

Figures:

The labels are sometimes too small. Besides, adding letters to identify each panel could help.

Done, most of the figures/labels have been enlarged

L. 138-140:

There are some words missing in this sentence.

The incriminated sentence has been modified as follows: “To ensure continuity, harmonization of TRMM-PR and GPM Ku-band (13.6 GHz) calibration, reducing systematic differences and improving long-term precipitation estimates.” ⇒ “In order to ensure continuity, harmonization of TRMM-PR and GPM Ku-band (13.6 GHz) calibration has been carried out, with the objective of reducing systematic differences and improving long-term precipitation estimates.”

L. 279-284: elementary convective structures

Could the authors clarify whether there is a physical justification for defining, for example, a 4×4 square as a single elementary convective structure rather than interpreting it as four 2×2 structures grouped together?

The continuity of the horizontal convective structure is inherited from the concept of ensembles of updrafts that are often observed aggregated to one another in radar observations, especially with relatively low spatial resolution (see early work by Houze and for instance Houze, 1997). This is also the case in airborne identification of updrafts over tropical oceans (see Zipser et al., 1994; Lemone et al., 1998). From this reflectivity (hydrometeors) or dynamical based definitions, a spectrum of continuous space scales has been observed and related to a dynamical core or a bulk plume. The homogeneous spatial distribution of the parameter is hence used to define the event. Details vary on how to define continuity (4 or 8 connectivity), on the threshold selection etc., but all techniques concur to identify a spectrum of horizontally spread objects with various sizes.

In this view, a continuous object would be a given convective entity. This is what we refer to as the Houze method in the manuscript. This classical approach does not imply any preferred morphology for the object. The present decomposition goes one step further by imposing a morphological constraint on the object delineation. A circular (square) shape is hence assumed and the most direct implementation of it is the method we have chosen: any continuous square object is associated with an individual core. That’s why the method identifies a single 4x4 pixel core instead of 4 2x2 pixel cores.

The last sentences of the paragraph is modified as follows: “This approach can be viewed as an upscaled, space-borne version of the analysis of precipitation core and updrafts/downdrafts joint occurrence analysis from high-resolution ground-based observations (e.g., Moroda et al., 2021; Lamer et al., 2023). In this perspective, we assume that the large cluster of continuous echoes (colored in green in Figure 3, middle) is composed of smaller coherent and compact convective features akin to a circular bulk updraft of varying diameter that we approximate

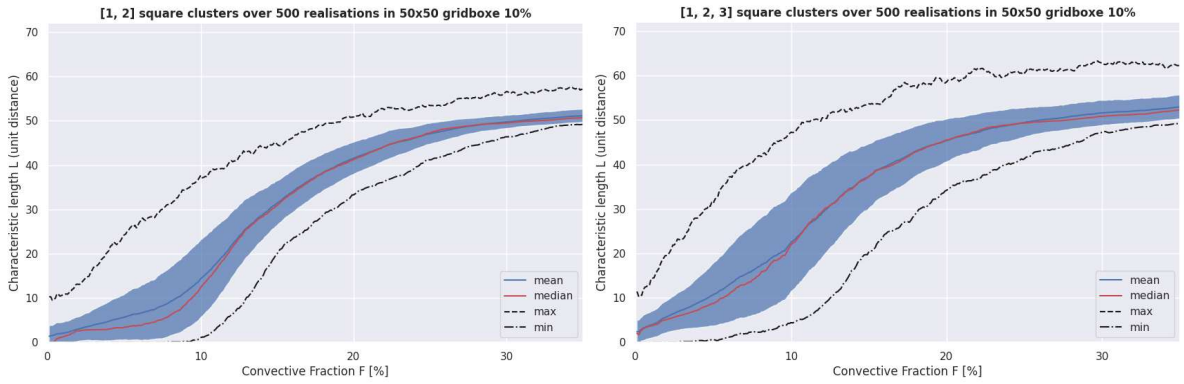
using size varying squares. We hence add a morphological constraint to the classical core segmentation."

Consequences of this assumption are further detailed below.

In particular, how does this assumption impact the calculation of the spatial organization metric (variable P)? It would be helpful to discuss whether this structural definition influences the interpretability of P.

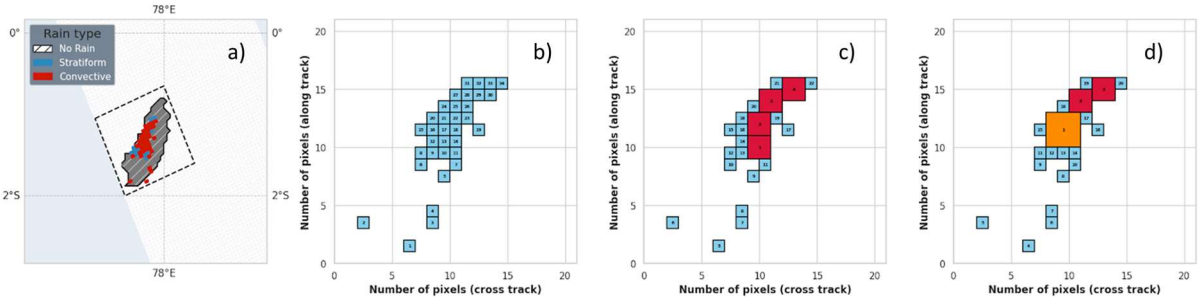
This point was also raised by the first reviewer, and we will structure our response around the shared concerns:

Firstly, autocorrelation is sensitive to dominant structures. Using structures composed of only one pixel is often not discriminating for low fractions (F). Figure 10 (left) shows the range of variation of L for a given F in a 50x50 grid with randomly positioned 1x1 structures. For low F values, the values are all confined until F exceeds a certain threshold. The same is true when 2x2 structures are also permitted, with the range of L expanding more rapidly with increasing F values. The same pattern continues up to 5x5 structures (Figure 10, right). See the figure below that complements the Figure 10, with generations with 1x1 and 2x2 permitted etc.

