Automated detection of frontal systems from numerical model-generated data

Conference Paper · August 2005

DOI: 10.1145/1081870.1081973 · Source: DBLP

10 authors, including:

Rahul Ramachandran
University of Alabama in Huntsville

Steven Greco
Simpson Weather Associates

Robert Atlas
National Oceanic and Atmospheric Administration

Automated Detection of Frontal Systems from Numerical Model-Generated Data

Xiang Li, Rahul Ramachandran, Sara Graves, Sunil Movva, Bilahari Akkiraju
Information of Technology and Systems Center
University of Alabama in Huntsville

David Emmitt, Steven Greco
Simpson Weather Associates

Robert Atlas
Goddard Space Flight Center, NASA

Joseph Terry, Juan-Carlos Jusem
Science Applications International Corporation

ABSTRACT
Fronts are significant meteorological phenomena of interest. The extraction of frontal systems from observations and model data can greatly benefit many kinds of research and applications in atmospheric sciences. Due to the huge amount of observational and model data available nowadays, automated extraction of front systems is necessary. This paper presents an automated method to detect frontal systems from numerical model-generated data. In this method, a frontal system is characterized by a vector of features, comprised of parameters derived from the model wind field. K-means clustering is applied to the generated sample set of the feature vectors to partition the feature space and to identify clusters representing the fronts. The probability that a model grid belongs to a front is estimated based on its feature vector. The probability image is generated corresponding to the model grids. A hierarchical thresholding technique is applied to the probability image to identify the frontal systems and a Gaussian Bayes classifier is trained to determine the proper threshold value. This is followed by post processing to filter out false signatures. Experiment results from this method are in good agreement with the ones identified by the domain experts.

Categories and Subject Descriptors
J.2 [Computer Applications]: Physical Sciences and Engineering.

General Terms
Design, Experimentation.

Keywords
Data mining applications, spatial data mining.

1. INTRODUCTION
Our daily weather is described by the atmospheric state variables and parameters such as temperature, pressure, wind, cloud amount and precipitations. However, it is the significant meteorological phenomena that catch most of our attention since these meteorological phenomena indicate weather changes and even the onset of severe weather that may affect our daily lives. These significant atmospheric phenomena of interests include hurricanes, tornados, and fronts. A front is a narrow transition zone at mid-latitudes where two air masses with different properties meet. A frontal system causes significant changes in temperature, wind, pressure, precipitation and even triggers severe weather. Global distribution of frontal systems can be obtained from satellite observations using cloud patterns and other measurements. Frontal systems information is also contained in the numerical model simulations that predict the atmospheric state and its evolution. Experienced domain experts can visually detect frontal systems in observational and model simulation data. However, with the advances in data collecting technology and computational power, the enormous volume of data collected by the observational platforms and generated by model simulations render such manual analysis impossible.

Data mining allows analysis and information extraction from large volumes of multidimensional data and has been successfully applied in the business domain. Scientific data mining has also developed as an application area of interest. Mining earth sciences data is one such example and is an important application area due to the huge volumes of data available and the spatial and temporal nature of the data. Over the years, data mining applications in Earth science range from mining framework development [6] to information extraction such as discovering ocean climate indices [7] and feature detection [2], [4], [5] using various kinds of data mining techniques. In this paper, we present an automated method to detect frontal systems in numerical model outputs using data mining techniques. The identified frontal systems will be used to evaluate the performance of a data assimilation system. Based on our knowledge, this is the first application of data mining techniques to detect frontal systems.

Section 2 provides a description of the numerical model data used in this study. Section 3 describes the methodology and data mining techniques used to extract frontal systems from the model data. Section 4 shows the results and discussions. Finally, the summary and future research work is presented in section 5.

2. DATA

The data used for the data mining experiment is the output from the Goddard Laboratory of Atmosphere's finite volume Community Climate Model (fvCCM). As part of the Observing System Simulation Experiments (OSSEs) research project, the fvCCM baseline model run, referred to as the Nature Run (NR), was generated for the time span from September 11 through December 31, 1999. The output fields have a horizontal resolution of 0.5 ° Latitude x 0.625 ° Longitude with a vertical resolution of 36 levels including the surface. The output fields were generated 4 times daily (every 6 hours) and included 3-D fields such as pressure, temperature, wind field, cloud amount and surface variables such as wind field, and temperature.

