Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models

Technical Report UCI-ICS 06-02
Bren School of Information and Computer Sciences
University of California, Irvine

Scott J. Gaffney[1], Andrew W. Robertson[2], Padhraic Smyth[1],
Suzana J. Camargo[2] and Michael Ghil[3,4]

[1]Department of Computer Science,
University of California, Irvine, CA, USA
[2]International Research Institute for Climate Prediction,
The Earth Institute at Columbia University, Palisades, NY, USA
[3]Department of Atmospheric and Oceanic Sciences and IGPP,
University of California, Los Angeles, CA
[4]Additional affiliation: Département Terre-Atmosphère-Océan
and Laboratoire de Météorologie Dynamique de CNRS/IPSL,
Ecole Normale Supérieure, Paris, France.

January 24, 2006

Abstract

A probabilistic clustering technique is developed for classification of wintertime extratropical cyclone (ETC) tracks over the North Atlantic. A regression mixture model is used to describe the longitude-time and latitude-time propagation of the ETCs. Tracks are obtained from a simple tracking algorithm applied to 6-hourly mean sea-level pressure fields from either a general circulation model (GCM) or an observed data set. Three clusters of ETC behavior are identified in both cases; they are characterized by predominantly south-to-north (S–N), west-to-east (W–E), and southwest-to-northeast (SW–NE) tracking cyclones. Quadratic curves are found to provide the best description of the data.

The results are broadly similar for the GCM and observed data. In the case of the latter, anomaly composites of the ambient 700-hPa geopotential height field resemble patterns of the positive and negative phases of the North Atlantic Oscillation for the SW–NE and W–E clusters respectively, while the S–N cluster is accompanied by a more transient geopotential trough over the western North Atlantic.

1 Introduction

1.1 Background and motivation

Wintertime extratropical cyclones (ETCs) are responsible for severe-weather events with high winds and/or flooding over North America and western Europe; they caused the second largest insurance loss due to weather (after hurricanes) during the period 1990–98 (Saunders 1999). On the other hand they are also the primary source of wintertime precipitation and total water resources for much of the western United States.

ETCs play a special role as intermediaries between large-scale climate dynamics and local impacts: they are crucial dynamical ingredients of the atmospheric circulation, while at the same time directly impacting local weather. ETCs constitute an important nexus between the potentially predictable large-scale components of climate, such as certain hemispheric or sectorial atmospheric teleconnection patterns associated with internal climate variability or with global warming, on the one hand, and societally important weather events, on the other. A better understanding of the behavior of ETCs in the context of climate variability and change could have important societal implications.

ETCs have localized coherent spatial structures that generally propagate toward the east and go through a well-defined lifecycle (Simmons and Hoskins 1978). Their population is thus most naturally described as a set of moving objects that follow various tracks and have differing individual lifecycle characteristics; this corresponds to a Lagrangian description in fluid dynamical-terminology. By contrast, most data analysis in the atmospheric sciences has been based on calculating Eulerian statistics on spatially fixed grids, often using principal component analysis (Preisendorfer 1988; von Storch and Zwiers 1999) to derive the leading patterns of spatio-temporal variability. These methods are poorly suited to cyclone trajectories.

The analysis of large sets of ETC trajectories from multi-decadal observed data sets and potentially much longer general circulation model (GCM) simulations requires a different approach. Cluster analysis provides a natural way to analyze sets of trajectories and their relationships with the larger-scale atmospheric circulation. In this paper, we use curve-based mixture modeling techniques to perform probabilistic clustering of ETC trajectories in latitude-longitude space. An identification and tracking methodology is developed to produce cyclone trajectories that are then clustered using a novel probabilistic technique based on mixtures of regression models.

We develop and test the tracking and clustering methodology using a 15-winter GCM-generated mean sea level pressure (MSLP) data set. The GCM we use simulates the characteristics of cyclones sufficiently well for tracking purposes. We then apply the methodology to a 44-year set of observed data.

1.2 Related work

Prior work on cyclone tracking has focused specifically on methods for automated identification and tracking of cyclones, usually from sea-level pressure data. Identification methods range from the relatively simple approach of finding minima in the surface pressure field (LeTreut and Kalnay 1990; König et al. 1993; Terry and Atlas 1996), or in the 5-point Laplacian thereof, to more complex approaches such as the use of image processing and computer vision techniques (Hodges 1994; Mesrobian et al. 1995; Hodges 1998), the latter often involving other atmospheric fields such

as vorticity (Hoskins and Hodges 2002). These algorithms are usually then coupled with a tracking scheme to produce a final set of trajectories. Methods proposed for tracking so far include a number of different schemes; for example, nearest-neighbor search (Blender et al. 1997; König et al. 1993), numerical prediction schemes with cost-minimizing optimizations (Murray and Simmonds 1991), and feature tracking methods from image analysis (Hodges 1994, 1995).

Blender et al. (1997) introduced the idea of using the K-means clustering algorithm to cluster ETC trajectories of fixed length. The K-means algorithm iteratively searches for compact clusters of multidimensional points in d-dimensional Euclidean space (Hartigan and Wong 1978); this algorithm minimizes within-cluster variance for a given number K of clusters. To apply the K-means algorithm to cyclone trajectory data, one must first convert the variable-length trajectories into fixed-dimensional vectors. To do this, Blender et al. (1997) constrained each of their storm trajectories to be exactly 3 days in length and then concatenated each of the latitude and longitude measurements to form the vectors pn which the K-means algorithm operates. Elsner et al. (2000) and Elsner (2003) also used the K-means algorithm to cluster cyclone trajectories based on the latitude and longitude locations of storms when they reach specific intensities.

This type of vector-based clustering has limitations when applied directly to trajectories. For example, the conversion of the time and space measurements into a fixed-dimensional vector-space loses spatio-temporal smoothness information related to the underlying dynamics of the ETC process. In addition it artificially constrains the trajectories to have fixed lengths. The regression-based clustering used in this paper has been shown to provide systematically better fit and more accurate predictions when used to cluster variable-length trajectory data, compared to vector-based clustering (Gaffney and Smyth 1999; Gaffney 2004)). Tracks of specific lengths may be of particular significance—for example, Simmons and Hoskins (1978) identified a lifecycle of about 10 days—much longer than that assumed by Blender et al. (1997). The approach we propose in this paper for ETC clustering, namely mixtures of regression models, directly incorporates spatio-temporal smoothness in the trajectories in the modeling process as well as accommodating cyclone trajectories of different lengths.

Hierarchical clustering could also be used in this context by defining a distance between pairs of trajectories. For example, dynamic time-warping techniques could be used to define a transformation distance between any curve and another (Wang and Gasser 1997). However, both hierarchical clustering and K-means clustering do not allow for a consistent and systematic approach to problems such as assessing the out-of-sample performance of a cluster model, model selection, or handling missing data. In contrast, the probabilistic regression-based approach to clustering, which we apply to variable length cyclone trajectories in this paper, provides a statistical basis within which all of these issues can be systematically addressed (Fraley and Raftery 1998; Smyth 2000; McLachlan and Peel 2000; Fraley and Raftery 2002).

The paper is organized as follows. Section 2 presents our cyclone identification and tracking methodology and describes the data sets used in this paper. Section 3 introduces a new curve-based methodology for ETC clustering in three parts: (i) a brief introduction to finite mixture models; (ii) the extension of such models to regression mixture models; and (iii) the integration of cyclones into this framework. Section 4 presents the clustering results with the GCM data, with the corresponding results for the observed data set in Section 5. Section 6 concludes the paper.

2 Data and Tracking Methodology

2.1 Data

The GCM data set used for this work was generated by the National Center for Atmospheric Research (NCAR) Community Climate Model (CCM3) (Hack et al. 1998), forced with observed sea surface temperatures specified at the lower boundary over the 1980/81–1994/95 period. For the tracking, we used 6-hourly MSLP fields on an approximate $2.8^\circ \times 2.8^\circ$ Gaussian grid for the extended winter months (1 November to 30 April) from 1980 to 1995; in each winter there are thus 181 days. In this paper we focus on North Atlantic ETCs in the area (30°N–80°N, 80°W–30°E).

The reanalysis data set used in Section 5 is the National Centers for Environmental Prediction (NCEP)–NCAR reanalysis of observed data. The reanalysis involves assimilating the data over the given 44-year period with a unique model and data assimilation method, rather than using the historical succession of NCEP models and methods, as they evolved over these decades (Kalnay et al. 1996). We use 6-hourly MSLP on a regular latitude-longitude grid of 2.5 x 2.5 degrees, over the same North Atlantic domain as for the GCM, but for the 44 extended winter seasons 1958/59–2001/02.