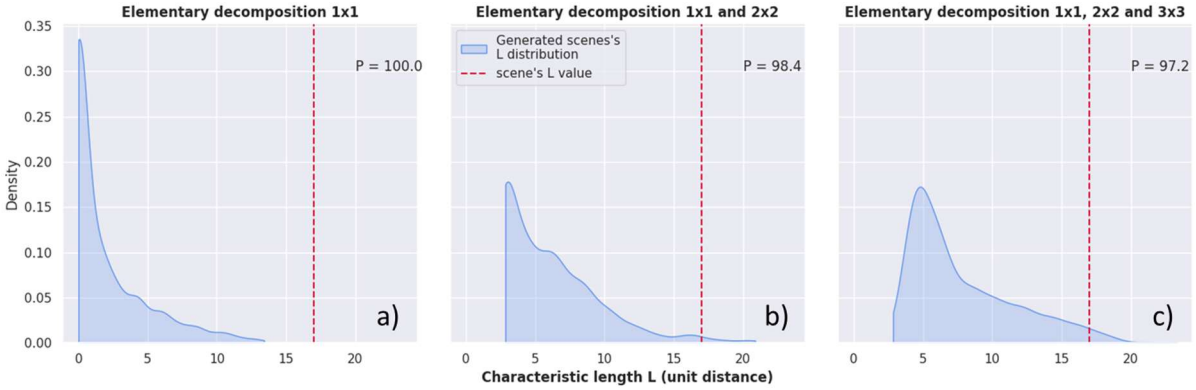


Same as figure 10 with Left: only square clusters of size 1x1 and 2x2 (MaxSquareSize = 2), Left: only square clusters of size 1x1 and 2x2 and 3x3 (MaxSquareSize = 3).

Moreover, it is impossible to break down all scenes using only structures that are 2x2 or larger. Almost always, structures that are 1x1 must be used in addition. Thus, the elementary decomposition impacts the calculation of L distribution for generated scenes, with the majority of the influence coming from the segmentation of dominant structures (larger squares or larger objects if a different decomposition is used). In fact, the more roughly the structures are cut up, the more sensitive L is to the main structures and their respective arrangements. Conversely, by cutting everything into finer elements, for example into 1x1 and/or 2x2 objects, which is also always possible, L will have a more confined range of values and will therefore struggle to express itself in a sufficiently broad distribution as a key variable for the rest of the characterization process. We therefore have a much greater impact on P than on L itself here, which we have quantified in the example using the figure 3:



Same as figure 3 (left and right only) with decomposition only with b) square clusters of size 1x1, c) size 1x1 and 2x2 (MaxSquareSize = 2), d) size 1x1, 2x2 and 3x3 (MaxSquareSize = 3)



Same as figure 9 (left only) with decomposition only with a) square clusters of size 1x1, b) size 1x1 and 2x2 (MaxSquareSize = 2), c) size 1x1, 2x2 and 3x3 (MaxSquareSize = 3).

Here, it is clear that when using larger structures, the L distribution shifts to higher values while P takes on lower values. Conversely, decomposing with either 1x1 or 1x1 and 2x2 structures from a scene that contains larger ones creates the impression that the organization is less random, as P reaches maximum values. This supports the fact that our approach is best suited to more random characterization when using dominant shapes within the scene.

It is acknowledged that more sophisticated decompositions could be proposed. However, it is asserted that the present decomposition is adequate, given its limited computational complexity and its capacity to discriminate adequately in the generated (empirical) distribution of L. This preliminary approach may be superseded in further work.

Additionally, would one expect such regularly shaped (e.g., circular or square) deep convective cores in environments with strong vertical wind shear? In such cases, convective elements may be elongated or tilted, which may not align with the chosen geometric representation. A short comment on the sensitivity of the method to these physical variations would be valuable.

The circular assumption relies on a vertically erected plume model which is a useful idealization of deep convection dynamics. In this simple case, the dynamics is indeed circular. Yet if the plume is slanted (due to wind shear), the cross section of the plume would look like an ellipse. The circular assumption (which translates into square pixels) should be understood as an idealized, first order approximation. To better convey this dimension of our work, the title has been modified.

Figure 4: “larger structures are associated with stronger and deeper convection”

It might be useful to mention that this relationship is consistent with observations. For instance, Moseley et al. (2019) show that larger convective cells tend to exhibit more intense precipitation.

Yes, it is also consistent with radar observations, starting with the GATE campaign (Betts and Houze, 1981). Note that Moseley et al. reports simulations-based results.

L. 354 355: “characteristics of the scene”

Consider referring to the “Scene characterization” section below to precise the variables that are retained for it.

OK done.

