Automated Cyclone Discovery and Tracking using Knowledge Sharing in Multiple Heterogeneous Satellite Data

Shen-Shyang Ho  
Jet Propulsion Laboratory  
California Institute of Technology  
4800 Oak Grove Ave 300-123  
Pasadena, CA 91109  
Shen.Shyang.Ho@jpl.nasa.gov

Ashit Talukder  
Jet Propulsion Laboratory  
California Institute of Technology  
4800 Oak Grove Ave 300-123  
Pasadena, CA 91109  
Ashit.Talukder@jpl.nasa.gov

ABSTRACT

Current techniques for cyclone detection and tracking employ NCEP (National Centers for Environmental Prediction) models from in-situ measurements. This solution does not provide true global coverage, unlike remote satellite observations. However, it is impractical to use a single Earth orbiting satellite to detect and track events such as cyclones in a continuous manner due to limited spatial and temporal coverage. One solution to alleviate such persistent problems is to utilize heterogeneous sensor data from multiple orbiting satellites. However, this solution requires overcoming other new challenges such as varying spatial and temporal resolution between satellite sensor data, the need to establish correspondence between features from different satellite sensors, and the lack of definitive indicators for cyclone events in some sensor data.

We describe an automated cyclone discovery and tracking approach using heterogeneous near real-time sensor data from multiple satellites. This approach addresses the unique challenges associated with knowledge discovery and mining from heterogeneous satellite data streams. We consider two remote sensor measurements in our current implementation, namely: QuikSCAT wind satellite measurements, and merged precipitation data from TRMM and other satellites. More satellites will be incorporated in the near future and our solution is sufficiently powerful that it generalizes to multiple sensor measurement modalities. Our approach consists of three main components: (i) feature extraction from each sensor measurement, (ii) an ensemble classifier for cyclone discovery, and (iii) knowledge sharing between the different remote sensor measurements based on a linear Kalman filter for predictive cyclone tracking. Experimental results on historical hurricane datasets demonstrate the superior performance of our approach compared to previous work.

Categories and Subject Descriptors

I.5.4 [Pattern Recognition]: Applications

General Terms

Algorithms, Design.

Keywords

Mining massive data stream, real-time data mining, heterogeneous data mining, multi-sensor data fusion, knowledge transfer, ensemble classifier, event tracking, event detection and prediction

1. INTRODUCTION

Tropical and extra-tropical cyclones are important components of the Earth climate system that exhibit variability at different temporal and spatial scales. Cyclone landfall causes great devastation, incurs fatality, and affects people’s livelihood. To identify and track tropical weather system, the Tropical Prediction Center/National Hurricane Center (TPC/NHC) uses conventional surface and upper-air observations and reconnaissance aircraft reports [1], and these are concentrated in the North American coasts and in Japan/Europe to some degree. Coverage on a global basis, especially in under-developed and developing nations such as large portions of Asia and Africa is limited or lacking which results in disastrous consequences in many of these regions. In recent years, some studies have used satellite images that are manually retrieved and analyzed to improve the accuracy of cyclone tracking; this procedure is currently slow, tedious, involves coverage of only local regions in North America, and requires close analysis by teams of experts.

In this paper, we describe a novel automated global cyclone discovery and tracking approach on a truly global basis using near real-time (NRT) (and historical) sensor data from multiple satellites. Our current implementation employs two types of satellite sensor measurements, namely: the QuikSCAT wind satellite data, and the merged precipitation data using TRMM and other satellites. We address the challenges of mining heterogeneous data from multiple orbiting satellites at different spatial and temporal resolutions (see Figure 1). In particular, knowledge sharing between the heterogeneous sensor measurements addresses the problem where some sensor measurements lack definitive indicator for cyclone events, and the spatial and temporal resolutions differ for different sensor. For instance, one cannot confidently identify cyclone based on TRMM precipitation data alone even though it has a finer temporal resolution than QuikSCAT. Through our knowledge...
sharing methodology, QuikSCAT wind data provides information to the TRMM precipitation data about the likely cyclone location so that the TRMM detector can focus its search on some local region and reduces false alarm for cyclone detection on TRMM data measurements (see Figure 2).