Before clustering ETC trajectories they must first be identified and tracked from the MSLP frames. Our identification and tracking scheme is based on methods already used in this context (Blender et al. 1997; König et al. 1993) and requires relatively few parameters to implement. As the procedure is standard by now, we only give a short description below; the full details can be found in Gaffney (2004).

2.2 Identifying and tracking cyclones

Cyclones are characterized as well-defined surface-pressure minima and their trajectories have lengths of a few thousand kilometers. We begin with a minimum-finding procedure to locate candidate centers of cyclones. In order to distinguish these minima more easily from larger-scale low-pressure areas, the gridded data were high-pass-filtered in space at each time so as to remove the largest planetary-wave scales (Hoskins and Hodges 2002; Anderson et al. 2003). Using bicubic interpolation a cyclone center that may be off-grid is then obtained. The algorithm includes criteria for the rejection of spurious minima.

Once the MSLP minima have been detected at each time, they are then linked together to form cyclone tracks. The previous 6-hour field is scanned for an MSLP minimum within a small neighborhood; if one exists, then the two centers are linked. Displacements of up to 7 degrees of longitude and 5 degrees of latitude in 6 hours are permited; these bounds correspond to a maximum velocity of approximately 129 km/h in longitude and 92 km/h in latitude. Our results indicate that these maximum displacement velocities are hardly ever reached. In the second step, we eliminate tracks shorter than 2.5 days. This removes many short noisy tracks that correspond to local small-scale weather disturbances not usually considered to be cyclones.

Application of this identification and tracking procedure to the CCM3 MSLP data produced 614 cyclones of different durations, each with a minimum of 10 observations (i.e., at least 2.5 days long). Figure 1 shows a sample of the resulting cyclone tracks, with starting positions indicated by circles.

Figure 2 contains three summary histograms describing the statistical characteristics of the entire set of trajectories. The cyclone tracks have typical durations of 2.5–4 days, typical velocities of 30–

60 km/hr (i.e. 8–16 m/s), and reach typical minimum intensities of –30 to –50 mb. These values are of the same order as the statistics derived from tracking observed cyclones in other studies (Hoskins and Hodges 2002). We will use this set of trajectories as input to our clustering algorithm in what follows.

3 Clustering methodology

3.1 Finite mixture models and model-based clustering

A finite mixture model is a probability density function (PDF) composed of a convex combination of component density functions (McLachlan and Peel 2000). In the standard mixture model framework, we model the d-dimensional vector $\mathbf{x}$, as a function of model parameters ϕ, by the mixture density

$$p(\mathbf{x}|\phi) = \sum_{k}^{K} \alpha_k \, p_k(\mathbf{x}|\theta_k), \tag{1}$$

in which α_k is the k-th component weight and p_k is the k-th component density with parameter vector θ_k. The mixture weights α_k sum to one and are nonnegative. Finite mixture models have been widely used for clustering data in a variety of areas (e.g., McLachlan and Basford (1988)) including atmospheric sciences (e.g., Smyth et al. (1999) ; Hannachi and O'Neill (2001)). The parameters for each density component $p_k(\mathbf{x}|\theta_k)$, as well as the corresponding weights α_k, can be estimated from the data using the Expectation-Maximization (EM) algorithm, a widely used technique for maximum-likelihood parameter estimation for mixture models (McLachlan and Krishnan 1997). The estimated density components $p_k(\mathbf{x}|\theta_k)$ are then interpreted as clusters.

A particular advantage of the probabilistic approach is the fact that the component PDFs, $p_k(\mathbf{x}|\theta_k)$ can be defined on non-vector data. This property is helpful in the present case, where the objects to be clustered are curves or trajectories of different lengths (i.e., each ETC track is given by a sequence of latitude-longitude tuples across time).

3.2 Cyclone regression mixture models

Regression mixture models extend the standard finite mixture model to conditional densities, where the data being modeled (i.e., cyclone tracks) are now a function of other input variables (i.e., time). Regression mixture models represent a mixture of K underlying functions (e.g., polynomials) which might have generated the observed data. The technique is quite general and can be adapted to many types of regression models including linear (DeSarbo and Cron 1988), binomial probit (Lwin and Martin 1989), kernel (Gaffney and Smyth 1999), and random effects (Lenk and DeSarbo 2000; Gaffney and Smyth 2003) regression models.

Suppose we have a set of n two-dimensional cyclone trajectories in latitude and longitude, measured over time. Each trajectory $\mathbf{z}_i$ is an $n_i \times 2$ matrix containing the sequence of n_i latitude-longitude measurements; note that n_i may be different for each trajectory $\mathbf{z}_i$. The associated $n_i \times 1$ vector of times at which the $\mathbf{z}_i$ measurements were observed is denoted as $\mathbf{t}_i$. Following Blender et al. (1997), each cyclone trajectory is referred to the origin in both space and time, so that each ($\mathbf{z}_i$, $\mathbf{t}_i$) begins at the relative latitude-longitude position of $(0, 0)$ and at a time of $t = 0$. Clustering is thus performed using only the *shape* of the trajectory, and initial starting positions are eliminated

as a source of variation.

We model the cyclone's longitudinal displacement with a polynomial regression model in which time is the independent variable, and likewise for latitude. The regression mixture model is then derived by substituting the conditional regression density components $p_k(\mathbf{z}|\mathbf{t}, \theta_k)$ in place of the unconditional density components $p_k(\mathbf{x}|\theta_k)$ of standard finite mixtures, cf. Eqn. (1). Details are provided in the appendix.

An EM algorithm for learning the component regression models and component weights for this conditional mixture can be defined in a similar manner to the EM algorithm for standard (unconditional) mixtures (McLachlan and Peel 2000; DeSarbo and Cron 1988; Gaffney and Smyth 1999). For example, the maximization (M) step consists of solving a weighted least-squares regression problem in which the weights are the membership probabilities calculated in the expectation (E) step. Implementation of the EM algorithm is described in Gaffney and Smyth (1999) and Gaffney (2004).

A graphical example of using EM to estimate the parameters of regression mixtures from simulated curve data is given in Figure 3. Four curves, here in a single space dimension for illustration, were generated from each of three different underlying quadratic polynomials (three clusters) for a total of 12 curves (Fig. 3a). Note that the cluster "labels," shown here using the solid, dashed, and dotted lines in Fig. 3a, were not given to the algorithm. Figure 3b shows the initial, randomly chosen starting trajectory of the algorithm for each of the three regression cluster models. The EM algorithm converges in 4 iterations and the final clustering is shown in Fig. 3c, along with the classification of each curve resulting from the clustering; the latter is 100% accurate in this example. The underlying true polynomials that generated the data are the dotted lines in Fig. 3c. The regression mixture methodology recovers the true cluster structure from the data, even though it is not visually apparent that the top two clusters in Fig. 3a can be separated.

[Fig. 4 near here, please.]

A quadratic polynomial is also used in our component regression models for the ETC tracks. This choice was based both on visual inspection of fitted-versus-actual trajectory data, see Fig. 4, as well as a quantitative cross-validation analysis. In the latter, we fitted regression mixture models with different orders of polynomial to randomly selected training sets of trajectories, and then computed the log-probability of unseen "test" trajectories under each model. This calculation is repeated $C = 10$ times over multiple train-test partitions of the data to generate average out-of-sample log-probability (log-p) scores (Smyth 2000; Smyth et al. 1999). The log-p score for regression mixtures is defined in Gaffney (2004). Table 1 shows the cross-validation log-p scores obtained on the cyclone data for polynomials from linear to cubic, for a range of K, the number of regression components. The highest score is achieved with quadratic polynomials across all values of K in the experiment.

4 Clustering of cyclones in GCM simulations

This section describes the results obtained from applying the clustering methodology (Sect. 3b) to the GCM cyclone trajectories of Sect. 2b. An important question is the selection of the number of cyclone clusters. Figure 5 shows the cluster-specific mean curves of each regression mixture model fitted to the cyclone data for $K = 2$, 3, 4, and 6 clusters. Each graph plots the cluster mean in relative latitude-longitude space, using trajectories referred to the origin. Blender et al. (1997) set the number of clusters to three based on various meteorological considerations. In a similar

manner, the three-cluster model in Fig. 5 provides a large-scale description of the North Atlantic cyclones. As the number of clusters is increased, the individual clusters tend to split into smaller refinements of the simpler cluster models, as seen in Fig. 5 for $K = 4$ and $K = 6$ (bottom panels).