Figure 6:

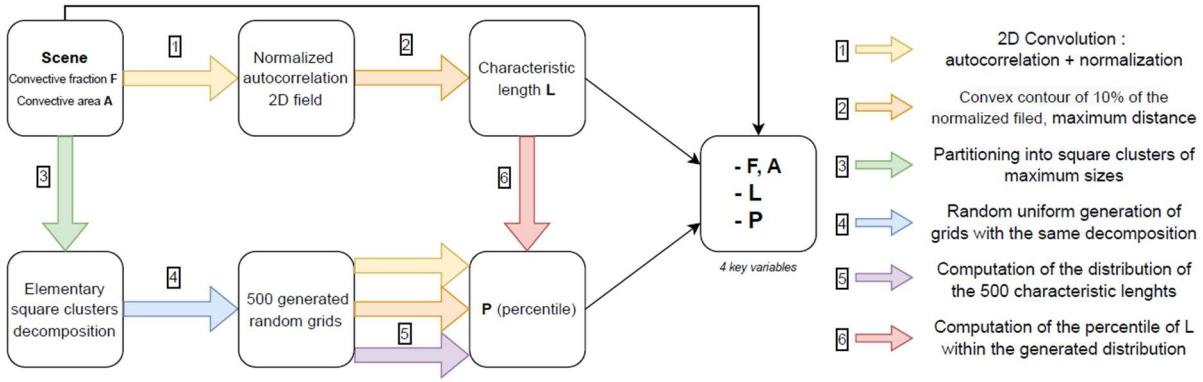
“length” → “length”

Yes done.

Some text are in bold font or in italics without obvious reason to me.

Consider adding a yellow and an orange line for step 5 as these processes are also applied to the random grids.

Yes, thank you, we’ve updated the incriminated figure as follows:



L 367: Scene area

Could the authors clarify how the scene area is precisely defined in the analysis? From Figure 3, it does not appear to correspond to the minimal rectangular bounding box enclosing the cloud shield.

Yes, indeed, the definition of the rectangular outline can be specified in more detail. It is the rectangle that most closely frames the outline of the cloudshield, but within the geometry of the TRMM/GPM swath, without the possibility of extending beyond it because the radar

information is missing there. We therefore retain a rectangle oriented in the direction of the satellite's orbit, with a margin of 1 pixel at the edges.

It is acknowledged that the TOOCAN-radar dataset is subject to certain biases, primarily concerning the delineation of this rectangular contour. Additionally, it has been observed that certain systems are not fully encompassed within the extent of the radar swath. However, we have limited this as much as possible by only retaining clouds for which at least 70% of their surface area is within the swath. It should be noted that these biases are statistically offset by the large number of cases contained within the database.

Since the scene area directly affects both the computed convective fraction and the generation of the reference random distribution for the spatial organization metric (P), its definition is critical.

Regarding the influence of the grid size chosen on the four variables, S is an absolute measure and L does not depend directly on the size of the rectangular contour, only on the number of pixels and their arrangement in the grid. On the other hand, for F, it is true that we introduce a more or less constant bias, tending to minimize F in comparison with the convective fraction with respect to the cloudshield (which is different). Here, it is the convective fraction in the rectangular scene, that is an algorithmic variable approximating the physical convective fraction. This is indeed a possible refinement for a future version that will need to be considered, but we show in the rest of the answer that F is not too sensitive in the final discrimination process when keeping <5-10% of error (see figures B3 answering a comment below, assessing the sensitivity to the four parameters in the classification), and F is kept constant for the generation of random scenes (same bias).

In particular, I am concerned about potential biases introduced by the shape of the cloud shield. For instance, in the case of an elongated DCS, the encompassing scene area may include large regions outside the actual cloud shield — areas that do not contain any DCCs. This could lead to an artificially low convective fraction and/or a misleadingly high organization score (e.g., P close to 1), depending if the DCS itself is densely populated with DCCs.Conversely, for more rectangular DCSs, the scene area might better reflect the actual cloud shield, resulting in more representative values of these metrics.

Regarding the variable P, it is true that the current process can artificially push P towards higher values, since random pixels are allowed throughout the rectangle and not within a more rigorous cloud mask. This will also be the subject of potential improvements in the future, even though we show here that the sensitivity of these choices does not require a revision of the major results of this study (see additional sensitivity figures from Monte-Carlo estimates). Thanks a lot for pointing this out.

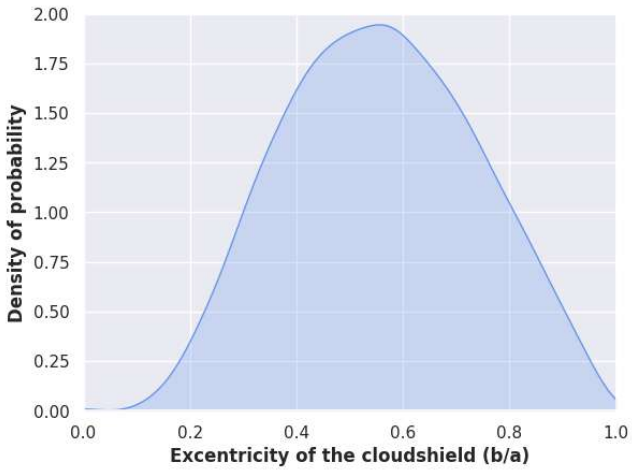
The DCS shown in Figure 14d is a good example of the limitations of the current definition of convective fraction based on the rectangular scene area. Although this system appears to have a high proportion of convective precipitation pixels relative to stratiform ones (possibly >50%), its computed convective fraction is only 7%. In contrast, the DCS in Figure 14c has a visibly lower ratio of convective to stratiform pixels, yet its convective fraction is twice as high. This discrepancy appears to stem from differences in cloud shield shape — with case c being more rectangular and thus more tightly filling the bounding box used as the scene area.

Here again, it is necessary to differentiate between the convective, stratiform precipitating part of the cloud (which is physical) and the convective fraction, which represents only the number of convective pixels within the rectangular grid (as a variable used for the algorithm described in this work). Since we have demonstrated that F could incur an error factor of 5–10% in the classification process without affecting the overall distribution of the classes, we believe that this approximation is sufficient without the need for more sophisticated methods (cloud masks, refined contouring, etc.).

