Scalable Feature Extraction and Tracking (SCAFET): A general framework for feature extraction from large climate datasets

Arjun Babu Nellikkattil¹,², Travis Allen O’Brien³,⁴, Danielle Lemmon⁵, June-Yi Lee¹,²,⁶, and Jung-Eun Chu⁷

¹Center for Climate Physics, Institute for Basic Science (IBS), Busan, South Korea, 46241
²Department of Climate System, Pusan National University, Busan, Republic of Korea, 46241
³Department of Earth and Atmospheric Sciences, Indiana University Bloomington, Indiana, USA 47403
⁴Climate and Ecosystem Sciences Division, Lawrence Berkeley National Lab, Berkeley, USA 95720
⁵Office of Clean Energy Demonstration, U.S. Department of Energy, Washington, DC 20585
⁶Research Center for Climate Sciences, Pusan National University, Busan, Republic of Korea, 46241
⁷Low-Carbon and Climate Impact Research Centre, School of Energy and Environment, City University of Hong Kong, Hong Kong, China

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

Abstract. This study describes a generalized framework, Scalable Feature Extraction and Tracking (SCAFET) to extract and track features from large climate datasets. SCAFET utilizes novel shape-based metrics that can efficiently identify and compare features from different mean states, datasets, and between distinct regions. Features of interest are extracted by segmenting the data based on a scale-independent bounded variable called shape index (SI). SI gives a quantitative measurement of the local geometric shape of the field with respect to its surroundings. To demonstrate the capabilities of the method, we illustrate the detection of atmospheric rivers, tropical and extratropical cyclones, sea surface temperature fronts, and jet streams. Cyclones and atmospheric rivers are extracted from the ERA5 reanalysis dataset to show how the algorithm extracts both locations and areas from climate datasets. The extraction of sea surface temperature fronts exemplifies how SCAFET effectively handles curvilinear grids. Lastly, jet streams are extracted to demonstrate how the algorithm can also detect 3D features. SCAFET can be implemented to extract and track most weather and climate features.

1 Introduction

The amount of climate data is growing exponentially owing to rapid expansions in both observational capabilities and computational power, driven by the need to observe and simulate ever-higher resolutions (Overpeck et al., 2011; Balaji et al., 2018). Frontier research like global cloud resolving and large ensemble simulations leads not just to increased volume but also to inflated velocity, variety, veracity, and value (5Vs) of climate data (Marr, 2015; Guo, 2017; van Gnderen et al., 2019) of climate data. This makes the detection of important atmospheric and oceanic features, such as atmospheric rivers (ARs), tropical and extratropical cyclones, sea surface temperature fronts (SSTFs), and jet streams, a daunting task. Although these features of interest influence regional and global weather and climate with immense societal, economic and ecological impacts, the amount of data representing these events and features would be a small percentage of the whole simulation. Thus, extraction
of features not only enables us to focus our analysis on high-impact rare events but can also considerably reduce the amount of data that needs to be stored, improving computational efficiency in analysing these features (Yang et al., 2016). Moreover, the mean, variability, and characteristics of features can be compared to observational data sets as a measure of bias within model simulations and various parameterizations (Sellars et al., 2013). Efficient and reliable extraction of these features is thus vital to climate data processing, analysis, and model development.

Despite the importance of feature extraction in climate data analysis and informed model development, there is little consensus on standard best practices for feature extraction. The simplest method for extracting a feature is to use a physical threshold for some climate variable (SST, precipitation, wind speed, humidity, etc.), or a combination thereof, to identify ARs, fronts, jet streams or tropical and extratropical cyclones (Bengtsson et al., 1982, 1995; Vitart et al., 1997; Hewson, 1998; Koch et al., 2006; Strong and Davis, 2007; Rutz et al., 2014; Guan and Waliser, 2015). The limitations of and discrepancies between these methods are linked to the somewhat arbitrary choice of physical thresholds in relation to the underlying spatio-temporal distributions of the climate variables. In other words, many studies choose a physical threshold that is not theoretically defined but rather a function of the location, timespan, and dataset used. Validation then unfortunately comes down to the intelligent but subjective human eye, or in other words tuning a threshold until it appears to have captured all the features of interest while leaving out the background noise (Zarzycki and Ullrich, 2017; Vishnu et al., 2020).

Choosing an absolute threshold from climate variables for feature extraction that is applicable to different climate models and spans multiple mean states and model scenarios is not straightforward. Even within the same model, a particular choice of threshold may be suitable for one region but not for another, given varying regional characteristics and topography. To account for these inter and intra-model discrepancies, thresholds are often calculated from the model- and/or simulation-specific distribution of basic climate variable fields in which the features are most visible, such as relative vorticity and sea level pressure anomalies for tropical cyclones (e.g., Vitart et al., 1997) or integrated water vapor transport for ARs (e.g., Guan and Waliser, 2015). Thus, before the actual detection process is applied, one must pre-process entire datasets just to calculate reasonable thresholds that will allow for comparison within and between models. This process becomes increasingly infeasible for higher resolutions and large ensemble data sets, highlighting the need for a method of feature extraction that is not empirically derived and thus less sensitive to the climate mean state.

Aside from the inter and intra-model discrepancies that arise from detecting features in present and historical model simulations, applying empirical present thresholds to detect features in future climate change scenarios is further untenable as the underlying spatio-temporal climate variable distributions change under global warming. Feature detection must be reconsidered when applied to variables with significant changes in their means and extremes in response to external forcings such as doubling or quadrupling carbon dioxide (CO₂) concentrations. It should be emphasized that applying different arbitrary thresholds can and does lead to contradictory conclusions regarding the response of these features to greenhouse gas warming (Horn et al., 2014; Zhao, 2020; O’Brien et al., 2022). To counter these uncertainties, methods based on topology, machine learning, ridge extraction, edge detection, and various other image processing techniques have been proposed over the years (Dixon and Wiener, 1993; Post et al., 2003; Molnos et al., 2017; Biard and Kunkel, 2019; Xu et al., 2020). While these meth-
ods offer an alternative for the extraction of features in datasets spanning different mean states, many of these methods were developed for detecting specific rather than general features.

In this study we introduce a novel method, Scalable Feature Extraction and Tracking (SCAFET), which is a general framework to detect and track features of various shapes, scales, and intensities. Simply put, SCAFET uses the curvature of a given scalar field to identify emergent shapes that correspond with distinct features of interest. The shape is calculated as a finite, bounded, and scale-independent quantity and can be tuned depending on the desired phenomenon. As this tuning relies on shape-based rather than physical thresholds (see Appendix A), the characteristics of the detected features are less sensitive to spatio-temporal and mean state variances. This also makes the feature extraction fully parallelizable along the time dimension, as the detection is carried out independent of the time information. SCAFET utilizes some of the modern python packages like xarray to handle NetCDF files and dask to parallelize the detection process in any machine with multiple cores. The code for this framework is fully open-source and written in Python in an easy-to-use package so that even beginner-level Python users can easily implement the algorithm.

