Supplement of Geosci. Model Dev., 17, 301–320, 2024
https://doi.org/10.5194/gmd-17-301-2024-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

Scalable Feature Extraction and Tracking (SCAFET): a general framework for feature extraction from large climate data sets

Arjun Babu Nellikkattil et al.

Correspondence to: Arjun Babu Nellikkattil (arjunbabun@pusan.ac.kr)

Contents

List of Figures

List of Videos

S1 Sensitivity Analysis of Parameters in Feature Extraction

To ensure the stability and reliability of the detection algorithm, we conducted a sensitivity analysis on various parameters used in SCAFET to detect Atmospheric Rivers (ARs) from ERA5 daily IVT data for the year 2019. This analysis assessed the impact of the parameter variations on key AR metrics, including the global mean frequency (Figure S1), total number (Figure S2), length (Figure S3), and area (Figure S4) distribution of ARs from the year 2019. In the sensitivity test, we systematically altered each parameter, while keeping all others consistent with the values specified in Table 1. The parameters subjected to sensitivity analysis includes:

- **Smoothing Scale:** This parameter determines the scale of smoothing applied to the data. The smoothing is applied as a preprocessing step to remove variability on scales much smaller than ARs.
- **Shape Index (SI):** The SI is used to extract specific geometric shapes from the dataset.
- **Coherence Angle:** This sets a threshold on the deviation between the local ridge direction and the local transport direction.
- **Minimum Length:** It sets the minimum length criterion for detected ARs.
- **Minimum Area:** This parameter defines the minimum area required for an AR to be considered valid.
- **Minimum Precipitation:** It establishes the threshold for minimum precipitation associated with the detected ARs.

The smoothing scale is a parameter that significantly influences the characteristics of the detected ARs. Specifically, when we increase the smoothing scale, which has the effect of smoothing the input field, we observe a corresponding increase in the global mean frequency of ARs (Figure S1a). This rise in the global mean frequency is coupled with an increase in the total number of detected ARs (Figure S2a). The reason behind these changes lies in the characteristics of the detected ARs. They tend to become larger both in terms of length (Figure S3a) and area (Figure S4a) as the smoothing parameter increases. This increase in mean area and length results in more AR-like objects that meet the filtering thresholds. Consequently, identifying a greater number of larger ARs on each time step leads to an overall increase in the global mean frequency of ARs.

The thresholds on the Shape Index (SI) determine the boundaries of the detected objects. Thus, as the chosen thresholds for SI increases, there is a corresponding decrease in

the global mean frequency of ARs (Figure S1b). This reduction in mean frequency is accompanied by a decrease in the total number of detected ARs, as illustrated in Figure S2b. The underlying reason for this decline becomes evident when analyzing the characteristics of the detected ARs. Higher SI thresholds result in the identification of smaller ARs, which is reflected in the length (Figure S3b) and area (Figure S4b) distributions. Consequently, many AR candidates fail to meet the stricter criteria and are filtered out, leading to the observed reduction in both mean frequency and total number of ARs.

As the permitted deviation of the local transport from the local ridge direction increases, a slight increase in both the global mean frequency and the total count of ARs is observed (Figure S1c, Figure S2c). These minor changes in mean frequency and counts are also mirrored in the sensitivity analysis of AR length and area (Figure S3c, Figure S4c).

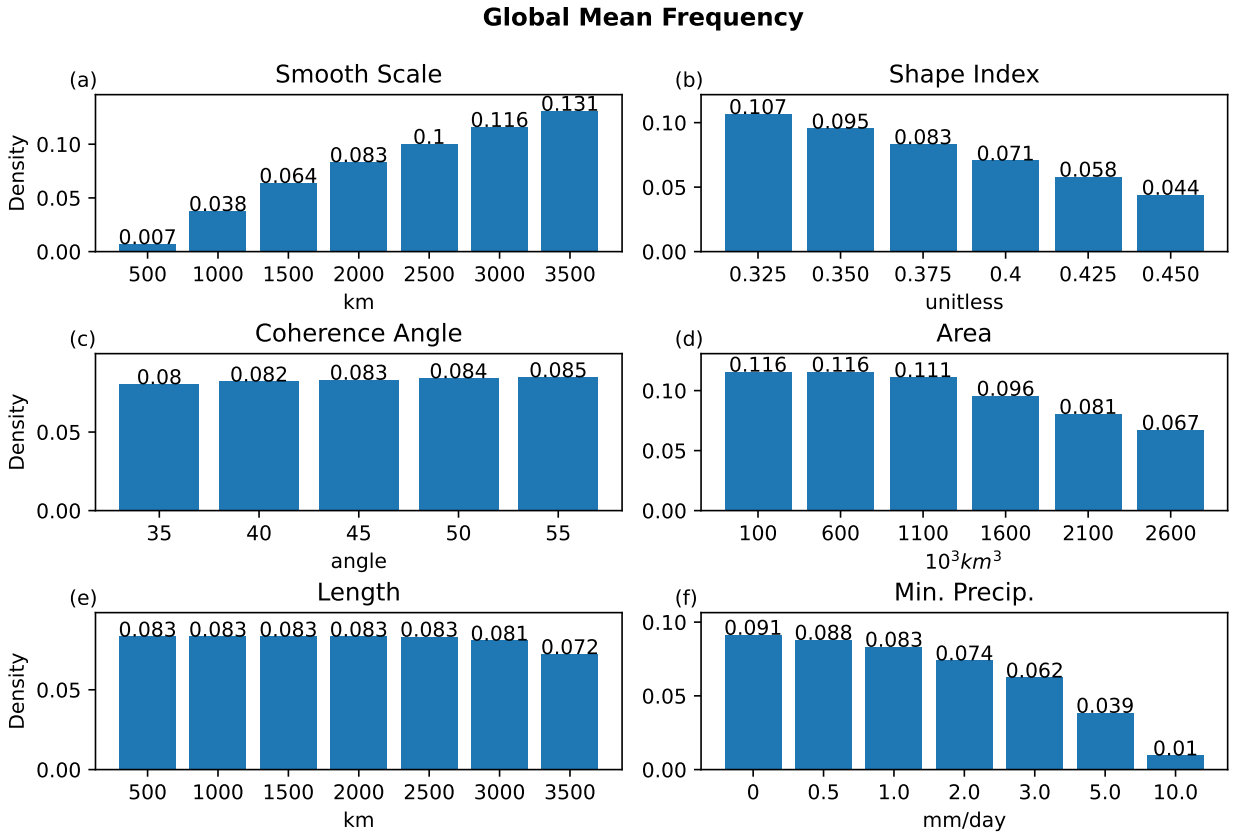
The sensitivity of mean AR characteristics to the minimum threshold on the area is relatively straightforward. As the threshold on area increases, both the global mean frequency and the total number of detected ARs decrease (Figure S1d, Figure S2d). This decrease occurs because only a smaller portion of the AR candidates can satisfy the filtering conditions. However, higher thresholds on the area would result in significantly increased mean area as well as length (Figure S3d, Figure S4d) for detected ARs. Notably, the minimum area thresholds have little impact if they are much smaller than the smoothing scale, as the smoothing process effectively removes smaller-scale variability.

