10.5 PEA: PHENOMENA EXTRACTION ALGORITHM

Rahul Ramachandran*, Xiang Li and Sara Graves
University of Alabama in Huntsville

Richard. D. Clark, and David Fitzgerald
Millersville University

1. INTRODUCTION

A phenomenon is defined as any state or process known through the senses rather than by intuition or reasoning, and thus is an observable event, especially something special or unusual. Many geophysical phenomena can be observed and monitored using space-based and ground-based sensors or simulated in numerical models. Research studies on geophysical phenomena rely on the data collected by these sensors or generated by the numerical models. A geophysical phenomenon in the context of the geoscience data can be characterized as a spatial region with the following:

- As a spatial region significantly different from the rest of the image.
- Having higher/lower than the average background intensity value **and/or** higher variation in intensity when compared to the remaining data points
- Having a spatial extent that is much smaller than the rest of the data
- Having a temporal extent, meaning the size and magnitude of the phenomenon and the variation within the region can change over time.

Examples of geophysical phenomena include the hurricanes characterized by the strong vorticity and weather fronts characterized by the significant variations of local wind patterns.

Data mining is a valuable tool in the analysis of the ever increasing volumes of observational and model geoscience data. Phenomena extraction is an important step within this process. Phenomena extraction algorithms typically used are either using domain knowledge based heuristic algorithms or application specific machine learning/image processing algorithms.

Rahul Ramachandran
Information Technology and Systems Center, University of Alabama in Huntsville,
Huntsville, AL 35899
Tel: 256-824-5157
email: rramachandran @itsc.uah.edu

For example, *Spencer and Braswell, 2001* used domain knowledge about temperature, wind speed and other physical parameters are used to detect tropical cyclones from the Advanced Microwave Sounding Unit-A (AMSU-A) measurements. Where as, *Li et al., 2005 utilized* highly specialized advanced machine learning and image processing methods to extract weather frontal systems from a model-generated dataset.

There are generic techniques such as thresholding and segmentation that can also be applied to detect phenomena on imagery data. These methods are based on the data statistics. The thresholding methods assume that there are distinct intensity distributions for the phenomena and the background and the extraction is based on locating the intensity that separates the two distributions optimally. The segmentation methods assume that the region of interest is homogeneous and distinct from the background.

In this paper, we present the Phenomena Extraction Algorithm (PEA), a novel, effective and efficient method for phenomena detection which is also based on the statistical properties of the dataset and therefore is data, domain and application independent. Unlike the thresholding methods, the PEA examines both the intensity and the variation in a region.

2. PHENOMENA EXTRACTION ALGORITHM (PEA)

The PEA algorithm has two components. The first component is a recursive decomposition algorithm to divide the data into smaller regions and the statistical tests to evaluate these regions forms the second component. The KD-tree (*Samet, 1984*) data decomposition strategy is used in the PEA algorithm to partition the original image into sub-regions hierarchically. The KD-tree stands for the k-dimensional binary tree. In each partitioning, a region is recursively split into two sub-regions, thus an image can be effectively represented as a number of sub-images hierarchically. Figure 1a shows an example of an image partitioned using the KD-tree structure and Figure

1b shows the image represented for this hierarchical partitioning. In this example, the original image is partitioned into five sub-regions and the five sub-regions are the leaf nodes in the hierarchical KD-tree. At this level of decomposition, a statistical F hypothesis test and Student's T test are applied to the data in the sub-region. These statistical tests determine if the region contains data points with abnormal intensities or large variations of intensities compared to the data statistics of the global image. If these statistical hypotheses are validated, the sub-region is selected as a candidate region representing the phenomena of interest and decomposed further. This recursive splitting of a region of interest proceeds until the size of a sub-region reaches six pixels or less. These remnant regions are the phenomena of interest. If the statistical hypotheses fail, the region does not contain the phenomena of interest and further decomposition on this region is terminated. Details of the PEA algorithm are given in the following pseudocode.

Algorithm *Phenomena Extraction Algorithm (PEA)*
Input: Two-dimensional multi-channel images *I(k), k=1 to N* where *N* is the total number of channels*; R*, a threshold ratio, used to determine the homogeneity of a region; and *L*, the significance level, used in the statistical hypothesis test.

