# Detecting Region Outliers in Meteorological Data

Jiang Zhao, Chang-Tien Lu, Yufeng Kou

Department of Computer Science
Virginia Polytechnic Institute and State University
7054 Haycock Road, Falls Church, VA 22043
Tel: 703-538-8373
[jzhao1, ctlu, ykou]@vt.edu

## ABSTRACT

Spatial outliers are the spatial objects with distinct features from their surrounding neighbors. Detection of spatial outliers helps reveal important and valuable information from large spatial data sets. In the field of meteorology, for example, spatial outliers can be associated with disastrous natural events such as tornadoes, hurricane, and forest fires. Previous study of spatial outlier mainly focuses on point data. However, in the meteorological data or other applications, spatial outliers are frequently represented in region, i.e., a group of points, with two dimensions or even three dimensions, and the previous point-based approaches may not be appropriate to be used. As region outliers are commonly multi-scale objects, wavelet analysis is an effective tool to study them. In this paper, we propose a wavelet analysis based approach to detect region outliers. We discuss the region outlier detection problem and design a suite of algorithms to effectively discover them. The algorithms were implemented and evaluated with a real-world meteorological data set.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining, Spatial Databases

## General Terms

Algorithms, Design

## Keywords

Outlier Detection, Spatial Data Mining, Meteorological Data

## 1. INTRODUCTION

Over the past decade, spatial database has become a significant area both in academia and in industry. From satellite observation system to urban planing, geography related spatial data are widely used; there are also other spatial data, such as medical image and gene maps, which are also important and useful. The applications of spatial information promote the development of the Spatial Database Management System(SDBMS). The research on spatial database [16, 19, 20] mainly focuses on spatial data modelling, spatial data access, spatial data query processing, spatial data visualization, and spatial data mining. Spatial data mining [3, 10, 11, 21] is the process of discovering implicit and useful spatial patterns or rules from large spatial data sets. Like traditional data mining [6], spatial data mining techniques can also be classified into classification, clustering, trend analysis, and outlier detection. Since the spatial data types have unique characteristics and spatial relations are more complicated than ordinary data, traditional data mining techniques may not be directly applied to mine spatial data [22].

Spatial data mining is very attractive to many research works as spatial data tend to be large in size, and it is very important to efficiently extract knowledge embedded in the spatial data sets. Spatial outlier detection is an essential part of spatial data mining. Outliers are the observations differing from the remainder of the whole data sets [1, 7]. Spatial outliers are those observations which are inconsistent with the surrounding neighbors. The outliers are frequently treated as the noise of the data sets. However, in some applications, outliers have real meaning and are essential components of the data as they reveal significant anomalous phenomena. Such abnormalities exist in traffic data, nature resource management data, quality control data, earthquake monitoring system, and weather and climate data.

In the research of the atmospheric sciences, huge amount of spatial data have been collected from both observation and modelling. Discovering useful patterns from these data sets would have great practical value and would help weather forecast, environment monitoring, and climate analysis. In the meteorological data, spatial outliers or anomaly patterns are often associated with severe weather events. Such events usually do not happen at a single point but in an area. That is to say, they are usually two dimensional region spatial outliers.

In this paper, we propose a wavelet analysis based spatial outlier detection method to detect region outliers and to discover meaningful anomalies in meteorological observation data. This work will help retrieve interesting and implicit information from the large volume of the spatial data sets. This paper is organized as follows. We provide the literature survey in Section 2. Section 3 discusses the problem and our proposed approach. Section 4 introduces the algorithms. Section 5 describes the meteorological data set and analyzes the experiment results. Finally, we summarize our work and address future directions in Section 6.


*GIS'03,* November 7–8, 2003, New Orleans, Louisiana, USA.

## 2. RELATED WORK

Many outlier detection algorithms have been proposed [1, 2, 8, 15,17,24]. These algorithms can be classified into multi-dimensional space based methods and graph based methods. Multi-dimensional outlier detection methods use distance, depth, or density functions in the multi-dimensional space to check if a point is different from the majority of the data [9,15,27]. For many multi-dimensional outlier detection approaches, the spatial relationship between neighboring objects are usually lost after clustering. Shekhar, et al. [22] pointed out that multi-dimensional outlier detecting methods have several limitations: the approaches do not consider the graph structure of the spatial data; and they do not exploit the a priori information about the statistical distribution of the data. In [23], a graph-based outlier detection approach is proposed to detect spatial outliers. In their work, for each point $x$, a difference function $S(x)$ is defined to represent the feature difference between a point and its neighboring points. Then normalized functions are computed to determine points which are spatial outliers.