The sensitivity of mean AR characteristics to length thresholds exhibits a pattern similar to that of area thresholds. When the minimum length threshold increases, both the global mean frequency and the total number of detected ARs decrease because fewer objects meet the length condition (Figure S1e, Figure S2e). Similar to the area thresholds, if the length threshold is smaller than the applied smoothing scale, the properties of the detected ARs show very low sensitivity to changes in length thresholds.

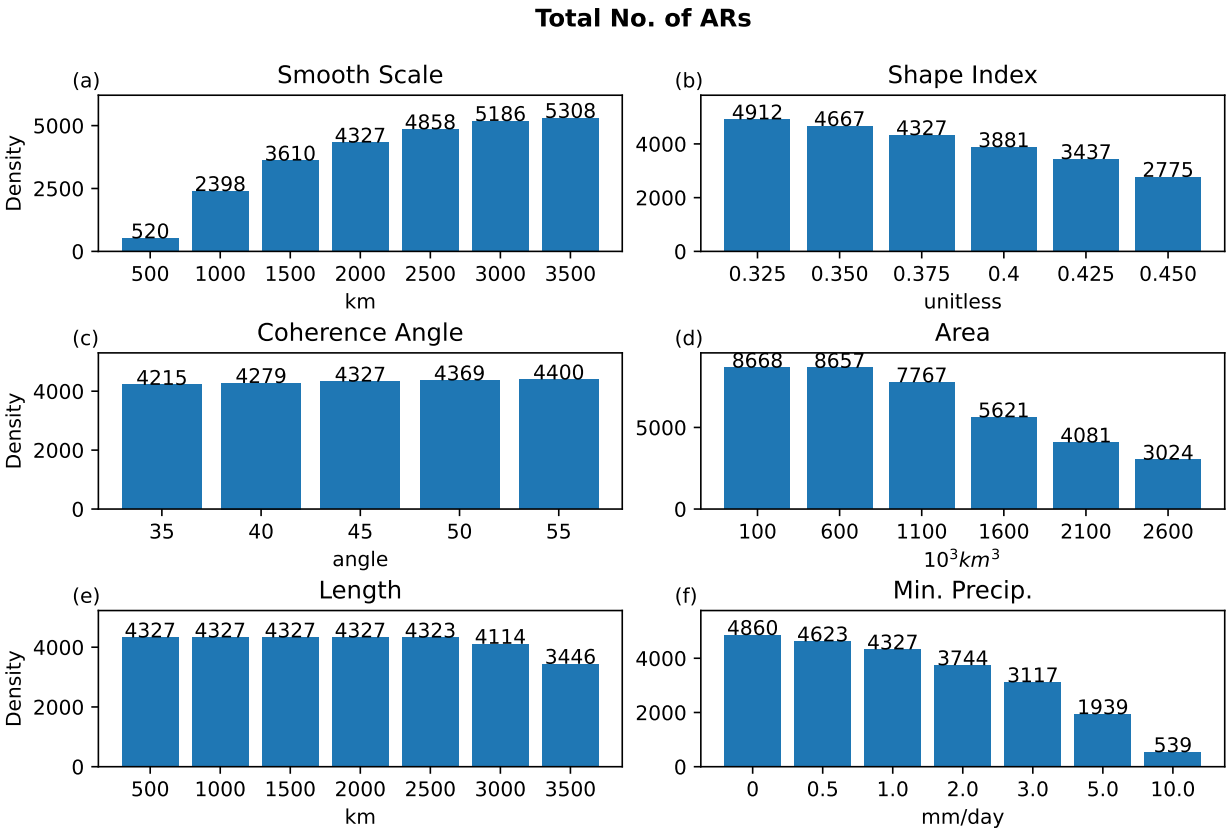
In SCAFET, mean AR precipitation intensity serves as a measure of AR strength, used to filter out weak AR candidates. When the condition on mean AR precipitation intensity increases, more objects are filtered out, resulting in a reduction in both the global mean frequency and the total number of detected ARs (Figure S1f and Figure S2f). In general, by focusing on strong ARs, we identify ARs that are longer and larger (Figure S3f and Figure S4).