[Fig. 5 near here, please.]

Based on this somewhat subjective analysis, three clusters are chosen to describe the cyclone dataset. These clusters are named "south-to-north" (S–N), "southwest-to-northeast" (SW–NE), and "west-to-east" (W–E); they are labeled according to their latitude-longitude orientation on the page: V ("vertical"), D ("diagonal"), and H ("horizontal"). The ETC trajectories assigned to each cluster are shown in Figure 6. The number of trajectories in each cluster is 220 for V, 215 for D, and 179 for H, out of a total of 614 cyclone tracks identified in the CCM3 simulation.

The characteristics of each cluster are given in Table 2. The *average acceleration* of a cyclone is calculated using the absolute rate of velocity change for each trajectory, over the 6-hour intervals at which data is available. The *curvature* of a cyclone is calculated by taking the average of the instantaneous curvature values along the trajectory. The *noisiness* of a cyclone estimates the degree of "erratic" departure from a smooth path calculated by the standard deviation of instantaneous curvature along a cyclone's trajectory .

The V-cluster consists of relatively short, south-to-north oriented cyclones with large curvature and noisiness. The cyclones in this cluster are fairly slow with many exhibiting relatively stationary behavior. The D-cluster consists of a large group of diagonally-oriented cyclones that generally cross the Atlantic travelling from south-west to north-east. These cyclones have the largest average velocity (59 km/h), intensity (-40 mb), duration (4.07 days), and the smallest noisiness (0.0078), as compared to those in the other two clusters. The H-cluster consists of cyclones that move west to east, across the western coastlines of Europe. These cyclones are the least intense on average (-34 mb) overall, but have the largest acceleration values (18.85 km/h^2) and curvature (0.0121); a large part of the curvature can be attributed to erratic behavior, as reflected by their large noisiness of 0.0191.

Figure 7 shows histograms of average speed, duration and minimum intensity, stratified by cluster. Cluster D contains the fastest cyclones in the overall set, with several having greater average speeds than 80 km/h. Cluster V contains the largest number of short-duration tracks, lasting 3 days or less, and only 6% of the cyclones in cluster V last longer than 5 days, as opposed to 11% and 18% for clusters H and D, respectively.

To evaluate the likelihood of cyclones of either type developing at a given epoch, we considered the relation between these clusters and the large-scale circulation regimes of CCM3, as we had done in Robertson and Ghil (1999) for large-scale weather and local precipitation in the western United States. The low-frequency variability of the CCM3 model, however, did not provide conclusive evidence for such associations.

5 Clustering of cyclones in reanalysis data

In this section we apply our new clustering methodology to observed cyclone trajectories. These were tracked in the same manner as the GCM trajectories, using the same region of 30N-80N and 80W-30E as used in the GCM analysis, resulting in a set of 1,915 observed trajectories. This is about 3 times larger than the number found from the GCM data, consistent with the reanalysis data having 3 times as many seasons (44 versus 15). Cyclones are active on approximately 75% of

the days in the 44-winter data set.

Summary histograms (not shown) of average velocity and minimum intensity are quite similar to those in Fig. 2 for the GCM data.

As in the GCM case we again selected K=3 clusters for analysis. The trajectories were first referred to their initial positions, but not otherwise normalized. Figure 8 shows the tracks from each of the three clusters.

As in the case of the GCM, the clusters show predominantly vertical (V), diagonal (D), and horizontal (H) track orientations on the page. The three observed clusters are quite equally populated, with 680, 604 and 631 trajectories, respectively. The 'V-cluster has meridional, recurving tracks, as in CCM3, but is much more heavily concentrated over the western North Atlantic than in the GCM; it contains the most members. The observed D-cluster also forms a much narrower diagonal swath of tracks from SW to NE across the Atlantic. Indeed, the GCM's cyclones are generally too zonal in their spatial distribution, extending excessively into Europe. The H-cluster has predominantly eastward oriented tracks but contains tracks that are more erratic than the GCM.

Compared to the three observed clusters of Blender et al. (1997, Fig. 3 there) in which higher-resolution ECMWF analysis data was used, our results do not include the "stationary" cyclones over Greenland and the Mediterranean. Our D-cluster can be equated with the "north-eastward" cluster of these authors, and our H-cluster with their "zonal" one. The largest difference from Blender and colleagues is our heavily populated V-cluster, with most trajectories close to the coastline of North America.

Figure 9 shows histograms of average speed, duration and minimum intensity for the observed trajectories, stratified by cluster, with summary statistics given in Table 3, analogous to Figure 7 and Table 2 for the GCM trajectories. The observed results are generally quite consistent with the GCM case. The D-cluster again contains markedly faster moving cyclones (mean velociy of 63 km/h), with the H-cluster containing the slowest (mean 38 km/h). The durations are more similarly distributed between the clusters than in the GCM. The observed H-cluster contains less intense cyclones, qualitatively similar to the GCM. The observed accelerations are slightly larger in all three cases, and lifetimes are slightly shorter and differ little between the observed clusters (3.5–3.6 days).

The largest differences between observed and GCM-simulated cyclones are in the curvature and stability. The observed cyclones exhibit much smaller curvature and noisiness, across all three clusters. Our results thus indicate that the data assimilation scheme used in the NCEP-NCAR reanalysis does not produce inherently noisy ETC trajectories. As in the GCM case, the observed D-cluster cyclones tend to have straighter and less-noisy tracks.

One of the motivations for clustering cyclone trajectories is to relate differing cyclone types to the larger-scale background flow. To this end, we have constructed composites of observed 10-day low-pass filtered 700-hPa geopotential height for the sets of days classified into each of the observed clusters. The three observed clusters account for 29%, 23%, and 23% of days respectively, with the remaining 25% assigned to the "quiescent" cluster, when no cyclones are active. The low frequencies are selected and the largest spatial scales retained, so as to focus on the component of the circulation that is not directly associated with the cyclones themselves.

Maps of the four composites based on anomalies from the grand mean of the 44 winters are plotted in Fig. 10. The quiescent composite shows a ridge over the climatological position of the storm track, consistent with reduced cyclone activity, and a weak trough west of Greenland. The V-cluster is

accompanied by a dipole, with a trough centered over Nova Scotia, and a ridge centered over Iceland. The D-cluster and H-cluster are accompanied by opposite phases of the North Atlantic Oscillation (NAO; e.g., Hurrell et al. (2003)), with a north-south dipole in geopotential height over the North Atlantic. The D-cluster corresponds to the positive phase of the NAO, with a trough over Greenland and a ridge to the south. In this phase, the NAO anomaly amplifies the climatological pressure gradients, steering cyclones to the northeast. In the NAO's negative polarity, the climatological gradients are weakened, and cyclones tend to be weaker and track more zonally. Our results are consistent with the NAO index regressed onto root-mean-square transient geopotential height in the 2–8 day band; see Hurrell et al. (2003, Fig. 15).

While the NAO is the most well known teleconnection pattern over the North Atlantic, the most-populated V-cluster is associated with a large-scale wave pattern that is less familiar from studies of low-frequency variability. This pattern is indeed more transitory than the D and H clusters, with a larger number of run-lengths shorter than 5 days (not shown). It shares certain features with ATL regime A2 of Kimoto and Ghil (1993), the Reverse W3 (RW3) wave train of Mo and Ghil (1988), and the eastern Atlantic ridge (AR) of Vautard (1990). The differences consist mainly in a zonal shift of the main features and might be due to differences in the domain of analysis and data set, even more so than to the difference in the compositing. Vautard (1990) notes, in fact, that the stormtrack is both shortened and displaced northward for his AR regime.

6 Summary and concluding remarks

Curve-based mixture models were used to perform probabilistic clustering of wintertime North Atlantic extratropical cyclone (ETC) trajectories in latitude-longitude space. In contrast to previous clustering methods, trajectories have varying durations and the clustering is performed directly in "trajectory-space" rather than in a fixed-dimensional vector space. Quadratic polynomials provide the best fits among the regression models we considered.

An identification and tracking procedure using MSLP fields was developed and applied both to an NCAR CCM3 simulation and the NCEP-NCAR reanalysis. The resulting cyclone trajectories were used as input to the clustering algorithm. For both data sets, we obtained three groups of tracks that are oriented predominantly south-to-north ("V"), southwest-to-northeast ("D"), and west-to-east ("H") respectively in latitude-longitude, over the North Atlantic. The characteristics of these three groups are remarkably similar between the GCM results and the observed data.