The need for a general framework in extracting and tracking features from large climate datasets has been raised in various climate science communities for the last several decades. In a pioneer study, Hodges developed a three-step general framework for extracting and tracking features from meteorological datasets. The first step is segmentation, where the field is split into distinct regions by applying a threshold and then labelling each of the connected regions as an object. Later, feature nodes are extracted by filtering out regions outlined in the first step based on the characteristics of each object. Finally, the feature nodes are tracked over time to produce the final output for dynamical analysis. This framework was developed further for cyclones, storm tracks, convective systems, ocean eddies, monsoon depressions, and more (Hodges, 1995; Hogg et al., 2005; Hodges et al., 2011; Burston et al., 2014; Hurley and Boos, 2014; Pinheiro et al., 2016; Priestley et al., 2020; Torres-Alavez et al., 2021; Karmakar et al., 2021). However, it is limited to the detection of points of local maxima in two-dimensional scalar fields, which do not always fully characterize various features.

In 2012 a team from the Lawrence Berkeley National Laboratory developed the Toolkit for Extreme Climate Analysis (TECA), integrating pre-existing, physical threshold-dependent detection methods and algorithms into a comprehensive software package that was parallelized to make the algorithms more suitable for large datasets (Prabhat et al., 2012). In a more recent effort, a team led by Paul Ullrich at the University of California-Davis created TempestExtremes (Ullrich and Zarzycki, 2017; Ullrich et al., 2021), another algorithm package that uses several core functions to detect a variety of features. These functions are being actively developed for extraction, characterization and uncertainty quantification of weather extremes. Both TECA and TempestExtremes have been widely implemented by the climate community and have been monumental in advancing scientific understanding of meso and synoptic scale processes and their contributions to long-term climate trends. Further discussion on the differences between SCAFET and other detection algorithms can be found in Appendix A. Even in the context of these recent advancements, SCAFET aims to upgrade the detection process to be less sensitive to the physical thresholds used while presenting a novel shape-based approach to feature extraction.

The novelty of SCAFET compared to pre-existing methods lies in the use of a comprehensive mathematical framework for extracting different features based on the overall “shape” of a climate variable field, rather than thresholding of that field. The
core methodology for the detection of any feature is the same and can be tuned using just two variables, one for the spatial scale and the other for the shape of features one is looking for. For example, between the two variables one can tune the difference between a long filament-shaped atmospheric river and a shorter round-shaped cyclone. The algorithm is applicable to both rectilinear and curvilinear grids and can also be extended to detect three-dimensional (3D) features. In a nutshell, SCAFET is mathematically comprehensive, computationally efficient, and easily implementable, and could prove to be a robust method for detecting a diverse set of features under different mean climate states.

The paper is organized as follows, section 2 describes the basics of SCAFET and how it is implemented in a two-dimensional (2D) field. Section 3 provides three SCAFET use cases for the detection of various climate features and different grid types. Extraction of 3D features will be discussed in section 4. Though the application of SCAFET is not limited to the features described here, this study showcases atmospheric rivers, cyclones, SST fronts, and jet streams to cover a broad range of phenomena through which users could learn to adapt SCAFET to their needs.

2 Description of Scalable Feature Extraction and Tracking

SCAFET follows the same three processes as discussed by Hodges - Segmentation, Filtering and Tracking (yellow boxes in Figure 1). Before starting the this process, SCAFET is initialized with information describing the datasets and the type of feature to be extracted (blue boxes in Figure 1). Primary inputs includes the following.

- **A primary field** ($\phi_p$): This is a gridded dataset in which the feature to be extracted is most clearly visible. Optionally, one or more secondary fields can be used to further constrain the detected features.

- **Grid Properties**: Information on the primary field including grid cell area/volume, grid distance, and coastlines are required for calculating derivatives of the basic field and identifying landfalling locations.

- **Object properties**: The algorithm requires information on the feature properties. This includes approximate spatial scale, shape, eccentricity (for 2D features only), minimum length, minimum area, minimum volume (for 3D feature only), minimum duration, and maximum distance per time step. (see section 3 and section 4 for examples)

As the segmentation, filtering, and tracking are mutually independent in the SCAFET scheme, users can replace any of them and still run the algorithm. After implementing all three steps, two outputs are obtained: one describing the properties of the detected objects and the other containing the labelled mask of the feature of interest in the input grid (pink boxes in Figure 1).

2.1 Segmentation

The core operation behind the extraction of features is to classify points in a scalar field into one of five shapes using the two principal curvature measurements derived as eigenvalues of the Hessian of the basic field. The five chosen shapes (see Figure 2) are an abridged version of the shapes described in previous studies (Koenderink and van Doorn, 1992). Depending on the feature of interest, one or more shapes are extracted from the primary field. Segmentation starts with scale-space selection

---

https://doi.org/10.5194/egusphere-2023-592
Preprint. Discussion started: 9 May 2023
© Author(s) 2023. CC BY 4.0 License.
of the field to remove smaller scales of variability that are background noise compared to the feature of interest. Next, the local geometric shape is calculated.

2.1.1 Scale-space Selection

Scale-space selection is a very common tool used in image and signal processing, as well as computer vision (Lindeberg, 2014). In this current work, the scale-space selection is limited to the application of a gaussian smoothing kernel to suppress variability less than the smooth scale ($\sigma$). Scale-space selection is mathematically implemented as a convolution of the primary field ($\phi_p$) with a gaussian function given as,

$$\phi_s = \phi_p(x, y, \ldots) * \frac{1}{2\pi\sigma} e^{-\left(x^2+y^2\right)/2\sigma^2}$$

(1)

In the context of the meso-synoptic scale processes explored in this study, scale-space selection will filter out smaller micro-scale features to isolate features like cyclonic vortexes or atmospheric rivers. However, this function can be adjusted to the
spatial scale of interest and could even theoretically be used to filter out synoptic scale features in isolating micro-meso scale processes. In climate datasets, the grids are not always uniformly spaced. To account for that, we adapt the above equation to be “grid-aware”. This is implemented by calculating the value of $\sigma$ along each circle of latitude. In future studies, one could experiment with other, more sophisticated scale-space selection methods.