In conclusion, this sensitivity analysis of various parameters in the SCAFET algorithm consistently produces reliable and comprehensible results. The choice of parameters significantly influences the characteristics of detected ARs. Increasing the smoothing scale leads to larger ARs, resulting in higher global mean frequencies. Conversely, stricter SI thresholds lead to smaller ARs and a decrease in both mean frequency and total count. Meanwhile, deviations in local transport from the ridge direction have minimal effects, as observed by slight changes in AR metrics. Minimum area and length thresholds show

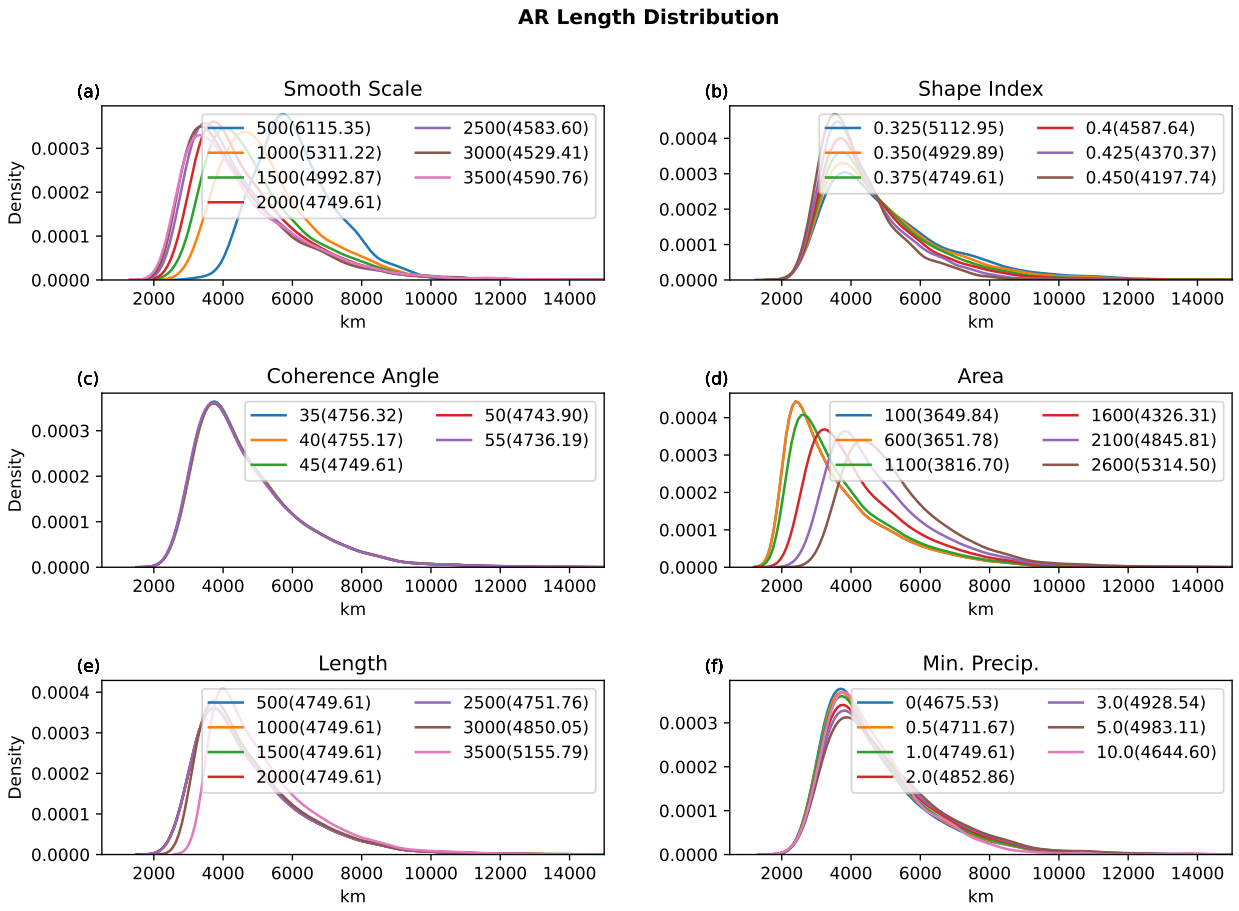
expected sensitivities: higher thresholds lead to fewer ARs meeting the criteria, but they exhibit a pronounced impact on AR size. Finally, filtering by mean AR precipitation intensity focuses attention on stronger ARs, which tend to be longer and larger. Our analysis provides a thorough validation of the algorithm's performance, enhancing confidence in its capacity to deliver consistent and meaningful results across various scenarios and datasets.