- Initialization
  - For each image channel, normalize pixel intensity values to 0 -1 range.
  - Calculate the global mean M_g and standard deviation σ_g for each of the *N* channels.
- For each of the N channels, use the entire image as the initial region and start recursively decomposing the region:
  - Termination Condition.
    - If the number of pixels in the region is less than 6, label as the region of interest and stop the recursion.
  - Otherwise, split the region into two sub-regions.
  - For each of the split sub-regions:
    - Calculate the local mean μ_{local} and local variance σ^2_{local}.
    - Calculate the upper bound of local variance $\sigma^2_{th} = (R \cdot \mu_{global})^2$.
    - Calculate variance ratio, $V = \sigma^2_{local} / \sigma^2_{th}$, and degree of freedom, dF
    - Apply statistical F-test, F_test(V, dF)
    - If the null hypothesis is true (variation is less or equal) then the region is homogeneous.
      - Apply statistical Student's T-test to determine if μ_{local} is significantly different from M_g for a given the significance level *L*.
      - If different, then split the sub-region.
      - Otherwise stop the recursion.
    - Otherwise, split the sub-region (the region has spatial variation)
- The data points retained at the end of the recursive splitting are the regions of interest (phenomena).

Two parameters that can be tuned to determine the result of the PEA algorithm are: *L*, the significance level for the statistical test; and *R*, the threshold ratio which sets the upper bound for homogeneity measure. Users can use these two parameters to achieve optimal performances. In general, a larger *R* value will cause the PEA to mark regions with larger intensity variances as homogeneous. As a result, fewer regions of interests will be retained as the phenomena. Significance level *L* sets the confidence level for the extracted regions of interest.

3. DATA DESCRIPTION

Performance of the proposed PEA algorithm was examined using two Earth science data sets. The first data set used was the model output from the North American Mesoscale (NAM) model data run. The spatial resolutions of the model output were 0.5º in both latitude and longitude. The accumulated precipitation data field of the model output was used to extract the target phenomenon: regions of significant precipitation over a study area. The area of study was 15ºN-60ºN in Latitude and 140ºW-50ºW in Longitude. This dataset contained a total of 28 images of accumulated precipitation field, corresponding to 28 time steps. The second dataset was the model output generated from the Goddard Laboratory of Atmosphere's finite volume Community Climate Model (fvCCM). The model outputs from September 11 to September 19, 1999 were used in this study to extract three weather phenomena: tropical cyclones, surface frontal systems and troughs. The output fields have a horizontal resolution of 0.5 ° Latitude x 0.625 ° Longitude. The surface U and V wind component fields were used for phenomena extraction. There was a total of 31 images in the data set, corresponding to 31 time steps.

4. INITIAL EXPERIMENTS

First, the impact of the threshold ratio *R* on the PEA performance was examined. The NAM model data was used in these experiments. The significance level *L* is set as 0.90. Figure 2a shows the accumulated precipitation for a selected NAM model data. Figures 2b, 2c and 2d show the results using three different *R* values: 0.25, 1.25 and 2.25, respectively. The global mean and standard deviation of the intensity of the image were 0.177 and 0.752, respectively. As expected, the larger the threshold ratio value, the fewer the identified regions of interest.

When *R* was set to 0.25, all of the precipitation regions were detected. When *R* was set as 2.25, only severe precipitation regions were detected. The same conclusion was drawn for the cyclone and frontal system extraction from the U and V wind fields in the second data set.

In the second experiment, PEA was compared against a thresholding and a segmentation algorithm. The fvCCM dataset was used to perform these comparisons. Initially, Otsu's thresholding algorithm (*Otsu, 1979*) was used. Otsu's algorithm is commonly used in image processing for thresholding and can automatically determine the optimal threshold. The optimal threshold is determined by finding the value that minimizes the inter-class variance between the background and an object. However, the results from Otsu's algorithm were poor. This can be attributed to the algorithm's requirement for a sufficient number of object data points to determine the optimal threshold. The phenomena in the fvCCM datasets were represented by only a small fraction of the total data size thus causing the poor results. Therefore, a simple global thresholding technique was used in this experiment for comparison. The threshold value used by this technique was the global mean plus twice the standard deviation. A graph-based image segmentation algorithm (*Pedro et al., 2004*) was also used in this experiment. This segmentation algorithm was selected since the source code for the algorithm was readily available.