The V-cluster consists of relatively short, slow-moving cyclones with S–N tracks, and intermediate curvature and noisiness. Cyclones in the observed V-cluster are much more geographically localized over the eastern seaboard of North America than in the GCM, where they tend to spread out more uniformly with longitude. The D-cluster consists of a large group of diagonally oriented cyclones that generally cross the Atlantic travelling SW–NE. These cyclones have the largest average velocity, intensity, and duration; their tracks are the straightest and smoothest of all cyclones. Their cross-Atlantic swath is again narrower in the observed data. The H-cluster consists of cyclones that generally move W–E, often across the western coastlines of Europe. These cyclones are relatively slow-moving and are the least intense on average, but have relatively large acceleration values and the largest values of curvature and noisiness in both data sets.

Two of the observed clusters, D and H, are closely related to the opposite phases of the well-known NAO teleconnection pattern; this result was less prominent in the GCM. Our V-cluster is associated with a less well-known circulation patterns, identified though by various names in different studies:

RW3 in Mo and Ghil (1988), AR in Vautard (1990), and A2 in Kimoto and Ghil (1993).

Blender et al. (1997) obtained a cluster of near-stationary cyclones, concentrated over the Mediterranean and near Greenland, largely absent from our analysis; these cyclones are likely to be misrepresented in our lower-resolution data set. Of Blender et al.'s two other clusters, the north-eastward one overlaps with our D- and V-clusters, and their zonal cluster is quite similar to our H-cluster.

The clustering performed in this paper is solely by trajectory shape, with the initial position of each cyclone subtracted at the outset. Unlike in Blender et al. (1997), the latitudinal and longitudinal displacements were not normalized, which would tend to emphasize meridional motions in that paper. The impact of this common origin of all trajectories is very apparent in the GCM clusters, which show little localization in space. It is interesting that the observed V- and D-clusters show a larger degree of localization, is spite of the clusters' being based only on trajectory shape.

Additional experiments (not shown) indicate that including initial position yields clusters with strong geographical centers of gravity. The associated cluster composites of geopotential height (not shown) are, however, less amenable to a physical interpretation, compared to those for the common-origin trajectories in Fig. 10.

We conclude that the SW–NE (D) and W–E (H) clusters of North Atlantic ETC tracks are robust across data sets and methodologies, and that they are associated with the positive and negative phase of the NAO. The recurving S–N (V) cluster is novel but, like the two previous ones, is reproducible between the NCEP/NCAR analysis and the CCM3 model; it seems to be associated with a less well-defined large-scale atmospheric pattern, marked by a trough over the western and a ridge over the eastern North Atlantic. It is remarkable that the CCM3 reproduces these observed ETC clusters fairly well, in spite of the fairly low T42 horizontal resolution.

Acknowledgments: We wish to thank Kevin Hodges for helpful discussions, and Jim Boyle and Peter Glecker for help in obtaining the NCAR CCM3 data. NCEP-NCAR Reanalysis data were provided by the NOAA CIRES Climate Diagnostics Center, Boulder, Colorado, from their Web site available online at http://www.cdc.noaa.gov. This work was supported in part by a Department of Energy grant DE-FG02-02ER63413 (MG and AWR), by NOAA through a block grant to the International Research Institute for Climate Prediction (SJC and AWR), and by the National Science Foundation under grants No. SCI-0225642 and IIS-0431085 (SJG and PS).

Appendix A Regression mixture models for ETCs

For illustration, consider a hypothetical trajectory $\mathbf{z}_i$ with $n_i = 4$ measurements, where the longitude and latitude measurements are in the first and second column, and have had their initial values subtracted, while $\mathbf{t}_i$ gives the time from initiation of the tracking:

$$\mathbf{z}_i = \begin{bmatrix} 0 & 0 \\ 1 & 0.2 \\ 2.5 & 0.4 \\ 3.3 & 0.7 \end{bmatrix}, \qquad \mathbf{t}_i = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}.$$

Note that this example represents a cyclone moving mostly in a zonal direction.

We model longitude with a polynomial regression model of order p, in which time $\mathbf{t}_i$ is the independent variable, and likewise for latitude. Both regression equations can be defined succinctly in terms of the matrix $\mathbf{z}_i$. The exact form of the regression equation for $\mathbf{z}_i$ is

$$\mathbf{z}_i = \mathbf{T}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \sim N(\mathbf{0}, \Sigma). \tag{A1}$$

Here $\mathbf{T}_i$ is the standard $n_i \times (p+1)$ Vandermonde regression matrix associated with the vector $\mathbf{t}_i$; $\boldsymbol{\beta}$ is a $(p+1) \times 2$ matrix of regression coefficients, which contains the longitude coefficients in the first column and the latitude coefficients in the second column; and ϵ_i is an $n_i \times 2$ matrix of multivariate normal errors, with a zero mean and a 2×2 covariance matrix Σ, cf. Eqn. A2 below. The covariance matrix Σ contains three distinct elements: σ_1^2 and σ_2^2, which are the noise variances for each longitude and latitude measurement, respectively, and the cross-covariance $\sigma_{12} = \sigma_{21}$ between any two longitude and latitude measurements. For simplicity, we make the assumption that $\Sigma = \text{diag}(\sigma_1^2,\ \sigma_2^2)$, so that latitude and longitude measurements are treated as conditionally independent given the model. The Vandermonde regression matrix $\mathbf{T}_i$ consists of $(p+1)$ columns of $\mathbf{t}_i$ so that the components of $\mathbf{t}_i$ in the m-th column are taken to the power of m for $0 \leq m \leq p$. For example, if $\mathbf{t}_i = (0, 1, 2, 3)'$, where $()'$ denotes the transpose, and $p = 3$ then

$$\mathbf{T}_i = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{bmatrix}.$$

The resulting conditional density model is often referred to as matrix multivariate normal, since the input $\mathbf{z}_i$ is a matrix. The conditional density for the i-th cyclone is defined as

$$\begin{aligned} p(\mathbf{z}_i|\mathbf{t}_i, \theta) &= f(\mathbf{z}_i|\mathbf{T}_i\boldsymbol{\beta}, \Sigma) \\ &= (2\pi)^{-n_i}\,|\Sigma|^{-n_i/2} \exp\left\{-\frac{1}{2}\left[(\mathbf{z}_i - \mathbf{T}_i\boldsymbol{\beta})\Sigma^{-1}(\mathbf{z}_i - \mathbf{T}_i\boldsymbol{\beta})'\right]\right\}, \end{aligned} \tag{A2}$$

where $\theta = \{\boldsymbol{\beta}, \Sigma\}$.

We can derive regression mixtures for the cyclones by substitution of the unconditional multivariate density components $p_k(\mathbf{x}|\theta_k)$ of standard finite mixtures, cf. Eqn. (1) with the conditional regression density components $p_k(\mathbf{z}|\mathbf{t}, \theta_k)$, defined above in Eqn. (A2). This results in the following regression mixture model for cyclones:

$$p(\mathbf{z}_i|\mathbf{t}_i, \phi) = \sum_k^K \alpha_k\ p_k(\mathbf{z}_i|\mathbf{t}_i, \theta_k) = \sum_k^K \alpha_k\ f_k(\mathbf{z}_i|\mathbf{T}_i\boldsymbol{\beta}_k, \Sigma_k). \tag{A3}$$

Note that in this model each trajectory, that is each ETC, is assumed to be generated by one of K different regression models, where each regression model has its own "shape" parameters $\theta_k = \{\boldsymbol{\beta}_k, \Sigma_k\}$. The clustering problem is to first (a) learn the parameters of all K models given data; and then (b) infer which of the K models are most likely to have generated each ETC.

If we let $\boldsymbol{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be the complete set of n cyclone trajectories and $\boldsymbol{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ be the set of associated measurement times, then the full probability density of $\boldsymbol{Z}$ given $\boldsymbol{T}$, also known as the conditional likelihood, is

$$p(\boldsymbol{Z}|\boldsymbol{T}, \phi) = \prod_{i}^{n} \sum_{k}^{K} \alpha_k \; f_k(\mathbf{z}_i|\mathbf{T}_i\boldsymbol{\beta}_k, \Sigma_k). \tag{A4}$$

The product form follows from assuming conditional independence of the $\mathbf{z}_i$'s, given both the $\mathbf{t}_i$'s and the mixture model, that is assuming ETCs do not influence each other. Strictly speaking this is not necessarily true since (for example) multiple ETCs can be present at the same time. However, to a first approximation, ETCs can be assumed to be conditionally independent of each other—this is certainly an adequate assumption for the purposes of clustering.