### 2.1.2 Local Shape Extraction

The local geometric shape of the field, $\phi_s$, is calculated as a function of the eigenvalues ($k_1$ and $k_2$) of the Hessian of the magnitude of the field ($\|\phi_s\|$), where the Hessian is given by,

$$
\mathcal{H}(\|\phi_s\|) = \begin{bmatrix}
\frac{\partial \|\phi_s\|}{\partial x^2} & \frac{\partial \|\phi_s\|}{\partial x \partial y} \\
\frac{\partial \|\phi_s\|}{\partial y \partial x} & \frac{\partial \|\phi_s\|}{\partial y^2}
\end{bmatrix}
$$

(2)

![Figure 2](https://doi.org/10.5194/egusphere-2023-592)

The abridged version of the shapes used in this study and the values of the shape index associated with each of them. The X and Y axis are a set of general axis while $Z(X,Y) = \sin(2X) + \cos(2Y)$. Regions within $Z(X,Y)$ satisfying conditions for different shapes are isolated to show the geometry associated with them.

From simple differential geometry, we know that if $k_2 \leq k_1 < 0$, the point under consideration is a local maximum, whereas if $k_1 \geq k_2 > 0$, the point is a local minimum. The applicability of such a criterion for feature extraction is limited to nodal features like tropical cyclones or monsoon depressions. To induce continuity in the shape extraction, we use shape index (SI) $(Koenderink and van Doorn, 1992)$, a quantitative measure of the local shape of the field defined as,

$$
SI = \frac{2}{\pi} \tan^{-1} \left[ \frac{k_2 + k_3}{k_2 - k_1} \right]
$$

(3)

Where $k_1$ and $k_2$ are the two eigenvalues, satisfying $k_1 \geq k_2$, for the Hessian matrix. The SI is used to classify the primary field into distinct shapes (see Figure 2). The values of the SI selected are set depending on the type of feature to be extracted.
For instance, caps and domes are selected to extract features such as atmospheric depressions or cyclones. Ridges, caps, and domes are selected to extract ARs and fronts.

SI is designed to be a bounded value (range -1 to 1) independent of the magnitude of the field (Figure 3). In simple terms, SI gives a continuous quantitative measurement of geometric shape of the field with respect to its immediate background field. This might be similar to how the trained eye of a climate scientist detects features from color/value contrast, though SI is arguably a more objective measure of geometric shape. These characteristics make it more suitable for feature extraction from datasets with varying mean states compared to traditional physical threshold-based methods. In addition to the two eigenvalues, the shape extraction provides us with corresponding eigenvectors. The eigenvector for $k_2$ points perpendicular to the local ridge direction while that of $k_1$ is parallel to it. This allows us to set further constrains on the local shape extraction if $\phi_s$ is a vector field, as demonstrated in the detection of ARs in subsection 3.1.

![Shape Index](https://doi.org/10.5194/egusphere-2023-592)

Figure 3. Sensitivity of shape index to eigenvalues $k_1$ and $k_2$. The X and Y axis represent values of the two eigenvalues used for calculating shape index while the color indicates the value of shape index.

### 2.2 Filtering

Once the right shapes are extracted, properties like area, location, mean, minimum, and maximum values of different properties are calculated for each of the objects. A series of filtering is carried out to remove objects which do not satisfy certain conditions regarding (a) grid properties like area, length, region masks, etc. (b) primary field properties like magnitude and direction, and (c) constraints from secondary field(s). The aim of the filtering process is mainly to remove small or weak objects. Since filtering is applied to the extracted objects rather than an entire field, computational cost is decreased relative to other methods.
2.3 Tracking

The extracted properties of each object include positional information for the center, maximum, and minimum values. To track objects through time, one of the pieces of the positional information is used. In the current study, a simple radius is defined and the closest object within the given radius to each object at time n is clustered and identified from time n+1 as the same object in motion. While this simple tracking method may not translate to micro-scale processes, it could be modified with more complexity if necessary.

3 Application to 2D Features

In this section, we exemplify SCAFET to detect cyclonic vortices, ARs, and SSTFs from various climate datasets. While the highlighted examples demonstrate SCAFET’s broader capabilities as applied to diverse features and grid types, all the examples go through the same general process shown in Figure 1. Each subsection has a table of parameters detailing the properties of the desired feature. The properties include typical spatial scales of the feature, shape index (SI), minimum length, minimum area, object eccentricity, and minimum duration of the track. The quantitative values for the properties are obtained from a consensus of previous studies referenced within each section. Apart from the results discussed in the sections below, videos for each feature are also attached in the supplementary section. As the aim of this work is to demonstrate the ability of SCAFET to detect various features, the results for long term climatology for each of the features presented for comparison with other published detection algorithms.

3.1 Atmospheric Rivers

According to the American Meteorology Society’s glossary of meteorology, atmospheric rivers (ARs) are “long, narrow, and transient corridors of strong horizontal water vapor transport that are typically associated with a low-level jet stream ahead of the cold front of an extratropical cyclone” (Ralph et al., 2018). Much of the precipitation and water vapor transport in midlatitudes occur within AR structures (Guan and Waliser, 2015). They are also responsible for over 50% of the extreme precipitation and wind events in the midlatitude region (Waliser and Guan, 2017; Nash et al., 2018). Detection and accurate projection of ARs are crucial for extreme weather preparedness as well as for water resource management in basins across the globe.

The ambiguity in the AR detection schemes stems from the quantitative definition of strength, length, narrowness, and other parameters used in detection. In comparison with other criterion, choosing how to fix the threshold for AR strength changes the inferences drawn between the detection schemes. Most of the AR detection algorithms empirically derive this threshold from dataset directly, making it sensitive to spatio-temporal and mean-state variances (Shields et al., 2018). SCAFET defines ARs as long (length > 2000km) narrow (eccentricity > 0.85) regions of strong water vapor transport (SI >0.375) and precipitation (minimum AR precipitation > 1mm/day) (see Table 1 for complete details). This makes the comparison of AR characteristics between different mean states less sensitive to arbitrary strength thresholds.
To demonstrate how ARs are detected using SCAFET, we used the daily mean European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis Version 5 (ERA5) data (Hersbach et al., 2020) for the period 2000 to 2019. The magnitude of daily mean integrated water vapor transport (IVT) is the basic field and the daily mean total precipitation is the secondary variable. All the datasets used have a spatial resolution of 0.25° × 0.25°. The vector field, IVT is calculated as,

\[ IVT_x = \frac{1}{g} \int_{1000hPa}^{300hPa} q \cdot U \, dp \]  
\[ IVT_y = \frac{1}{g} \int_{1000hPa}^{300hPa} q \cdot V \, dp \]  
\[ \|IVT\| = \sqrt{IVT_x^2 + IVT_y^2} \]  