The data used in this study are the surface wind and potential temperature fields of nature run outputs from September 12, 1999 to September 19, 1999. A total of 31 data sets, corresponding to 31 time steps during this time period were used. The area of study is 20°N - 70°N Latitude and 100°W - 10°E Longitude. Figure 1 shows the surface wind field at September 12, 00Z Universal Time (UT). The solid lines in the figure represent the locations of frontal systems identified by the domain experts and the numbers are the front indices. These frontal system locations are considered as "ground truth" information and are used to evaluate the performance of the automated frontal system detection methodology developed here. The ground truth information covered 8 time steps between September 12, 00Z and September 13, 18Z.

3. METHODOLOGY

The method of automated frontal system detection contains four major steps: 1) identifying features to characterize frontal systems, 2) partitioning feature space to identify front clusters and to generate probability image for frontal systems, 3) detecting frontal systems based on hierarchical thresholding, region merging and shape classification, and 4) post processing to filter out false frontal systems detected and to smooth and skeletonize the frontal regions. These steps are described in details in each of the following sub-sections. An overview of the methodology is shown in Figure 2. The horizontal dash lines in the diagram separate the flow diagram into the four steps and the vertical dash line separates the flow diagram into the development phase and operational phase.

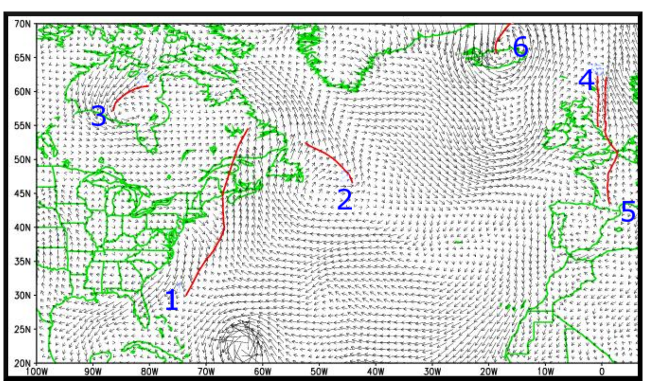

Figure 1. Surface wind field from fvCCM model output at 09/12/00Z, 1999. Solid lines indicate identified front locations.

3.1 Identifying features to characterize frontal systems

A front represents a narrow transition zone between two air masses. Due to the different natures of the two air masses, atmospheric state variables change dramatically in the zone, including pressure, temperature, and moisture content. Also accompanying a frontal system is the formation of clouds, precipitation and wind direction changes. The two of the most significant characteristics of a frontal system are the large temperature gradient across the zone and the significant changes in wind direction. When a cold front passes by, temperature can drop more than 10°C and the wind direction will always veer. Since temperature variation also occurs across coastal zones, it is difficult to determine whether the significant temperature variation is due to a frontal system or due to land-water transition. Thus, the features used to characterize a frontal system are derived only from the model wind fields. Two of the features used are vorticity and confluence. These two features characterize different aspects of wind variations in a zone. Vorticity is helpful