References

Anderson, D., K. I. Hodges, and B. J. Hoskins, 2003: Sensitivity of feature-based analysis methods of storm tracks to the form of background field. *Mon. Wea. Rev.*, **131**, 565–573.

Blender, R., K. Fraedrich, and F. Lunkeit, 1997: Identification of cyclone-track regimes in the North Atlantic. *Quart J. Royal Meteor. Soc.*, **123**, 727–741.

DeSarbo, W. S. and W. L. Cron, 1988: A maximum likelihood methodology for clusterwise linear regression. *J. Classification*, **5**, 249–282.

Elsner, J. B., 2003: Tracking hurricanes. *Bull. Amer. Meteor. Soc.*, **84**, 353–356.

Elsner, J. B., K. b Liu, and B. Kocher, 2000: Spatial variations in major US hurricane activity: Statistics and a physical mechanism. *J. Climate*, **13**, 2293–2305.

Fraley, C. and A. E. Raftery, 1998: How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578–588.

—, 2002: Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Stat. Assoc.*, **97**, 611–631.

Gaffney, S. and P. Smyth, 1999: Trajectory clustering with mixtures of regression models. *Proc. Fifth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, S. Chaudhuri and D. Madigan, eds., ACM Press, N.Y., 63–72.

Gaffney, S. J., 2004: *Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models*. Ph.D. Dissertation, Department of Computer Science, University of California, Irvine.

Gaffney, S. J. and P. Smyth, 2003: Curve clustering with random effects regression mixtures. *Proc. 9th Internat. Workshop on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, eds., Key West, FL.

Hack, J. J., J. T. Kiehl, and H. J. W., 1998: The hydrologic and thermodynamic characteristics of the NCAR CCM3. *J. Climate*, **11**, 1179–1206.

Hannachi, A. and A. O'Neill, 2001: Atmospheric multiple equilibria and non-Gaussian behaviour in model simulations. *Quart. J. Royal Meteor. Soc.*, **127**, 939–958.

Hartigan, J. A. and M. A. Wong, 1978: Algorithm AS 136: A K-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.

Hodges, K. I., 1994: A general method for tracking analysis and its applications to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586.

—, 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–3465.

—, 1998: Feature-point detection using distance transforms: Application to tracking tropical convective complexes. *Mon. Wea. Rev.*, **126**, 785–795.

Hoskins, B. J. and K. I. Hodges, 2002: New perspectives on the northern hemisphere winter storm tracks. *J. Atmos. Sci.*, **59**, 1041–1061.

Hurrell, J. W., Y. Kushnir, G. Ottersen, and M. Visbeck, 2003: An overview of the North Atlantic Oscillation. *Geophys. Monogr.*, **134**, 2217–2231.

Kimoto, M. and M. Ghil, 1993: Multiple flow regimes in the northern hemisphere winter. Part II: Sectorial regimes and preferred transitions. *J. Atmos. Sci.*, **16**, 2645–2673.

König, W., R. Sausen, and F. Sielman, 1993: Objective identification of cyclones in GCM simulations. *J. Climate*, **6**, 2217–2231.

Lenk, P. J. and W. S. DeSarbo, 2000: Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, **65**, 93–119.

LeTreut, H. and E. Kalnay, 1990: Comparison of observed and simulated cyclone frequency distribution as determined by an objective method. *Atmosfera*, **3**, 57–71.

Lwin, T. and P. J. Martin, 1989: Probits of mixtures. *Biometrics*, **45**, 721–732.

McLachlan, G. and T. Krishnan, 1997: *The EM Algorithm and Extensions*. John Wiley and Sons, New York, NY.

McLachlan, G. J. and K. E. Basford, 1988: *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McLachlan, G. J. and D. Peel, 2000: *Finite Mixture Models*. John Wiley & Sons, New York.

Mesrobian, E., R. Muntz, E. Shek, C. R. Mechoso, J. Farrara, J. Spahr, and P. Stolorz, 1995: Real time data mining, management, and visualization of GCM output. *Supercomputing '94, IEEE Computer Society*, 81–87.

Mo, K. and M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. *J. Geophys. Res.*, **93D**, 10927–10952.

Murray, R. J. and I. Simmonds, 1991: A numerical scheme for tracking cyclone centres from digital data Part I: Development and operation of the scheme. *Aust. Meteor. Mag.*, **39**, 155–166.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, Amsterdam, 425 pp.

Robertson, A. W. and M. Ghil, 1999: Large-scale weather regimes and local climate over the western united states. *J. Climate*, **12**, 1796–1813.

Saunders, M. A., 1999: Earth's future climate. *Phil. Trans. Royal Soc. Lond. A*, **357**, 3459–3480.

Simmons, A. J. and B. J. Hoskins, 1978: The life cycles of some nonlinear baroclinic waves. *J. Atmos. Sci.*, **35**, 414–432.

Smyth, P., 2000: Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**, 63–72.

Smyth, P., K. Ide, and M. Ghil, 1999: Multiple regimes in northern hemisphere height fields via mixture model clustering. *J. Atmos. Sci.*, **56**, 3704–3723.

Terry, J. and R. Atlas, 1996: Objective cyclone tracking and its applications to ERS-1 scatterometer forecast impact studies. *15th Conf. Weather Analysis & Forecasting*, Amer. Meteor. Soc., Norfolk, VA.

Vautard, R., 1990: Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. *Mon. Wea. Rev.*, **45**, 2845–2867.

von Storch, H. and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, MA, 484 pp.

Wang, K. and T. Gasser, 1997: Alignment of curves by dynamic time warping. *Annal. Stat.*, **25**, 1251–1276.

Table 1: Log-probability (log-p) scores using cross-validation on the GCM cyclone data for K-values of 1 to 4 with linear, quadratic, and cubic polynomials.

K	Linear	Quadratic	Cubic
1	-3.6024	-3.5902	-3.5929
2	-3.4380	-3.4169	-3.4198
3	-3.3279	-3.3016	-3.3051
4	-3.2355	-3.2120	-3.2170

Table 2: Cluster-wide average measures for various GCM cyclone statistics. Both means (μ) and standard deviations (σ) are given for each cluster.

Cluster-specific statistics	V		D		H	
	μ	σ	μ	σ	μ	σ
Minimum intensity (mb)	-39.93	8.9	-40.00	8.2	-33.79	7.4
Average velocity (km/h)	42.49	11.54	59.36	13.75	42.82	15.86
Average acceleration (km/h^2)	15.40	5.66	16.52	5.65	18.85	7.48
Lifetime (days)	3.61	0.10	4.07	1.17	3.75	1.17
Curvature	0.0121	0.0220	0.0048	0.0052	0.0152	0.0178
Noisiness	0.0191	0.0484	0.0078	0.0131	0.0235	0.0358

Table 3: Observed cluster-wide average measures for various cyclone statistics. Both means (μ) and standard deviations (σ) are given for each cluster column.

Observed cluster-specific statistics	V		D		H	
	μ	σ	μ	σ	μ	σ
Minimum intensity (mb)	-39.12	8.49	-39.22	8.09	-35.41	7.75
Average velocity (km/h)	43.87	10.49	62.57	12.93	37.77	12.76
Average acceleration (km/h^2)	17.96	7.06	19.58	6.89	19.52	7.99
Lifetime (days)	3.48	0.93	3.62	1.14	3.54	1.15
Curvature	0.0029	0.0031	0.0018	0.0019	0.0062	0.0055
Noisiness	0.0046	0.0069	0.0032	0.0038	0.0086	0.0096