This illustrates that the current definition of F is highly sensitive to cloud shield geometry, particularly for elongated or irregular systems, and may not reflect the true convective content of the DCS. I suggest that the authors more explicitly discuss this limitation and consider whether an alternative scene definition — e.g., based on the actual cloud shield contour, the area of precipitation, or a convex hull — might reduce this bias.

In conclusion, we are aware of this limitation, which is inherent in the diversity of cloud shield shapes observed in this satellite database. Indeed, there are several orders of magnitude in size, and clouds are sometimes elongated, or have parts outside the swath of the TRMM/GPM satellite.

F is not a direct physical variable and remains difficult to estimate visually because scale and edge effects must be taken into account depending on the size of the cloud. However, we have opted for a rectangular outline in the geometry of the radar swath for the sake of efficiency and generalization in the calculations for the subsequent steps in the algorithm, assuming that edge or distortion effects are not predominant. It should be noted that most clouds are on average fairly circular, sometimes slightly elongated, but rarely with a width-to-length ratio greater than 3 (see figure below that illustrates the distribution of the eccentricity (length of the minor axis to the length of the major axis) of the >60 000 cloudshields within the TOOCAN-radar dataset.



Empirical probability density function (estimated via KDE) of the eccentricity (length of the minor axis to the length of the major axis) of all the sampled cloudshields within the TOOCAN-radar dataset

It should also be noted that these limitations are much less pronounced for the TOOCAN-RCE dataset, which has rounder and less regularly elongated clouds, and which has no swath or grid geometry limitations. The consistency of the results between the TOOCAN-radar and

TOOCAN-RCE datasets shows that for this study, our approximation is sufficient, although refinements are possible for future studies.

L. 394: “convex”

The use of a "convex" contour to define the central region of the autocorrelation field is somewhat unclear. In Figure 7, the contour shown does not appear strictly convex or minimal. Could the authors clarify how this contour is derived and to what extent its geometry affects the estimation of the typical aggregation distance?

I am concerned that applying a convex hull to potentially irregular or elongated shapes may artificially increase the value of L. Since L is defined as the maximum internal distance within the convex hull — rather than within the original (possibly non-convex) shape — the metric may become sensitive to shape distortions, particularly for highly anisotropic or fragmented patterns. This could introduce a systematic overestimation of L for some systems but not others, depending on the complexity of the spectral power field. I suggest that the authors clarify how often such distortions occur in practice, and whether they have assessed the sensitivity of L to the convex hull approximation.

Firstly, it represents convexity at pixel level, so it cannot be perfectly convex. This precaution was originally included to ensure that L could be approximated by the major axis of an ellipse encompassing the 10% contour. This illustrates that we are looking for a characteristic length within a simple shape. The value of L is not modified by the convex contour; it is only modified by choosing the central contour from the threshold mask T, and all sensitivity depends on T. This has been addressed in the paper. This is an over-engineered process that could possibly be removed in future versions, as it is useful only in very rare cases and does not modify the estimation of L or the classification process at all.

Fig. 7:

Consider representing the metric L on this figure.

Done with the green arrows on Figure 7 (+ updated caption).

L. 408: “anisotropy”

The mention of anisotropy feels somewhat out of place, given that it is not used in the method. That said, it raises an interesting point. One could imagine that a variable quantifying the anisotropy of the DCC distribution might complement the current set of metrics used to characterize spatial organization. While I understand the motivation to limit the number of variables for clarity and robustness, this example illustrates that additional descriptors could be considered in future work for a more refined or context-specific characterization of convective organization.

Yes, we are also interested in this notion. It should be noted here that part of this anisotropy is taken into account in the autocorrelation by definition and by extension in L, but that it is indeed a potential refinement that would require decorrelating the anisotropy factor from the other components of the autocorrelation 2D. This will be potentially included in future work as it does not align with our primary objectives for the presented paper.

L. 415-416:“spatial morphology” vs. “condensed structural information”

I find the distinction between “spatial morphology of the raw field” and “condensed structural information captured” somewhat unclear. Could the authors clarify what is meant by this difference?

Yes, we agree this needs to be reformulated. We wanted to illustrate that L is a measure of physical distance as well as condensed information about the spatial arrangement and periodicity.

“As a result, L encodes both the spatial morphology of the raw field and the condensed structural information captured by the autocorrelation”

⇒ “As a result, L encodes both the perception of the spatial extent of the main structures of the raw field and the condensed structural information regarding the spatial arrangement of these structures captured by the autocorrelation.”

L. 420:

What is “C”?

Replaced by F.

L. 424, 427, 459, and elsewhere: Use of “probability” to describe P

In several places, P is referred to as the “probability of the scene being randomly organized” or “probability of deviation from a random distribution.” However, as I understand it, P is more precisely defined as the percentile rank of the scene’s characteristic length (L) within a reference distribution derived from randomly generated (uniform) DCC patterns. It does not represent a statistical probability in the formal sense.

For clarity and consistency, I suggest revising the terminology throughout the manuscript to reflect this. Referring to P as e.g. a percentile-based measure of deviation from randomness would more accurately describe its meaning and avoid potential confusion with probabilistic frameworks.

Yes, indeed, all the occurrences of "probability" have been modified or contextualized throughout the text as suggested by the referee.