In the previous work, the outliers are usually points. In the real world, there are many cases in which the outliers exhibit in other spatial forms such as line or region. Such outliers exist in weather and climate data. For these two dimensional outlier detection, the determination of the region and their neighbors would be crucial. For the points enclosed in a region, the features should be rather similar, while for the outside points surrounding the region, the feature would be distinctly different. we cannot detect the points within outlier region as outliers using traditional point outlier detection, since a point surrounded by points with similar feature is not a spatial outlier. In order to determine the region and the surrounding neighbor, we must determine the shape and the localization of the regions. In the field of image processing, many works have been done to detect the edge of image objects and various edge detectors have been developed [28]. However, the image objects usually have clear edges in those works, the change of the features is sharp at the edge and will be detected by analyzing the feature gradients or spectral/wavelet transforms [26]. Whereas for the meteorological data, the feature changes are usually not as sharp as a clear edge. Therefore, it is difficult to detect the edge of the outlier regions. That is to say, we may not be able to get the shape or coverage of the outlier regions by using image edge detection methods.

In recent years, wavelet analysis methods are widely used in many science and technology fields, such as signal processing and image processing [4, 13]. In the following sections, we will discuss the advantages of the wavelet analysis over the traditional Fourier analysis. In the data mining area, wavelet analysis techniques have been used in clustering, classification, regression, forecasting, and visualization [12]. Sheikholeslami, et al. developed the *WaveCluster* approach which uses wavelet transform to cluster the spatial data in the frequency with different resolutions [18]. Their clustering method takes advantage of the multi-scale, multi-resolution properties of wavelet analysis and can effectively identify arbitrary shape clusters at different degrees of accuracy. However, they used wavelet transform in the feature space rather than in the real spatial domain.

## 3. REGION OUTLIER DETECTION PROBLEM AND OUR APPROACH

In the real atmosphere, the anomalies emerge at different spatial scales and may appear in different spatial shapes. This makes the outlier detection a challenging task. Figure 1 shows an image of the water vapor distribution over the east coast of the USA, Atlantic Ocean, and the Gulf of Mexico. As can be seen, there is a hot spot located at the left portion of the image, i.e., a hurricane at the Gulf of Mexico, and this spot is totally different from its surrounding neighbors. It is also clear that this outlier spot are not the single point but a group of the points or a region. Here, we define *a region outlier as a group of adjoining points whose feature is inconsistent with that of their surrounding neighbors*. And the hurricane is a region outlier in Figure 1. There exist other region outliers in this Figure; the number of region outliers detected will be determined by the pre-defined threshold provided by domain experts. The problem is to design an efficient and practical approach to detect region outliers (could be in irregular shapes) from spatial data sets. In the real application, such approaches can help identify spatial anomalies such as hurricanes, forest fires, tornado, thunder storm, and other severe weather events from the observation data.

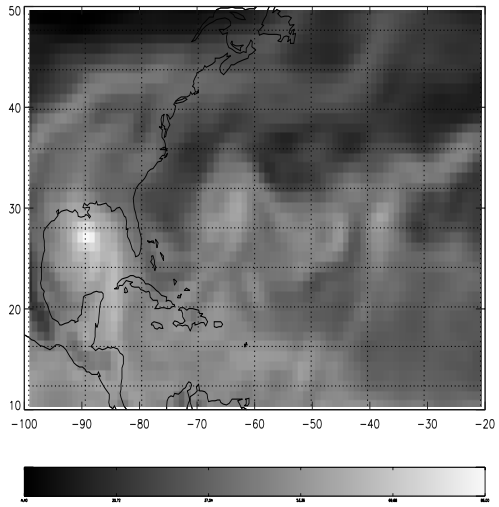


**Figure 1: A region outlier (hurricane) in meteorological data.**

For meteorological data, it is preferable to decompose the original observation data into different spatial scales and treat them separately. That is widely used in meteorological analysis to simplify the problem and centralize the target object. Wavelet analysis method provides the capability for us to achieve this because of its multi-resolution character and the localization of variation in the frequency domain. First, wavelet analysis can decompose the original spatial variation of the data into different scales and focus on only certain scales of interest. Second, the localization of variation in the frequency domain is very helpful in determining the spatial location of the potential spatial outliers.

In our work, we will use wavelet transform in the real spatial domain, then analyze the wavelet transformation data for a particular set of scales. The data will be re-composed back into the spatial domain and the outlier detection algorithm can be used to verify the region outliers at these scales. Technically, we can apply this procedure to all scales at which meaningful meteorological anomalies may exist. However, as spatial outliers are usually small in size compared with the environment, we focus only on small scale weather systems such as hurricanes or tornadoes. We will use wavelet transform on the data along single latitude lines. Since the wavelet transform power indicates the strength of the variation along the latitude, the localization of the high value reveals the place where anomalies exist. After integrating the localiza-

tion of the wavelet power at all latitude, the center and the coverage of each suspect region outlier can be determined. It is called suspect region outlier because: (a) the selected regions are areas with prominent spatial variation at certain scales, we need to verify them according to the outlier definition: are they really distinct from their neighbors? (b) we only perform wavelet transform on latitude dimension and derive the variation along latitude($x$). For most weather system, it usually also implies the similar variation along longitude($y$), but we should verify it in case that false region outliers are detected. Therefore, we will apply outlier detection algorithm on the data set to verify these suspect region outliers.