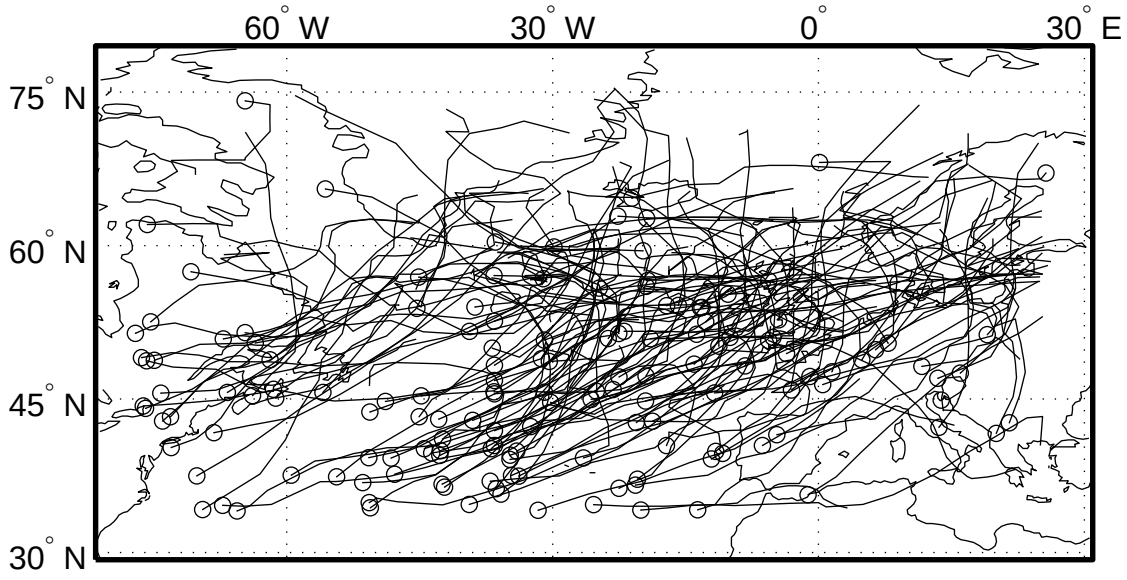

Figure 1: Random sample of 200 CCM3 cyclone trajectories tracked over the North Atlantic domain of interest. The circles indicate initial starting position.

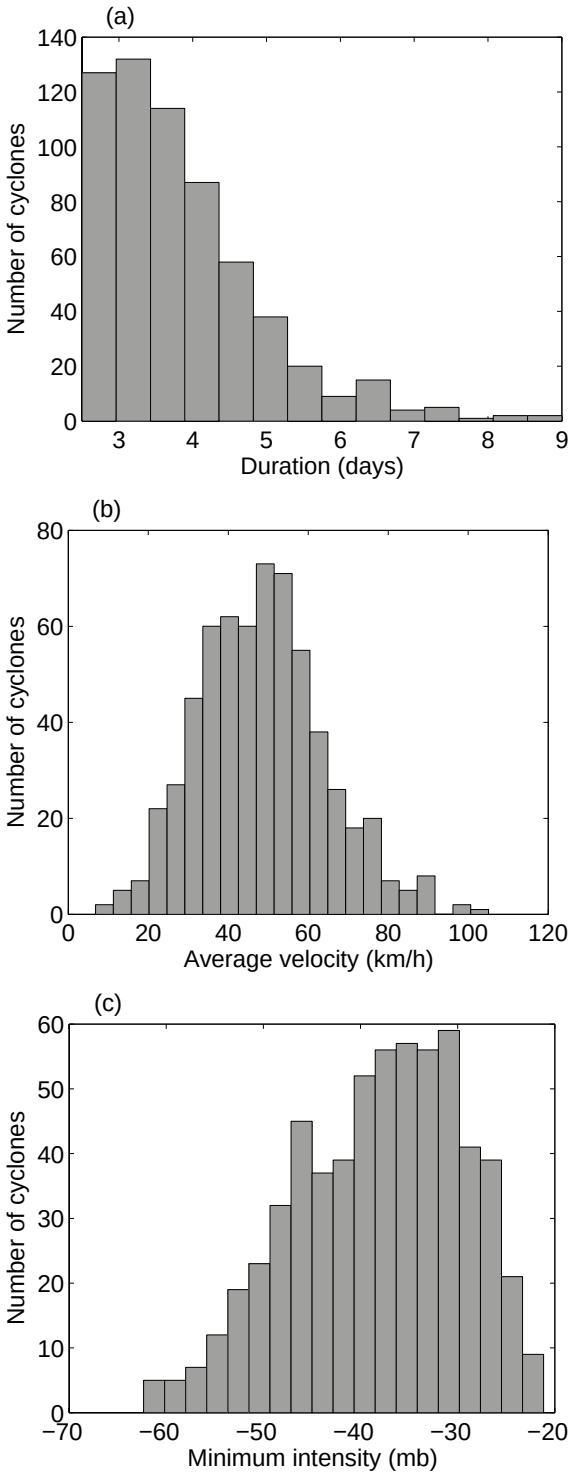

Figure 2: Summary histograms for GCM cyclone data set: (a) cyclone duration, (b) average velocity, and (c) minimum intensity (MSLP).

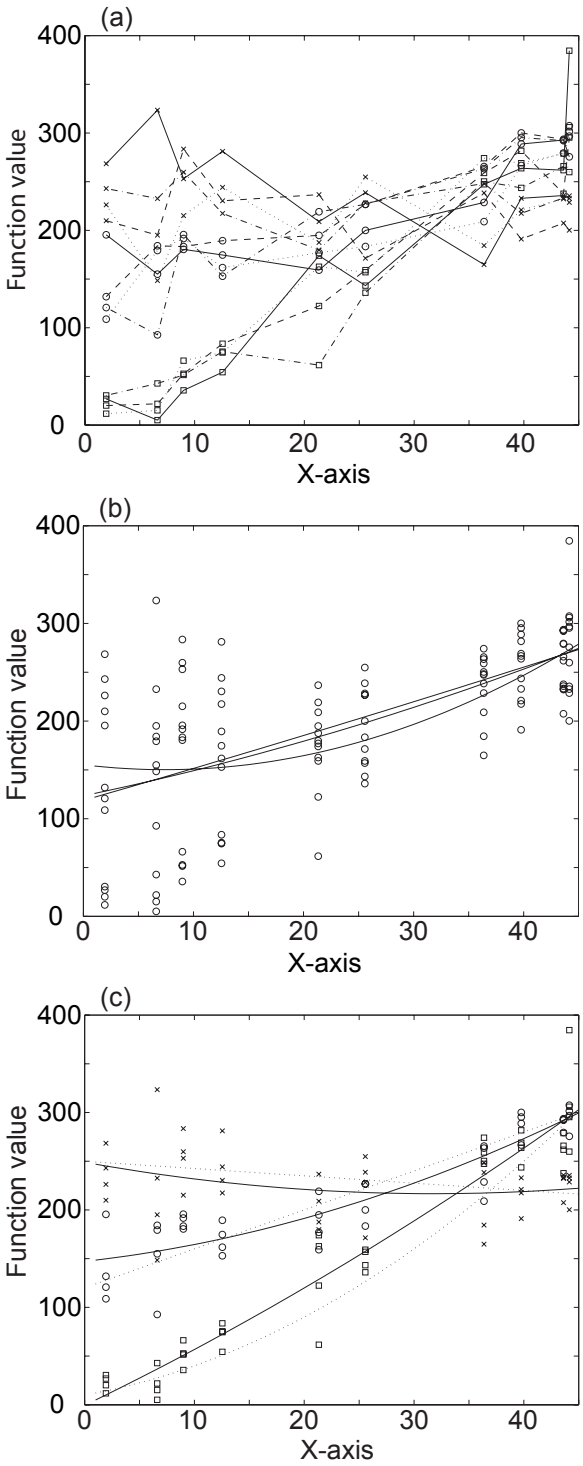

Figure 3: Performance of the EM algorithm as applied to synthetic trajectories, generated from polynomial regression mixture model: (a) set of synthetic trajectories presented to the algorithm (solid, dashed, dotted lines denote 3 generating models); (b) initial random starting curves (solid) for the three clusters; (c) cluster locations (solid) after EM convergence (iteration 4) as well as the locations of the true data-generating trajectories (dotted).

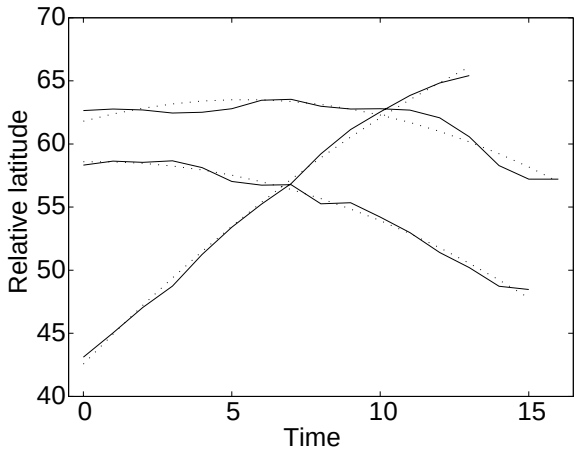

Figure 4: Quadratic polynomial regression models (dotted) fitted to three GCM cyclone trajectories (solid) in latitude-time plane.