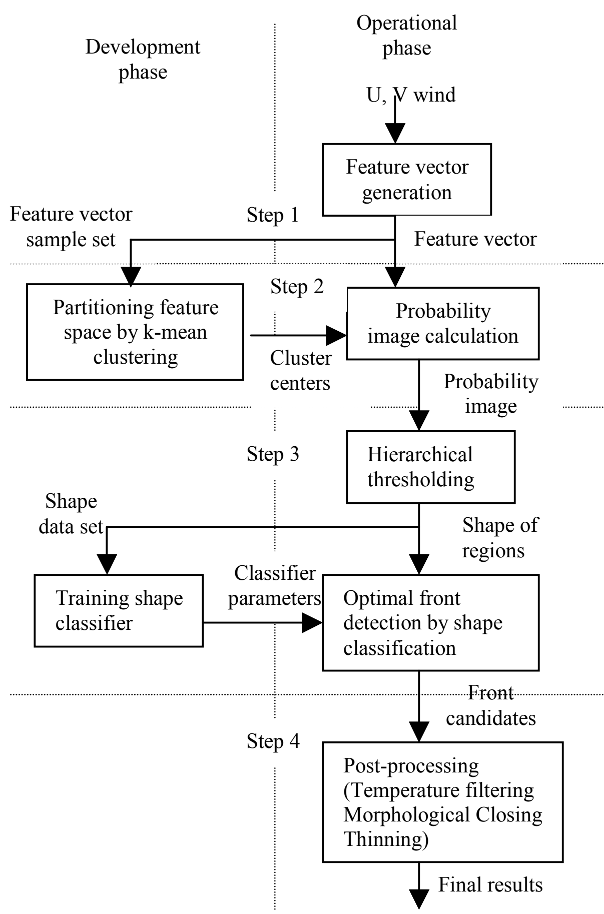

Figure 2. Flow diagram of the automated front detection method

in identifying the veering of wind direction change where as confluence characterizes the convergence zone of wind field and is appropriate for describing the stationary fronts. The third feature is the wind direction variation (WDV) and characterizes the wind direction change at a grid. It is calculated using the wind parameter of its eight neighboring grids using equations 1 and 2 for grid point (i,j):

$$WDV(i,j) = \frac{1}{N}\sum_{k,m} \Delta wdir(k,m), \quad (1)$$

where k,m=[-1,1] and N=8 and

$$\Delta wdir(k,m) = \left| wdir(k+i, m+j) - wdir(i,j) \right|, \quad (2)$$

where *wdir*(i,j) indicates wind direction at grid (i,j). If Δ*wdir*(i,j) is larger than 180, then Δ*wdir*(i,j) is set as 360 - Δ*wdir*(i,j).

Combined with the wind speed, a total of four features are used to characterize a frontal system. These features are calculated for each model grid and are grouped as a feature vector. Each feature vector is a point in the 4-dimensional feature space.

3.2 Partitioning feature space using K-means clustering method

Each model grid is represented with a feature vector for front detection purpose. A grid point may be a part of a frontal system or not, depending on the values of the feature vector. It is assumed that the distribution of feature vectors for front and non-front grids are significantly different. Therefore, they will form clusters at different regions of feature space. A feature vector data set is generated using the wind field at 09/18/00Z and 09/19/00Z. The two time steps are randomly taken and the generated data set is assumed to be representative of all feature vectors. The feature vector at each model grid point for the two time steps contributes to the data set. Consequently, the feature vector data set contains both front and non-front information in these samples. The K-means clustering [8] is applied to the generated data set to partition the feature space into individual clusters. The purpose of applying K-means clustering is to identify the clusters in feature space that represent frontal system. Since the number of cluster k has to be supplied to the clustering algorithm, the value of k is determined using following heuristic approach. The K-Means clustering algorithm is applied iteratively to the data set with k values ranging from 3 to 25. The overall root-mean-square (RMS) error is calculated for each k value. The overall RMS error decreases monotonically with increasing k value. The final value of k is the one at which the decrease of relative RMS error is less than 1%. Experiment shows that the heuristic k value for the data set is 19 and is used in the K-means clustering. The 19 cluster centers from K-means algorithm are shown in Table 1. A frontal system is normally characterized by moderate to high values in wind speed, vorticity, confluence, and wind direction variation. As a result, the clusters 6, 7, 8, 11, and 18 are labeled as the ones representing the features of frontal systems based on the domain knowledge. The rest of the centers in Table 1 represent the features of non-frontal systems.

Similar to the membership function used in the Fuzzy C-Mean clustering [8], a membership vector M is calculated for a feature vector based on the distance of the feature vector to each of the cluster centers. The size of the membership vector is K and each vector element M_i represents the probability that the feature vector belongs to the ith cluster center. The membership vector M for feature vector V is calculated using equations 3 and 4:

Table 1. Cluster centers from K-Means clustering algorithm, K=19. In the table, Conflu stands for confluence. ID represents cluster ID.

ID	Wind Speed (m/s)	Conflu ($x10^{-3}$)	WDV (°)	Vorticity ($x10^{-6} s^{-1}$)	Cluster Labels
1	5.9	-5.2	7.9	28.5	No front
2	1.9	-6.4	35.5	-13.3	No front
3	1.3	4.9	70.5	10.3	No front
4	4.2	-12.2	9.3	-10.7	No front
5	3.9	-0.7	4.7	-1.5	No front
6	3.4	19.4	28.3	33.2	Front
7	5.4	52.4	35.3	105.4	Front
8	29.9	160.6	46.4	397.8	Front
9	4.8	13.8	9.2	6.2	No front
10	11.2	-6.2	2.0	-7.0	No front
11	9.2	21.5	8.0	54.7	Front
12	7.7	6.4	3.1	-13.3	No front
13	5.7	-29.9	13.2	-39.8	No front
14	6.0	1.4	6.5	-32.6	No front
15	20.1	-25.4	5.3	9.2	No front
16	8.3	2.4	2.1	-0.4	No front
17	2.7	1.5	17.5	-12.9	No front
18	24.0	52.9	11.9	63.0	Front
19	12.9	7.7	2.8	10.9	No front