Only a qualitative analysis was performed in this study to compare the results from three algorithms. Human beings tend to have the innate ability to see complete forms present in the data (the so called gestalt effect). It is extremely difficult to replicate this ability in a feature extraction algorithm. Therefore, in this analysis, the results from algorithms were considered correct if they picked out only part of the phenomena.

The actual data with the truth labeled by the domain experts and the results from the three algorithms are presented in Figures 3 a-d. Figure 3a contains a tropical cyclone (labeled C), a stationary trough (labeled T) and six surface fronts (labeled from 1 to 6). The six surface fronts are of varying intensity, shape and orientation. PEA extracts the cyclone and the stationary trough correctly (Fig 3b). It also extracts all six surface fronts. Fronts 1, 6 and 4 are captured properly by the PEA. It only gets part of fronts 2, 3 and 5. It also extracts a region (labeled F in Fig 3b) incorrectly that is deemed as a false signature. The results from the global thresholding algorithm are presented in Fig. 3c. The global thresholding algorithm extracts the stationary trough and part of the cyclone. It only picks up three of the fronts (2, 4 and 6) and also extracts a false signature. The results from the graph-based segmentation algorithm are presented in Fig. 3d. The segmentation algorithm extracts the cyclone and part of the stationary trough. It does extract all the six surface fronts. However, it does produce five false signatures.

5. SUMMARY AND FUTURE WORK

The results of the first experiment demonstrate that the PEA can be tuned for different data sets and applications by the users. Even though the analysis in the second experiment is qualitative, the differences in results produced by the three algorithms can be clearly seen in Fig 3. The global thresholding algorithm does not extract all the phenomena. The combined use of threshold value (T-test) and the variance (F-test) allow the PEA to extract all of the phenomena. Thus, the use of spatial variance of the data value to characterize the phenomenon is just as important as the magnitude of the data value. The segmentation algorithm results are similar to PEA. It does extract all the phenomena but also produces more false signatures.

The initial results from these experiments are encouraging and work is underway to further refine the PEA algorithm. Additional tests are required to compare and quantify the differences between the PEA and other techniques such as global thresholding and segmentation. These tests are planned as part of future work.

6. ACKNOWLEDGEMENTS

This work was conducted as part of the LEAD project that is funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594, ATM-0331591, ATM-0331574, ATM-0331480, ATM-0331579, ATM03-31586, ATM-0331587, and ATM-0331578.

7. REFERENCES

Li, X., R. Ramachandran, S. Graves, S. Movva, B. Akkiraju, D. Emmitt, S. Greco, R. Atlas, J. Terry, and J. C. Jusem, 2005: Automated detection of frontal systems from numerical model-generated data. *The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-05)*, Chicago, IL, USA.

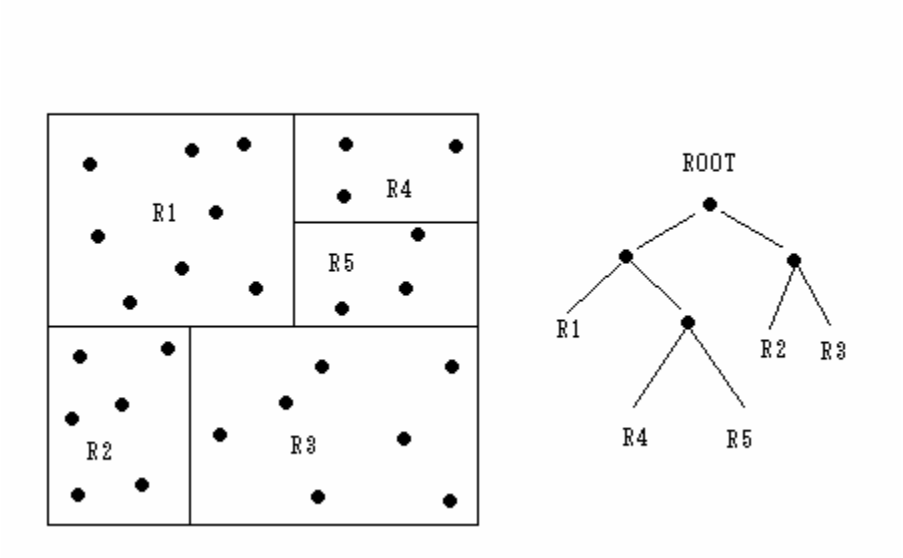
Otsu, N., 1979: A threshold selection method from gray-level histograms. *IEEE Trans. Sys. Man. Cyber.*, **9,**

62-66.