![Image](http://podaac.jpl.nasa.gov/hurricanes)

**Figure 1.** Data availability timeline from TRMM (3B42 data), QuikSCAT (L2B data) and Aqua (MODIS) on 18 Aug 2007 for Hurricane Dean

Besides the challenges posed by mining heterogeneous remote satellite data, in general there are challenges related specifically to the problem of detection and tracking of cyclones. First, cyclone events are dynamic in nature i.e., they evolve rapidly in shape and size over time. Second, there is a lack of annotated negative (non-cyclone) examples by experts; this makes training of classifiers for cyclone detection a difficult one. Third, a single satellite sensor may miss a cyclone event due to a pre-defined orbiting trajectory. Our approach addresses and provides an effective solution to each of these challenges and we demonstrate its effectiveness on some recent hurricane events.

The paper is organized as follows. Section 2 provides a brief review on previous work on cyclone detection and tracking. Section 3 describes the data used in our implementation for cyclone discovery and tracking. In Section 4, we discuss some pragmatic issues and research challenges for cyclone discovery and tracking using heterogeneous data from multiple satellites. In Section 5, we described our approach for cyclone discovery using an ensemble classifier and knowledge sharing between QuikSCAT and TRMM data for cyclone tracking. In Section 6, the experimental results on some historical hurricane occurrences are presented. In Section 7, we discuss lessons learned from our prototype implementation.

## 2. PREVIOUS WORK

No solution currently exists that uses heterogeneous sensor measurements to automatically detect and track cyclones. In a few partially successful studies, visible and infrared images from geostationary satellites are analyzed manually, together with other data sources, using the Dvorak technique [2] to classify the tropical cyclone development stage. Different intensity and track forecast models are computed based on the identified hurricane location and related information. The models are analyzed manually to eliminate the unlikely predictions. Automated tropical cyclone forecasting system provides an organized framework for forecaster to access information such as cyclone data and numerical weather prediction (NWP) model data has been developed [3, 4]. Forecasters, however, have to make their own conclusions based on the available information. These work focus on detecting and tracking hurricanes that are likely to landfall only in North America, and they involve human interference and decisions.

Prior techniques proposed for automated storm or cyclone identification and tracking use aerial reconnaissance aircraft data and local radar data that have limited coverage and do not measure parameters on a global scale. An improved algorithm for the Weather Surveillance Radar, 1988, Doppler (WSR-88D) has been proposed for storm identification and tracking [5]. Sinclair [6] noted the importance of good features for cyclone identification and proposed a variant of the vorticity feature. Lee and Liu [7] proposed an automated approach for the Dvorak technique using an elastic graph dynamic link model based on elastic contour matching. Lakshmanan et al. [8] proposed a hierarchical K-means clustering method to identify storms and their motions at different scales. These approaches are concerned with storm or cyclone tracking which require manually locating the initial cyclone tracks and tedious data retrieval.

There are existing and developing web-based information systems which systematically archive satellite measurements for

---

Peta-bytes of Earth science remote sensor measurements acquired by the NASA satellites are publicly available for analysis and knowledge discovery. These data consist of both archived historical (science products) unlabeled data and near real-time (NRT) data streams, much of which is also not analyzed. There are a number of challenges pertaining to mining data from orbiting satellites. For example, each orbiting satellite (such as QuikSCAT, AVHRR, MODIS) typically cannot monitor a region continuously and the measurements are instantaneous. While these challenges cannot be completely overcome, one can minimize their effects by using data from multiple satellites. However, different satellites provide different measurements. Moreover, different satellite sensors acquire measurements at different spatial and temporal resolutions. These problems make mining heterogeneous data from multiple orbiting satellites extremely challenging, and remains as of now primarily an unsolved problem.
hurricanes\textsuperscript{1} or (more generally) tropical cyclones\textsuperscript{2} for scientific purposes. However, these information systems are based on track information from TPC/NHC. Two products of JAXA (Japan Aerospace Exploration Agency) related to our research are the “AMSR-E Typhoon Real-Time Monitoring” for the Western Pacific region and a global real-time monitor using the TRMM satellite\textsuperscript{3}. Again, these products involve human detection and tracking. One interesting development in event monitoring is the Autonomous Science-craft Experiment (ASE) which automatically prioritizes and schedules observations on regions of interest \textsuperscript{9}. Currently, this technology is used for NRT monitoring of events such as volcano activities\textsuperscript{5} and floods\textsuperscript{6}.

3. DATA DESCRIPTION

In this section, we describe the two types of remote sensing data used in our cyclone discovery and tracking implementation: QuikSCAT wind data from a polar orbiting satellite (Section 3.1), and the merged high quality/infrared (HQ/IR) precipitation data from the TRMM orbiting satellite and other Geostationary Operational Environmental Satellites (GOES) (Section 3.2).