L. 431 and elsewhere: “bootstrapping”

As I understand it, the authors generate randomized spatial patterns of DCCs by redistributing a fixed number of DCCs uniformly within the scene. Since this process does not involve resampling from the observed data with replacement, it does not correspond to a formal “bootstrapping” procedure. Rather, it is more accurately described as a Monte Carlo approach for generating a reference distribution under spatial randomness.

I recommend updating the terminology accordingly to avoid confusion with statistical bootstrapping methods.

Again, all the occurrences of "bootstrap" or equivalent have been modified throughout the text to meet the referee's suggestion.

L. 443: "500"

I wonder whether 500 samples are sufficient to robustly estimate the distribution of L under the assumption of a uniform DCC distribution. While some sensitivity analysis is provided in the appendix, the figures are somewhat limited in really showing the accuracy or convergence of the Monte Carlo approximation in each case.

Since P is a percentile rank estimated from a finite sample, its uncertainty can be approximated as $\sqrt{(P(1–P)/n)}$, which implies a standard error of about ±1–2% for n = 500. This level of uncertainty is likely acceptable for the broad classification presented in this manuscript.

However, it may become problematic if users wish to compare two scenes with similar P values. I suggest the authors consider including a more detailed justification of this sample size or provide confidence bounds on P where relevant.

This point is partially addressed in Figure B2 in the Annex section. As demonstrated by Figure B2 (where, even for 100 realizations, the distributions' parameters are highly comparable to the one with 500 realizations) and the numerous experiments conducted during the course of this study (not shown), 500 realizations are a satisfactory compromise between computational time and the quality of the generated L distribution across the two datasets. The estimation of P is highly consistent regarding the M factor, be it 100 or 1000.

Moreover, we also carried out an integrated sensitivity study, the details of which are provided below, at the end of this document. This study was conducted to address the P sensitivity; see Annex B3 (added). Please see also the detailed response to one of the final comments.

L. 447: "spatial organization" → "spatial aggregation"?

Done ⇒ "spatial arrangement"

L. 465: "such a specific spatial arrangement (or its equivalent) almost never occurs in the randomly generated scenes"

This statement could be made more precise. A scene being rare under the assumption of randomness does not, in itself, indicate whether the spatial distribution is clustered or regular. Since the main goal is to assess clustering or aggregation of DCCs, I suggest rephrasing this to highlight that P captures deviation from randomness, and that high P values specifically indicate increased clustering, rather than rarity alone.

We don't use the clustering notion to illustrate the meaning of the value of P. Besides, rarity is somehow a vague notion that can lead to ambiguous interpretations. We don't mention it directly. Instead, we express the notion of how many times an event occurs within the total number of attempts in the context of the generated scenes, as the percentile computation. Besides, as described in the paper, our objective is broader than simply distinguishing between aggregated and clustered, as with metrics such as lorg/Lorg, for example. Rather,

we aim to quantify the deviation from randomness and the structuring of convective cores. Therefore, we believe that the original phrase is appropriate in this context.

L. 472-473: "… the generation of the ensemble of scenes with a given convective fraction (F) is constrained by a few parameters of the stochastic model …"

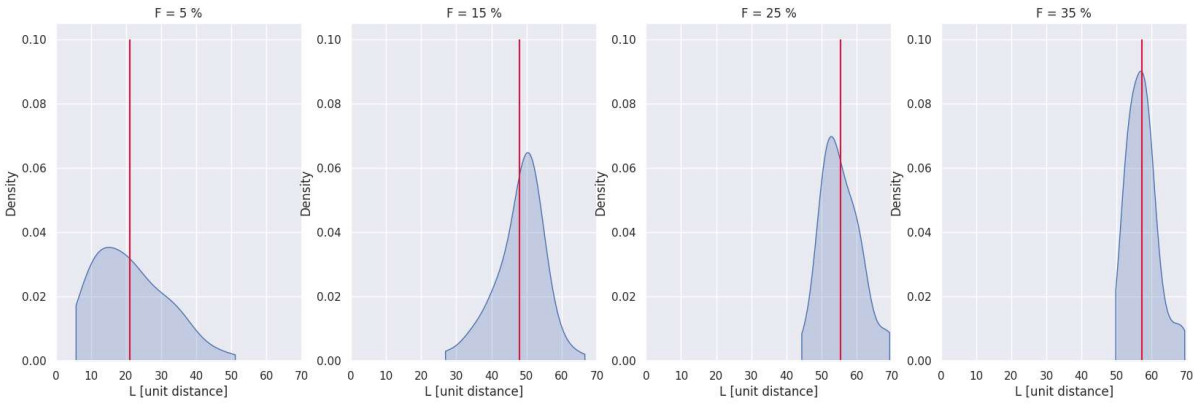
I would like to point out that the scene area is also a critical parameter influencing the metrics (see my comment above on scene area).

Modified as follows ⇒ "Indeed, the generation of the ensemble of scenes given a convective fraction (F) and convective area (A) is constrained …"

L. 475-477: "sensitivity of F"

The current analysis of the sensitivity of the L distribution to the convective fraction F (Figure 10, B1, B2) is informative but could be presented in a more visual or intuitive way. For example, plotting histograms or kernel density estimates of L across several discrete values of F could help the reader more clearly understand how F shapes the reference distribution and, by extension, the behavior of P. This might be more accessible than the current presentation.

Thank you for this feedback. The current representation aims above all to show that the evolution is non-linear as a function of F. Furthermore, the distribution of L is often almost symmetrical or pseudo-Gaussian (see figure below that is the L distribution for several F values, with the mean value represented in red), regardless of the value of F, but has activation thresholds that depend on the dominant second-order structures. We consider our original figure to be more informative and appropriate, as the main message is more about the activation of unconstrained data ranges for certain values of F and not the centering of the distribution in L (already visible with the blue shaded areas representing the +/- 1σ.



