Cluster Analysis of Typhoon Tracks. Part I: General Properties

Suzana J. Camargo and Andrew W. Robertson

International Research Institute for Climate and Society, The Earth Institute at Columbia University, Palisades, New York

Scott J. Gaffney and Padhraic Smyth

Department of Computer Science, University of California, Irvine, Irvine, California

Michael Ghil*

Department of Atmospheric and Oceanic Sciences, and Institute for Geophysics and Planetary Physics, University of California, Los Angeles, Los Angeles, California

(Manuscript received 6 January 2006, in final form 28 August 2006)

ABSTRACT

A new probabilistic clustering technique, based on a regression mixture model, is used to describe tropical cyclone trajectories in the western North Pacific. Each component of the mixture model consists of a quadratic regression curve of cyclone position against time. The best-track 1950–2002 dataset is described by seven distinct clusters. These clusters are then analyzed in terms of genesis location, trajectory, landfall, intensity, and seasonality.

Both genesis location and trajectory play important roles in defining the clusters. Several distinct types of straight-moving, as well as recurving, trajectories are identified, thus enriching this main distinction found in previous studies. Intensity and seasonality of cyclones, though not used by the clustering algorithm, are both highly stratified from cluster to cluster. Three straight-moving trajectory types have very small within-cluster spread, while the recurving types are more diffuse. Tropical cyclone landfalls over East and Southeast Asia are found to be strongly cluster dependent, both in terms of frequency and region of impact.

The relationships of each cluster type with the large-scale circulation, sea surface temperatures, and the phase of the El Niño–Southern Oscillation are studied in a companion paper.

1. Introduction

Typhoons have a large socioeconomic impact in many Asian countries. The risk of landfall of a typhoon or tropical storm depends on its trajectory. These trajectories, in turn, vary strongly with the season (Gray 1979; Harr and Elsberry 1991), as well as on interannual (Chan 1985) and interdecadal time scales (Ho et al. 2004). However, current knowledge is largely qualitative, and the probabilistic behavior of tropical cyclone trajectories needs to be better understood in order to isolate potentially predictable aspects of landfall. Well-calibrated probabilistic seasonal predictions of landfall risk could form an important tool in risk management.

Tropical cyclogenesis over the tropical northwest (NW) Pacific takes place in a broad region west of the date line, between about 8° and 25°N. South of 15°N, most of these tropical cyclones (TCs) follow rather straight west-northwestward tracks. About one-third of them continue in this direction and make landfall in southeast Asia and southern China. Most of the remainder "recurve," that is, slow down, turn northward, and then accelerate eastward as they enter the midlatitude westerlies (e.g., Harr and Elsberry 1995). Another fraction of TCs track northward over the ocean, posing no threat to land.

The large-scale circulation of the atmosphere has a

* Additional affiliation: Département Terre-Atmosphére-Océan, and Laboratoire de Météorologie Dynamique du CNRS/IPSL, Ecole Normale Supérieure, Paris, France.

Corresponding author address: Dr. Suzana J. Camargo, International Research Institute for Climate and Society, Monell 225, 61 Route 9W, Palisades, NY 10964-8000.
E-mail: suzana@iri.columbia.edu

DOI: 10.1175/JCLI4188.1

predominant role in determining a TC's motion through the steering by the surrounding large-scale flow (e.g., Chan and Gray 1982; Franklin et al. 1996; Chan 2005). The cyclone and the environment interact to modify the surrounding flow (Wu and Emanuel 1995), and the vortex is then advected (steered) by the modified flow. One important dynamical factor is the beta drift, involving the interaction of the cyclone, the planetary vorticity gradient, and the environmental flow. This leads TCs to move northwestward even in a resting environment in the Northern Hemisphere (Adem 1956; Holland 1983; Wu and Wang 2004). Other effects can also be important: the interaction of tropical cyclones with mountain ranges leads to significant variations in tracks, as often occurs in Taiwan (Wu and Kuo 1999).

This two-part study explores the hypothesis that the large observed spread of TC tracks over the tropical NW Pacific can be described well by a small number of clusters of tracks, or TC "regimes." The observed TC variability on seasonal and interannual time scales is then interpreted in terms of changes in the frequency of occurrence of these TC regimes. In this paper, we explore the basic attributes of the underlying clusters by applying a new clustering technique to the best-track dataset of the Joint Typhoon Warning Center (JTWC). The technique employs a mixture of polynomial regression models (i.e., curves) to fit the geographical "shape" of the trajectories (Gaffney and Smyth 1999, 2005; Gaffney 2004). Camargo et al. (2007, hereafter Part II) examine relationships between the clusters we describe in the present paper and the large-scale atmospheric circulation, as well as the El Niño–Southern Oscillation (ENSO).

In midlatitude meteorology, the concept of planetary circulation regimes (Legras and Ghil 1985), sometimes called weather regimes (Reinhold and Pierrehumbert 1982), has been introduced in attempting to connect the observations of persistent and recurring midlatitude flow patterns with large-scale atmospheric dynamics. These midlatitude circulation regimes have intrinsic time scales of several days to a week or more and exert a control on local weather (e.g., Robertson and Ghil 1999). Longer time-scale variability of weather statistics (TCs in our case) is a result of changes over time in the frequency-of-occurrence of circulation regimes. This paradigm of climate variability provides a counterpart to wave-like decompositions of atmospheric variability, allowing the connection to be made with oscillatory phenomena (Ghil and Robertson 2002), such as the Madden–Julian oscillation.