3.1 QuikSCAT Wind Data

The QuikSCAT (Quick Scatterometer) mission provides important high quality ocean wind data set. QuikSCAT is a polar orbiting satellite with 1800 km wide measurement swath on the Earth surface. Generally, this results in twice per day coverage over a given geographic region. The specialized microwave radar (SeaWinds instruments) on the QuikSCAT satellite measures wind speed and direction under all weather and cloud conditions over Earth oceans. Near real-time wind data is available to weather forecasting agencies from NOAA within three hours of observation. The ocean wind vectors in the measurement swaths have a spatial resolution of 12.5 and 25 km. The ocean wind data is used for global weather forecasting and modeling. It is also used to understand environmental phenomena such as El-Niño, tropical cyclones, and the effects of winds on ocean biology.

The SeaWinds Processing and Analysis Center (SeaPAC) at JPL is responsible for the reception of the telemetry data from the satellite, raw data processing and analyzing. The processed data is then delivered to the Physical Oceanography Distributed Active Archive Center\textsuperscript{2} (PO.DAAC) for public distribution. More information about QuikSCAT science data product is found in \textsuperscript{10}.

Recent research showed that QuikSCAT data is useful for early identification of tropical depression \textsuperscript{11} and early detection of tropical cyclones \textsuperscript{1, 12}. Moreover, QuikSCAT data has been used in the three-dimensional variational (3D-var.) data assimilation technique for better cyclone track and intensity forecasting \textsuperscript{13}. Our recent work \textsuperscript{14} showed the feasibility of using QuikSCAT wind measurements for automated cyclone identification.

3.2 Precipitation Data from TRMM satellite

The Tropical Rainfall Measurement Mission (TRMM) is a joint mission between NASA and the JAXA designed to monitor and study tropical rainfall\textsuperscript{8}. The TRMM satellite carries five remote sensing instruments onboard, namely: Precipitation Radar (PR), TRMM Microwave Imager (TMI), Visible Infrared Scanner (VIRS), Clouds and Earth Radiant Energy Sensor (CERES), and Lightning Imaging Sensor (LIS).

The TRMM satellite orbits between 35 degrees north and 35 degrees south of the equator. It takes measurements between 50 degrees north and 50 degrees south of the equator. The real-time processing and post-processing of the TRMM science data is performed by the TRMM Science Data and Information System (TSDIS). All TRMM products are archived and distributed to the public by the Goddard Distributed Active Archive Center (GES DISC DAAC)\textsuperscript{9}.

The (Level) 3B42 TRMM data product used in this paper is produced using the combined instrument rain calibration algorithm using an optimal combination of (Level) 2B-31 data (vertical hydrometeor profiles using PR radar and TMI data), (Level) 2A-12 data (vertical hydrometeor profiles at each pixel from TMI data), SSMI (Special Sensor Microwave/Imager\textsuperscript{10}), AMSR (Advanced Microwave Scanning Radiometer on board the Advanced Earth Observing Satellite-II (ADEOS-II)\textsuperscript{11}) and AMSU (Advanced Microwave Sounding Unit on NOAA geostationary satellites) precipitation estimates, to adjust IR estimates from geostationary IR observations. Near-global estimates are made by calibrating the IR brightness temperatures to the precipitation estimates. The 3B-42 data quantifies rainfall for 0.25°×0.25° degree grid boxes every 3 hours and the precipitation measurements range from 0.0 to 100mm/hr.

4. ISSUES AND CHALLENGES

Before describing our approach in Section 5, we discuss the challenges that we encounter and how they are addressed. The main issue of satellite measurements is that they are instantaneous. For example, a QuikSCAT measurement does not quantify “sustained” wind which defines a cyclone. However, the measurement provides important markers to allow detection of a cyclone or a hurricane-force wind \textsuperscript{16}. Moreover, the QuikSCAT data only measures wind 10m above the ocean surface. The TRMM Level 3B42 data is known to underestimate rainfall. These characteristics do not affect their usefulness for cyclone analysis and understanding by climatologist. It has little effect on our automated detection and tracking system since a cyclone is an event that is characterized by strong wind and substantial rainfall.