$$Dist(i) = \sqrt{\sum_{j=1}^{N} (V_j - C_{j,i})^2} \quad (3)$$

$$M_i = \left(\sum_{k=1}^{K} \frac{Dist(i)}{Dist(k)}\right)^{-1}, \text{ M = 1, 2, …, K} \quad (4)$$

where N is the dimension of the feature space, K is the number of clusters, and M_i is the ith element of the membership vector.

Based on the membership vector information, the probability image is generated corresponding to the model grid. The value of an image pixel is the probability that it belongs to a frontal system and is determined as the summation of the membership elements corresponding to the front clusters, that is, clusters 6, 7, 8, 11, and 18. Figure 3 shows the probability image at 09/12/00Z. Dark pixels indicate high probabilities that they belong to a frontal system. It is seen that regions identified as frontal systems by the experts coincide with the grid points with high probabilities in the image.

3.3 Detecting frontal systems using hierarchical thresholding technique

Frontal systems are detected by applying thresholding technique to the generated probability image. Figure 4 shows the binary image with threshold value of 0.15 at 09/12/00Z. Clearly, the frontal systems identified by the domain experts are observed in

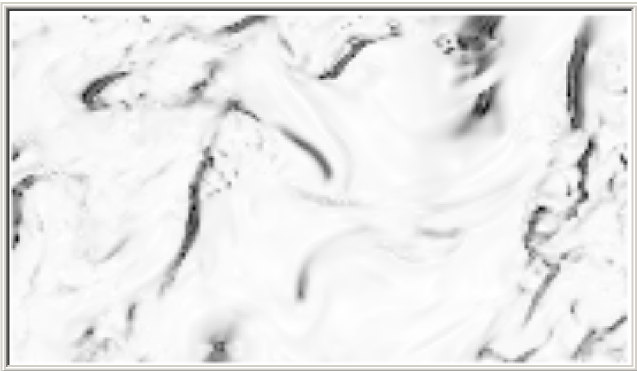

Figure 3. Probability image at 09/12/00Z. The value at each grid represents the probability that it belongs to a frontal system.

this binary image. Applying a region growing [1] technique to the binary image, we obtain a number of individual regions. Since the probability values are different for different frontal systems due to their intensity differences, it is not possible to properly detect all frontal systems using a single threshold. If threshold value is too low, the detected region may be too large to accurately represent the true frontal system. If threshold is too high, a frontal system may be broken into pieces or not be detected. For instance, the region detected over Iceland in Figure 4 is too large for the front identified and labeled as index 6 in Figure 1. Higher threshold is needed to detect this front. As a result, instead of using a global thresholding technique, a hierarchical thresholding is used in this method. It is important to note that all the detected regions are not the true frontal systems.

Fronts normally appear as elongated strips or curves due to their narrow transition zones. Therefore, a region can be identified as frontal or non-frontal regions based on its shape. Many kinds of shape descriptors are available for shape description. In this study, the following eight shape descriptors based on the region and its convex hull [1] are used: 1) perimeter of the region; 2) area of the region; 3) perimeter of the convex hull of the region; 4) area of the convex hull of the region; 5) aspect ratio of the region, defined as the square of the region perimeter divided by the 4 times the area of the region; 6) the ratio of the two principle axes of the region [3]; 7) ratio of the perimeter of the region to the perimeter of its convex hull; and 8) ratio of the area of the region to the area of its convex hull. A Gaussian Bayes classifier is trained to

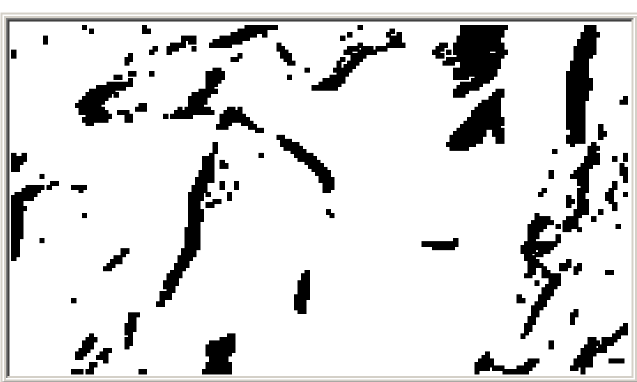

Figure 4. Thresholded binary image of the probability image at 09/12/00Z. The threshold is set as 0.15.