# 4. WAVELET ANALYSIS AND VERIFICATION ALGORITHMS

In this section, we introduce the wavelet analysis and our proposed algorithms. In our approach, we first perform wavelet transform on image data along all latitudes. We select a set of scales which correspond to the sizes of the region outliers of interest, record the spatial indices of the location whose wavelet power is greater than the pre-defined threshold, and perform inverse wavelet transform to reconstruct the data back to the original domain. Note that we choose only the scales of interest into reconstruction. We then group the location indices recorded. Those grouped locations are the areas of suspect region outliers. Finally, a outlier detection algorithm is applied to the reconstructed image data to discover region outliers from those suspect areas.

## 4.1 Wavelet Analysis

Wavelet analysis is a practical tool to study the subjects in signal analysis and image processing. Traditional Fourier transform can also transfer the signal into frequency domain and separate the scales, but wavelet analysis has special attractive features: (1)**Multi-resolution**: wavelet analysis examines the signal at different frequencies with different resolutions. That is to say, it uses wider window for low frequency and uses narrower window for high frequency analysis. This feature especially works well for signals whose high frequency components have short durations and low frequency components have long durations [14]. In the real world, such signals are typical. The changes of the signal at different scales may be studied with different focuses, they can be separated and recomposed at will. This feature makes wavelet an effective tool to filter the signal and focus on certain scales. (2)**Localization of the frequency**: In traditional Fourier transform, the frequency domain has no localization information. In other words, if the frequency changes with time in the signal, it is hard to tell which frequency happens within what time range although all the frequencies may be detected. In the nature world, signals are usually complicated and non-stationary. If we want to know the exact information of a variation, the frequency and the location of a certain variation, or the strength of the variation at certain location, wavelet analysis has advantages over Fourier transform. For a wavelet function $\Psi(t)$, the continuous wavelet transform of a discrete signal $X_i(i = 0, N-1)$ is defined as the convolution of X with scaled and translated $\Psi$:

$$W(n,s) = \sum_{i=0}^{N-1} x(i)\Psi^*\left[\frac{(i-n)\delta t}{s}\right] \quad (1)$$

where (*) indicates the complex conjugate, $n$ is the localization of the wavelet transform and $s$ is the scale. The wavelet transformation can also be inversely transformed to (or reconstruct) the original data set :

$$x_i = \frac{\delta j \delta t^{1/2}}{C_\delta \Psi_0(0)} \sum_{j=0}^{J} \frac{RealW(n,s_j)}{s_j^{1/2}} \quad (2)$$

**Table 1: Scale Table for Mexican Hat Wavelet**

| Index | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Scale | 2 | 2.83 | 4 | 5.65 | 8 |
| Period | 7.95 | 11.23 | 15.9 | 22.47 | 31.79 |

**Table 2: Scale Table for Morlet Wavelet**

| Index | 1 | 2 | 3 | ... | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Scale | 2.83 | 4 | 5.65 | ... | 16 | 22.6 | 32 |
| Period | 2.92 | 4.13 | 5.84 | ... | 16.52 | 23.4 | 33.05 |

where $C_\delta$ is a constant for each wavelet function, $\Psi_0$is the normalized wavelet function, and $J$ is the maximum scale index. For the details of wavelet transform, please refer to [5,25]. We may not include all scales of the wavelet transform into the reconstruction to filter out the variation of no interest, and the reconstructed data will be composed by the scales interesting to us. For example, if the low frequency of the variation in the data set is of interest, a low pass data set may be reconstructed to filter out the high variation and make low variation more visible. Many functions can be used as base or mother function for wavelet analysis. We use two of the most widely used bases: Mexico hat base and Morlet base. The base function for Morlet wavelet is:

$$\Psi_0(\eta) = \pi^{-1/4} e^{\omega_0 \eta} e^{-\eta^2/2} \quad (3)$$

And the Mexico hat function is:

$$\Psi_0(\eta) = \frac{(-1)}{\sqrt{\Gamma(21/2)}} \frac{d^2}{d\eta^2}\left(e^{-\eta^2/2}\right) \quad (4)$$