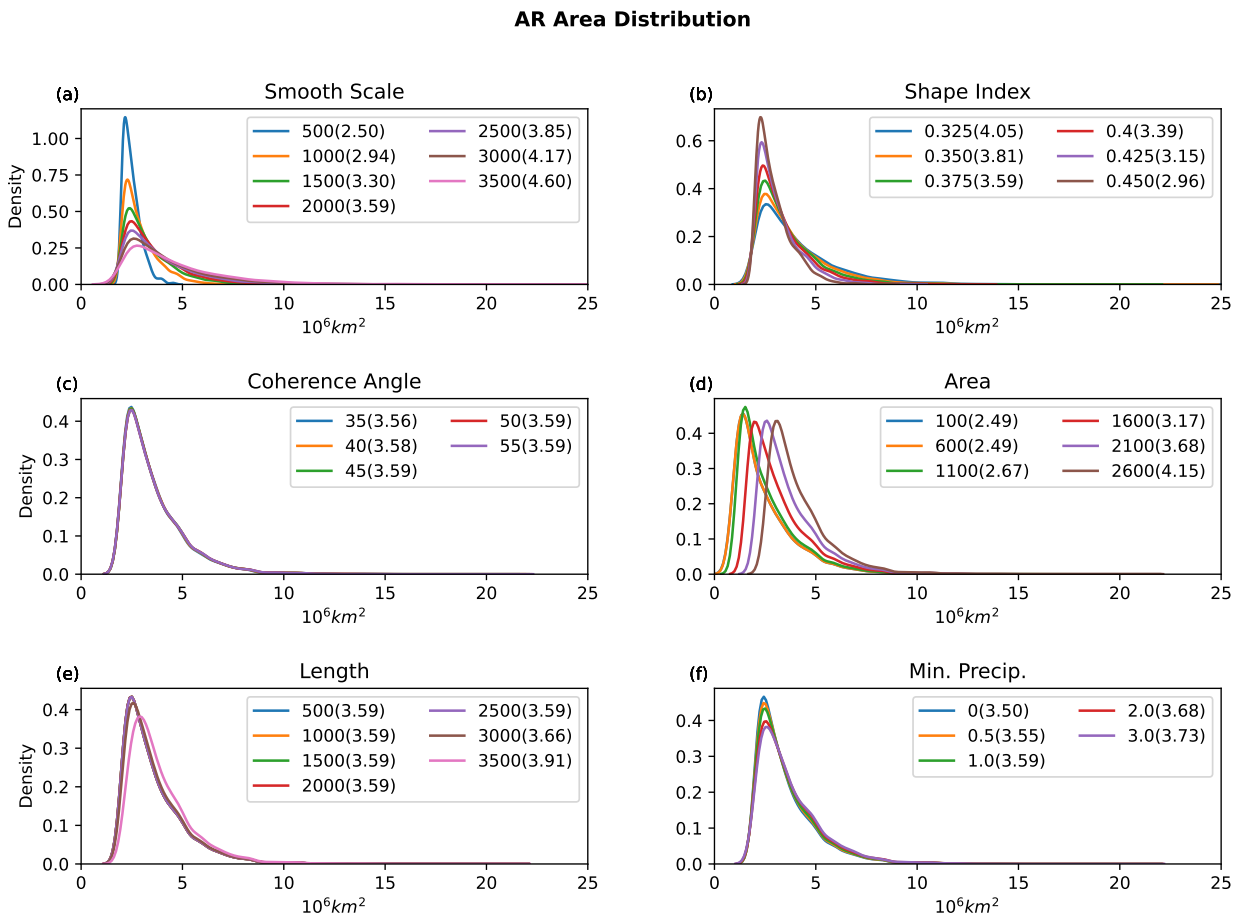
Supplementary Figure S1: Sensitivity of parameters used for Atmospheric River (AR) detection to the global mean frequency of ARs for the year 2019. Parameters such as (a) smooth scale, (b) shape index (SI), (c) coherence angle, (d) minimum area, (e) minimum length, and (f) minimum precipitation are considered for sensitivity analysis. For further details regarding each of the parameters, refer to the supplementary section and the main manuscript. The X-axis of each subplot shows the variations applied to each parameter during the sensitivity analysis while the Y-axis shows the probability density.



Supplementary Figure S2: Same as in Figure S1 but shows the sensitivity of each parameter to the total number of detected ARs for the year 2019.



Supplementary Figure S3: Sensitivity of the parameters used in AR detection to the length distribution of detected ARs for the year 2019. Probability distribution function changes to variations in (a) smoothing scale, (b) shape index (SI), (c) coherence angle, (d) minimum area, (e) minimum length, and (f) minimum precipitation are presented. Each curve in the plot represents distinct values used for these parameters, which correspond to the X-axis in Figure S1. The value of each parameter is specified in the legend of each plot, with mean values indicated within parentheses. The mean values, as well as the X-axis of the plots, display AR length in kilometers, while the Y-axis illustrates the probability density.



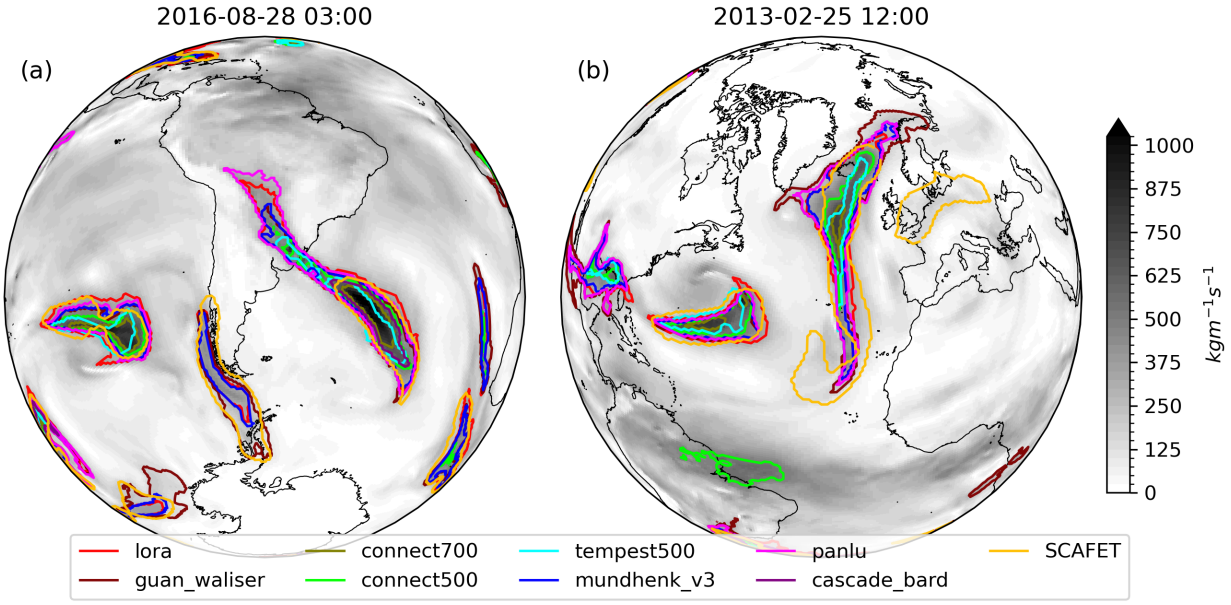
Supplementary Figure S4: Same as Figure S3, but shows the sensitivity analysis for the changes in AR area distribution. The legend in each plot specifies the values of the corresponding parameter, with the mean value indicated within parentheses. The units for the mean values are $\times 10^6 km^2$. The X-axis of the plot is presented in the same units, while the Y-axis displays the probability density.