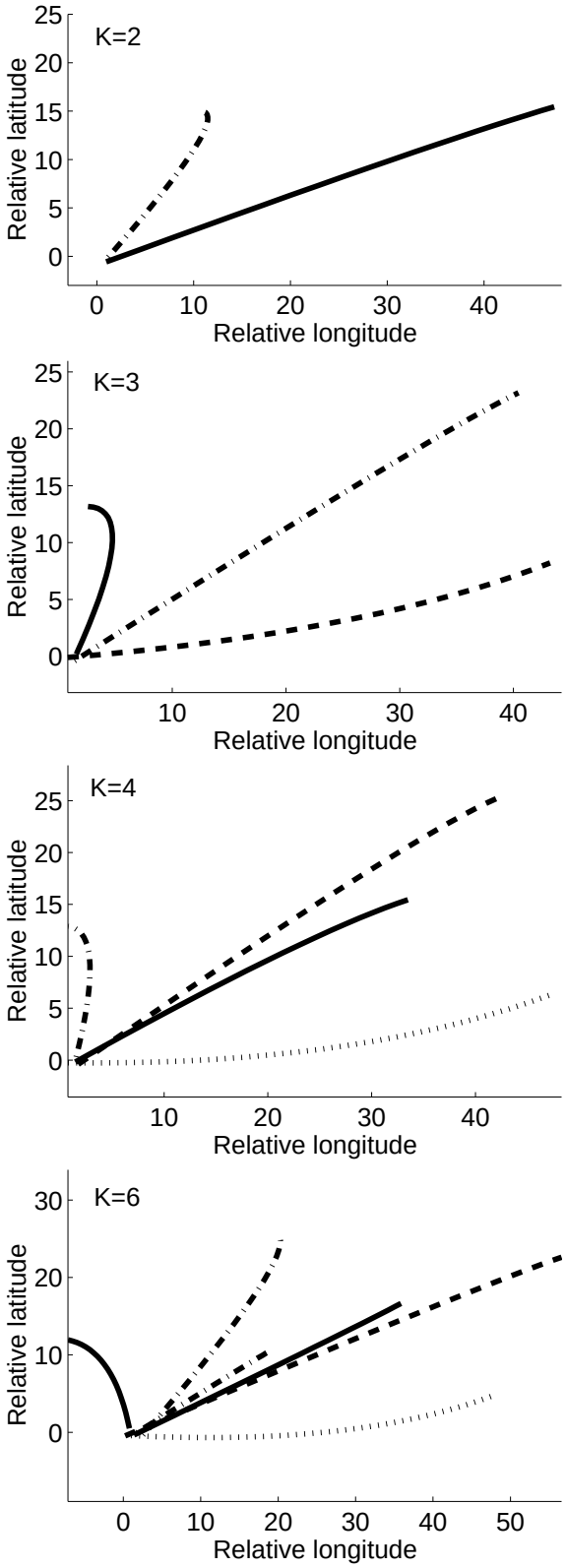

Figure 5: GCM cyclone cluster models when $K = 2, 3, 4$, and 6.

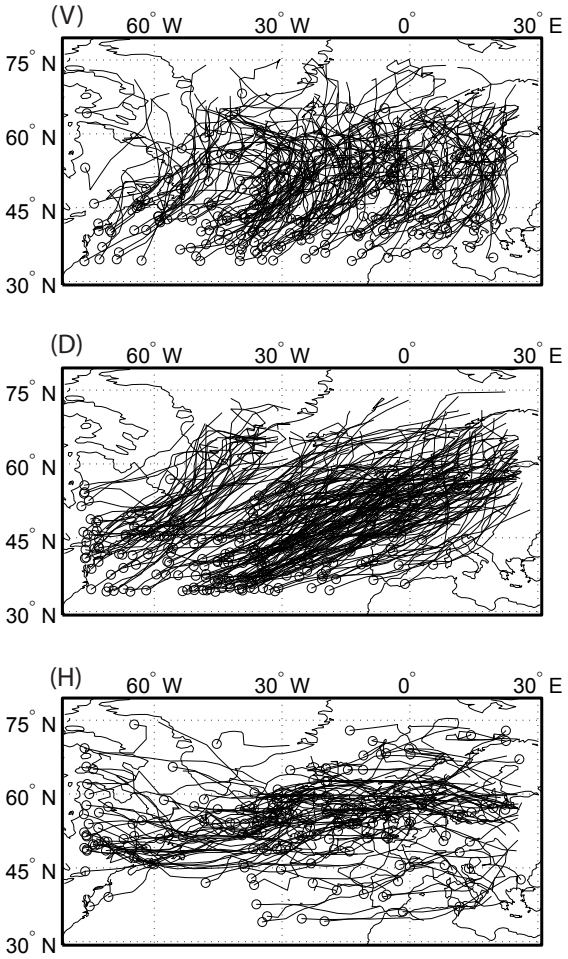

Figure 6: (V) South-to-north, (D) southwest-to-northeast, and (H) west-to-east oriented clusters from GCM data. All trajectories are shown.

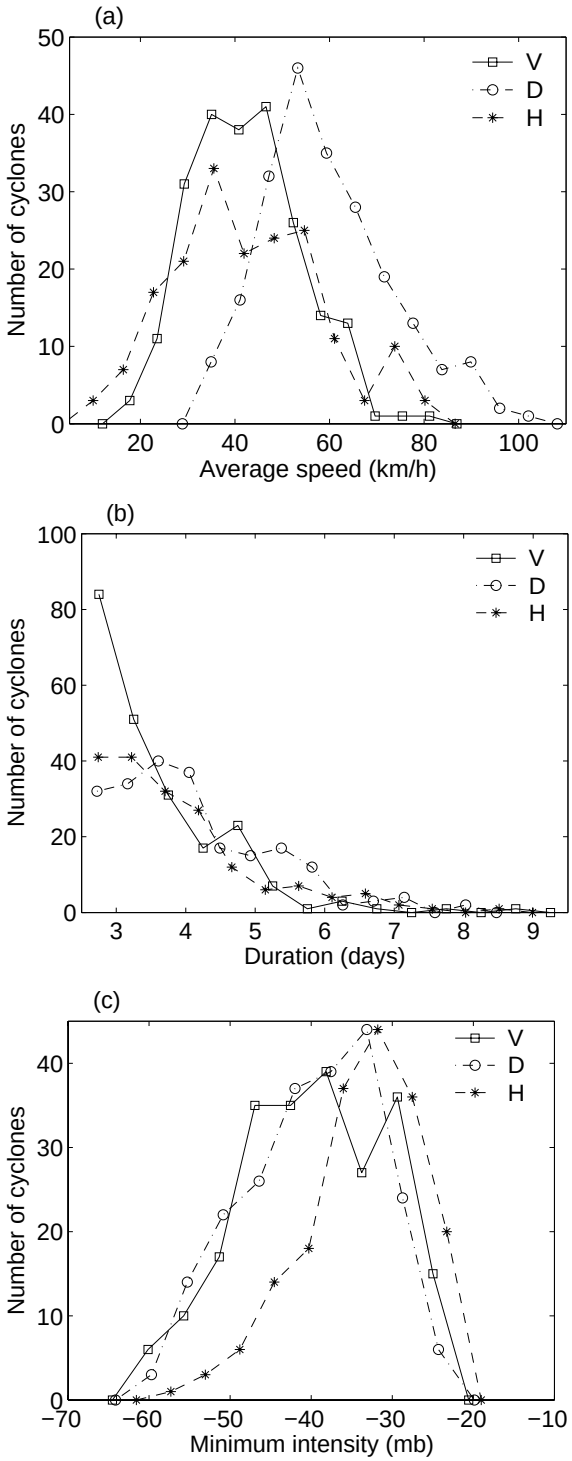

Figure 7: Histograms stratified by GCM cluster: (a) average velocity, (b) cyclone duration, and (c) minimum intensity (MSLP). The histograms are plotted as line graphs for clarity.