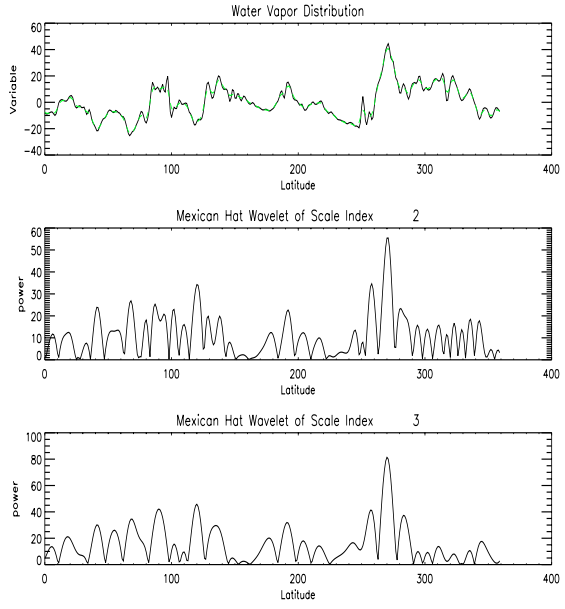


**Figure 2: A sample output of Mexican hat wavelet (a:top, b: center, c:bottom).**

When we perform the wavelet analysis, the scales are selected by $S_0 * 2^{j/2}(j = 0, 1, J)$, and $J$ is the maximum scale index which satisfies : $J \leq 2\log_2(\frac{N}{2})$, where $N$ is the length of the signal. In this case $S_0 = 2\delta x$, $N = 360$. We use $j$ as the scale index. When we say scale 2, we mean the real scale is $S_0 2^{0.5*2} = 4$. Tables 1

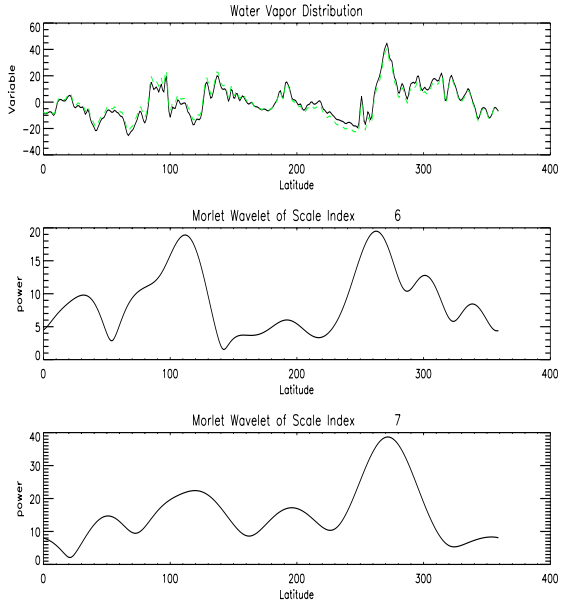


**Figure 3: A sample output of Morlet wavelet (a:top, b: center, c:bottom).**

and 2 provide the relationship between scale index, real scale, and the corresponding period of Fourier transform (here since we do wavelet analysis on spatial domain, it is in fact the wavelength of the spatial variation) for Mexican Hat wavelet and Morlet wavelet. From the tables, it can been seen that as scale increases, the period (or wavelength) of the real object the wavelet focuses on grows as well. However, the growth rates are different for the two wavelets. For Morlet wavelet, the period grows slower than the Mexican hat wavelet. That is to say, Morlet wavelet has a better frequency resolution than Mexican hat wavelet. This also implies that it has a poorer localization resolution.

The Morlet wavelet is a complex wavelet and the Mexico hat wavelet is a real wavelet. Mexican hat captures both the positive and negative variation as separate peaks in wavelet power. Morlet wavelet power combines both positive and negative peaks into a single broad peak. Figure 2 and Figure 3 are examples of the two wavelet transforms. Figure 2(a) is the original data water vapor distribution along a latitude. Figure 2(b) and (c) are the wavelet transform power at two different scales. Figure 3 uses Morlet wavelet at higher scale indices. From Figures 2 and 3, we can see that the power of wavelet transform can depict the distribution or localization of the variation at certain scales. As Mexican hat wavelet provides a better localization (spatial resolution), we will mostly use Mexican hat wavelet to perform the analysis.

## 4.2 Algorithm

We propose two algorithms, Wavelet Analysis and Verification. The Wavelet Analysis Algorithm will apply wavelet transformation to image data in order to discover regions with prominent spatial variation at certain scales. Detailed steps are described in Algorithm 1. As these regions are suspect region outliers, they need to be verified in case false spatial outliers are selected. Verification is based on Z-value approach. Z-value means standardization of the attribute difference between an object and its neighboring objects. It can be used for detecting irregular data objects. We need to reconstruct the original data set to "big points," where each point

**Algorithm 1** Wavelet Analysis

**Input**:
$D$ is the given data set;
$S$ is a set of selected scales;
$\alpha_1$ is the beginning latitude(or longitude);
$\alpha_n$ is the ending latitude(or longitude);
$\theta_w$ is the pre-defined threshold of wavelet power;
$idxSet$ is a set of location indices;
**Output**:
$D'$ is the reconstructed data set
$O_s$ is the set of suspect outlier regions

*/* wavelet transform along all latitudes or longitudes */*
for(i=$\alpha_1$;i $\leq \alpha_n$ ;i++) {
wDomain = WaveletTransform(D,$S$,i); }
*/* record points with prominent wavelet power */*
for every point p in wDomain {
if ( p $> \theta_w$ ) {
AddToLocationSet(p,$idxSet$);} }
$D'$ = inverseTransform(wDomain,$S$)
*/* group points to regions */*
$O_s$ = GroupIndices($idxSet$);
Output($D'$,$O_s$); /*output result*/

denotes a region. Z-value approach can then be used to detect region outliers. Algorithm 2 provides detailed steps for verification.