S2 Comparison with Other Detection Algorithms

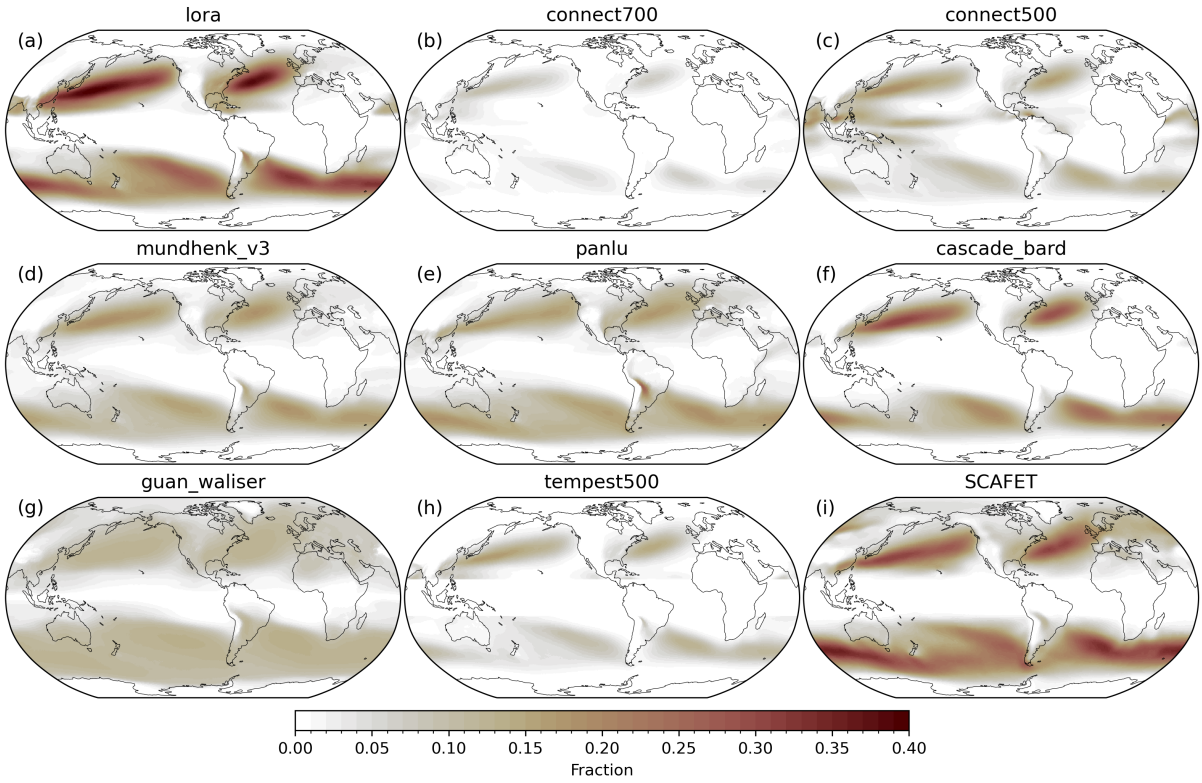
We compare the fundamental characteristics of the detected ARs with the results obtained from other AR Detection Tools (ARDTs). Specifically, we present the ARs detected using SCAFET and various other ARDTs as part of the Atmospheric River Tracking Method Intercomparison Project (ARTMIP). This comparison is conducted in two main aspects. First, we assess how effectively SCAFET detects individual ARs and determines mean AR frequencies in comparison with other detection algorithms (subsection S2.1). Second, we evaluate how the computational characteristics of SCAFET compare with those of select other tools (subsection S2.2). Additionally, this section addresses potential areas for further improvements that could enhance the algorithm's usability and efficiency.

S2.1 Detected Features

In our evaluation of SCAFET's performance in identifying ARs compared to other ARDTs, we present the detected masks from two randomly selected time steps from MERRA-2 reanalysis data (Figure S5). Our findings consistently demonstrate that SCAFET produces results comparable to those generated by other ARDTs participating in ARTMIP. Remarkably, SCAFET displays a tendency to identify more ARs than other algorithms. This is primarily attributed to SCAFET's independence from reliance on just physical thresholds. This characteristic is also reflected in the mean AR frequency depicted in Figure S6, which is notably higher on a global scale when compared to the majority of other ARDTs. The mean frequency, as illustrated in Figure S6, is calculated from three-hourly MERRA-2 reanalysis data spanning the period from 1980 to 2017.



Supplementary Figure S5: Intercomparison of the detected AR masks from SCAFET and 8 other ARDTs. Each ARDT is represented by uniquely colored contour lines, as indicated in the legend. Two random time snapshots, denoted as (a) 2016-08-28 03:00 and (b) 2013-02-25 12:00, were chosen for this comparative analysis of detected ARs. The shading corresponds to the magnitude of integrated vapor transport. Notably, SCAFET effectively identifies all of the prominent AR structures and more.



Supplementary Figure S6: Intercomparison of annual mean frequency of ARs obtained from SCAFET to that produced by 8 other ARDTs. The datasets used in this comparison were sourced from the Climate Data Gateway at NCAR (https://www.earthsystemgrid.org/search.html?Project=ARTMIP&q=&page=1&rpp=20). Additional information about each of the detection algorithms can be found at https://www.cgd.ucar.edu/projects/artmip.

S2.2 Computational Characteristics

In this section, we undertake a comparison of the computational characteristics of the feature detection method SCAFET with two other commonly utilized ARDTs. To ensure a fair and comparable environment for these comparisons, all three detection tools are executed on a macOS 12.6.9 system equipped with a Quad-Core Intel Core i5 processor clocked at 2.8 GHz and 16 GB of RAM. It is important to note that the computational measurements presented here should be considered as approximations, as actual processing speeds may vary depending on concurrent background processes running on the system.

