A novel web-based system for tropical cyclone analysis and prediction

Yee Leung, Man Hong Wong, Ka Chun Wong, Wei Zhang, Kwong Sak Leung

Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
Institute of Environment, Energy and Sustainability The Chinese University of Hong Kong, Shatin, Hong Kong, China

Available online: 16 Jan 2012

To cite this article: Yee Leung, Man Hong Wong, Ka Chun Wong, Wei Zhang & Kwong Sak Leung (2012): A novel web-based system for tropical cyclone analysis and prediction, International Journal of Geographical Information Science, 26:1, 75-97

To link to this article: http://dx.doi.org/10.1080/13658816.2011.572075
A novel web-based system for tropical cyclone analysis and prediction

Yee Leung\textsuperscript{a,c,*}, Man Hong Wong\textsuperscript{b}, Ka Chun Wong\textsuperscript{b}, Wei Zhang\textsuperscript{a} and Kwong Sak Leung\textsuperscript{b}

\textsuperscript{a}Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong; \textsuperscript{b}Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong; \textsuperscript{c}Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Shatin, Hong Kong, China

(Received 25 May 2010; final version received 10 March 2011)

A web-based system is developed for the analysis and prediction of tropical cyclones, particularly their landfalls and recurvatures. To facilitate accessibility to the system, its development is based on Google Maps application programming interface (API), Java and client/server architecture. In addition to the construction of a powerful query system for the multi-source, multi-scale and multi-level tropical cyclone database, data mining approach and dynamic modelling approach have been implemented and integrated for effective and efficient analysis, prediction and visualization of tropical cyclone movements. The system can be accessed worldwide by researchers, professionals and the general public. It is thus a powerful system for research, real-life application and knowledge dissemination. Its extensibility and user-friendliness pave the road for further development and enable more in-depth analysis and real-time operation.

Keywords: data mining; landfall; recurvature; tropical cyclones; web-based system

1. Introduction

Tropical cyclone (TC) is one of the most notorious hazards that cost human lives and tremendous economic loss, especially in the coastal areas of the world (Simpson and Riehl 1981, Pielke and Pielke 2000, Chan 2005). It has been reckoned that TCs cause more deaths and inflict heavier losses than any other natural disasters (Murnane and Liu 2004). If the destructive weather associated with a TC could be well predicted, then the benefit to humankind would be enormous (Chan 2005).

Therefore, it is of considerable importance to correctly predict TC tracks (Harr and Elsberry 1991, Aberson and Sampson 2003, Chan 2005) and intensities (Velden and Leslie 1991, Wang and Wu 2004, Wong and Chan 2004) for effective disaster analysis, response, mitigation and management (Simpson and Riehl 1981). Landfalls (e.g. Wu et al. 2003, Lyons 2004, Yang et al. 2008) and recurvatures (e.g. Chan et al. 1980, Krishnamurti et al. 1992, Hodanish and Gray 1993, Cheung 2006) of TCs are particularly critical problems since most damages are caused during TC landfalling and recurvature when affected places are caught unprepared.

Over the years, a number of TC systems have been implemented at various places. They can be grouped into several categories:
Stand-alone systems for professional use only. These systems have limited capability for TC visualization and query. They are designed for professional use at various meteorological agencies and are not made available to the general public. Furthermore, system developments are only based on stand-alone edition workstations. These systems, however, do not provide case-by-case visualization and query of historical TCs and do not have facilities for web-based services. The Canadian Hurricane Centre Forecaster’s Workstation (MacAfee 1997), the Automated Tropical Cyclone Forecasting System (Sampson and Schrader 2000) of the US Department of Defence and the National Weather Service, as well as the Australian Tropical Cyclone Workstation (Woodcock 1995), are typical examples.

Web-based systems for visualization and information dissemination. Contrary to the systems in (1), these TC systems are mainly developed for the dissemination of information or visualization of the historical TC database via the web, but they lack the capabilities for TC analysis and prediction. Examples are ‘Eye of the Storm’, ‘Global Tracks’, ‘Hurricane Watch 2000’, ‘HurrTrak’, ‘Merlin’, ‘Personal Hurricane Center’, ‘Storm’, ‘Tempest Hunter’ and ‘Tracking the Eye’ on the webpage of the National Oceanic and Atmospheric Administration (NOAA). Those at the Joint Typhoon Warning Center, Japan Meteorological Agency (JMA), National Hurricane Center and Hong Kong Observatory have simple function for TC tracking and analysis or display of prediction results.

GIS-based system for simple TC analysis. There are a number of popular GIS software, such as ArcGIS server, ArcIMS, MapExtreme, MapX, which provide useful interfaces for system development. For example, a desktop version of a GIS-based TC system is built for the simple analysis of historical and active TCs (Kong et al. 2008). A GIS-and-remote-sensing-based system is developed to monitor the TC movement (Kumar et al. 1998). The fundamental spatial database and application programming interface (API) are, however, relatively expensive to access for general users.

To recapitulate, there are essentially two types of TC systems: one mainly for access and visualization of TC information and the other mainly for analysis and prediction of TC track and intensity. It should also be noted that very few TC forecasting systems are built by web-based GIS software. Existing web-based GIS systems for TCs are mainly built for information dissemination and map distribution via the Internet, with only a few having limited data analysis capability. As to the accessibility of GIS, by virtue of its technical complexity and cost, GIS has traditionally been made available only to the government and public administration, where it plays a vital role in the majority of its daily operations (Masser et al. 1996).

