Scalable feature extraction and tracking (SCAFET): A general framework for feature extraction from large climate datasets
https://doi.org/10.5194/egusphere-2023-592
EGU Sphere
July 10, 2023

1 Summary

A feature extraction algorithm is presented that combines elements of shape recognition with existing feature extraction methods based on more traditional atmospheric and oceanic variables. The utility of the method is demonstrated on 2D problems that include atmospheric rivers, tropical cyclones, and oceanic SST fronts. A 3D demonstration (jet stream analysis) is also presented.

The shape recognition procedure is based on an analysis of a smoothed data surface field using properties of the second order multidimensional Taylor series expansion, which in includes the local Hessian matrix of the surface. Such a technique has the potential to introduce an objectively defined qualitative "shape" into a data-based algorithm for feature detection, which is a significant advance. The variety of the chosen demonstration applications shows the high potential of the method for the broad field of climate modeling.

Unfortunately, the paper includes seemingly contradictory statements, statements that are not justified by the presented work, and omits some qualifications related to the method's practical use. For example, the introduction lists many challenging pieces of the climate data analysis problem and suggests that the presented method is a solution to them. These challenges include such things as objective identification criteria that are independent of specific model configurations, a reduced need for preprocessing data, and the challenge of dealing with enormous data output from high resolution simulations. When the results are presented in subsequent sections, however, the reader discovers that the filtering kernel must be tuned to a particular grid and that several of the standard preprocessing steps are still necessary. The computational performance of the method, relevant for the large data discussion, is not demonstrated. I therefore recommend major revisions.

The presented work has significant potential to become a new standard for climate data analysis, and I encourage the authors to continue its development.

2 General comments

1. Figure 1: The smoothing scale σ is defined as a function of the grid and the length scale of the feature of interest; however, this function is not defined in the paper. In the sentence beginning on line 131, the statement "This is implemented by calculating the value of σ along each circle of latitude," is particularly uninformative.

2. The introduction spends a lot of time and effort making a case for objective detection methods that do not rely on a human that will work well despite the challenge of "inter- and intra-model discrepancies" (line 38). Having to adapt the smoothing kernel to the grid spacing of specific resolution configurations seems to undercut those primary goals.

3. Koenderink and van Doorn (1992) advocate the use of pair of measures for shape recognition, "curvedness" and "shape index." The present work seems to discard "curvedness" but does not mention why. Furthermore, it does not discuss significant considerations related to the use of these measures for meteorological applications.

   The Hessian of a surface, $z(x, y)$, is a second-order term in the surface's local Taylor series expansion; its use here is only applicable if the first-order term (the gradient of $z(x, y)$) is zero. Indeed, even the

cited reference that the authors rely upon, Koenderink and van Doorn (1992), contains the strong caution that the interpretation of shape from the Hessian matrix (emphasis theirs) "is *only* valid in representations where the magnitude of the gradient of z vanishes." This would seem to suggest that a preprocessing step to find critical points (where the gradient is zero) is necessary, but I did not see such a step mentioned.

4. The discussion of the method's extension to 3D needs more detail. Given the horizontal-vertical splitting common to climate data sets, this is a nontrivial task. All of the shape index mathematics are formulated for 2D problems (with only 2 eigenvalues); for the 3D demonstration, the choice to use k_1 and k_3 and exclude k_2 seems significant. What is its physical interpretation? Perhaps this is justifiably outside the scope of the paper, but I wonder how would the definition of shape index change to account for all three eigenvalues? What are the analogous "shapes?"

5. The notation in equation (2) is confusing. It is customary for the numerator to contain the maximum order of each derivative; here, in all four terms the superscript 2 is missing. In the mixed terms, the symbol ∂ is missing between the x and y. The same comment applies to the indices.

6. I prefer to see the functional dependence of newly introduced methods defined explicitly; for example in Equation (3), $SI(k_1, k_2)$ is more informative than simply SI, and $\phi_s(x, y)$ should match $\phi_p(x, y, \dots)$ in equation (1).

7. What to the dots $(\dots)$ represent in equation (1), and in Figure 1's definition of ϕ_s in the exponent?

8. The EGU audience is interdisciplinary, and some may not be as familiar with the properties of the Hessian matrix and its eigenvectors. Adding their illustration to the dictionary of shapes in Figure 2 would be very helpful.

9. It's not clear how distinct and/or subjective the boundaries between different SI regions are; for example, how different is a "Rut" with SI = -0.374 from a "Saddle Point" with SI=-0.375? What about other boundaries, e.g., Ridges and Saddles, Ridges and Caps/Domes? How does this affect the various features that are sought — how easily could a "cap" be misclassified as a "ridge," and how significant might that be to the results of a study?

10. A validation study comparing SCAFET to existing methods (such as TECA, which the authors mention) would be helpful — given the same criteria, do they detect the same tropical cyclones? If there are differences, what are the characteristics of the storms that appear in one but not the other. Similarly, given the emphasis of the work on large data sets produced by high resolution models, how does the computational performance of the proposed method compare with previous methods? Does the method achieve faster processing times? Both variety of studies were performed in [1], which should also be cited here.

11. The paragraph beginning on Line 35 suggests that defining thresholds for particular features is challenging and can vary between and even within individual models. The implied suggestion is that the presented method, SCAFET, would solve this problem; however, the remainder of the work relies on the same expert analysis (for example, Table 1) that the paper claims to avoid elsewhere. Similarly, Table 2 presents a set of well-defined criteria for tropical cyclones that contradicts the Line 35 paragraph.

12. Consider re-drawing Figure 7 with a white background for hard-copy readers.

3 Specific comments

1. Figure 1: "Hessian" should be capitalized.

2. There are numerous grammatical and typesetting errors. An incomplete list includes:

   (a) The inverse tangent function in equation (3) should not be italicized; so should sin and cos in the Figure 2 caption.

   (b) Period missing after equation (3).

   (c) The symbols n and $n+1$ in Section 2.3 should be italicized.

   (d) "circ" in Lines 202 and 236 should be $^\circ$.

   (e) km in lines 203, 204 should not be italicized

   (f) Please use either $\tan^{-1}$ or arctan, not both.

References

[1] P. A. Bosler, E. L. Roesler, M. A. Taylor, and M. R. Mundt, 2016, Stride Search: A general algorithm for storm detection in high-resolution climate data, *Geoscientific Model Development* 9:1383–1398.