In Algorithm 1, first, a set of scales of interest should be provided from domain experts. Then wavelet transformation is performed on input data set along all latitudes or longitudes. $\alpha_1$ denotes the beginning latitude(or longitude) and $\alpha_n$ denotes the ending latitude(or longitude). *wDomain* is the domain of wavelet power values transformed from the original data set $D$. The algorithm selected the points with wavelet power greater than threshold $\theta_w$ and records their location indices in a set named $idxSet$. Next, inverse wavelet transformation is performed to reconstruct data set from *wDomain* to original domain $D'$. Note that as only specific scales of interest are selected to transform the wavelet domain back to the original domain, $D'$ is not the same as $D$. $D'$ will be used in Algorithm 2, Verification, to check wether the recorded locations really have abnormal attribute values. The algorithm groups points with adjacent location indices in $idxSet$ to regions and store the regions in $O_s$. These regions are suspect region outliers. Finally, $D'$ and $O_s$ are output to be used in Algorithm 2. In the next step, each region will be viewed as a single point and outlier detection algorithm will be applied to verify wether the suspect regions are true outliers. We describe the verification process in Algorithm 2.

In Algorithm 2, we apply Z-value approach. Z-value approach is a spatial outlier detection algorithm which uses the standardized attribute difference between a point and its neighbors to detect outliers. First, the reconstructed data set is converted to groups of points. Each group is viewed as a "region," whose attribute value is the average of all points in this group. Note that when grouping $D'$, we first take the suspect regions out, then group the remainder of the data set by imposing grids on $D'$. The grid size is defined as the size of the smallest suspect region. This size definition achieves more accurate result as it can well proximate the neighborhood relationship. Each region is viewed as a "big point." Function *createRegions* implements above functionality to get a set of regions $X$=$\{x_1, x_2, \ldots, x_n\}$. Next, function *getAdjNbr* is used to get the adjacent neighbors of each region. Then, the neighborhood function $g$ of a region $x_i$ is taken to be the average attribute value of its neighbors. $|AN(x_i)|$ denotes the number of neighbors of region

**Algorithm 2** Verification

**Input**:
$D'$ is the reconstructed dataset
$O_s$ is a set of suspect regions from wavelet analysis
$AN(x_i)$ is the adjacent neighbors of a region $x_i$
$f(x_i)$ is a function to get attribute value of region $x_i$
$g(x_i)$ is a function to get average attribute value
of the neighbors of $x_i$
$h(x_i)$ is the difference between $f(x_i)$ and $g(x_i)$
$\theta_z$ is the threshold of Z-value

**Output**:
$rSet$ is the set of regions defined as outliers

X = createRegions($D'$,$O_s$);
For each region $x_i \in$ X {
*/* get adjacent neighbor set of $x_i$ */*
$AN(x_i)$ = getAdjNbr(X,$x_i$);
*/* compute difference with neighbors*/*
$g(x_i) = \frac{1}{|AN(x_i)|} \sum_{x \in AN(x_i)} f(x_i)$
$h(x_i)$ = $f(x_i)$ - $g(x_i)$ }
*/*get mean and standard deviation*
*of* $\{h_1, h_2, \ldots, h_n\}$**/*
$\mu_h$ = getMean($\{h_1, h_2, \ldots, h_n\}$);
$\sigma_h$ = getSTD($\{h_1, h_2, \ldots, h_n\}$);
*/* stanardize h value of all regions */*
For each region $x_i \in$ X {
$z_i = |\frac{h_i - \mu_h}{\sigma_h}|$;
if ( $z_i > \theta_z$ ) {
AddToSet($rSet$,$x_i$); } }
Output($rSet$); /*output region outliers*/

$x_i$. Each region may have different number of neighbors. Function *h* is used to calculate the difference between *f* and *g* for each region. Then we can get a sequence of h value $H$=$\{h_1, h_2, \ldots, h_n\}$ for all regions. By standardizing set $H$, the algorithm produces a sequence of Z values, $z_i = |\frac{h_i - \mu_h}{\sigma_h}|$. It is clear that the attribute value of a region $x_i$ is extreme iff $h_i$ is extreme in the standardized data set. Finally, we choose the top ranking regions whose Z value is greater than threshold $\theta_z$ as region outliers. These outliers are stored in $rSet$. Notice that there are two thresholds in Algorithm 1 and Algorithm 2, $\theta_w$ and $\theta_z$. They have different meaning. $\theta_w$ is used for judging if a point's wavelet power is abnormal, whereas $\theta_z$ is used for checking whether the Z-value of a region is above normal values.