Therefore, a web-based and free-access GIS system should be built to bring, at little or no cost, GIS technology and TC knowledge to the researchers who would like to perform TC analysis and prediction, and the general public who might just need TC information. Thus, it should have functionalities for the visualization of TCs for the general public, but also complex data analysis and information distribution for research and professional use. A comprehensive TC analysis and prediction system with all the above features as well as powerful data mining (DM) and dynamic modelling capabilities has yet to be developed.

The overall objective of the present study is to develop a powerful web-based system with platform-independent interface for the tracking, prediction and visualization of
TC movements, landfalls and recurvatures via dynamic modelling and DM in assimilated multi-source, multi-scale and multi-level TC database. It is to be realized through the accomplishments of the following interrelated sub-goals using innovative design concepts. First, a web-based platform with user-friendly interface and effective visualization capability is built with advanced GIS technology and scientific speed-up techniques. Second, the system can run on a standard personal computer. Dynamic models based on physical processes in meteorology are utilized to generate data for tracking typhoons, particularly on the dynamic changes of landfalls and recurvatures. Third, novel DM algorithms are developed to discover rules and regularities for TC landfalls and recurvatures. The DM algorithms and the dynamic model will mutually enrich each other in a complementary and integrated manner. Fourth, a multi-source, multi-scale and multi-level remote sensing database consisting of atmospheric and meteorological data such as wind fields in different levels, humidity, sea-level temperature and pressure, geopotential height, rainfall, wind shear and best-track data is constructed for effective and efficient typhoon track analysis and forecasting. Lastly, using state-of-the-art software tools such as Adobe Flash and Sun’s Java in conjunction with Google Maps API, the system is developed for public consumption on the Internet.

In what follows, the overall architecture of this system is first discussed in Section 2. System functionality and implementation are scrutinized in Section 3. Prediction of TCs with DM algorithms and dynamic model is elaborated in Section 4. The article is then concluded with a summary and outlook for further research.

2. Overall architecture of the system

2.1. Overall system architecture

The present system uses Google Maps as a basis for development since it is web-based, flexible and efficient and is open and free in terms of geographic data management and API. The tropical cyclone analysis and prediction system (TCAPS) is constructed on the basis of the client/server architecture. The overall system architecture is depicted in Figure 1.

When clients access the system through the World Wide Web, a request will be submitted to the server for onward submission to the Apache Web Services. The Apache Web Services will then transfer these messages to relevant services or applications. Apache Web Services act as the web gateway to be shown on the World Wide Web.

Specifically, Apache Tomcat Services provide a gateway for different applications to transfer interactions and communications to Java Server Page application or Java Servlet. The Apache Web Services, on the other hand, provide a web gateway for various services or applications. Flash Applications render dynamic graphical user interface for the visualization of geographical and meteorological data and connection to Google Maps Services, Open Action Message Format (OpenAMF) and Geographical Data Plotting Program, as well as display of TC best tracks on Google Maps. The Geographical Data Auto-Insertion Program can be employed to extract the downloaded TC-related meteorological data, transform the geographical data to specified format and insert the meteorological data into the database. The Meteorological Data Plotting Program transforms the meteorological data as images in JPEG formats used as images on superimposed layers of the TC Track Viewer. Google Maps Services connect the background world map data from the Internet to other applications. The Hyper Text Markup Language applications supply static Webpages through Apache Web Services for user navigation and are used mostly for
organization purpose, that is, linking of different applications and Webpages. Meanwhile, Java Servlet/Java Server Page applications offer database query services and Java Web Applications through the Apache Tomcat Services. The OpenAMF manages Java-to-Flash Remote Services and provides linkages for Flash Action Script and Java Class, aiming at retrieving data from the database. The Oracle Database stores geographical data and TC-related data in a relational database. The Result Extraction Program can extract results obtained by the Fifth-Generation NCAR/Penn State Mesoscale Model (MM5).

2.2. General functionality of the TC track viewer

The TC Track Viewer of TCAPS provides the map view, terrain view and satellite view through Google Maps API. The link to the system is http://pc89075.cse.cuhk.edu.hk:8080/myapp2/NewFlashGIS.html.
TC analysis and prediction are hidden by default in order to provide more operational space in the interface for analysis and decrease the loading time of the TC track viewer. However, users can right click the mouse to show the options.

The TC track viewer is developed by Flash Action Script 3.0 and cannot be directly connected to the Oracle database. Therefore, Java programming is used to connect the TC track viewer to the database. Java Servlet, OpenAMF and Java programs form a linkage between the database and the TC track viewer. The viewer is built on Flash and is embedded into the Hyper Text Markup Language. It not only provides a high-level graphical user interface for the visualization of TCs and their related attributes, but also contains different kinds of multi-criteria spatial queries in terms of TC track, recurvature and landfall.

Basic functionalities of the TC Track Viewer are summarized as follows:

(1) Animation of TC tracks
This tool makes a TC recur in 6-hour increments. The animation starts at a predefined time and from there users can play, pause, go backward or forward to view the TC tracks. The viewer can be set for any year starting from 1951.