To detect AR-like structures, SCAFET looks for shapes such as ridges, caps, and domes (see Figure 2). Following Figure 1, shape index (SI) is calculated after applying a grid aware smoothing that suppresses variability smaller than 1000 km (Figure 4(a)). Once SI is calculated for \(\|IVT\|\) (Figure 4(b)), regions where SI > 0.375 is passed on to the next stage for filtering. To maximally utilize the vector qualities of the primary field, we ensure that the local transport direction (arrows in Figure 4(a)) and local ridge direction (arrows in Figure 4(b)) do not deviate by more than 45°. The local ridge direction is defined as the eigenvector corresponding to the larger eigenvalue \(k_1\). Filtering based on the grid properties removes small (length < 2000 km and area < \(1 \times 10^{12} \) km²), and wide (eccentricity < 0.75) candidates. The secondary field, total precipitation within each object is used as to filter out weak (precipitation < 1 mm/day) AR-like structures. Precipitation is used to assess the strength of AR as it is the most socio-economically relevant. Following other AR detection methods, we have also imposed a regional mask to get rid of AR-like structures along the equatorial belt. All the previously mentioned steps can be applied in parallel along the time axis, making it computationally fast. Each time step will identify AR-like structures similar those shown in Figure 4(c). Once all ARs are identified, a simple tracking algorithm is implemented on the daily data to filter out ARs that last less than two days. Tracking can be implemented based on one of the location parameters, i.e., the center, maximum, or minimum points of each detected object. For ARs, we use the centroid of each detected object to track it. Closest objects within a distance of 4000 km between two time steps are considered the same object progressing in time (Figure 4(d)). The annual mean frequency of the detected AR objects and their seasonality is shown in Figure 4(e), (f), and (g). The spatial distribution can be found to be within the uncertainty induced by other detection algorithms from the Atmospheric River Tracking Method Intercomparison (ARTMIP) catalog (Lora et al., 2020).
Figure 4. Major steps in the detection of Atmospheric Rivers. (a) is the smoothed primary field, which is the vertically integrated water vapor transport (IVT). The smoothing removes variability smaller than 1000 km from the IVT. The arrows in (a) represent the direction of unsmoothed IVT. (b) shows the magnitude (shading) of shape index (SI) and direction of the local ridge (arrows) calculated from (a). In the next step ridges, caps, and domes are extracted from (b) and weak and small AR candidates are filtered out. The AR objects after this filtering is shown in (c). Finally all the objects in (c) are tracked as shown in (d) to obtain tracks as well as other properties like mean IVT and precipitation. Objects that does not last more than one day is removed in this step. AR annual mean frequency for the period 2000 to 2019 is shown in (e), the anomaly for November to March (f), and May to September (g) relative to the annual mean are also plotted.
Table 1. The table shows the values of all the different parameters used in the detection of ARs. Rows for each step—segmentation, filtering, and tracking—are grouped together and labeled.

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Segmentation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Smooth Scale</td>
<td>2000</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Angle Coherence</td>
<td>45</td>
<td>degrees</td>
</tr>
<tr>
<td>3</td>
<td>Selected Shape</td>
<td>(0.375,1.0]</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>Filtering</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Minimum Length</td>
<td>2000</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Minimum Area</td>
<td>$2 \times 10^5$</td>
<td>$km^2$</td>
</tr>
<tr>
<td>3</td>
<td>Eccentricity</td>
<td>[0.75, 1.0]</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Minimum Precipitation</td>
<td>1</td>
<td>mm/day</td>
</tr>
<tr>
<td>5</td>
<td>Latitude Mask</td>
<td>(-20, 20)</td>
<td>degrees</td>
</tr>
<tr>
<td></td>
<td><strong>Tracking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Minimum Duration</td>
<td>24</td>
<td>hours</td>
</tr>
<tr>
<td>2</td>
<td>Maximum Distance per timestep</td>
<td>1000</td>
<td>km</td>
</tr>
</tbody>
</table>
3.2 Tropical and Extratropical Cyclones

Cyclones are defined in the literature as large (500–4000 km) regions of strong cyclonic circulation with low pressure at the center and extremely high winds around it (Emanuel, 2003; Schultz et al., 2019; Encyclopaedia, 2022). The dynamics and characteristic of cyclones will be slightly different based on the genesis location, translations speed etc. For instance, cyclones formed near the equator (tropical cyclones) are in general smaller in area compared to that of the cyclones formed in midlatitudes (extratropical cyclones). Independent of this, they produce extremely high rainfall, and winds along the track and flooding, landslides, and severe damage to infrastructure along the coastlines where it makes landfall (Knutson et al., 2010; Mendelsohn et al., 2012; Ranson et al., 2014). With the rise in sea level and enhanced intensity of cyclones in response to warming, identification and future projection of cyclones is gaining a lot of attention from the climate community (Woodruff et al., 2013).

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Smooth Scale</td>
<td>1500</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Selected Shape</td>
<td>0.625</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>Minimum Length</td>
<td>20</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Minimum Area</td>
<td>$10^4$</td>
<td>km$^2$</td>
</tr>
<tr>
<td>3</td>
<td>Eccentricity</td>
<td>[0.0,1.0]</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Minimum Vorticity</td>
<td>$10^{-6}$</td>
<td>s$^{-1}$</td>
</tr>
<tr>
<td>5</td>
<td>Minimum Max. Windspeed</td>
<td>10</td>
<td>ms$^{-1}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Minimum Duration</td>
<td>48</td>
<td>hours</td>
</tr>
<tr>
<td>2</td>
<td>Maximum Distance per timestep</td>
<td>500</td>
<td>km</td>
</tr>
<tr>
<td>3</td>
<td>Net Minimum Displacement</td>
<td>1000</td>
<td>km</td>
</tr>
</tbody>
</table>

Table 2. Same as in table Table 1, but for detecting tropical and extratropical cyclones

Once again, discrepancies between the different detection algorithms can be traced back different choices of physical thresholds or functions on size, wind speeds, vorticity or surface pressure anomaly. Even though most studies converge on the conclusions for present and future characteristics of cyclones, ironing out the details such as the changes in genesis rate, dura-
tion are hindered by the uncertainties in the detection methods (Ulbrich et al., 2009; Neu et al., 2013; Horn et al., 2014; Walsh et al., 2015). In this study, SCAFET identifies cyclones as regions of strong local maxima of cyclonic circulation (SI > 0.625) with maximum wind speeds greater than 10m/s. This definition is able to identify strong cyclonic vorticities all over the globe including, but not limited to, tropical and extratropical cyclones. The basic field used for cyclone detection is the absolute value of cyclonic relative vorticity (\(\zeta\)) defined as,

\[
\zeta = \nabla \times \mathbf{U}
\]  

Where \(\mathbf{U}\) is the 6 hourly wind speeds at 10 meters from surface obtained from ERA5 reanalysis with a spatial resolution of 0.25° x 0.25° (Hersbach et al., 2020). The magnitude of wind speed at 10 meters is used as the secondary field. Other cyclone related variables like the surface pressure anomaly, and potential temperature can also be used as secondary field to identify/classify cyclones.