## 5. EXPERIMENT RESULTS

In the experiment, we used NOAA/NCEP global reanalysis data sets, which is a multiple-parameter data with a horizontal resolution of 1 degree by 1 degree. The data covers the whole earth and is updated 4 times a day. We used the data of water vapor, as water vapor is a good indicator to depict the weather system. Figure 4 is the image of global water vapor distribution on October 3, 2002. Generally, the tropical region is covered by high value of water vapor. There are also some "hot spots" scattered over the earth. What we want to verify is: Are these spots really outliers? Are there other outliers hidden somewhere?

First we did Mexican hat wavelet analysis on data over all latitudes. Figure 5 is the water vapor data for latitude $26^o$ North and its transformed wavelet power. In Figure 5(a), the solid line is the original data and the dashed line is the filtered (reconstructed with scales 2 and 3) data. Figure 5(b) is the plot of the wavelet power

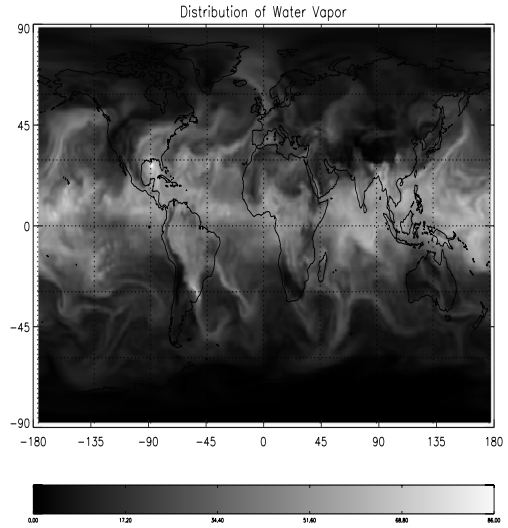


**Figure 4: Global distribution of water vapor.**

of the original data. Figure 5(c) is the plot of the wavelet power of the filtered data. Figure 5 shows that the variation exists on all scales and the power of variation changes at different locations. This figure also shows that Mexican hat wavelet has a satisfactory localization resolution.

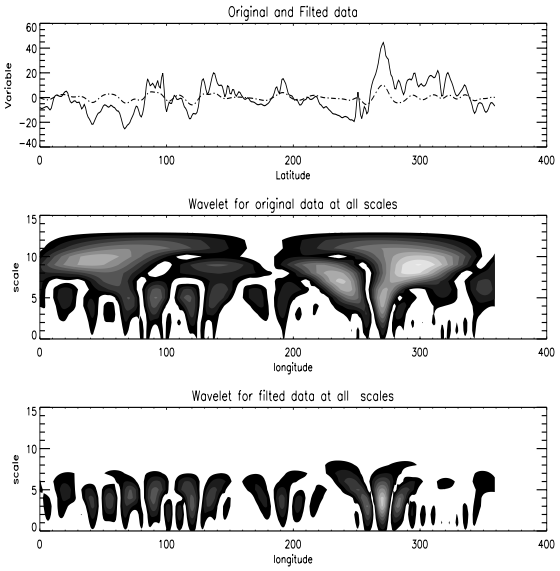


**Figure 5: Mexican hat wavelet power with locations and scales (a:top, b:center, c:bottom).**

We mainly focused on the anomalies with sub-weather scales approximately 1000km or 10 degrees in longitude at mid-latitude region. Figure 6 is the global map of wavelet transform power with scale index 3. Clearly, there are some areas where the power is especially high. In these areas the spatial variation with scale index 3 is prominent and these areas are suspect region outliers.

Comparing Figure 6 with Figure 4, the area with high value in Figure 4 over the Gulf of Mexico also has a high wavelet power.