(2) Layered display of features
To facilitate visualization, atmospheric features can be shown in layers. Sea-level pressure (SLP), sea-surface temperature (SST), cloud formation, humidity, wind field, wind shear on land surface and high altitude can be individually displayed as image layers or can be superimposed onto one another. It runs in parallel with the animation of TC tracks to give the atmospheric and oceanic conditions throughout the TC movement.

(3) Data analysis
The viewer provides facilities for users to incorporate different attributes into the analysis, and results can be displayed on the line chart which is movable and with changeable opacity. It efficiently depicts the correlation between selected attributes.

(4) Information display via cursor
TC information of each point along a track can be retrieved and displayed by clicking the mouse cursor at the point. It can also show its atmospheric and oceanic background information. Such information can be hidden by moving the cursor away from it.

(5) Data search
For easy manipulation, the viewer can show or hide a wide range of data-searching panels such as TC classification. Users can set the visibility of these panels.

(6) Result Selector Panel and Result List
Results obtained from the query are shown on the Result Selector Panel and user can select ‘proceed’ or ‘show details’ on the panel. After clicking ‘show details’, the Result List will be shown automatically, giving brief descriptions of the results. User can further select the results on the Result List.

(7) TC prediction
DM algorithms are put in store for TC track prediction via finite mixture model (FMM), spatial similarity and temporal similarity and dynamic analysis by MM5.

(8) TC genesis analysis
Since TCs originated at different locations in the Pacific Ocean and the South China Sea will have different characteristics in terms of track, intensity and landfall, analysis of the initial condition, called genesis analysis, at the point where TC gains its intensity to become tropical storm is important for the prediction of its
subsequent development. This analysis focuses on the number of TCs formed in a specific time interval at a latitude–longitude grid cell. Each latitude–longitude grid cell is coloured according to the frequency of TC occurrences at that location over the years.

3. System functionality and implementation

3.1. Building the TC database

In TCAPS, the TC archives are organized in terms of TC tracks, meteorological variables and satellite cloud images. TC tracks display TC points constituting the polyline in 6-hour intervals. The track data are downloaded from the JMA. Meteorological variables (e.g. wind field, wind shear, SLP, SST, rainfall and humidity) are important factors affecting TC track, intensity, genesis, landfall and TC-triggered disasters (e.g. flood, land slide, storm surge). Therefore, these variables are collected from authoritative meteorological agencies and can be superimposed in TCAPS. Satellite cloud image can not only display the TC structure but also indicate TC intensity and motion. In the present system, the study area mainly covers the western North Pacific and the South China Sea, that is, 0°–45°N in latitude and 100°E–180°E in longitude. Therefore, the meteorological variables are extracted within the specified study area. A description of the TC database is given in Table 1.

The comprehensive TC database in TCAPS integrates TC best-track database and other meteorological variables (such as SST, SLP, humidity, wind field in different atmospheric levels) acquired from NOAA, National Aeronautics and Space Administration (NASA), National Centres for Environmental Prediction/National Centre for Atmospheric Research (NCEP/NCAR) and JMA. A tool kit is constructed to facilitate data retrieval.

The TC track data are derived from the JMA best track in 6-hour intervals. It is organized by TC number, intensity, date, central pressure, latitude and longitude. The SST data and images are downloaded from the remote sensing systems under the sponsorship of NOAA and NASA. The data have a spatial resolution of 0.25° by 0.25° with a daily interval for the years starting from 1951. SLP data and images have a 6-hour interval. The wind at sea level of 10 m, wind shear and humidity are collected from NCEP/NCAR

Table 1. Description of the TC database.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
<th>Time resolution</th>
<th>Spatial resolution (latitude × longitude)</th>
<th>Image range a</th>
<th>Raw data range b</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC track</td>
<td>JMA</td>
<td>Every 6 hours</td>
<td>N/A</td>
<td>1951–2009</td>
<td>1951–2009</td>
</tr>
<tr>
<td>Wind at sea level</td>
<td>NCEP/NCAR</td>
<td>Every 6 hours</td>
<td>2.5° × 2.5°</td>
<td>2003–2007</td>
<td>2003–2007</td>
</tr>
<tr>
<td>level of 10 m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind shear</td>
<td>NCEP/NCAR</td>
<td>Every 6 hours</td>
<td>2.5° × 2.5°</td>
<td>2000–2007</td>
<td>2003–2007</td>
</tr>
<tr>
<td>SLP</td>
<td>NOAA</td>
<td>Every 6 hours</td>
<td>2° × 2°</td>
<td>2003–2007</td>
<td>None</td>
</tr>
<tr>
<td>SST</td>
<td>NASA/NOAA</td>
<td>Daily</td>
<td>0.1° × 0.1°</td>
<td>2003–2007</td>
<td>None</td>
</tr>
<tr>
<td>Rainfall</td>
<td>NASA</td>
<td>Every 3 hours</td>
<td>0.25° × 0.25°</td>
<td>2006–2007</td>
<td>2006–2007</td>
</tr>
<tr>
<td>Satellite cloud images</td>
<td>NASA</td>
<td>Every 3 hours</td>
<td>1/6° × 1/6°</td>
<td>About 10 days</td>
<td>N/A</td>
</tr>
<tr>
<td>Humidity</td>
<td>NCEP/NCAR</td>
<td>Every 6 hours</td>
<td>2.5° × 2.5°</td>
<td>2003–2008</td>
<td>2003–2008</td>
</tr>
</tbody>
</table>

a Image range stands for the temporal data range for the image layers superimposed in the system.

b Raw data range stands for the temporal data range for the raw data stored in the background database.
reanalysis (Kalnay et al. 1996). SLP is downloaded from the website of NOAA with a 2° × 2° rectangular grid resolution in latitude and longitude. Satellite cloud images in a 3-hour interval are obtained from NASA. Due to the large storage requirement, about 10 days of satellite cloud images are temporarily collected for preliminary analysis.