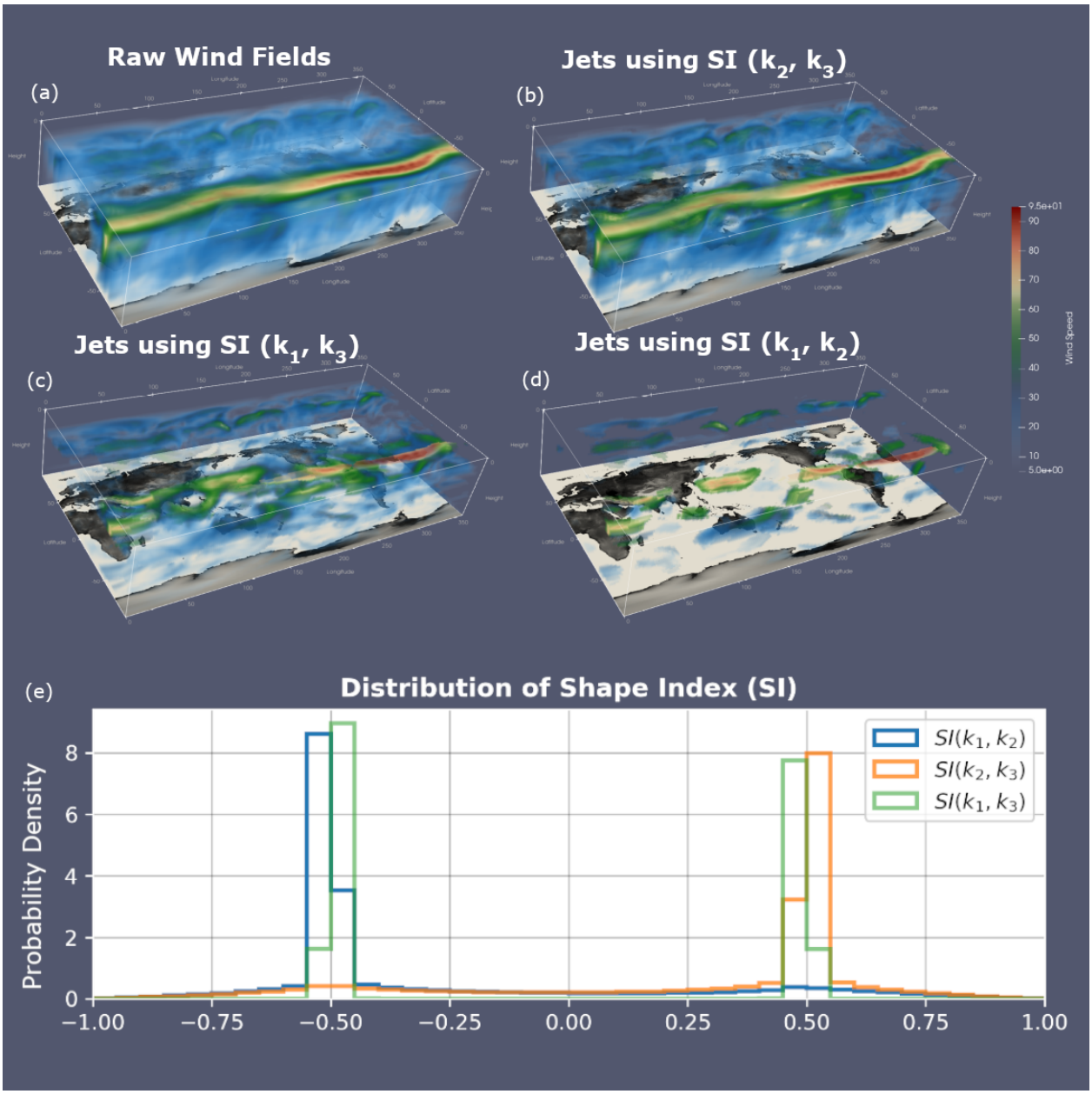
Upon comparing the results, it becomes evident that SCAFET requires more computational time compared to other tools. This disparity can be attributed, in part, to the inherent slowness of the programming language used for its implementation (https://scipy.github.io/old-wiki/pages/PerformancePython.html). Additionally, SCAFET is designed to extract an extensive list of properties from each detected object, which con-

tributes to the overall computation time. It is important to note that, in line with the other detection algorithms, the AR detection process in SCAFET is executed without tracking enabled. Consequently, SCAFET involves only two major steps: segmentation and filtering. Our analysis reveals that, of these two steps, the filtering process consumes approximately 98% of the computational time on average. As part of our ongoing efforts, we are committed to enhancing the computational efficiency of this step in the future.

Algorithm	Memory Usage (%)	CPU Usage (%)	Time (mins)
Guan_2015	4.9	208.0	66:53.43
Ullrich_2017	3.4	94.7	6:12.30
SCAFET	35.3	103.9	532:43.78