In contrast with ARs, cyclones are detected using a scalar field, and in this case cyclonic relative vorticity. A grid-aware gaussian smoothing is applied to suppress spatial variability smaller than 750 km (Figure 5(a)). The smoothing scale was chosen so that we can identify both tropical and extratropical cyclones. Caps and dome shapes (SI > 0.625) from the smoothed relative vorticity is identified as potential cyclones (Figure 5(b)). The next step is to filter out objects with area less than \(10^4\) km², and diameter less than 20km. Once the aforementioned spatial characteristics are fulfilled, we can further filter out weak cyclonic vorticities (relative vorticity < \(10^{-6}\) s⁻¹), and maximum wind speed < \(10\) ms⁻¹) giving us all the strong cyclonic systems identified for a given time step (Figure 5(c)). All the processes described can be parallelized along the time dimension as in the AR example. Once all potential cyclones are identified, they are tracked like AR tracking algorithm. However, the radius for search is limited to 1000km as we are using 6 hourly data and translations speeds of cyclones are much lower than 150km/h. A minimum duration of 48 hours and a minimum total displacement of 500km is applied to isolate propagating cyclonic circulations from stationary ones. An example of a tracked cyclone, commonly known as cyclone Dorian (Lixion A. Avila and Hagen, 2020) is compared with the observed track from IBTrACS dataset (Knapp et al., 2010, 2018) Figure 5(d)). In comparison to the observed track, SCAFET’s track is much longer as we use a more relaxed condition on \(\zeta\) and winds speed thresholds. Another reason for the longer track is that, SCAFET does not distinguish between tropical and extratropical cyclones and would end up following the vorticity while it transitions from tropical to midlatitude storms. The long-term averages for cyclone frequency and its seasonal variability are comparable with other studies (e.g., Ullrich and Zarzycki, 2017). Unlike other cyclone detection algorithms, SCAFET does not identify cyclones as a point object, but as a surface encompassing the point of maximum \(\zeta\). This will help us study the nuanced properties of cyclones like the maximum and minimum values of wind speed, precipitation within the whole cyclone structure.
Figure 5. Major steps in the detection of cyclones. (a) Smoothed primary field, which in this case is the absolute value of cyclonic relative vorticity ($\zeta$). The smoothing removes variability smaller than 750 kms from $\zeta$. (b) Magnitude of SI calculated from (a). In the next step, caps, and domes are extracted from (b) and weak and small cyclone candidates are filtered out. Cyclonic vorticities after this filtering are shown in (c) with the background color representing the unsmoothed values of $\zeta$. Finally, all the objects in (c) are tracked as shown in (d) to obtain tracks as well as other properties like wind speeds and vorticity. Track obtained for cyclone "Dorian" from SCAFET is compared with that of the track from IBTrACS dataset (d). Objects that do not last more than 48 hours are removed in this step. The annual mean frequency cyclone occurrence for the period 2000 to 2020 is shown in (e), the anomalous cyclone frequencies for JJA (f), and DJF (g) relative to the annual mean are also plotted.
3.3 Sea Surface Temperature Fronts

SST fronts are the confluence regions of different water masses. They are often manifested as having strong horizontal gradients in temperature, salinity, density, and other characteristics (Bowman, 1978; Legeckis, 1978; Fedorov, 1986; Yoder et al., 1994). Frontal structures are often observed in much smaller spatio-temporal scales than the other features described in this study. Accurate identification of SSTFs are important as these features are often associated with strong upwelling, and high biogeochemical productivity (Clayton et al., 2014, 2021; Nagai and Clayton, 2017). Identification of SSTFs also demonstrates how SCAFET can be used to detect features in curvilinear grids.

Most previous frontal detection algorithms use edge detection algorithms and the gradient of sea surface temperature and/or height and to identify fronts (Canny, 1986; Castelao et al., 2006). We use the magnitude of daily mean SST horizontal gradient as the primary field for the detection of SST fronts. The SSTs were obtained from a fully coupled, ultra-high-resolution (\( \approx 10 \, \text{km} \)) CESM v1.2.2 simulation of present day mean climate (Small et al., 2014; Chu et al., 2020). The data is fed into SCAFET in the tripolar POP grid. To demonstrate the detection process, the analysis is confined to the Kuroshio frontal and extension domain for the last 10 years of the simulation.

Frontal structures, ridges, caps, and domes are extracted very similarly as in the detection of ARs. A spatial smoothing of approximately 30 km is applied before extracting the frontal structures. From the extracted SSTF candidates, objects with a mean SST gradient lower than \( 10^{-4} \, \text{K/m} \) is removed. So are circular (eccentricity < 0.5) and small (area < 1000 km\(^2\)) objects. Unlike AR detection, SSTFs are not tracked as ocean fronts are stationary rather than transported into other regions. The detected frontal frequency shows familiar patterns and seasonality as in previous studies (Xi et al., 2022).

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Smooth Scale</td>
<td>30</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Selected Shape</td>
<td>(0.375, 1.0]</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>Minimum Length</td>
<td>500</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Minimum Area</td>
<td>( 10^3 )</td>
<td>km(^2)</td>
</tr>
<tr>
<td>3</td>
<td>Eccentricity</td>
<td>(0.5,1.0]</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Minimum SST Gradient</td>
<td>( 10^{-4} )</td>
<td>Km(^{-1})</td>
</tr>
</tbody>
</table>

Table 3. Same as in table 1 but for detecting Sea Surface Temperature Fronts (SSTFs)
Figure 6. Major steps in the detection of sea surface temperature fronts (SSTFs). (a) Magnitude of shape index (SI) calculated from the smoothed primary field, which is the horizontal gradient of sea surface temperature ($\nabla$SST). The smoothing removes variability smaller than 15 km from $\nabla$SST. In the next step, ridges, caps, and domes are extracted from (b) and weak and small SSTF candidates are filtered out. SSTFs after this filtering are shown in (b) with the background color representing the unsmoothed values of $\nabla$SST. The annual mean frequency SSTF occurrence for the 10 years of the present climate simulation is shown in (c), the anomalous frontal frequencies for JJA (d), and DJF (e) relative to the annual mean are also plotted.