The SQL Loader, Ferret and Grid Analysis and Display System are employed to transform the original raw data into the format suitable for TCAPS operations.

3.2. System requirements

The construction of the TCAPS platform involves Windows XP workstation and Linux Ubuntu workstation. The Windows XP workstation is set up as web server, while the Linux Ubuntu workstation is set up to run the MM5 for TC prediction (Table 2).

Besides the software listed in Table 2, we also use other state-of-the-art software, for example, Adobe Flash, SQL Loader, gfortran, OpenAMF, Grid Analysis and Display System, Statistical Product and Service Solutions (SPSS) and Ferret. The system architecture and functionalities are detailed in the discussion to follow. We first discuss the query system in Section 3.3. TC prediction is then examined in Section 4.

3.3. Spatial query

TC query plays a key role in TC analysis in TCAPS. The purpose of TC query here is to retrieve TC tracks from the TC database with respect to the multiple spatial or non-spatial criteria, to mine and visualize useful information hidden in the historical TC database by powerful algorithms and to provide useful reference for TC prediction in terms of genesis, track, intensity and landfall. The spatial query system is depicted in Figure 2. The TC query system efficiently finds the required TC tracks with respect to multiple criteria (e.g. spatial, temporal, GIS-based) via the effective query outlined in Figure 3.

Among these query methods, Query by similar path, Query by key area and Query by landfalling locations find TC tracks passing through a buffer zone with certain radius or through one or two arbitrarily specified areas (i.e. rectangle, circles and arbitrary polygon) and coastal areas (e.g. Guangdong Province, Hong Kong). Query by genesis obtains useful information on the frequency of occurrence of historical TCs in a certain region and unravels the TC hot spots in western North Pacific and the South China Sea. Query by direction and Query by turning angle search TCs moving in a certain direction and turning in a certain angle, respectively.

During each query, users can specify location, intensity (e.g. tropical depression, tropical storm), duration (e.g. 10 days, 12 days), period (e.g. from 1951 to 2007) and month of the TC. Therefore, each query is multi-criteria in space and time. Various queries are further explained in the discussion to follow.

Table 2. The operating systems and functions of the TCAPS platform.

<table>
<thead>
<tr>
<th>Operation system</th>
<th>Windows XP</th>
<th>Linux (Ubuntu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installed application(s)</td>
<td>Apache HTTP Server 2.0.48</td>
<td>MM5 prediction model</td>
</tr>
<tr>
<td></td>
<td>Tomcat 6.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oracle 11g Release 2</td>
<td></td>
</tr>
<tr>
<td>Functions</td>
<td>Storing TC data</td>
<td>TC track prediction</td>
</tr>
<tr>
<td></td>
<td>Hosting the system (GIS)</td>
<td></td>
</tr>
</tbody>
</table>
3.3.1. Similar path query

Similar path query is a kind of case-based reasoning (Kolodner 1993, Aamodt and Plaza 1994, Kurbalija et al. 2009). It aims at the retrieval of all TC tracks that passed through a specified buffer zone. The buffer zone treats the selected TC track segment as the buffer centre. This algorithm is divided into two sub-functions, namely ‘Locating TC tracks passing through a rectangular region’ and ‘Finding tracks with similar path’.
3.3.1.1. Locating TC tracks passing through a rectangular region. The query aims at the retrieval of all TC tracks that pass through an area specified by the user. As shown in Figure 4, the semi-transparent blue rectangle represents the region of interest on the map. By dragging the two purple points freely at the corners of the rectangle, the user can move or resize the region on the map. The query algorithm then finds and displays all TC tracks passing through the rectangular region on the viewer.

3.3.1.2. Finding tracks with similar path. The purpose of this query is to retrieve all tracks going through a path specified by the user. Users can draw points by left-clicking the mouse on the screen bounded by the semi-transparent rectangle and then draw the subsequent points to form the shape of the path. A TC track is regarded as having a similar path if it falls within the specified path or if its starting and ending points locate at the two ends of the specified path.

The similar path algorithm can be modified to predict the future path of an active TC track by using the subsequent paths of the tracks to which it is similar. The integration of this query with seasonal similarity and clustering results for TC movement prediction is discussed in Section 4.1.

3.3.2. TC genesis query

Genesis of a TC is the point at which the TC first reaches the intensity level of tropical storm. TCs occurring at different locations might possess different characteristics in terms of shape, intensity, landfall and recurvature in their subsequent developments (Harr and Elsberry 1995). Thus, this query might be instrumental in the prediction of TCs based on their initial conditions.
TC genesis deals with the counting of TCs that started within a specified latitude–longitude range, duration and time (month) range. The records of TCs are retrieved from the database. The starting latitude and longitude of a TC is defined as the earliest record of the TC. Since the records can be ordered by latitude and longitude according to the query, no sorting of the records is required. The counting result will be passed to a function for visualization in colour. Grids with different number of occurrences are displayed in different colours. TC tracks that start from individual grids can also be displayed (Figure 5).