\textsuperscript{1}http://disc.sci.gsfc.nasa.gov/hurricane/HurricaneArchiveGallery.html (only North America hurricanes)
\textsuperscript{2}http://tropicalcyclone.jpl.nasa.gov/hurricane/main.jsp
\textsuperscript{3}http://sharaku.eorc.jaxa.jp/TPY_DB/index_e.shtml
\textsuperscript{4}http://www.eorc.jaxa.jp/TRMM/NRTtyphoon/index_j.htm (Japanese Only)
\textsuperscript{5}http://modis.higp.hawaii.edu/
\textsuperscript{6}http://www.dartmouth.edu/~floods/index.html
\textsuperscript{7}http://podaac.jpl.nasa.gov/
\textsuperscript{8}http://trmm.gsfc.nasa.gov/
\textsuperscript{9}http://disc.sci.gsfc.nasa.gov/
\textsuperscript{10}http://www.ncdc.noaa.gov/oa/satellite/ssmi/ssmiproducts.html
\textsuperscript{11}http://sharaku.eorc.jaxa.jp/AMSR/index_e.htm
4.1 Non-Continuous Region Monitoring
Unlike Geostationary Operational Environmental Satellites (GOES)\textsuperscript{12} which monitors a specific region continuously, an orbiting satellite (such as QuikSCAT or TRMM) does not measure a specific region continuously. While USA has at least four GOES, most countries do not have such local weather monitoring satellites. They have to rely on publicly available data from orbiting satellites.

For an orbiting satellite, one can only have data on a particular region at some time instances per day. Sometimes, the orbiting satellite may just miss the region completely (See the white region between any two side-by-side swaths in Figure 3). This problem can be alleviated to a large extent by using sensor measurements from multiple satellites.

![Figure 3. Data from ascending passes of the QuikSCAT satellite](http://manati.orbit.nesdis.noaa.gov/quikscat/)

4.2 Event Occlusion
Very often a swath of an orbiting satellite captures partial measurement of the (cyclone) event of interest since one has no control above the satellite trajectory. The worst case scenario is a complete miss. One sees from the top right image that the swath only captures a “small” region of the hurricane in Figure 4. The bottom left image, which is the next swath after 1 orbit (101 minutes), contains a partial hurricane. The bottom right image contains a larger hurricane region but is missing the cyclone eye.

The situations described above make cyclone discovery and tracking especially difficult using only data from a single orbiting QuikSCAT satellite. Again, using measurements from multiple satellites can alleviate such situations if a solution can be found to handle the mining of data with varying spatial and temporal resolutions between different measurement types.

4.3 Varying Temporal and Spatial Resolution
Measurements from different sensor instruments on a satellite have different temporal resolution. This is due to the orbit time (QuikSCAT: 101 minutes; ~14 orbits/day; TRMM: 92.5 minutes, ~16 orbits/day), the swath width of sensor instrument (SeaWinds: 1800 km; TRMM (PR): 247 km, TRMM (TMI): 878 km, TRMM (VIRS): 833 km), and geographical coverage (QuikSCAT: Global; TRMM: 50N-50S). For QuikSCAT, it is usually about twelve hours that measurements are taken again on the same region.

The spatial resolution depends on the sensor instruments (TRMM (TMI): 5.1km, TRMM (PR): 5.0km, TRMM (VIRS): 2.4km), the satellite orbital altitude (TRMM Pre-boost (350km) (TMI): 4.4km to 5.1km (Post-boost (403 km)), and the processing algorithms (operational QuikSCAT data has spatial resolutions of 12.5km and 25km). Different data products from the same sensor satellite can also have different spatial resolution. Level 3 data products usually have lower spatial resolution than Level 1 and 2 data products. For example, TRMM 3B4X where X = 0; 1; 2 and 3 data products have spatial resolution of more than 50km.

![Figure 4. Swaths capturing Hurricane Dean on Aug 16 2007 at 2156 (Top Left), Aug 17 at 0900 (Top Right), Aug 17 at 1041 (Bottom Left) and Aug 17 at 2310 (Bottom Right).](http://sharaku.eorc.jaxa.jp/TYP_DB/index_e.shtml)

4.4 Lack of Annotated Negative Examples
In the study of tropical cyclone, scientists are only interested in tropical cyclones and related weather events. Hence, tropical cyclone images are annotated, organized and well-documented\textsuperscript{13}. One can easily download satellite data that contain relevant cyclone events to analyze and visualize. On the other hand,
interesting non-cyclone events are not organized and documented like the cyclone events. They are hidden in the massive amount of unlabeled data in the archival databases. There are no existing well-defined negative examples.

In this paper, we select random regions in the QuikSCAT images from days where no cyclone is detected as negative examples. One area of future work is to incorporate technology in semi-supervised learning [17] to cyclone detection.