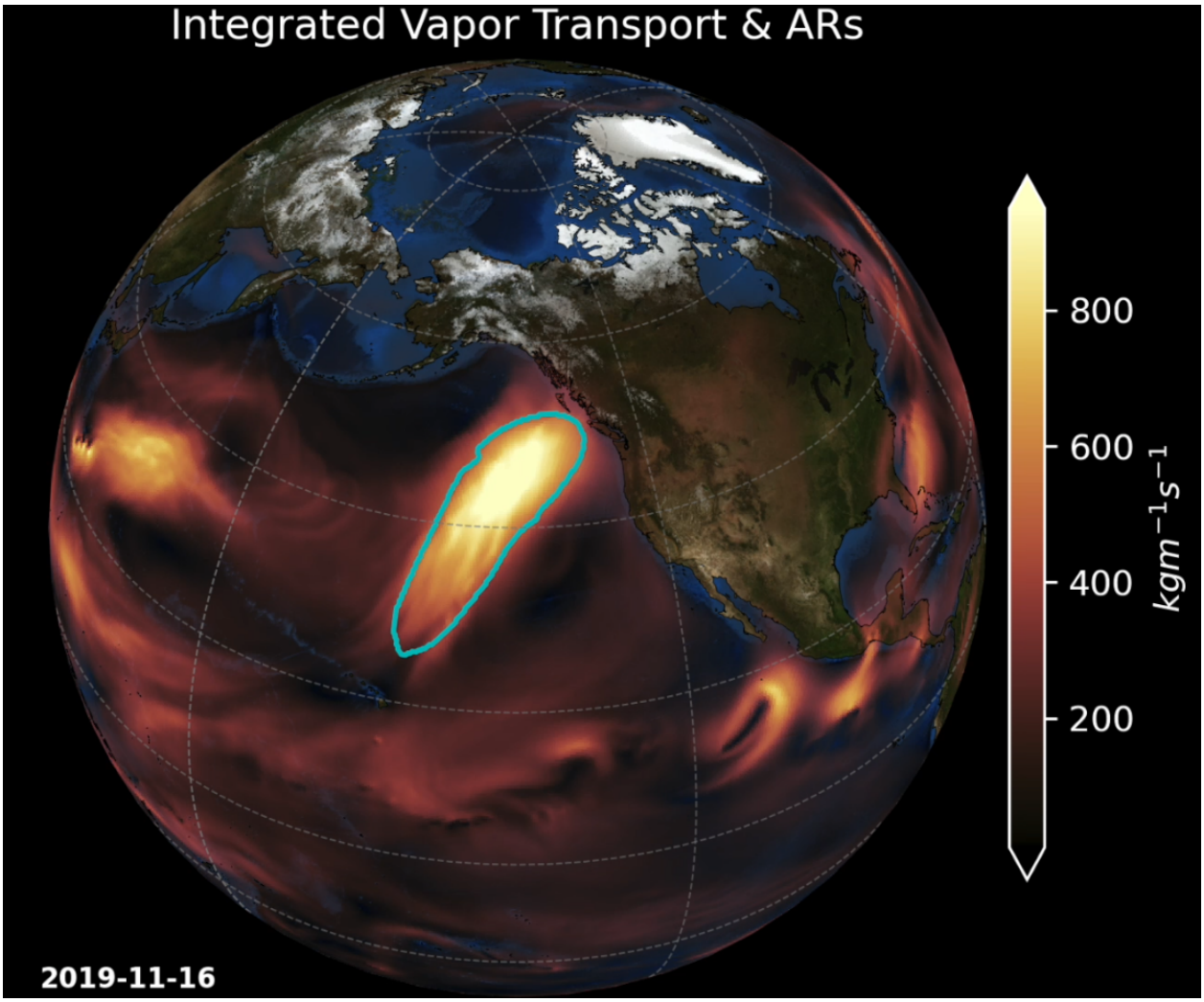
Table 1: Comparison of various computational characteristics between SCAFET and two other frequently utilized AR detection tools. Each row, from left to right, denotes the algorithm's name or reference, the mean memory usage, the mean CPU utilization, and the total time required to process a year of high-resolution ERA5 reanalysis data. The mean usage values are calculated throughout the detection process.