3.3.3. Query by key areas
This function allows users to retrieve TCs passing through one or two areas of interest. Single area query enables users to analyse the degree to which the area is subjected to TC attacks. By specifying another area, the system can reveal all TCs that have passed through the previously specified area and the current area. The query can be done with respect to time and other options.

3.3.4. Query by direction
Each TC moves along certain direction throughout its life. Directions of TCs in the western North Pacific and the South China Sea exhibit different characteristics in terms of landfall, recurvature and intensity. This algorithm helps to find the movement directions of TCs.

Users can query the database in eight directions: east, north-east, north, north-west, west, south-west, south and south-east and can select as a basis of query the threshold, percentage of the TC path that lies in the selected direction.

3.3.5. Query by turning angle
Affected by various factors, TCs might change course in their movements.
Sudden turning of a TC always causes considerable damages to areas lying on its path if they are caught unprepared. The query is to retrieve TC tracks that change courses at certain angles and at certain points in time.

To retrieve relevant TC tracks, users can specify turning angles, number of turns and minimum distance between two adjacent segments at the turning point. The turning angle is determined by using the vector dot product $\vec{a} \cdot \vec{b} = |\vec{a}| |\vec{b}| \cos \theta \implies \theta = \cos^{-1}\left(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}\right)$ to find the angle between two sequential segments (Figure 6). The user can choose a threshold value to see whether the turning angle of a TC track is larger than the threshold angle.

Figure 7 depicts TC tracks retrieved for the period from 1 January 1951 to 1 January 1965 with turning angles $\geq 70^\circ$, one single turn and minimum distance 0.6° in latitude or longitude.
Table 3. The intensity level of TC by the Chinese Meteorological Administration.

<table>
<thead>
<tr>
<th>Grades of tropical cyclone</th>
<th>Strength (knots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super typhoon</td>
<td>$\geq 100$</td>
</tr>
<tr>
<td>Severe typhoon</td>
<td>81–100</td>
</tr>
<tr>
<td>Typhoon</td>
<td>64–81</td>
</tr>
<tr>
<td>Severe tropical storm</td>
<td>48–61</td>
</tr>
<tr>
<td>Tropical storm</td>
<td>34–48</td>
</tr>
<tr>
<td>Tropical depression</td>
<td>$\leq 33$</td>
</tr>
</tbody>
</table>

3.3.6. **Query by landfalling location**

TC landfalling (Ching et al. 2000, Lyons 2004) always inflicts severe damages to coastal areas. TCs that make landfall in different coastal areas may be similar in track shape and trend. Query by landfalling location can help to analyse all TCs that landed in certain coastal area in selected months and periods.

3.3.7. **Additional query options**

When users are performing any of the above queries, they could set additional options for TC retrieval. The additional options are about period, location, length and duration. Users can specify the query period in terms of year, month and day; and/or location in terms of coastal area, region or a city; and/or strength according to, for example, Table 3. Query results are tabulated in the result list showing TC IDs, names and start dates. Selected TCs are then plotted onto the Google Maps for visualization.

4. **Prediction of tropical cyclones**

4.1. **Mining of historical tracks for TC prediction**

Among other things, prediction of TC track is a main area of research in TC analysis and is of public concern. To facilitate analysts to make prediction, the present web-based system renders the DM approach and the dynamic model approach to TC prediction. It also offers the integrated use of both approaches to enhance prediction.

The DM approach, at present, offers TC-pattern clustering by the FMM and TC recurvature prediction by the classification and regression tree (CART). The dynamic model, MM5, on the other hand, is incorporated into the current system for TC track prediction.

The movement of a current TC may be predicted by the movements of the past TC tracks exhibiting similar characteristics. Thus, DM methods can be employed to unravel the characteristics of relevant TC patterns in the past and to shed light on the prediction of the current TC. TC track clustering through FMM, for example, has been performed over the years (Gaffney and Smyth 1999, 2007, Camargo et al. 2004, 2007a, b, 2008, Gaffney 2004). Camargo et al. (2004, 2007a, b, 2008) discovered 7 clusters in historical TC track archive of western North Pacific, while Gaffney (2004) found 10 clusters in similar TC track database.

Integrating spatial similarity, temporal similarity and the clustering procedure of Camargo et al. (2008), we constructed a TC track prediction scheme, which is similar to the case-based prediction scheme and climatology and persistence (Neumann 1972) scheme.