Empirical probability density function (estimated via KDE) of the generated distribution of characteristic length L for several F values, with the same condition of experiment as in Figure 10, right. Average value is shown in red.

L. 484: "challenging to compute a meaningful probability"

The phrase "meaningful probability" is somewhat vague, and it's not clear from Figure 10 what specifically supports this statement. If the issue is uncertainty in L and P for extreme

values of F, it would strengthen the analysis to move beyond the empirical filtering approach (e.g., excluding scenes with F < 8% or F > 25%).

Instead, I suggest a more quantitative uncertainty-based filtering method. For instance, the authors could:

1. Estimate or infer the uncertainty in L,
2. Combine this with the Monte Carlo sampling error (as discussed above) to deduce a total uncertainty in P,
3. And exclude scenes for which the total uncertainty in P exceeds a given threshold (e.g., ±5%).

This would provide a clearer and more defensible basis for filtering scenes, and would strengthen the methodological transparency of the approach.

Again please find the detailed answer at the end for the final comment that addresses those points in detail, completing the original sensitivity study.

Figure 10, B1, B2:

Please indicate what the shaded area represents.

OK added in the caption.

L. 499: “grid size” → “scene size”?

OK done.

L. 518: “consistent with the unfiltered distribution”

Please indicate whether this is shown in the paper or not.

OK done (“not shown” added).

L. 534:

See comment for line 465.

Again the sentence in question appears to be correct from our perspective. See the answer given for the precedent comment.

L. 535: “distribution” → “distributions”

OK done.

L. 549: “Empircal metrics … show”

The wording here is slightly misleading. The metrics are quantitative and not empirical in nature. It is rather the interpretation of these metrics that involves empirical reasoning or subjectivity.

OK done ⇒ “Classical clustering metrics”.

L. 551: “Figure 12” → “Figure 13”?

OK done.

L. 559: “more randomly distributed” → “less aggregated”

Modified ⇒ “less clustered”.

L. 562: “This indicates that F alone is not a discriminating variable …”

I do not find that the last two sentences indicate that F is not a discriminating variable (but maybe more the end of this sentence), please reformulate.

⇒ L562 : “This indicates that F alone is not sufficient to fully discriminate between organization types, as the spatial arrangement strongly influences the other three key parameters.”

L. 563: “very unlikely to be randomly spatially distributed”

This could also be the case of regular patterns in the other end of the distribution. Please specify.

OK modified ⇒ “very unlikely to be randomly spatially distributed with high L values”

L. 564: “do not necessarily have high F values”

The statement that “highly organized scenes do not necessarily have high F values” is phrased as if it were surprising. If this is indeed counterintuitive or contrasts with previous assumptions, it would be helpful for the authors to provide citations or context to support why this is the case. Otherwise, I suggest rephrasing the sentence to avoid implying that this decoupling is unexpected.

OK modified removing “necessarily”.

L. 566: “(Figure 12)” →”(Figure 13)”?

OK thank you, we modified it.

L. 578: “Figure 11” → “Figure 13”?

OK thank you, we modified it.

L. 546-585:

I find it difficult to form a synthetic view of the organizational classes after reading this paragraph, as the presentation feels somewhat interwoven and comparative from the start. I suggest restructuring the paragraph to improve clarity: the authors could first systematically describe the main characteristics of each class individually, and only after that proceed to

comparisons across classes. This would help readers better understand and differentiate the classes before being asked to contrast them.

The classification from random to organized is based on the P distribution (and secondary on the A and L variables for which the order is the same in terms of means and modes of distribution), that is more and more close to 1 as the class number increases. Note that class 0 and class 1 are not much closer to each other than any other Class.

We do not see the point in refining the details on each class individually before defining them as a physically sound family of convective patterns (which is done at the end of the paragraph), considering that in such clustering process, each class is distinguished by its major differences from the others, as detailed in the original text.

Nevertheless, we included a modification in the dedicated paragraph to complete the original description before comparing the classes: “As P increases (and secondary L and A too), so does the level of organization. The order of the classes follows this progression, starting with class 0 as the least organized and ending with class 3 as the most structured.”

L. 621-631:

I think it is also important and interesting that the authors compare their percentile-based organization metric (P) to the organization index Lorg introduced by Biagioli and Tompkins (2023). While Lorg is based on nearest-neighbor distances and does not require scene-level resampling, P is derived from the maximum spatial extent of the autocorrelation field and relies on Monte Carlo sampling.

It would be helpful if the authors could briefly discuss how these differing foundations may affect the quantification of the spatial aggregation of DCC and under what conditions one metric might outperform or complement the other.

Yes, thank you for the suggestion. Extending our comparison to other existing metrics would probably complement the work already carried out in the Discussion with COP, ABCOP and ROME metrics. It is important to note though that Iorg and Lorg are point process-based metrics differ from the present object-oriented approach. Iorg/Lorg offer a completely different way of quantifying *organized* convection, as they do not account for the areas of the convective objects involved, only their centers (e.g. their number). Actually, both these methods focus on the *clusteriness* of the center of the convective cores as explained in the Introduction.

Furthermore, Iorg is notably sensitive to the number N of objects (with a high negative correlation between Iorg and N) and their relative distances. This necessitates a stable number of objects (greater than 35) for reliable measurements. (Mandorli et al., 2024; Semi & Bony, 2020), which is often not possible in our data, even when using our decomposition process.