5. HETEROGENEOUS REMOTE SATELLITE-BASED DETECTION AND TRACKING APPROACH

In Section 5.1, we describe the QuikSCAT features used in our ensemble approach for cyclone detection described in Section 5.2. Our cyclone tracking solution based on knowledge sharing between heterogeneous TRMM and QuikSCAT data is described in Section 5.3.

5.1 QuikSCAT Feature Selection

In our automated cyclone identification and tracking approach, features which characterize and identify a cyclone are selected and extracted from the QuikSCAT satellite data. We utilize the QuikSCAT Level 2B data which consists of ocean wind vector information organized by full orbital revolution of the satellite as it is very similar to the NRT wind data. One satellite full polar orbiting revolution takes about 101 minutes. The Level 2B data are grouped by rows of wind vector cells (WVC) which are squares of dimension 25 km or 12.5 km. A complete coverage of the earth circumference requires 1624 WVC rows at 25 km spatial resolution, and 3248 rows at 12.5 km spatial resolution. The 1800 km swath width amounts to 72 25 km WVCs or 144 12.5 km WVCs. Occasionally, the measurements lie outside the swath. Hence, the Level 2B data contains 76 WVCs at 25 km spatial resolution and 152 WVCs at 12.5 km spatial resolution to accommodate such instances.

There are 25 fields in the data structure for the Level 2B data. We are, however, only interested in the latitude, longitude, and the most likely wind speed and direction for the WVCs. The fields that are of interest to us are summarized in the Table 1.

After the Level 2B data is received, it needs to be interpolated on a uniformly gridded surface. This is due to the non-uniformity in the measurements taken by the QuikSCAT satellite on a spherical surface. The nearest neighbor rule is used for this pre-processing procedure for both wind speed and direction.

<table>
<thead>
<tr>
<th>Field</th>
<th>Unit</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>WVC latitude</td>
<td>Deg</td>
<td>-90.00</td>
<td>90.00</td>
</tr>
<tr>
<td>WVC longitude</td>
<td>Deg E</td>
<td>0.00</td>
<td>359.99</td>
</tr>
<tr>
<td>Selected speed</td>
<td>m/s</td>
<td>0.00</td>
<td>50.00</td>
</tr>
<tr>
<td>Selected direction</td>
<td>Deg from North</td>
<td>0.00</td>
<td>359.99</td>
</tr>
</tbody>
</table>

Histograms are constructed to estimate the underlying probability density of the wind speed (WS) and wind direction (WD) within a predefined bounding box extracted from a QuikSCAT image. Let $WS(l_f, j)$ and $WD(l_f, j)$ be the wind speed and wind direction at location $(l_f, j)$. One defines the direction to speed ratio (DSR) at $(l_f, j)$ as:

$$DSR(l_f, j) = \frac{WD(l_f, j)}{WS(l_f, j)}$$

When there is a strong wind with wind circulation, the DSR at a wind vector cell (WVC) will be small. In particular, a histogram constructed to estimate the underlying probability density of DSR in a region will have a skewed distribution towards the smaller value. When there is weak (or no) wind with no circulation, DSR histogram does not have the skewed characteristics. We use a bin size of 4, 30, and 5 for WS, WD, and DSR, respectively according to [14]. One notes that there is a marked difference between a cyclone event and a non-cyclone event in their WS, WD and DSR estimated probability density using histogram. These histogram features are helpful in discriminating between the two events. When a region contains a cyclone, the WS histogram shows a density estimate that skewed towards the larger values. Furthermore, WD histogram shows a “near uniform” distribution.

According to the National Oceanic and Atmospheric Administration (NOAA), a cyclone is defined to be a “warm-core non-frontal synoptic-scale” system, with “organized deep convection and a closed surface wind circulation about a well-defined center”. To discriminate between cyclone and non-cyclone events based on this circulation property, we use two additional features: (i) a measure of relative strength of the dominant wind direction (DOWD) [14], and (ii) the relative wind vorticity (RWV).

Let $u(l_f, j)$ and $v(l_f, j)$ be the u-v components of the wind direction $WD(l_f, j)$ at location $(l_f, j)$ with $1 \leq m \leq 144$ and $1 \leq n \leq 1624$. One constructs a $(m \times n)$-by-2 matrices $M$ of the form

$$M = \begin{bmatrix} u(1,1) & v(1,1) \\ u(1,n) & v(1,n) \\ u(n,1) & v(n,1) \\ u(n,m) & v(n,m) \end{bmatrix}$$

Let $\lambda_1$ and $\lambda_2$ be the eigenvalues of matrix $M$ such that $\lambda_1 < \lambda_2$. The eigenvalue ratio of a bounding box $B$ of dimension $m$ by $n$ is

$$ER_2 = \frac{\lambda_2}{\lambda_1}$$

$ER_2$ is used to quantify the relative strength of the dominant wind direction (DOWD) [14] within the region of interest (box) B. If there is circulation (i.e., a cyclone in B), $ER_2$ will be near to 1. If the wind is unidirectional (i.e., no storm or cyclone in B), $\lambda_2$ will be much greater than $\lambda_1$. As a result, $ER_2$ is much larger.