In the present system, the FMM-based clustering model is implemented within the Matlab platform (Gaffney and Smyth 1999, Camargo et al. 2004, 2007a, 2008, Gaffney
The parameters \( \beta, \sigma^2 \) and \( \nu^2 \) in Equation (1) (Gaffney 2004), characterizing the shape of each cluster can be obtained as follows:

\[
p_k(y_i) = \int p_k(y_i, t_i) dt_i = \int p_k(y_i|t_i) p_k(t_i) dt_i = \int N(y_i|X_i\beta_k + t_i, \sigma_k^2 I) N(t_i|0, \nu_k^2) dt_i = N(y_i|X_i\beta_k, \nu_k^2 I + \sigma_k^2 I)
\]

where \( y_i \) is the \( i \)-th curve of length \( n_i \) (it denotes the \( i \)-th TC track), \( p_k(y_i) \) is the probability of track \( y_i \) belonging to the \( k \)-th cluster, \( t_i \) is a curve-specific translation scalar for the \( i \)-th track, \( I \) is the identity matrix, \( \beta_k \) is the \( p \times 1 \) mean coefficient vector of the \( k \)-th cluster and \( \sigma_k^2 \) and \( \nu_k^2 \) are variances of \( y_i \) and \( t_i \) in the Gaussian model for the \( k \)-th cluster, respectively. Here, according to Gaffney (2004),

\[
\hat{\nu}_k^2 = \frac{1}{n} \sum_i g(\hat{t}_{ik}),
\]

\[
\hat{\sigma}_k^2 = \frac{1}{N} \sum_i \hat{f}(\hat{t}_{ik}),
\]

\[
\hat{\beta}_k = \left[ \sum_i X_i' X_i \right]^{-1} \sum_i X_i' (y_i - \hat{t}_{ik})
\]

where \( \hat{\nu}_k^2, \hat{\sigma}_k^2, \hat{\beta}_k \) are estimates of \( \nu_k^2, \sigma_k^2 \) and \( \beta_k \); \( n \) and \( N \) are the total number of sample TC tracks; \( \hat{t}_{ik} \) is the estimation of \( t_{ik} \); \( t_{ik} \) is the curve-specific translation scalar of the \( i \)-th track for the \( k \)-th cluster; \( X_i \) is the Vandermonde matrix evaluated at the \( i \)-th TC track; and \( X_i' \) is the transpose of \( X_i \).

Using the longitude and latitude of each point of a new TC track as input of the clustering model, we can calculate the probability of the current active track belonging to each TC class using Equation (1). Essentially, the clustering procedure employs the TC position orders (e.g. 1–3) as a basis in the polynomial regression. The shapes of the TC tracks (curves) are depicted by the parameters of the polynomial regression model. Different shapes are integrated into the FMM according to the distribution of the error terms. Then, the current TC will be assigned to the class with the largest posterior probability. The characteristics of the TC class and the historical TC tracks belonging to this class can be used as a reference for forecasting. Once an active TC track has been assigned to a cluster, all the historical tracks of that cluster will be retrieved and employed to predict the shape and movement of an active track.

Besides the clustering results, spatial similarity or case-based reasoning via the ‘Query by similar path’ can also be employed to find useful historical tracks for prediction. In general, TCs occurring in a certain season have special or similar characteristics according to the position and strength of subtropical high and monsoon trough (Lehodey et al. 1997, Lu and Dong 2001, Ho et al. 2004). Thus, historical TC tracks that occurred in the same season as the current active track can be used to predict the movement of the latter. This is called ‘temporal similarity’ or ‘seasonal similarity’. Thus, TC tracks unravelled satisfy both the clustering results and the spatial and temporal similarity query results. The tracks
obtained from these three procedures will then be cross-compared and integrated, and the subsequent points of movement of the integrated track can be employed to make more reliable prediction of the current track.

Figure 8 shows the flow of the above prediction procedure, and Figure 9 depicts a reasonably accurate experimental result of such prediction. The blue buffer is the buffer zone of the selected TC track segment. The yellow line with blue dots is the predicted 72-hour track. The other is the real TC track.

4.2. Prediction scheme based on MM5

In place of making prediction through the method of clustering, the direct modelling of TC dynamics has been attempted over the years. Among others, MM5 is one of the most widely used numerical weather-prediction systems (e.g. Deng et al. 2004) developed partly for such purpose. Its construction is based on the physical processes and mechanisms plausibly governing TC development and dynamics and it has been employed to predict TC movements (Juneng et al. 2007, Mandal et al. 2007, Ramsay and Leslie 2008). Due to high computational cost and complexity, we have implemented MM5 into the 32 × 4 parallel computation environment, and it is run using a quadcore.

There are many atmospheric datasets available on the Internet, and these datasets have different characteristics. The two major datasets used by MM5 in the present system are ‘FNL’ and ‘GFS data’, both of which are published by NCEP/NCAR. The weblink to ‘FNL’ dataset is http://dss.ucar.edu/datasets/ds083.2/, whereas the link to ‘GFS’ dataset

![Diagram of the flow of the scheme for TC track prediction.](image-url)
Figure 9. TC track predicted by FMM. (The blue buffer is the buffer zone of the selected TC track segment. The yellow line with blue dots is the predicted 72-hour track, and the other is the real TC track.)

is http://www.nco.ncep.noaa.gov/pmb/products/gfs/. ‘FNL’ is the abbreviation for ‘Final analyses’. This dataset contains historical data in high resolution. By using this dataset, we can use MM5 to perform past TC simulations. ‘GFS’ is the abbreviation for ‘Global Forecast System’. The ‘GFS’ database contains high-resolution data for future prediction. The dataset can be used by MM5 to perform TC prediction.

Figure 10 shows the experimental result of a MM5 prediction of the sudden recurving of a TC. The TC in this case is ‘CHANCHU’ whose lifespan is from 9 May 2006, 20:00, to 19 May 2006, 17:00, and its serial number is ‘200601’. The 12 red dots in Figure 10 are the points in time of the predicted track of the active TC after 20:00 on 14 May. For comparison, the real TC track (in rainbow colours) is also shown in Figure 10. It can be observed that MM5 successfully predicted the sudden recurvature of this TC.