Pedro, F. F. and D. P. Huttenlocher, 2004: Efficient Graph-Based Image Segmentation. *Inter. Journal of Computer Vision*, **59,** 167-181.

Samet, H., 1984: The Quadtree and Related Hierarchical Data Structures. *ACM Computing Surveys (CSUR)*, **16,** 187-260.

Spencer, R. W. and W. D. Braswell, 2001: Atlantic Tropical Cyclone Monitoring with AMSU-A: Estimation of Maximum Sustained Wind Speeds. *Monthly Weather Review*, **129,** 1518-1532.



(a) (b)

Figure 1: (a) A conceptual illustration of the KD-tree partitioning of an image, (b) The associated hierarchical tree structure representation.

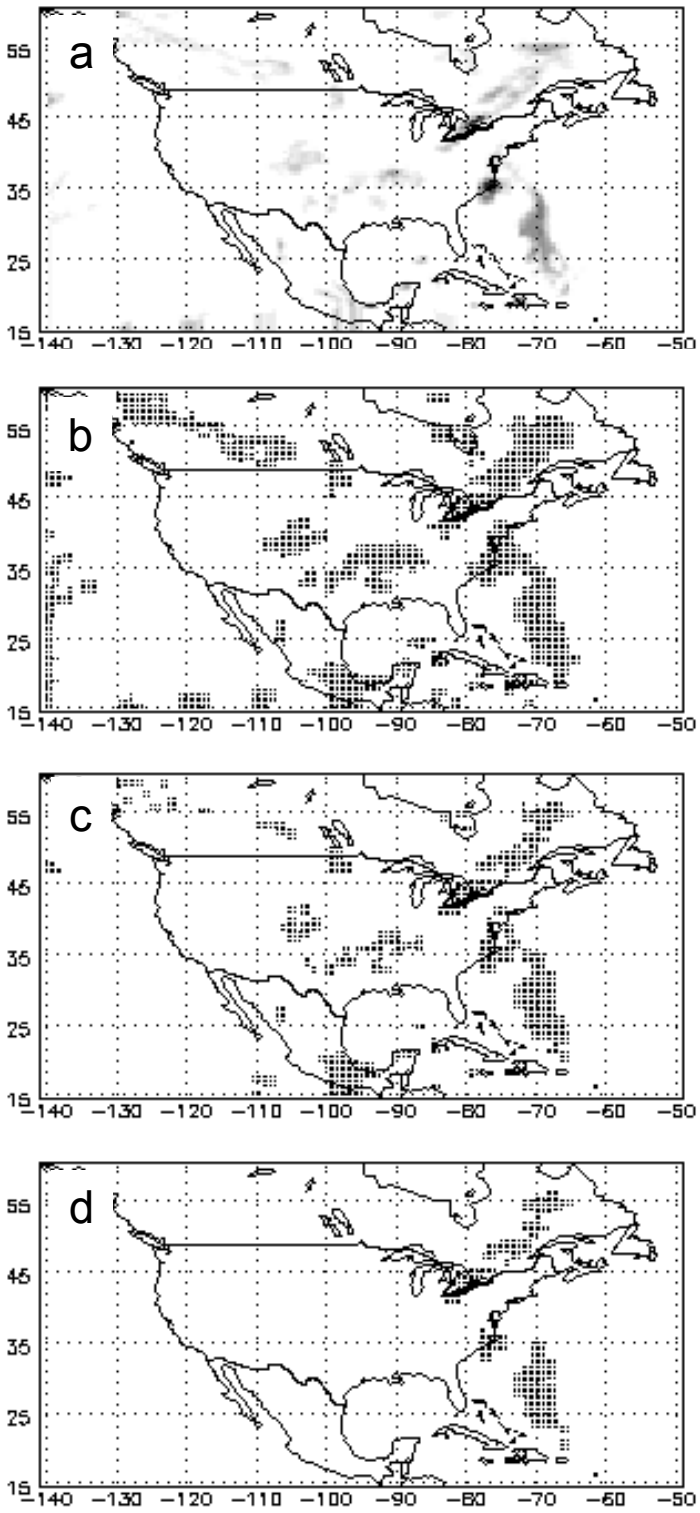


Figure 2: (a) Accumulated Precipitation Field from the NAM (b) PEA result using a threshold ratio of 0.25 (c) PEA result using a threshold ratio of 1.25 (d) PEA result using a threshold ratio of 2.25

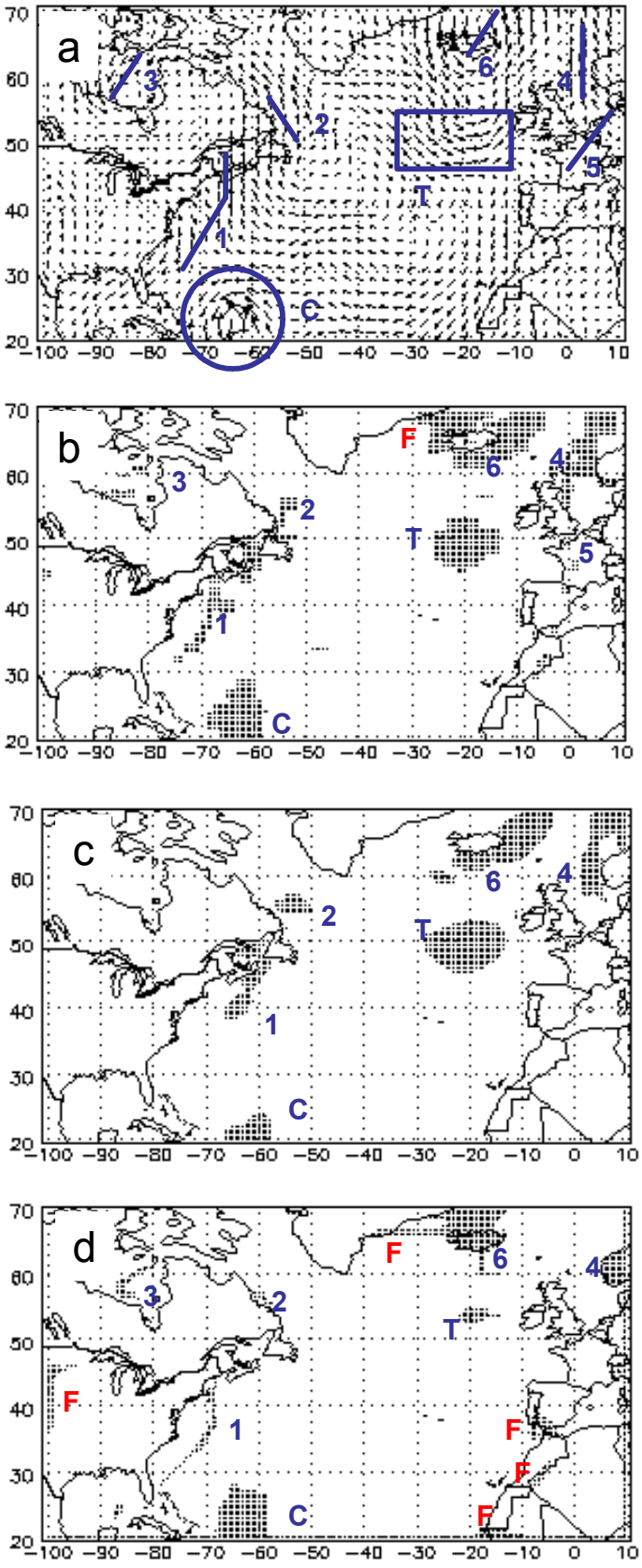


Figure 3: (a) Wind field from the fvCCM (b) PEA result (c) Global thresholding result (d) Segmentation result