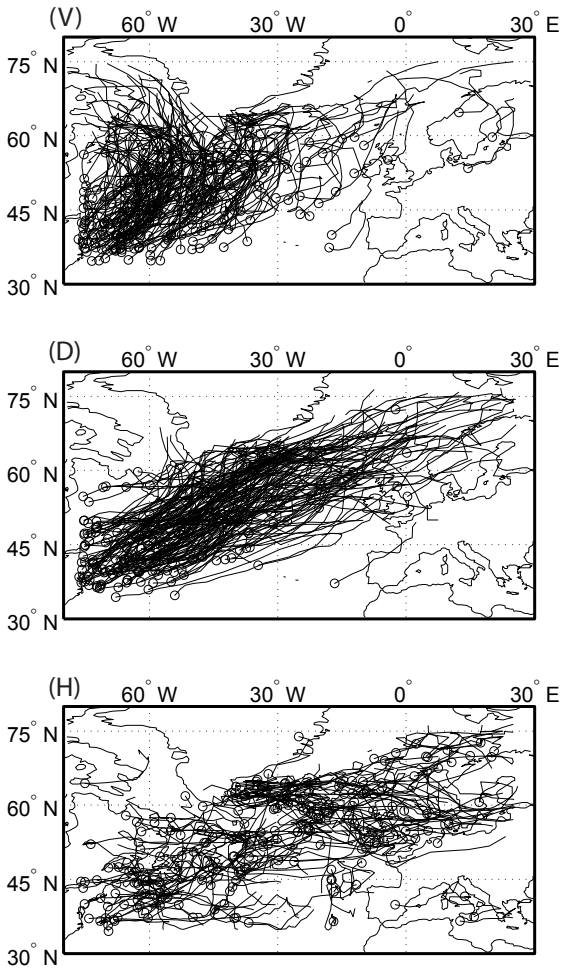

Figure 8: Clusters derived from observational data: (V) South-to-north, (D) southwest-to-northeast, and (H) west-to-east. Only 200 random tracks are shown in each cluster for clarity.

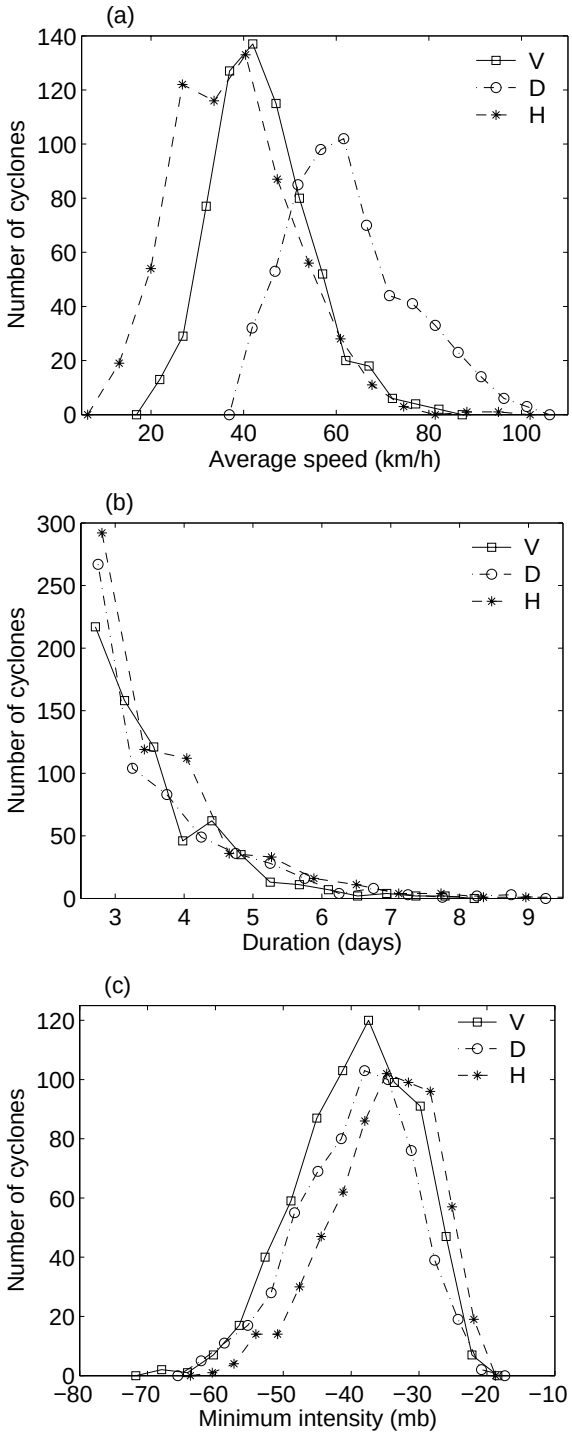

Figure 9: Histograms of observed trajectories stratified by cluster: (a) average velocity, (b) cyclone duration, and (c) minimum intensity (MSLP). The histograms are plotted as line graphs for clarity.

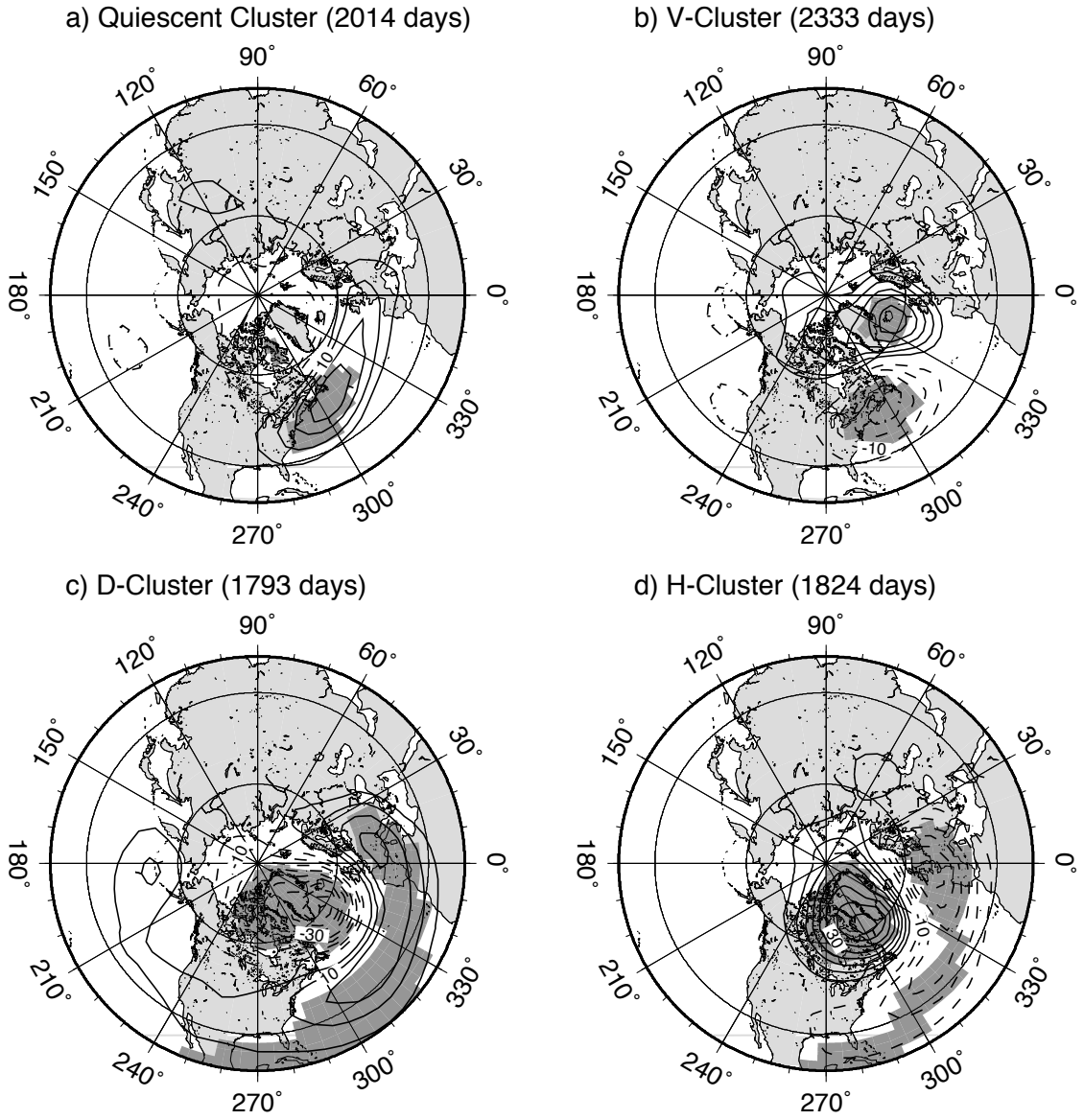

Figure 10: Composites of low-pass filtered 700-hPa geopotential height anomalies for the days assigned into each cluster. In each case the 44-winter time average has been subtracted. The shaded regions are significant at the 99% level according to a two-sided Student t-test with 120 degrees of freedom; this number is smaller than the number of days in each composite divided by 10. Contour interval: 5 meters.