Figure 5. Identified frontal systems at 09/12/00Z using hierarchical thresholding technique.

classify a region as front or non-front based on its shape descriptor. The sample data set of shape descriptors used to train this classifier is generated by hierarchical thresholding of probability images. Each of the detected regions is labeled as front or non-front based on the truth data. Frontal regions are identified at each threshold level. If, over the same geographical region, more than one fronts are detected at different threshold levels, the front that shows maximum likelihood based on the Bayes classifier is considered as the final front.

The generated data set had a total of 1045 samples with 386 samples are labeled as fronts. The generated sample data set was randomly split into two subsets, one for training classifier and the other for testing classifier performance. Training subset had total of 563 samples while testing subset had total of 482 samples. The accuracies of training and testing set are 90.2% and 91.5%, respectively. The overall accuracy of the classifier is 90.8%. Figure 5 shows the detected front regions at 9/12/00Z using hierarchical thresholding and Bayes classifier.

3.4 Post processing

The potential temperature field from the model output is used to further filter out false signatures detected during the hierarchical thresholding process. Since significant temperature gradient is a prominent feature of a frontal system, the standard deviation of potential temperature (STD_PT) of an identified region is calculated and is used to determine if it is a frontal region or not. If the STD_PT value of a region over water surface is less than 0.9 K or the region over land surface is less than 1.5 K, the identified frontal region from hierarchical thresholding is tagged as a non-frontal system. The differences in potential temperature criteria over land and water are because the temperature over oceans is more uniform than that over land surfaces. Figure 6 shows the result of applying the temperature filters. Compared with Figure 5, a couple of regions erroneously labeled as frontal systems have been filtered out.

A morphological closing operation [1] is further applied to clean up the frontal systems detected by smoothing the boundaries of the fronts and filling the holes in the front regions. The template size used for the closing operation is 3 x 3. The smoothed frontal regions are then thinned using the Zhang-Suen algorithm [9] to obtain the skeletons of the frontal systems. The skeleton of a frontal system indicates the location center of the system.

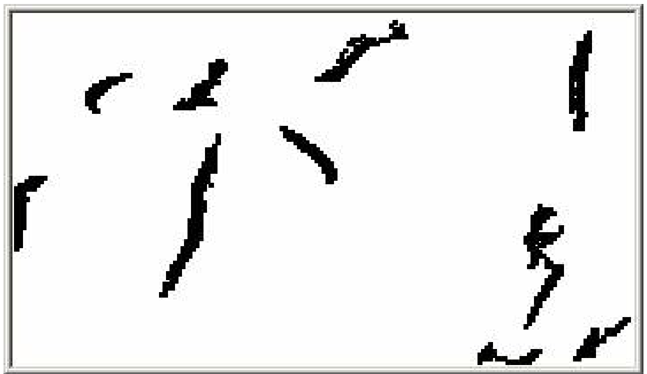

Figure 6. Frontal systems identified at 09/12/00Z after noise filtering using potential temperature parameter.

4. RESULTS AND ANALYSES

The frontal system detection methodology is applied to all 31 data sets, of which the ground truth of the frontal systems is available from 09/12/00Z to 09/13/18Z. Figures 7a, 7c, 7e, and 7g show the wind field and expert labeled frontal systems at the time steps of 00Z and 12Z of 09/12, and 00Z and 12Z of 09/13, respectively. Figures 7b, 7d, 7f, and 7h show the detected frontal systems at these corresponding time steps.