In summary, reconciling and comparing the P and Iorg/Lorg metrics may not prove easy and possibly not even interpretable due to the strong differing assumptions between the two methods. In particular we are not suggesting that P should be used as a standalone, continuous metric like Iorg, so the comparison may not be very meaningful. In any case, such a study is beyond the scope of this paper.

L. 642: “However, the fact that ABCOP distinguishes the classes more effectively than ROME suggests that while organization is more influenced by the total convective area (S), it cannot be fully captured by this single variable or its combination with F”.

I may be misunderstanding the logic here, but the sentence seems to suggest that because ABCOP outperforms ROME, this implies that spatial organization cannot be fully captured by S combined with F. It would be helpful for the authors to clarify the reasoning behind it.

OK modified ⇒ “However, the fact that ABCOP distinguishes the classes more effectively than ROME suggests that organization is more influenced by the total convective area (A) than the mean convective object area. Nevertheless, the overlaps within ABCOP distributions illustrate that the complexity of spatial arrangements cannot be fully captured by only A or its combination with F.”

L. 657: “In this study, a new method is introduced”

This sentence would benefit from a clearer link to the utility or purpose of the method, as it currently appears disconnected from the previous discussion. I suggest rephrasing to briefly restate what the method is intended to achieve.

OK modified ⇒ “In this study, a new method is introduced to quantify and characterize spatial arrangement of convective areas within a specific grid encompassing the instantaneous cloudshield of a deep convective system (DCS).”

L. 663: “establishing a probabilistic distance from a random spatial organization”

Not really (see previous comments). Please also note that the Lorg index (Biagioli and Tomkins, 2023) is also a measure of the deviation from a random spatial organization.

Here, the referee refers to the work of Biagioli and Tomkins, 2023, but to our knowledge, it is the OII index that measures this deviation, not Lorg directly (see citations below). Furthermore, Lorg is used in point-processes, not in object-oriented processes, unlike our approach.

> “To summarize the departure from randomness, we introduce a second index, the organization irregularity index (OII)” (Biagioli and Tomkins, 2023)
>
> “the organization irregularity index (OII), which is an integrated measure of the departure of convection from randomness across all spatial scales. Thus, Lorg and lorg give two integrated assessments of the mean organization, while the OII is an integrated measure of the variance of organization.” (Biagioli and Tomkins, 2023)

However, we have reformulated as follow: “establishing a probabilistic measure of the deviation from a random spatial arrangement”

L. 700:

Missing closing parenthesis.

OK done: “[...] for instance).”

Conclusion:

As I understand it, the TOOCAN algorithm defines Deep Convective Systems (DCSs) such that only one convective seed is permitted within each cloud shield. Given that deep convective cores (DCCs) typically coincide with the coldest cloud-top temperatures, this definition could influence the resulting spatial organization of DCCs within individual DCSs. In particular, constraining the number of convective seeds may limit the diversity of spatial arrangements that the method can capture.

We thank the referee for this remark. The TOOCAN algorithm does indeed identify convective seeds, but this procedure is based on a region-growing technique implying an iterative process of detection of convective seeds at several thresholds followed by a watershedding procedure applied on a spatio-temporal volume of infrared brightness temperature. Deep convective systems (DCSs) and their associated cloud shield are then identified throughout their full life cycle. However, these IR-based DCS definitions are not directly comparable to the deep convective cores (DCCs) detected in our study using spaceborne radar, which provides direct information on the convective precipitation cores rather than the extent of the cloud shield. The sensitivity of the TOOCAN seed definition has been carefully evaluated in dedicated publications, and it does not affect the collocation procedure applied here. More importantly, it has no impact on the identification of DCCs in the radar-derived precipitation fields used in this work.

While this does not undermine the value of the proposed approach, I suggest briefly discussing this methodological dependency in the conclusions—especially regarding the sensitivity of the characterization framework to the specific MCS tracking algorithm used.

The TOOCAN algorithm has been used in intercomparison exercises (Prein et al., 2024; Feng et al., 2025) and shown to perform well in identifying the life cycle of MCS. Yet here we are not dealing with MCS only but with the full spectrum of DCS for which such an intercomparison does not exist. Instead TOOCAN has been used in many physical studies showing its good performance (e.g., Elsaesser et al., 2022 among others) revealing its ability to depict a physically sound cloud shield life cycle for the wide spectrum of DCS. This gives us some confidence in using TOOCAN for this purpose. This said, our results are only valid for TOOCAN segmentation and only relevant in this specific context. We hope we have described in enough detail our technique that any interested readers shall be able to implement it for any other tracking algorithms and assess the relevance of the results in that case.

L. 888-889: “a sensitivity analysis revealed that this non-uniqueness has no significant effect on the subsequent methodology”

Please indicate that this is not shown in this paper.

OK done.

L. 905: “Figure B2” → “Figure B1”?

Yes thank you, we modified it.