4 Application to 3D Features

In this section, we show how to extend SCAFET to detect features from three-dimensional (3D) primary fields. The scale-space selection is carried out by applying gaussian smoothing along the three dimensions separately. Also, a 3D basic field would yield three eigenvalues ($k_1 \geq k_2 \geq k_3$) rather than two. Here, SI can be calculated by combining the eigenvalues in three different ways. The SI calculated using $k_1$ and $k_2$ is used for the extraction of jet streams as it provides a more conservative estimate for the jet domain (see Supplementary Figure S1).
4.1 Jet Streams

Independent of the dynamics, jet streams are manifested as narrow regions in the upper atmosphere with relatively high wind speeds compared to its surroundings (Koch et al., 2006). Apart from the obvious direct impact on aviation, the location and characteristics of jet streams strongly influence surface weather conditions. For instance, a persistent jet in boreal summer can lead to extreme heat and flooding events while a meandering jet in the winter induces extreme cold spells in the midlatitudes (Petoukhov et al., 2013; Coumou et al., 2014; Kretschmer et al., 2016). Further, the northward movement of jet streams in response to greenhouse warming leads to the poleward propagation of tropical cyclones (Studholme et al., 2021). Thus accurate and robust detection of jet streams are fundamental in the prediction and projection of mean and extreme weather systems. Similar to the detection of other weather phenomenon discussed in this study, most previous studies use a physical threshold in identifying jet locations. Moreover, most of these studies except Limbach et al. and Kern et al. identifies jet streams as either a one- or two-dimensional features. Here we intend to demonstrate the capability of SCAFET to detect jet streams as a three-dimensional structure. Since the scope of this section is limited to the validation of the detection method, we have only shown jet detection in three selected time steps. A video showing the results for a longer period can be seen in the supplementary section.

The primary field used in the extraction of jet streams is the six-hourly, three-dimensional wind speeds obtained from ERA5 reanalysis data with a spatial resolution of $1^\circ$ with 37 vertical levels Hersbach et al. (2020). The magnitude of wind speed is calculated as:

$$W = \sqrt{U^2 + V^2}$$

Where, $U$ and $V$ are the zonal and meridional wind velocities.

Comparable to detection of 2D features, the detection process starts by applying a gaussian smoothing to remove wavelengths less than $6000 \text{km}$ in the horizontal dimensions. No smoothing is applied along the vertical dimension. Next, SI is calculated using the two largest eigenvalues $k_2$ and $k_3$. Regions corresponding to ridges, caps, and domes (SI > 0.375) are isolated for filtering. Filtering removes objects with volume less than $1000 \text{km}^3$, horizontal length less than $5000 \text{km}$, and maximum wind speed within each object less than $50 \text{m/s}$. In the current version of SCAFET, the tracking algorithm is not applied on jet detection (see Figure 7).
Figure 7. 3D jet streams extracted using SCAFET. (a), (c), (e) shows the magnitude 3D wind speed for 2022-08-25 00:00, 2022-08-28 06:00, and 2022-08-31 18:00 respectively. The 3D jet streams extracted for the corresponding time period is show in (b), (d), and (e) respectively.
<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Smooth Scale</td>
<td>6000</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Selected Shape</td>
<td>(0.375,1.0]</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum Length</td>
<td>5000</td>
<td>km</td>
</tr>
<tr>
<td>2</td>
<td>Minimum Height</td>
<td>5</td>
<td>km</td>
</tr>
<tr>
<td>3</td>
<td>Minimum Volume</td>
<td>500</td>
<td>km$^3$</td>
</tr>
<tr>
<td>4</td>
<td>Minimum Area</td>
<td>$10^5$</td>
<td>km$^2$</td>
</tr>
<tr>
<td>5</td>
<td>Minimum Max. Wind speed</td>
<td>50</td>
<td>m/s</td>
</tr>
</tbody>
</table>

**Table 4.** Same as in Table 1 but for Jet Streams.
5 Conclusions

In this study we introduced a new framework and algorithm package for extracting and tracking meso-synoptic scale features from large climate datasets, called Scalable Feature Extraction and Tracking (SCAFET). As the volume and diversity of observational and model climate data grow, an alternative method to physical threshold-dependent feature detection is necessary to compare features within and between observational and model data sets. Furthermore, a novel shape-based approach for feature extraction will give us further insights into detection method discrepancies in projections and aid the community in building scientific consensus. To demonstrate the ability of SCAFET in advancing these goals, we illustrated 2D detection of atmospheric rivers (ARs), tropical and extratropical cyclones, sea surface temperature fronts, and 3D detection of jet streams. Each application was intended to give characteristic examples from which users can customize SCAFET for their own research purposes.

Apart from the obvious benefits like a more generalized framework and parallelized implementation, SCAFET, more importantly, provides a novel perspective on how we could define various features in climate datasets covering large periods of time, in which the mean climate varies significantly. Instead of extracting features from physical thresholds, we can identify them based on their local shape in the field and refine the analysis by optionally applying a minimum threshold on the extracted objects. This approach provides a view of the continuous changes in feature properties that account for mean state changes. Since results of meso-synoptic scale studies are sensitive to thresholds in a varying mean state, conclusions inherently depend on the feature extraction method. Such varying conclusions are noticed while studying the response of ARs to greenhouse warming Zhao (2020); O’Brien et al. (2022). Thus physical threshold-independent algorithms such as SCAFET may be crucial in furthering scientific understanding and climate model development.

Undertaking a more fundamental level research into differential geometry and mathematical derivation of the relationship between SI and local geometric shape could transform the way we identify extreme events from large datasets. Due to its design, SCAFET does not require a priori climate information to identify features. This property can be utilized to develop simple web-based solutions for identifying and warning public against presence of extreme weather systems.

Author contributions. ABN wrote the software package and prepared the manuscript draft with inputs from co-authors. TAO and DL was involved in developing a mathematical framework for the algorithm. JEC provided input and guidance on the detection and tracking of tropical and extratropical cyclones. JYL, TAO, DL, and JEC contributed equally on the manuscript revisions.