There are total of 34 frontal systems identified by the experts in these 8 time steps. Fronts with indices 4 and 5 in Figures 7a and 7c are counted once at each time step since they are too close to each other and are detected as a single frontal system by our method. Our method detects 27 of the 34 frontal systems and the detection rate is 79.4%. The major frontal system at the east coast of the U.S., which is indexed as 1 in these figures, is precisely detected by the method at each of the time steps. This method also detects the other major frontal system, indexed as frontal system 9 in the figures, which moves across from the West to the East of the U.S. In addition, this method detects this frontal system at the early stage of its development at 09/12/00Z that was not identified by the domain experts. Experts did not identify this front at this time step because it is located at the boundary of the area of the study. Front with index of 7, located at southwest of the Greenland, is missed at 09/13/00Z, though part of the frontal system is detected at its two adjacent time steps, 09/12/18Z and 09/13/06Z (Figures are not shown due to space limitation). The method does not identify the front at 09/12/00Z over the north of the Iceland, which has front index of 6 in Figure 7a. Also, the front over the northeast of Canada, which is indexed as 8, is not detected at time steps 09/12/06Z, 09/12/12Z, and 09/13/00Z. However, it is identified at time steps 09/12/18Z and 09/13/06Z. This is a weak front compared to fronts 1 and 9. The frontal systems that are missed by the classifier are probably because of the weak shape descriptors defining these systems. This method also persistently detects regions over Spain and northwest of Africa, even though there are no weather fronts identified by domain experts over these regions. The wind conditions due to orographical effects are most likely causing this detection. Special treatment and additional information is needed to filter out these false frontal systems identified over the regions.

5. SUMMARY

An automated frontal detection methodology is developed in this study that utilizes a number of data mining techniques. Three feature components in addition to wind speed are derived from the model wind field and are used to characterize frontal systems. A K-Means clustering algorithm is used to partition the feature space into clusters and the clusters for the frontal systems are identified. The probability that each grid point belongs to a frontal system is calculated based on its feature components and is used to create a probability image. A hierarchical thresholding technique is applied on this image to identify the frontal systems and a Gaussian Bayes classifier is trained to help determine the proper threshold value for a particular frontal system. Model data fields such as temperature are used to help filtering out some of the false frontal systems detected. Experiments results show that our method identifies most of the frontal systems labeled by the domain experts. It also identifies some of the systems that were missed by the domain experts. Some of the frontal systems identified by the automated method are false signatures.

ACKNOWLEDGEMENT

This work was funded by Ames Research Center and Goddard Space Flight Center, NASA grant NAG5-13263.

REFERENCES

[1] Gonzalez, R.C., and Woods, R.E. Digital image processing, *Addison-Wesley Publishing Company, 1992, 716pp.*

[2] Li, X., Ramachandran, R., Rushing, J., Graves, S., Kelleher, K., Lakshmivarahan, S., Kennedy, D., and Levit, J. Mining NEXRAD Radar Data: An Investigative Study**,** *80th AMS Annual Meeting, 20th International Conference on Interactive Information and Processing Systems (IPPS) for Meteorology, Oceanography, and Hydrology, Seattle, WA, Jan. 11 - 15, 2004.*

[3] Luo, T., Kramer, K., Goldgof, D.B., Hall, L.O., Remsen, A., and Hopkins, T. Recognizing plankton images from the shadow image particle profiling evaluation recorder, *IEEE Trans. On Sys. Man and Cyber., part B:Cybernetics, Vol., 34, No. 6, 2004.*

[4] Nair, U., Rushing, J., Ramachandran, R., Kuo, K.-S., Graves, S.J., and Welch, R. Detection of Cumulus Cloud Fields in Satellite Imagery**,** *The International Symposium on Optical Science, Engineering and Instrumentation, Denver, 1999.*

[5] Peak, J.E., and Tag, P.M. Segmentation of Satellite Imagery Using Hierarchical Thresholding and Neural Networks, *Journal of Applied Meteorology: Vol. 33, No. 5, 1994, pp. 605–616*

[6] Rushing, J., Ramachandran, R., Nair, U.J., Graves, S.J., Welch, R., and Lin, A. ADaM: A Data Mining Toolkit for Scientists and Engineers, *Computers and Geosciences, Nov. 1, 2004*

[7] Steinbach M., Tan P.-N., Kumar V., Klooster S., and Potter, C. Data mining for the discovery of ocean climate indices, *Proc of the Fifth Workshop on Scientific Data Mining at 2nd SIAM International Conference on Data Mining, 2002.*

[8] Theodoridis, S., and Koutroumbas, K. Pattern recognition, *Academic Press, 1999, 625pp.*

[9] Zhang, T.Y., and Suen, C.Y. A fast parallel algorithm for thinning digital patterns, *Comm. ACM, vol. 27, no. 3, 1984, pp236-239.*