The relative wind vorticity (RWV) [15] at location $(l_f, j)$ is

$$RWV_{l_f, j} = \frac{\partial v(l_f, j)}{\partial x(l_f, j)} - \frac{\partial u(l_f, j)}{\partial y(l_f, j)}$$

where $u$ and $v$ are the two wind vector components in the west-east and south-north directions, and $d$ is the spatial distance between two adjacent QuikSCAT measurements in a uniformly
ensures a high detection rate for cyclones while maintaining a fine temporal resolution during cyclone tracking. Our automated cyclone tracking using knowledge sharing is shown in Figure 6. Initially, QuikSCAT data is retrieved from the database or from real-time streaming information, and is input into the cyclone discovery module (Figure 5) to locate/identify possible cyclones. The cyclone location is then used to predict the regions that are likely to contain a cyclone at the next incoming data stream retrieved using a linear Kalman filter predictor. If the next data stream is the 3B42 TRMM data, a constrained search is carried out around the region most likely to contain the cyclone as identified by the Kalman Filter predictor. This constrained tracking via the Kalman Filter predictor is especially important for the 3B42 TRMM precipitation data as it is not a definitive indicator of cyclones and is susceptible to high false alarms. The estimated search region localizes the region that is most likely to contain cyclone based on past cyclone tracks and hence the incidence of false alarms is minimized by a large margin. A cyclone is localized by applying a threshold to the TRMM precipitation rate measurement (T6 = 0). After a cyclone is located in the TRMM data, the Kalman filter measurement update (“correction”) is applied to obtain an estimate of the new state vector or the predicted location of the cyclone in the next TRMM (or QuikSCAT) observation cycle after 3 hours.

The system equations used in the Kalman filter are

\[ x_{k+1} = A x_k + w_k \]

\[ z_k = H x_k + v_k \]

where \( x_{k+1} \) is the state vector at time instance \( k+1 \), \( z_k \) is the observation vector at time instance \( k \), \( A \) is the state transition matrix, \( H \) is the observation matrix, \( w_k \) and \( v_k \) are Gaussian noise at time instance \( k \). The matrix form of the above system equations are as follows.

\[
\begin{bmatrix}
X_{k+1} \\
\end{bmatrix} = \begin{bmatrix}
1 & \Delta t \\
0 & 1 \\
\end{bmatrix} \begin{bmatrix}
X_k \\
\end{bmatrix} + \begin{bmatrix}
\Delta x_k \\
\end{bmatrix} + \begin{bmatrix}
w_k \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
Z_k \\
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\end{bmatrix} \begin{bmatrix}
X_k \\
\end{bmatrix} + \begin{bmatrix}
v_k \\
\end{bmatrix}
\]
where is the time difference between the next satellite image and the current satellite image. This is a known parameter between two consecutive TRMM satellite images (3 hours), and between a current QuikSCAT image and the next TRMM satellite image. As mentioned earlier, since the spatial resolution varies for different satellite data, we use the longitude and latitude coordinates as the fixed x-y reference frame for the tracking computation.

A cyclone is a dynamic event and its size evolves rapidly over time. Typical tracking and prediction techniques use the center of an object as the single point to track and predict over time. This model works well for rigid objects that do not change shape with time. However, modeling and predicting the evolution of a cyclone in space over time using only the cyclone center will be grossly inadequate since a cyclone often increases in size as it evolves from a depression to a storm to a hurricane, and then decreases rapidly in size after hitting landfall. We therefore model the cyclone using the maximum and minimum latitude/longitude of the bounding box spanned by the cyclone. Our hypothesis is that the bounding box that is described by the (x,y) spatial span of the cyclone evolves linearly in space over time. We expand (or contract) the estimated bounding box based on the estimated Kalman error covariance to define a search region for the cyclone in the TRMM image. This modeling approach significantly improves the quality of knowledge sharing between heterogeneous satellites as compared to using a predictor/tracker using only the center coordinates of the cyclone.