Code and data availability. The latest version of the Scalable Feature Extraction and Tracking (SCAFET) algorithm can be downloaded from https://github.com/nbarjun/SCAFET. The version of the codes used for feature extraction and creating relevant figures in this manuscript can be downloaded from https://doi.org/10.5281/zenodo.7767301. A sample dataset for the curvilinear SST data is also included in the repository. The directory also includes sample outputs for various features discussed in the manuscript. The ERA 5 reanalysis data with varying resolutions can be downloaded from https://cds.climate.copernicus.eu/cdsapp. Single-level variables like 10m wind are obtained.
from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form, while three-dimensional variables can be extracted from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form. To see the exact codes used for downloading ERA5 data, readers could refer to ERA5Data folder in the zenodo repository. For any further details on code and data, feel free to contact the first author.

**Competing interests.** TAO is a member of the editorial board of the journal Geoscientific Model Development. The peer-review process was guided by an independent editor, and the authors have no other competing interests to declare.

**Acknowledgements.** The research was supported by the Institute for Basic Science (IBS), Republic of Korea, under IBS-R028-D1. O’Brien’s contributions were supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and by the Environmental Resilience Institute, funded by Indiana University’s Prepared for Environmental Change Grand Challenge initiative. The authors would like to thank Prof. Axel Timmermann, Dr. Karl Stein, and Dr. Pavan Harika Ravi from IBS Center for Climate Physics for their comments on shape based feature extraction. We would also like to thank the ARTMIP community for their feedbacks on SCAFET. Part of the analysis of the detected features were conducted on the IBS/ICCP supercomputer “Aleph” 1.43 peta flops high-performance Cray XC50-LC Skylake computing system with 18,720 processor cores, 9.59 PB storage, and 43 PB tape archive space. We also acknowledge the support of KREONET for the fast and reliable data transfers. Special thanks to early users of the algorithm.

**Appendix A: Shape Based Feature Extraction on Simple Datasets**

The aim of this section is to demonstrate how shape-based matrices can be used to extract features from simple mathematical functions. This section is indented to provide readers with more insights into the basic principle behind shape-based feature extraction and how it differs from other published methods. We have also tried to demonstrate some properties of shape-based feature extraction methods like its insensitivity to mean state changes and linear trends.

**A1 Application to 1D datasets**

In this section, we are constructing an analogy between application of SCAFET on a 2D dataset and shape-based feature extraction from a one-dimensional dataset. The purpose of this discussion is not to advocate for a shape-based extraction of features from 1D datasets but to help the readers understand its strengths and weaknesses.

Conventionally, for any differentiable curve \( C \), the curvature is measured as the instantaneous rate of change of direction of \( a \) along the curve. Simply put, the curvature is measured as the rate of change of the unit tangent to the curve at any given point. An osculating circle can be used to intuitively represent the curvature of a surface or a curve (see Figure A1). At any point \( P \), the curvature, \( k \) is the reciprocal of the radius \( (R) \) of the circle. The sign of \( k \) determines if the curve has a concave
Figure A1. A schematic representation of measuring curvature of a curve C at point P. At P the curvature is the reciprocal of the radius R of the osculating circle. In differential geometry, an osculating circle is defined as the circle passing through the point P and a pair of additional points infinitesimally close to P.

or a convex curvature. More information and mathematical proof for these concepts can be found in any standard differential geometry textbooks.

Following the derivation of Shape Index (SI) for 2D datasets, we could normalize the curvature for a function f to give the shape parameter as,

\[ K = \frac{2}{\pi} \tan^{-1}(f'') \]  

(A1)

Values of K closer to 1 can be identified as regions of local minima while K closer to -1 are regions of local maxima (black curve in Figure A2). Depending on the severity of the extreme event, one could choose a value for K to get regions of local maxima (red caps in Figure A2) and local minima (green caps in Figure A2). The curvature of the function is insensitive to linear trends and mean state changes as the application of the same shape thresholds identifies identical regions as local maxima and minima in a simple trigonometric curve (blue curve in Figure A2) and the same curve with an added linear trend (orange curve in Figure A2). The values of K for both curves are represented by the black line in Figure A2. Thus, it can be used to identify extreme events from datasets without being affected by the background state changes.

A2 Application of SCAFET to simple Geostrophic Motion

In this section we demonstrate the application of SCAFET to a simple geostrophic rotational motion. The goal of this discussion is to see how shape-based extraction of 2D features differ from other conventional methods used. The calculation of SI involves the computation of the two eigenvalues, \(k_1\) and \(k_2\) of the hessian of any gridded dataset. As discussed in the previous section, the measurement of curvature, \(k_1\) and \(k_2\) can be visualized as the reciprocal of the radius of two osculating circles orthogonally intersecting at a point in the surface. Large negative eigenvalues represent surfaces with strong convex curvature while positive
Figure A2. Shape extraction for a simple one-dimensional curves given by $f = \sin 2x + 3\cos 5x$ and $f + 0.5x = \sin 2x + 3\cos 5x + 0.5x$. The first one is a simple trigonometric curve while the second curve includes a linear trend as is evident from their functional forms. The magnitudes of both the functions are shown in the left Y axis while the right Y axis indicates the values of the shape parameter ($K$). Note that $K$ is same for both the functions. The green and red highlighting on the curves indicates the regions where $K > 0.99$ and $K < -0.99$. The red highlighted are tagged as extreme maxima events and those highlighted with green are be characterized as extreme minima events.

values are identified as troughs or cups. To demonstrate the characteristics and strengths of feature detection based on Shape Index (SI) let’s consider a simple rotation wind field (see Figure A3(a) vectors) given by

$$u_g = -\Omega y \quad \text{(A2)}$$

$$v_g = \Omega x \quad \text{(A3)}$$

Where $\Omega$ is a constant and $x,y$ represents the grid. In the current example the value of $\Omega$ is set as $10^5 \text{rad/s}$. The geopotential height ($h$) of the field (see Figure A3(a) shading) is used as our primary field to identify features using shape index (SI). $h$ is estimated as,

$$h = \frac{\Omega f}{2g} (x^2 + y^2) \quad \text{(A4)}$$

$f$ and $g$ are the Coriolis parameter and the gravitational constant respectively. SI is calculated from the eigenvalues of the hessian of $h$ using the formula,

$$SI = \frac{2}{\pi} \arctan \left[ \frac{k_2 + k_1}{k_2 - k_1} \right] \quad \text{(A5)}$$
Where the eigen values $k_1$ and $k_2$ are given by,

$$k_{12} = \frac{f\zeta_g}{2g} \pm \sqrt{\left(\frac{f}{2g}\right)^2 - \left(\frac{f}{g}\right)^2 \frac{\partial v_g}{\partial x} \frac{\partial u_g}{\partial y} + \frac{\partial u_g}{\partial x} \frac{\partial v_g}{\partial y}}$$

(A6)