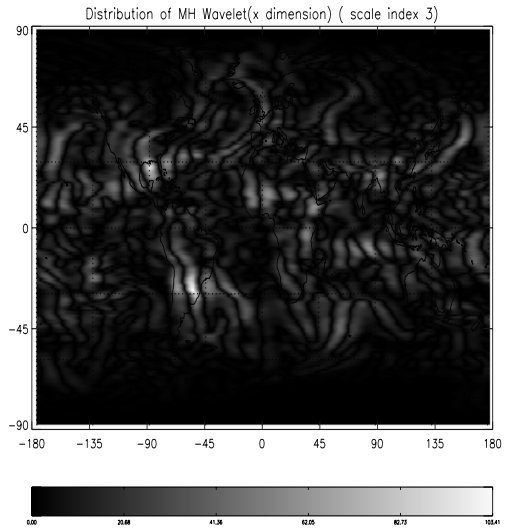


**Figure 6: Wavelet power distribution at scale index 3.**

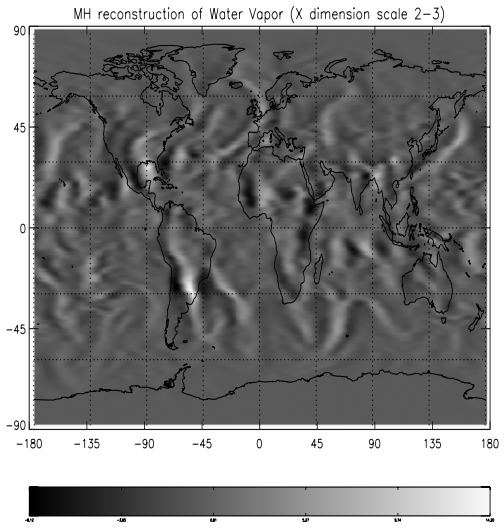


**Figure 7: Reconstructed data based on scale index 2-3.**

However, the high vapor value areas near $160^oW$ in tropic region do not show strong wavelet power in Figure 6, and the low value areas in South America show high wavelet power in Figure 6. Therefore, a high value does not necessarily guarantee a high wavelet power. Wavelet power mainly represents the variation of the signal on the spatial domain. From the outlier point of view, we should focus on the spatial variation, not the value of the variable. The wavelet transform provides a better description of the variation which makes it an effective tool for detecting region outlier. Another advantage of using wavelet transform is its multi-scale capability as we mentioned earlier: we can focus on only the scale of interest. For the multi-scale data such as meteorological data, this makes the complicated variation easier to be studied. Keep in mind that wavelet transform can be used to reconstruct the data. That is the inverse transform from frequency domain back to the original spatial domain. Figure 7 is the reconstructed water vapor distribution using inverse wavelet transforms with scale indices 2 and 3. The new reconstructed data only include the variation of scales we are interested in and filtered out the other variations. Comparing Figure 7 with Figure 6, their spatial patterns are similar, but Figure 7 shows the real physical value distribution.

We performed wavelet transform on the $X$ dimension because for the weather system the scale is usually represented on latitude. For the basic atmospheric parameter distribution, there is a strong variation with latitude, such as the difference between tropics and high latitude areas. This variation is the normal pattern of the general atmosphere and is not the anomalous feature. So when we detect the spatial variation, we focus on the variation along the latitude($X$). Technically, we can also perform wavelet transform along longitude ($Y$). Figure 8 shows the reconstructed water vapor distribution using inverse wavelet transform along latitude and longitude ($X$ and $Y$). Figure 8 reveals more patterns than Figure 6 and Figure 7. These patterns are caused by the normal variation along the longitude $Y$ and are noises in most cases.

After we did the wavelet transform and the inverse transform, we can identify the suspect areas which are the candidates for the region outliers. We used thresholds on wavelet power to determine which points fall into the candidate group and record the spatial indices ((i,j) or (latitude, longitude)) of the points. After grouping the indices, we obtained the suspect outlier regions which are the spatial outlier candidates. And we used the approach described in Algorithm 2 to verify the suspect outlier regions. The results of the $Z$-value verification on the reconstructed data set (Figure 7) identify two outlier regions, one over south America (center at $27^oS$ and $55^oW$) and one over the Gulf of Mexico region(center at $27^oN$ and $90^oW$), both with distinct $Z$-values. Those two regions are detected as region outliers at sub-weather scale. Meteorological records confirm that the region over the Gulf of Mexico area is a hurricane and the region over south America is a storm.

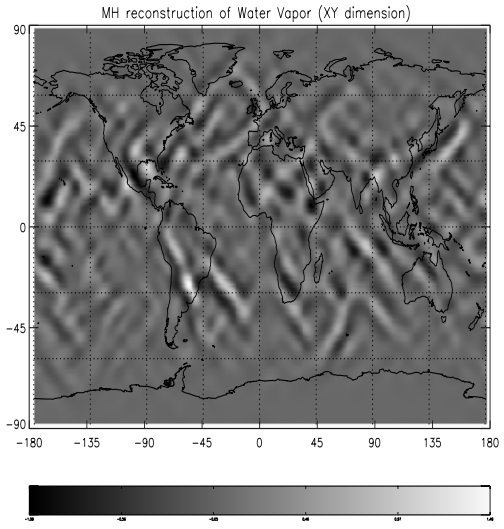


**Figure 8: Reconstruction on both X-Y dimensions.**

## 6. CONCLUSION

In this paper, we propose a wavelet analysis based approach to detect region outliers. Using wavelet analysis on the global meteorological data, we take the advantage of the multi-scale capability of the wavelet transform and inverse transform to focus on certain scales of spatial variation. This will help to identify distinct spatial patterns. Some of those patterns may be visible in the original data set, whereas some may be hidden in the original data and might be ignored if not using wavelet analysis. By analyzing the results from the wavelet analysis, we identify the suspect region outliers which could be the spatial outliers. We then propose a spatial outlier detection approach to verify whether these regions are statistically different from their surrounding neighbor and are true region outliers. Our experiment results demonstrate that our approach can effectively discover anomalies corresponding to severe weather events.