6. EXPERIMENTAL RESULTS

Experiment results show the competitive performance of our proposed ensemble classifier. We show that the classifier based on the new RWV feature is better than the DOWD feature in terms of classification performance. We also demonstrate the cyclone discovery and tracking system using both Level 2B QuikSCAT data and 3B42 TRMM data.

Our training data consists of 191 QuikSCAT images of tropical cyclones (i.e. tropical depression, tropical storms, and hurricanes) occurring in the North Atlantic Ocean in 2003. We also randomly collected 1833 unlabeled examples from four days in 2003 when no tropical cyclone occurred. These examples, labeled as negative examples, are included into the training set. Our testing set consists of 84 cyclone events in the North Atlantic Ocean in 2006 and 1822 non-cyclone events, collected from four days in the same year when there was no tropical cyclone.

First, we determine the thresholds for the DOWD and RWV features. One notes that from Figure 7 if a threshold value of 1.958 is used, an optimal training accuracy for the DOWD classifier is only 0.59. This corresponds to a true positive rate of 0.7958 and a true negative rate of 0.3840. One can increase this threshold value if a higher true positive rate is desirable, or to decrease the threshold if a higher true negative rate is preferred. For the RWV classifier, one sees that from Figure 8 if a threshold value of 1.51 is used, an optimal training accuracy of 0.85 can be achieved. This corresponds to a true positive rate of 0.8010 and a true negative rate of 0.8910. One can increase this threshold if a higher true negative rate is desirable, or to decrease the threshold if a higher true positive rate is preferred. The ROC curves in Figure 9 show that RWV is a more robust feature than DOWD in discriminating cyclone and non-cyclone events.

The performance of the DOWD classifier [14], the RWV classifier, the Cyclone Discovery Module (CDM), the cyclone identification system (CIS) [14] and the SVM ensemble [14] are compared in Table 2. We also constructed two smaller ensemble classifiers, one with RWV classifier removed from CDM (SVM+DOWD ensemble) and the other with DOWD classifier removed from CDM (SVM+RWV ensemble) to compare the performance among different ensembles.

The SVM ensemble uses identical parameters to the SVM classifiers in the CDM. The CIS parameters are found in [14]. The DOWD and RWV classifiers use the thresholds we determined earlier. The parameters for CDM are set as follows: T1=12m/s, T2=400 pixels, T3=1.510, T4=1.958, and T5 = 2.

From Table 2, one sees that CDS is a significant improvement from the CIS. Moreover, RWV classifier by itself is also a powerful classifier with both a high true positive rate (TPR) and a extremely favorable true negative rate (TNR) compared to the
other classifiers. However, for RWV classifier to achieve the TPR of CDS (by lowering the threshold value), its TNR becomes less than 0.7.

Table 2. Comparison of various classifiers on the testing data (TPR: True Positive Rate; TNR: True Negative Rate)

<table>
<thead>
<tr>
<th></th>
<th>SVM Ensemble</th>
<th>DOWD</th>
<th>CIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TPR</strong></td>
<td>0.8810</td>
<td>0.8452</td>
<td>0.7262</td>
</tr>
<tr>
<td></td>
<td>(74)</td>
<td>(71)</td>
<td>(61)</td>
</tr>
<tr>
<td><strong>TNR</strong></td>
<td>0.7261</td>
<td>0.4232</td>
<td>0.5521</td>
</tr>
<tr>
<td></td>
<td>(1323)</td>
<td>(771)</td>
<td>(1006)</td>
</tr>
</tbody>
</table>

Table 2

From Table 2, one observes that even though DOWD classifier and SVM+DOWD ensemble are weaker classifiers on negative examples, the inclusion of DOWD classifier into the SVM+RWV ensemble classifier to become the CDM improves the TNR by more than 0.07.

7. LESSONS LEARNED

We briefly discuss two salient lessons learned from our experience in designing and implementing the automated cyclone detection and tracking prototype. First, even though we can easily access the publicly available data, it is important to work closely with the data processing team to understand the satellite data, its

Figure 10. Two days tracking of Hurricane Isabel (2003) using the QuikSCAT and the TRMM data (Latitude: 0°N-50°N; Longitude: 30°W-80°W). Predicted cyclone is enclosed by a box.