S3 3D Jet Detection using SCAFET

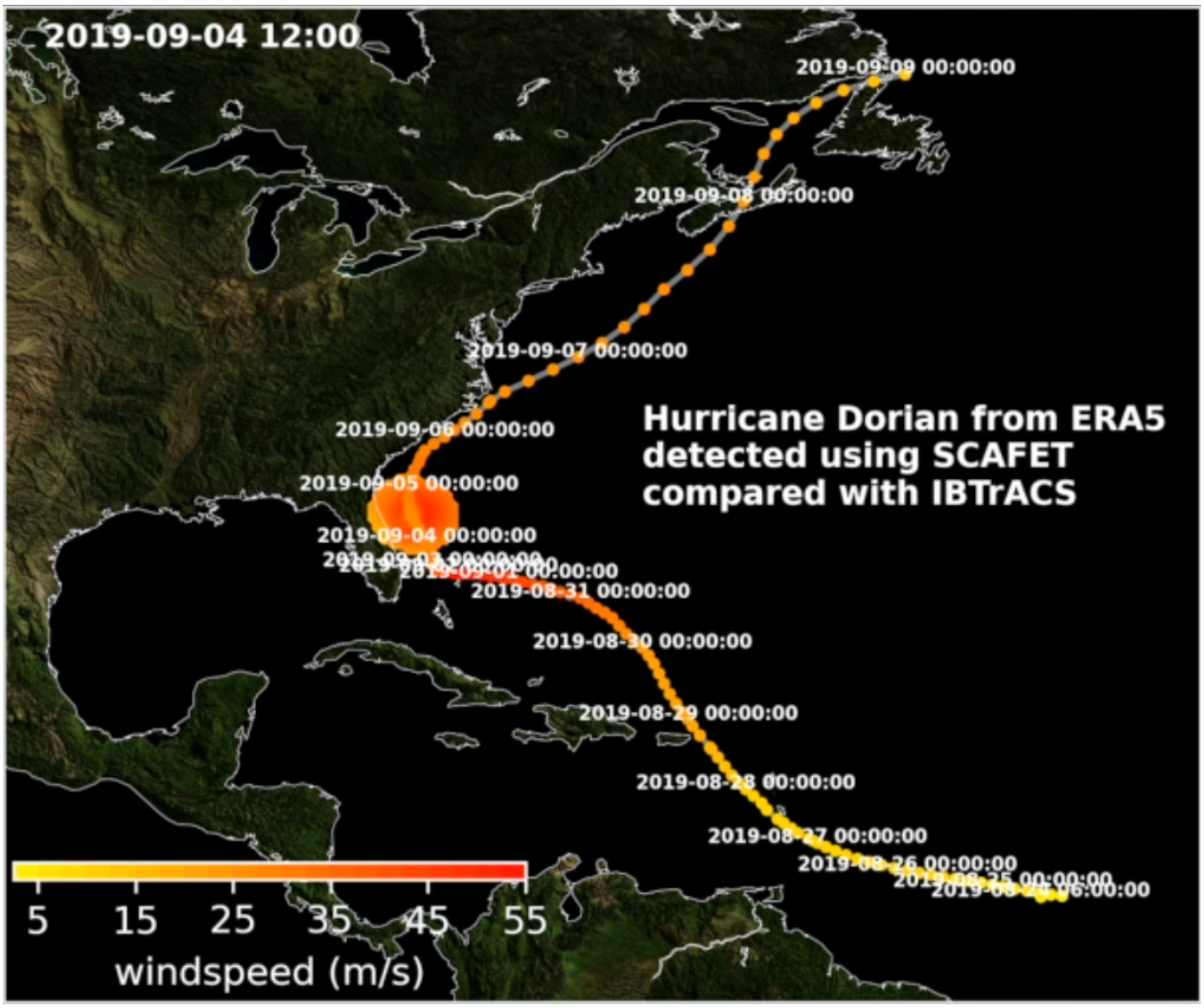


Supplementary Figure S7: Comparison of the three-dimensional jet features extracted using shape indexes (SIs) derived from multiple combinations of the three eigenvalues (k_1, k_2, and k_3). (a) Shows the three-dimensional wind speed which is the input to the detection algorithm for identifying jet streams. The jet locations are extracted by selecting regions where the SI is greater than 0.375. (b) The regions identified as jets by calculating SI as a function of k_2, and k_3. (c) The regions identified as jets by calculating SI as a function of k_1, and k_3. (d) Jets detected using SI as a function k_1, and k_2. (e) The probability distribution of the three different SI calculations for (a). The distribution indicates that $SI(k_1, k_2)$ gives the most conservative estimate of regions with convex curvature.

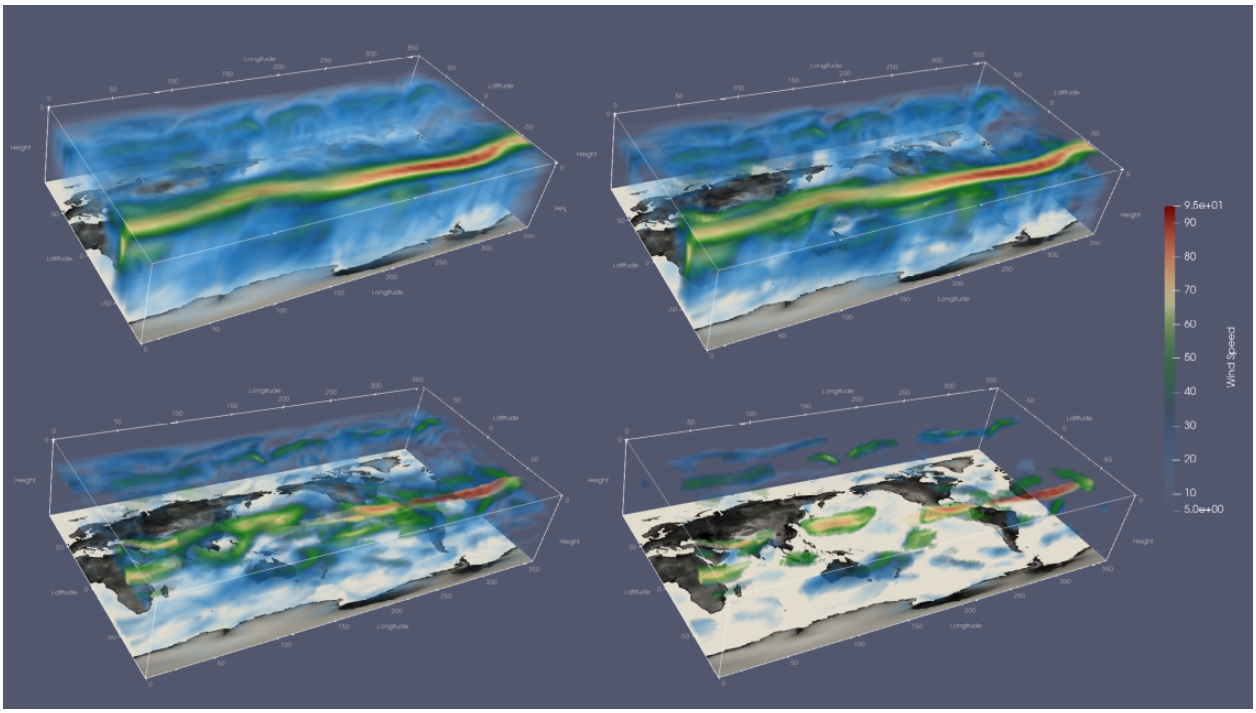
S4 Supplementary Videos



Supplementary Video S1: The video shows an Atmospheric River (AR) landfalling on the eastern coastline of North America detected using SCAFET. The dates for the plots are shown in lower left corner. The shading indicates the magnitude of integrated water vapor transport (IVT). The cyan outline is the detected AR.



Supplementary Video S2: The video shows the evolution of tropical cyclone "Dorian" identified by SCAFET from ERA5 reanalysis compared against the International Best Track Archive for Climate Stewardship (IBTrACS) track. The region identified as cyclone by SCAFET is plotted as the northward moving circular object with the shading indicating the windspeed. IBTrACS track is shown as the grey line with the dots representing the maximum sustained wind speeds.



Supplementary Video S3: The video shows the detection of three dimensional jet streams from windspeed data. The video shows the three dimensional windspeed field (left), used as the primary input field for jet detection and the jet objects detected using SCAFET (right). The shading indicates the magnitude of the windspeed. The represents 6 hourly data for the month of August 2022 (124 time steps).