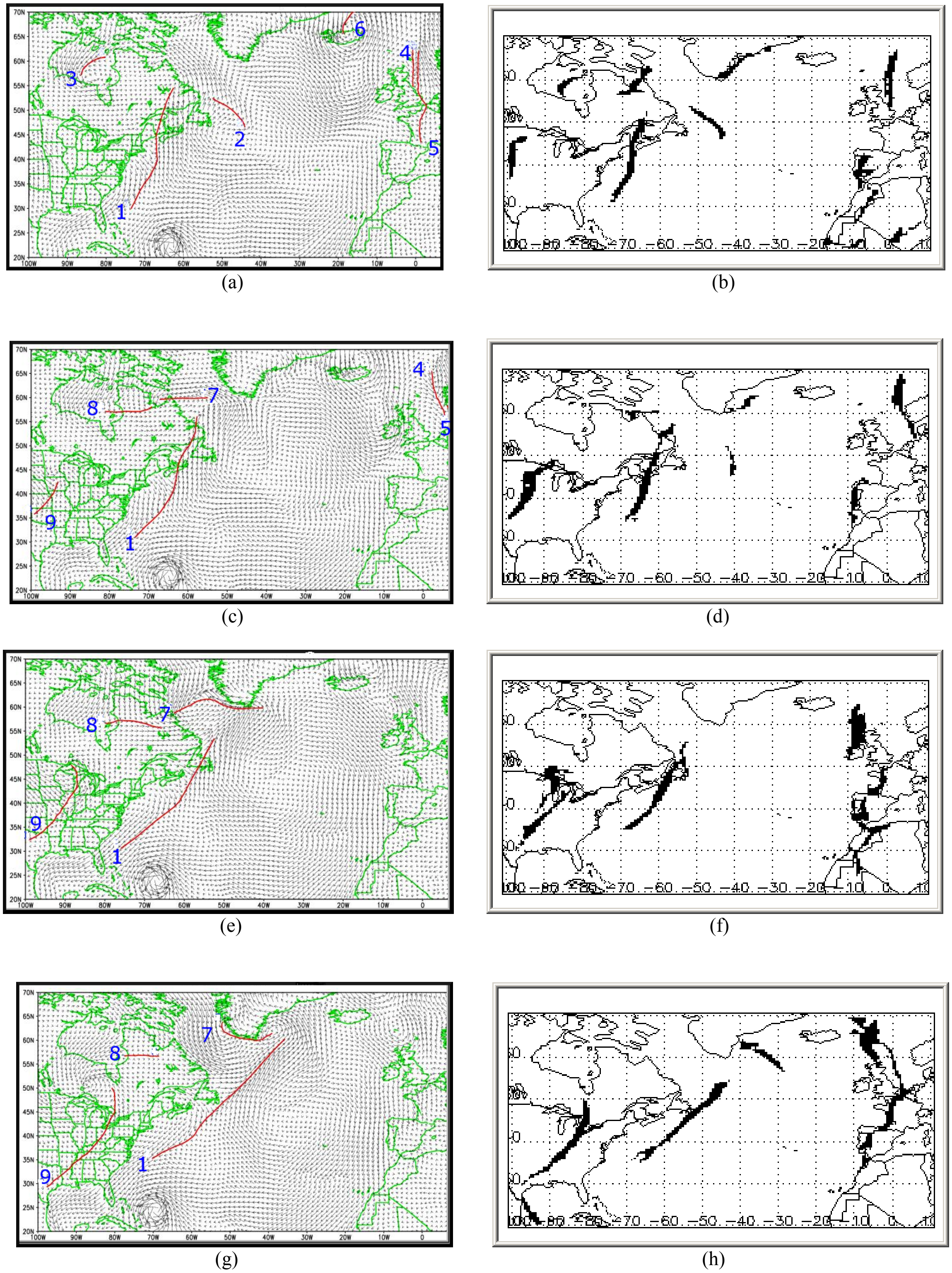

Figure 7 (a) – (h) Comparisons of frontal systems detected using the automated data mining method with the ground truth frontal systems identified by the domain experts. (a), (c), (e), and (g) show the ground truth fronts at 09/12/00Z, 09/12/12Z, 09/13/00/Z, and 09/13/12Z, respectively. (b), (d), (f), and (h) show the automatically detected frontal systems at 09/12/00Z, 09/12/12Z, 09/13/00Z, and 09/13/12Z, respectively.