Figure 10 demonstrates the feasibility of tracking methodology using both Level 2B QuikSCAT data and 3B42 TRMM data for Hurricane Isabel in North Atlantic Ocean in 2003 for two days. We include the detection and tracking sequence for the 2007 Hurricane Gonu (reaching Category 5 wind speed level), the strongest tropical cyclone since record keeping begun in 1945 for the North Indian Ocean and the Arabian Sea, as additional material to support our tracking methodology. Hurricane Gonu is an interesting event as tropical cyclones developed in the Arabian Sea very rarely exceed the tropical storm intensity, i.e. becoming a hurricane.

7. LESSONS LEARNED

We briefly discuss two salient lessons learned from our experience in designing and implementing the automated cyclone detection and tracking prototype. First, even though we can easily access the publicly available data, it is important to work closely with the data processing team to understand the satellite data, its

Figure 10

Figure 10 demonstrates the feasibility of tracking methodology using both Level 2B QuikSCAT data and 3B42 TRMM data for Hurricane Isabel in North Atlantic Ocean in 2003 for two days. We include the detection and tracking sequence for the 2007 Hurricane Gonu (reaching Category 5 wind speed level), the strongest tropical cyclone since record keeping begun in 1945 for the North Indian Ocean and the Arabian Sea, as additional material to support our tracking methodology. Hurricane Gonu is an interesting event as tropical cyclones developed in the Arabian Sea very rarely exceed the tropical storm intensity, i.e. becoming a hurricane.

7. LESSONS LEARNED

We briefly discuss two salient lessons learned from our experience in designing and implementing the automated cyclone detection and tracking prototype. First, even though we can easily access the publicly available data, it is important to work closely with the data processing team to understand the satellite data, its

Figure 10

Figure 10 demonstrates the feasibility of tracking methodology using both Level 2B QuikSCAT data and 3B42 TRMM data for Hurricane Isabel in North Atlantic Ocean in 2003 for two days. We include the detection and tracking sequence for the 2007 Hurricane Gonu (reaching Category 5 wind speed level), the strongest tropical cyclone since record keeping begun in 1945 for the North Indian Ocean and the Arabian Sea, as additional material to support our tracking methodology. Hurricane Gonu is an interesting event as tropical cyclones developed in the Arabian Sea very rarely exceed the tropical storm intensity, i.e. becoming a hurricane.

7. LESSONS LEARNED

We briefly discuss two salient lessons learned from our experience in designing and implementing the automated cyclone detection and tracking prototype. First, even though we can easily access the publicly available data, it is important to work closely with the data processing team to understand the satellite data, its
characteristics, and its limitations. Many different science products are available from each satellite sensor. One needs to know which science product is the most relevant for the problem. For example, a Level 3 QuikSCAT product has too low resolution for good detection performance even though it is gridded. Moreover, we need to understand the features of the data product that could be useful for cyclone identification. Good communication with the data processing team shortens our time in understanding the data and implementing our prototype system.

Second, to build up a system deploying heterogeneous data, one should include the different measurement data into the system one at a time, starting from the measurement data that ensures a reasonably good performance. Then one includes other measurement data incrementally while maintaining or improving the current system performance one at a time.

8. CONCLUSIONS

Knowledge sharing between different satellites is a challenging and as yet unresolved problem, and an efficient solution such as ours that taps the information from such disparate sources will greatly improve science data understanding in various domains in environmental and space science. To date, we have tested our technique on isolated hurricanes. We aim to test our implementation over a longer time scale in a region where there are occurrences of multiple cyclones that will allow us to demonstrate the global tracking capability of our implementation.

Due to page limitations, some issues such as (i) the effect of different preprocessing algorithms for historical and NRT data, and (ii) utilizing gridded and non-gridded satellite data, are not discussed. These (open) issues together with those discussed in Section 4 are important factors that will help us to better understand how one can improve near real-time knowledge mining of satellite data and design new data mining techniques for satellite data.

Autonomous knowledge discovery from massive heterogeneous satellite data is extremely desirable for advance scientific understanding of the global climate, environmental science, space science, and Earth science. Yet, conventional methods cannot handle such massive unlabeled high-dimensional heterogeneous data. These data remain largely unexplored and under-utilized due to the lack of human resources to manually analyze such data using science experts, and inadequate data mining techniques to process these data. This works provide a glimpse into the many research opportunities in mining and data understanding from such publicly available satellite data.

9. ACKNOWLEDGMENTS

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology with funding from the NASA Applied Information Systems Research (AISR) Program. The first author is supported by the NASA Postdoctoral Program (NPP) administered by Oak Ridge Associated Universities (ORAU) through a contract with NASA. The authors thank Andrew Bingham and Eric Rigor for their help in processing the Level 2B QuikSCAT data.

10. REFERENCES