4.3. TC recurvature analysis through CART

Different from dynamic modelling, the DM approach aims at the discovery of the mechanisms/rules characterizing the conditions under which TC might recurve. Instead of using process models, the DM approach tries to unravel from TC data rules governing the recurvature of TCs throughout their movements. DM, in general, is the process of extracting hidden and useful patterns and information from data (Fayyad and Stolorz 1997, Han and Kamber 2006, Leung 2010). In recent years, a number of DM algorithms have been employed to unravel TC tracks and intensities in TC database (e.g. Harr and Elsberry 1991, 1995, Gaffney and Smyth 1999, Camargo et al. 2004, 2007a, 2008, Gaffney 2004, Gaffney et al. 2007, Lee et al. 2007, Cheng et al. 2008). However, only a few DM-derived results have been successfully applied to predict TC movement. The DM approach and dynamic
modelling can actually complement each other to achieve more reliable and comprehensive prediction of TC movements. DM can uncover useful information for the enrichment of the dynamic models, and the mechanisms underpinning the dynamic models can be used to guide the DM process.

For TCs in the northern hemisphere, recurvature is the change from a north-westward to a north-eastward track moving into the mid-latitude westerly with prevailing west wind (JTWC 1988, Dobos and Elsberry 1993). TC recurvature in the Northern Hemisphere is sensitive to the change speed of the zonal and meridional wind in the 200 and 500 mb layers within the octant 1, 2, 3 (north and west) of the 6–8° radial belt (Hodanish and Gray 1993). Chan et al. (1980) demonstrated that zonal and meridional wind (u and v wind) in the 200, 300, 400, 500, 700, 800, 850, 900 and 1000 mb layers (especially the 200, 500 and
700 mb layers) in the 5–7° radius circle play an important role in deciding the recurvarture. Mean values of geopotential height (Leftwich 1979, Lage 1982) on 500 mb layer for 5–7° radius circle are important factors in TC recurvature. TC intensity can also affect TC recurvature. However, the physical mechanism leading to TC recurvatures still needs to be exactly determined (Bao and Sadler 1983, Evans and McKinley 1998, Knaff 2009).

The present system provides a DM procedure to discover conditions under which TCs may change their directions of movement. The idea is to unravel from data mechanisms affecting TC recurvatures. The ultimate purpose is to integrate the dynamic modelling and the DM approaches to achieve more in-depth and accurate prediction. The data processing methods of Hodanish and Gray (1993) are chosen to calculate the attributes relevant to TC recurvatures. They are, for example, zonal and meridional wind in 200, 500 and 700 mb layers, the geopotential height in 200 and 500 mb layers, in the octant 1, 2, 3 (north and west) for 6–8° radial belt and maximum sustained wind.

CART is chosen as a method for DM because of its simplicity and interpretability (Breiman et al. 1984, Leung 2010). The purpose is to unravel rules for TC recurvature. The sample data of the initial experiment were taken from 2000 to 2009. The potential factors that influence TC recurvature are categorized into three groups: variables relating to large-scale circulation, variables measuring the circulation surrounding TC and variables describing the characteristics of TC (see Table 4).

In Table 4, the variables are displayed in abbreviations. In the group ‘Circulation surrounding TC’, ‘uwnd_200’ and ‘vwnd_200’ are, respectively, the average zonal and meridional winds of 6–8° radial belt at 200 hPa level. The other variables in the same group are defined likewise. Five variables chosen to measure the strength and position of ‘large-scale circulation’ (i.e. subtropic high (STH), East Asian summer monsoon (EASM), westerlies) that largely control TC recurvature (e.g. Hodanish and Gray 1993, Harr and Elsberry 1995) are area index (area_IndexSTH), intensity index (inten_IndexSTH) and westward extension index of STH (west_extSTH) in WNP (for measuring the strength of STH and position of STH ridge), EASM index by Wang and Fan (1999) (Monsoon_WF) (demonstrated by Wang et al. (2008) to be nearly identical to the leading principal component of the EASM and greatly facilitates real-time monitoring), and westerly index (W_Westerly) proposed by Rossby (1939). For the characterization of TCs, we use the

<table>
<thead>
<tr>
<th>Groups</th>
<th>Potential variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulation surrounding TC</td>
<td>uwnd_200, uwnd_300, uwnd_400, uwnd_500,</td>
</tr>
<tr>
<td></td>
<td>uwnd_600, uwnd_700, uwnd_800, uwnd_850,</td>
</tr>
<tr>
<td></td>
<td>uwnd_1000, vwnd_200, vwnd_300,</td>
</tr>
<tr>
<td></td>
<td>vwnd_400, vwnd_500, vwnd_600, vwnd_700,</td>
</tr>
<tr>
<td></td>
<td>vwnd_800, vwnd_850, vwnd_1000</td>
</tr>
<tr>
<td>Large-scale circulation</td>
<td>area_IndexSTH, inten_IndexSTH</td>
</tr>
<tr>
<td></td>
<td>west_extSTH, Monsoon_WF</td>
</tr>
<tr>
<td></td>
<td>W_Westerly</td>
</tr>
<tr>
<td>Characteristics of TC</td>
<td>Lon, Lat, Pressure (central pressure of TC centre)</td>
</tr>
</tbody>
</table>

Note: The area_IndexSTH, inten_IndexSTH and west_extSTH are the area index, intensity index and westward extension index of subtropical high, respectively. Monsoon_WF is the monsoon index proposed by Wang and Fan (1999), and W_Westerly is the westerly index.
longitude (Lon) and latitude (Lat) of TC centre, as well as central pressure of TC (Pressure) in the CART mining process.