Where $\zeta_g$ is the geostrophic vorticity. Which gives SI as,

$$SI = \frac{2}{\pi} \tan^{-1} \left( \frac{-\zeta_g}{-2\sqrt{\left(\frac{\zeta_g}{2}\right)^2 - \frac{\partial u_g}{\partial x} \frac{\partial u_g}{\partial y} + \frac{\partial u_g}{\partial x} \frac{\partial v_g}{\partial y}}} \right)$$

(A7)

A detailed derivation of the above equation can be found in Appendix B. Plugging in the values for the rotational motion, we get

$$\zeta_g = \nabla^2 h = \Omega f/g$$

(A8)

$$\frac{\partial u_g}{\partial x} = \frac{\partial v_g}{\partial y} = 0$$

(A9)

$$\frac{\partial v_g}{\partial x} \frac{\partial u_g}{\partial y} = \Omega^2$$

(A10)

Therefore,

$$SI = \frac{2}{\pi} \tan^{-1} \left( \frac{\Omega^2}{-\sqrt{\Omega^2 - \Omega^2}} \right) = -1$$

(A11)

Thus, SCAFET classifies the whole domain with the anticlockwise rotational motion as a trough with $SI \approx -1$ regardless of the absolute value of the field or $\Omega$. A traditional method that uses thresholding directly on the geopotential height would identify regions depending on the value of the threshold on $h$. However, the value of the threshold must be adjusted depending on the mean (time) and background (space) state. Another widely used practice is to define a threshold on the smallest eigenvalue. The intention of such methods is to identify extreme features based on the strength of the curvature rather than the actual value of the field. TempestExtremes, a feature extraction framework previously discussed, follow this method to identify Atmospheric Rivers from gridded datasets. In the current example, this would correspond to thresholding on $f\Omega/g$. In other words, TempestExtremes would only detect the trough if the value of $\Omega$ is greater than the predetermined threshold. Contrastingly, SCAFET would identify the trough as a trough regardless of the actual value of the field or $\Omega$. Hence, we see that feature extraction using SI and other published methods can give us different results depending on the input data as they are looking at different properties of the field.
Figure A3. Comparing different feature extraction techniques on synthetic rotational wind field. (a) The geopotential height ($h$) (shading) of the rotational wind field (arrows). $h$ is defined as $\frac{\Omega f(x^2 + y^2)}{2g}$, where $f = 10^5 - 4)1/s$, $g = 9.805m/s$ and $\Omega = 10^5 rad/s$. (b) The magnitude of the smallest eigen value. From the equation, the smallest eigen value can be derived as $f\Omega/g = 1.0199$. (c) Shows the value of the shape index (SI). From the equations and the plot, we can see that the value of $SI = -1$ throughout the domain.
Appendix B: Derivation of Shape Index for Geostrophic Motion

The complete derivation of the SI for geostrophic wind fields are shown in this section. The result from the derivation is used in the previous section. Let $h$ be the geopotential height at a certain level. The hessian of $h$ is given by.

$$\mathcal{H}(h) = \begin{pmatrix} \frac{\partial^2 h}{\partial x^2} & \frac{\partial^2 h}{\partial x \partial y} \\ \frac{\partial^2 h}{\partial y \partial x} & \frac{\partial^2 h}{\partial y^2} \end{pmatrix}$$  \hspace{1cm} (B1)

The eigen values of the symmetric matrix $\mathcal{H}$ is calculated by solving the quadratic equation.

$$\left( \frac{\partial^2 h}{\partial x^2} - \lambda \right) \left( \frac{\partial^2 h}{\partial y^2} - \lambda \right) - \left( \frac{\partial^2 h}{\partial x \partial y} \right)^2 = 0$$  \hspace{1cm} (B2)

Which can be expanded as;

$$\lambda^2 - \lambda \left( \frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} \right) + \frac{\partial^2 h}{\partial x^2} \frac{\partial^2 h}{\partial y^2} - \left( \frac{\partial^2 h}{\partial x \partial y} \right)^2 = 0$$  \hspace{1cm} (B3)

$$\lambda^2 - \lambda \nabla^2 h + \frac{\partial^2 h}{\partial x^2} \frac{\partial^2 h}{\partial y^2} - \left( \frac{\partial^2 h}{\partial x \partial y} \right)^2 = 0$$  \hspace{1cm} (B4)

NOTE: The geostrophic vorticity ($\zeta_g$) is defined as

$$\zeta_g = \frac{g}{f} \nabla^2 h$$  \hspace{1cm} (B5)

The geostrophic velocities are defined as

$$u_g = -\frac{g}{f} \frac{\partial h}{\partial y} = -\frac{\partial \psi}{\partial y}$$  \hspace{1cm} (B6)

$$v_g = \frac{g}{f} \frac{\partial h}{\partial x} = \frac{\partial \psi}{\partial x}$$  \hspace{1cm} (B7)

Where $\psi$ is the geostrophic stream function. This implies.

$$\frac{\partial^2 h}{\partial x^2} = \frac{f}{g} \frac{\partial^2 \psi}{\partial x^2} = \frac{f}{g} \frac{\partial v_g}{\partial x}$$  \hspace{1cm} (B8)

$$\frac{\partial^2 h}{\partial y^2} = \frac{f}{g} \frac{\partial^2 \psi}{\partial y^2} = -\frac{f}{g} \frac{\partial u_g}{\partial y}$$  \hspace{1cm} (B9)

Adding the abovementioned relationships to equation (3)

$$\lambda^2 - \frac{\lambda f}{g} \zeta_g - \frac{f^2}{g^2} \frac{\partial v_g}{\partial x} \frac{\partial u_g}{\partial y} + \frac{f^2}{g^2} \frac{\partial u_g}{\partial x} \frac{\partial v_g}{\partial y}$$  \hspace{1cm} (B10)

Solving for $\lambda$ we get
\[ \lambda_{12} = \frac{f \zeta_g}{2g} \pm \sqrt{\left( \frac{f}{2g} \right)^2 - \left( \frac{f}{g} \right)^2 \partial_v \partial_u + \partial_u \partial_v} \]  
(B11)

\[ \lambda_{12} = \frac{f}{g} \left( \frac{\zeta_g}{2} \pm \sqrt{\left( \frac{\zeta_g}{2} \right)^2 - \partial_v \partial_u + \partial_u \partial_v} \right) \]  
(B12)

Thus the shape index for \( h \)

\[ SI = \frac{2}{\pi} \arctan \left[ \frac{\zeta_g}{-2 \sqrt{\left( \frac{\zeta_g}{2} \right)^2 - \partial_u \partial_v + \partial_u \partial_v}} \right] \]  
(B13)
References