In this work, we focus on two-dimension region outlier. In the future, we are planning to explore region outlier in three-dimension space. In three-dimension space, it will require new definition of region and neighborhood relationship. Currently, our approach detect region outliers at a fixed time frame, and it would be interesting to trace the moving of region outliers at different time frames. In addition, we would like to extend our algorithm to process region outliers with multiple features, such as the combination of pressure, rain fall, wind, cloud, and temperature, and to apply our algorithms to other spatial data sets, such as medical images, gene data, and nature resource data.

## 7. REFERENCES


[1] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, New York, 1994.

[2] M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. OPTICS-OF: Identifying Local Outliers. In *Proc. of PKDD '99, Prague, Czech Republic, Lecture Notes in Computer Science (LNAI 1704), pp. 262-270, Springer Verlag*, 1999.

[3] S. Chawla, S. Shekhar, W.-L. Wu, and U. Ozesmi. Modelling spatial dependencies for mining geospatial data: An introduction. In *Harvey Miller and Jiawei Han, editors, Geographic data mining and Knowledge Discovery (GKD)*, 1999.

[4] G. Erlebacher, M. Hussaini, and L. Jameson. *Wavelet Theory and its Application*. Oxford University, 1996.

[5] E. Foufoula-Georgiou and E. P. Kumar. *Wavelets in Geophysics*. Acadamic Press, 1995.

[6] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[7] D. Hawkins. *Identification of outliers*. Chapman and Hall, Reading, Massachusetts, 1980.

[8] E. Knorr and R. Ng. A Unified Notion of Outliers: Properties and Computation. In *Proc. of the International Conference on Knowledge Discovery and Data Mining*, pages 219–222, 1997.

[9] E. Knorr and R. Ng. Algorithms for mining distance based outliers in large datasets. In *Proceedings of 24 th VLDB Conference*, 1998.

[10] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. In *Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'96)*, pages 1–10, Montreal, Canada, 1996.

[11] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases, Proc. of 4th International Symposium, SSD'95*, pages 47–66, Portland, Maine, USA, 1995.

[12] T. Li, Q. Li, S. Zhu, and M. Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explorations*, 4(2):49–67, 2002.

[13] Y. Meyer. *Wavelet and Operators*. Cambridge University Press, 1992.

[14] R. Polikar. *The Wavelet Tutorial*. Internet resources, http://engineering.rowan.edu/

[15] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 427–438, 2000.

[16] P. Rigaux, M. Scholl, and A. Voisard. *Spatial Database: With Application to GIS*. Morgan Kaufmann, 2002.

[17] I. Ruts and P. Rousseeuw. Computing Depth Contours of Bivariate Point Clouds. In *Computational Statistics and Data Analysis, 23:153–168*, 1996.

[18] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of 24th VLDB Conference*, 1998.

[19] S. Shekhar and S. Chawla. *A Tour of Spatial Databases*. Prentice Hall, 2002.

[20] S. Shekhar, S. Chawla, S.Ravada, A. Fetterer, X. Liu, and C. Lu. Spatial database: Accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):45–55, 1999.

[21] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. In *Proc. Spatio-temporal Symposium on Databases*, 2001.

[22] S. Shekhar, Y. Huang, W. Wu, C. Lu, and S. Chawla. What's special about spatial data mining: three case studies. In *Data Mining for Scientific and Engineering Applications. V. Kumar, R. Grossman, C. Kamath, R. Namburu (eds.)*, 2001.

[23] S. Shekhar, C. Lu, and P. Zhang. Detecting graph-based spatial outliers. *International Journal of Intelligent Data Analysis (IDA)*, 6(5):451–468, 2002.

[24] T. Johnson and I. Kwok and R. Ng. Fast Computation of 2-Dimensional Depth Contours. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 224–228. AAAI Press, 1998.

[25] C. Torrence and G. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, January 1998.

[26] Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.

[27] D. Yu, G.Sheikholeslami, and A. Zhang. Findout: Finding outliers in very large datasets. In *Development of Computer Science and Engineering. Tech Report 99-03*, SUNY Buffalo, 1999.

[28] D. Ziou and S. Tabbone. Edge detection techniques: An overview. *International Journal of Pattern Recognition and Image Analysis*, 8(4):537–559, 1998.