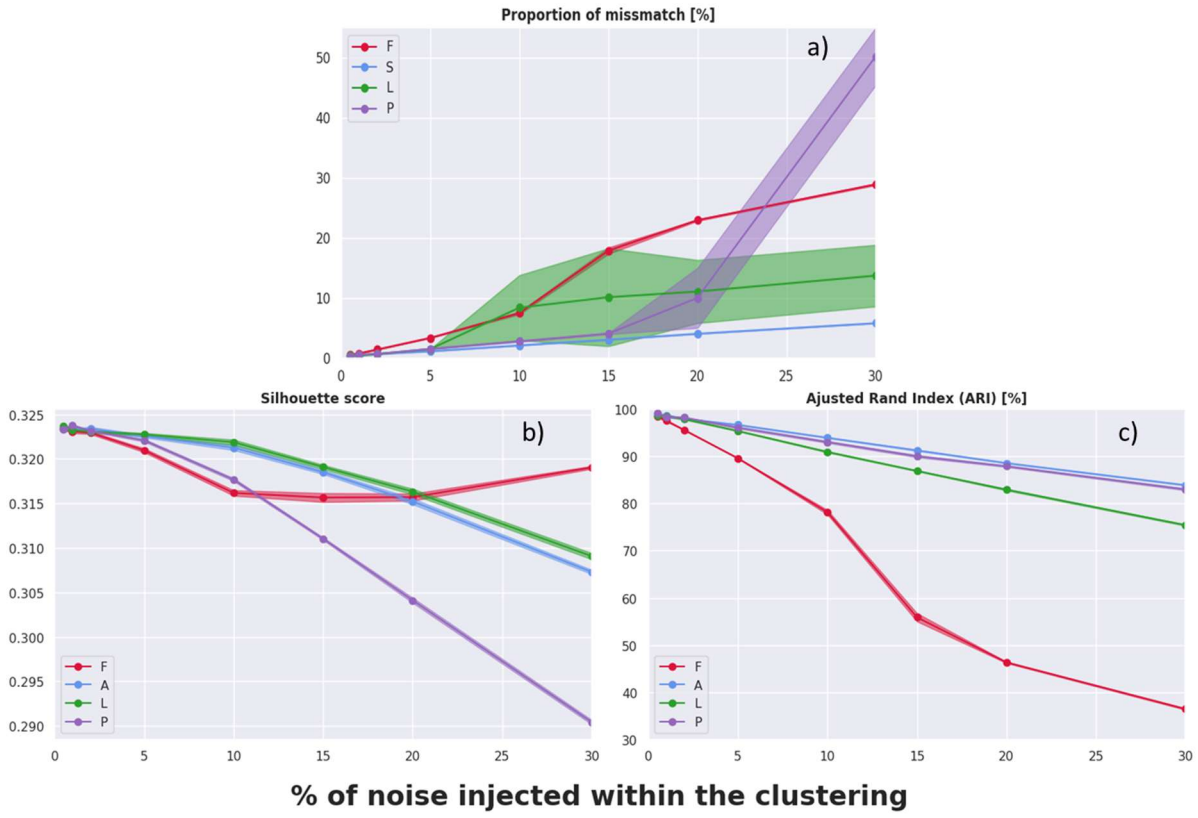
L. 909-912: “the most suitable”,”reasonable”

The use of terms like “most suitable” and “reasonable” in reference to scene selection feels vague, especially when based solely on empirical observation.

As mentioned earlier, it would strengthen the methodological rigor of the study to define selection criteria based on a quantitative threshold — for instance, by estimating the uncertainty on P for each scene and excluding those with uncertainty above a predefined limit. This would provide a more transparent and reproducible basis for filtering.

To answer this comment and some others that came before it, we present here several elements that complete the sensitivity study that has been already carried out in the manuscript:

A Monte Carlo-style study was conducted on K-means clustering, adding noise to each parameter in the radar dataset (~60,000 points). For each parameter, we added uniform noise with amplitude [-alpha/100*(Perc_90-Perc_10),+alpha/100*(Perc_90-Perc_10)], with alpha in % for values [0.5, 1, 2, 5, 10, 15, 20, 30]. Then we reclassify using the same Kmean algorithm (under the same conditions), and compare the class labels, calculating the number of points that change classes (or their proportion), the change in the silhouette score, and the Adjusted Rand Index (ARI), which quantifies the agreement between two partitions, correcting for random fluctuation. An ARI close to 1 means a very high correspondence between the two partitions, and when it approaches 0, we are close to an agreement corresponding to random draws. We perform this process 10 times for each value of alpha to estimate uncertainty on the mean values of each of the metrics, for each of the parameters. The results are presented in the figure below (that corresponds to the added Figure B3).



We summarized here several key findings that are described in the dedicated section in Annex B:

⇒ The order of importance regarding parameter sensitivity in classification is F, L, P, S

⇒ Up to +/- 5% noise on all parameters, the classification remains stable (with ARI>0.9, and even up to >10% for S, P, and L. Only F shows greater sensitivity above 10%

⇒ The uncertainty estimated with 10 experiments for each of the selected set of parameters/alpha combination also shows that the method is statistically robust.

We do not have the complete error model with error propagation in L and P (as it is difficult to access), we acknowledge that. However, thanks to these experiments, we have the consequences of a possible systematic disturbance/error on one or other of the parameters, and we note that within a range of approximately +/- 5-10% error, which is relatively significant, the classification is very stable and therefore considered robust.

For example, for 500 realizations, using the formula given by the referee on the uncertainty of P: $\sqrt{(P(1–P)/n)}$, we have a maximum error of approximately 2% (for values of P close to 0.5, but this drops to around 1% for values >0.9 (which is the majority of cases in this dataset), so it has no impact on the classification (see figures a, b, c of Figure B3).

L. 914: “We prove (not shown) that it would be interesting …”

Please reformulate.

Modified as follows: “We suggest that it could be interesting in such cases to increase the contour threshold up to 15% to better capture a proper range for L values.”

L. 924: “histograms” Figure B1 and B2 do not show histograms (... although it would be helpful to see histograms here, see comment on L. 475–477).

⇒ “histograms” replaced by “distributions”

Figures B1, B2: “Same as figure 9 ...” → “Same as figure 10”

⇒ modified in figures’ caption