From the result, we can see that Lon, Lat, Pressure, uwnd_200, uwnd_1000, vwnd_800, vwnd_850, vwnd_1000, area_IndexSTH, west_extSTH and Monsoon_WF are chosen by the CART algorithm to build the decision tree (Figure 11) stipulating 18 unravelled rules governing TC recurvature. We can observe that the average accuracy of TC recurvature prediction by CART is 84.364%. In the figure, ‘1’ means recurvature and ‘0’ means non-recurvature. The rectangles are leaf nodes, while ellipses or circles are parent nodes. Taking leaf node ‘0(263.0/12.0)’ as an example, ‘0’ in front of the bracket means non-recurvature and ‘263.0’ and ‘12.0’ indicate that among the 275 samples of the leaf node, there are 263 non-recurvature samples and 12 recurvature samples. A path from the root node to the leaf node represents a rule, which can provide reference for TC recurvature prediction. Each rule can be justified by meteorological and TC theories. Taking the rule formed by the path from root node to leaf node ‘1(617.0/26.0)’ as an example, it can be stated as follows:

‘If the longitude of a TC centre is to the east of 130°E, the central pressure is less than 1006 hPa, the number of grids within the 5880 gpm contour of 500 hPa geopotential height is less than 314, the zonal wind at 200 hPa is mainly from west to east and the western ridge of STH is to the east of 133°E, then the TC will recurve’. According to this rule, TCs will recurve under the conditions of weak and retreating STH and moderate westerly. Geopotential height and deep-layer mean wind vector calculated by the TC samples belonging to this rule are all in line with the analysis (Figure 12). The contour 5880 of Figure 12 is the region with nearly the highest geopotential height, indicating the status of STH. The deep-layer mean wind layer is, in general, the pressure-weighted mean wind from 850 to 200 hPa. Therefore, the deep-layer mean indicates the steering current that determines the TC movement to a large extent. It has been reckoned that deeper layer means, for example, 850–200 hPa, are best for forecasting (Holland 1993).
Figure 12. The composition of wind fields and geopotential height of the samples in the CART node (0(263.0/12.0)). (The wind arrows are the composite wind fields. Typhoon symbol at coordinate (0,0) marks the TC centre. The 5880 contour in 500 hPa layer indicates the subtropical high.)

The above CART result is only for the purpose of pedagogy. A more comprehensive study will be carried out to discover rules governing TC landfalls and recurvatures within the complex monsoonal system.

To recapitulate, the present system provides facilities for the prediction of TC movements by the DM and dynamic modelling approaches. The initial implementations of FMM and CART for DM, and MM5 for dynamic modelling have shown the effectiveness of both approaches. The next step is to perform more in-depth investigation of these methods and to take advantages of both approaches to achieve more accurate prediction through their effective integration.

5. Conclusion

In this study, novel methods and advanced information technologies have been employed to build a TC analysis and prediction system suitable for academic research, practical application, and public education. For technical research, the system provides facilities for process models and DM algorithms. Although they are made transparent to users, a more flexible interface can be built in the future for models management and algorithm implementation to facilitate researchers to perform TC analysis and prediction. For the general public, the system provides user-friendly interface for TC query and visualization. It is thus a web-based system with high academic and educational value and has great potential for technology transfer.
The significance of this system lies in its efficient and effective visualization of historical TC archives and the retrieval of data for analysis and forecasting in the web-based environment. Its user-friendly interface and efficient query system greatly facilitate public consumption and professional use. Its facilities for the employment of dynamic models and DM algorithms make it a powerful system for TC analysis and prediction, particularly on TC landfalls and recurvatures. Furthermore, the TC database is extensible for future applications.

The present system is a prototype that can be extended into a full-fledged system in which TC dynamics can be studied within the monsoonal system affecting the development of TCs (Harr and Elsberry 1995, Liu and Chan 2003, Elsberry 2004, Harr and Chan 2005). The DM function of the system, in particular, can be extended to unravel rules stipulating the interplay between TC and Madden–Julian oscillation (Madden and Julian 1971, 1972, 1994), quasi-biennial oscillation (Ebdon 1960, Reed et al. 1961) and El Niño-southern oscillation including El Niño and La Niña occurrence (Chan 1985, 2000, Wang et al. 2003). To increase its function and applicability, it can be further developed into a real-time system with monitoring and mitigation capabilities.

Acknowledgements
The project was supported by the research grant (6106c-SF) of the Chinese University of Hong Kong. The authors thank Chan Chun Sing, Chan Wing Kwan, Chu Man Hin, Ho Ho Fai, Hung Yuet Mui, Li Wing Yin and Mak Seng Hin for their participation and programming in various development phases of the system. The authors thank the editor and the anonymous reviewers for their valuable comments and suggestions.

References