Circulation regimes have most often been defined in terms of clustering, whether fuzzy (Mo and Ghil 1987) or hierarchical (Cheng and Wallace 1993), in terms of maxima in the probability density function (PDF) of the large-scale, low-frequency flow (Molteni et al. 1990; Kimoto and Ghil 1993a,b), as well as in terms of quasi stationarity (Ghil and Childress 1987; Vautard 1990) and, more recently, using a probabilistic Gaussian mixture model (Smyth et al. 1999).

In the case of TC trajectories, the *K*-means method (MacQueen 1967) has been used to study western North Pacific (Elsner and Liu 2003) and North Atlantic (Elsner 2003) TCs. In those studies, the grouping was done according to the positions of maximum and final hurricane intensity (i.e., the last position at which the TC had hurricane intensity). In both basins, three clusters were chosen to describe the trajectories. The *K*-means approach has also been used to cluster North Atlantic extratropical cyclone trajectories, where 6-hourly latitude–longitude positions over 3 days were converted into 24-dimensional vectors suitable for clustering (Blender et al. 1997).

The *K*-means method is a straightforward and widely used partitioning method that seeks to assign each track to one of *K* groups such that the total variance among the groups is minimized. However, *K*-means cannot accommodate tracks of different lengths, and we show this to be a serious shortcoming for TCs. On a different approach, Harr and Elsberry (1995) used fuzzy cluster analysis and empirical orthogonal functions to describe the spatial patterns associated with different typhoon characteristics.

The finite mixture model used in this paper to fit the geographical shape of the trajectories allows the clustering to be posed in a rigorous probabilistic framework and accommodates tropical cyclone tracks of different lengths. These characteristics provide advantages over the *K*-means method used in previous studies. The main novelty here is to use an objective method to classify the typhoon tracks based not only on a few points of the trajectory, but on trajectory shape and location.

The clustering methodology is briefly described in section 2 and applied to the JTWC best-track dataset in section 3. The two main trajectory types identified by the cluster analysis correspond to straight movers and recurvers; additional clusters correspond to more detailed differences among these two main types, based on location and track type. We study several characteristics of the TCs in each cluster, including first position, mean track, landfall, intensity, and lifetime, and compare them with previous works in section 4. Discussion and conclusions follow in section 5. In Part II, we study how the large-scale circulation and ENSO affect each cluster.

2. Data and methodology

a. Data and definitions

The TC data used in this paper were based on the JTWC best-track dataset available at 6-hourly sampling frequency over the time interval 1950–2002 (Joint Typhoon Warning Center 2005). The tracks were studied over the western North Pacific, defined such that the latitude–longitude of the TCs are inside the "rectangle" (0°–60°N and 100°E–180°) during at least part of their lifetimes. The clustering technique and the resulting analysis were applied to a total of 1393 cyclone tracks. We included only TCs with tropical storm intensity or higher: tropical storms (TSs), both category 1 and 2 typhoons (TYs) as defined by the Saffir–Simpson scale (Saffir 1977; Simpson and Riehl 1981), and intense typhoons (ITYs; categories 3–5). Tropical depressions are not included in the analysis.

The observed data quality is thought to be considerably poorer during presatellite years (pre-1970). We assume that although some of the TCs may be missing in the JTWC (2005) database for the presatellite data, especially those that remain over the ocean, the tracks for those that do appear in the dataset are reliable, even if their intensity is not. We repeated the cluster analysis for the time interval 1970–2002 and found that the types of tracks obtained in each cluster are essentially the same. This verification lends credence to the data in the earlier part of the record and demonstrates the robustness of our results.

b. Clustering methodology

We present here a brief summary of the clustering methodology (details are given in the appendix). A more complete discussion is given by Gaffney (2004), with an application of the clustering method to extratropical cyclones over the North Atlantic (Gaffney et al. 2007; a Matlab toolbox with the clustering algorithms described in this paper is available online at http://www.datalab.uci.edu/resources/CCT).

Our curve clustering method is based on the finite mixture model (e.g., Everitt and Hand 1981), which represents a data distribution as a convex linear combination of component density functions. A key feature of the mixture model is its ability to model highly non-Gaussian (and possibly multimodal) densities using a small set of basic component densities. Finite mixture models have been widely used for clustering data in a variety of areas (e.g., McLachlan and Basford 1988), including the large-scale atmospheric circulation (Smyth et al. 1999; Hannachi and O'Neill 2001).

Regression mixture models extend the standard mixture modeling framework by replacing the marginal component densities with conditional density components. The new conditional densities are functions of the data (i.e., cyclone position) conditioned on an independent variable (i.e., time). In this paper, the component densities model a cyclone's longitudinal and latitudinal positions versus time using quadratic polynomial regression functions, as discussed in Gaffney (2004). The latitude and longitude positions are treated as conditionally independent given the model, and thus the complete function for a cyclone track is the product of these two. Other models, such as higher-order polynomials and splines can also be used within the mixture framework, but the simple quadratic model appears to offer the best trade-off between ease of interpretation and goodness-of-fit.

Each trajectory (i.e., each cyclone track) is assumed to be generated by one of K different regression models, each having its own shape parameters. The clustering problem is to (i) learn the parameters of all K models given the TC tracks, and (ii) infer which of the K models are most likely to have generated each TC track. Each track can be assigned to the mixture component (and thus the cluster) that was most likely to have generated that track given the model. In other words, the assigned cluster has the highest posterior probability given the track. An expectation maximization (EM) algorithm for learning these model parameters can be defined in a manner similar to that for standard (unconditional) mixtures (DeSarbo and Cron 1988; Gaffney and Smyth 1999; McLachlan and Krishnan 1997; McLachlan and Peel 2000). The resulting EM algorithm is straightforward to implement and use, and its computational complexity is linear in the number of observations.

Certain preprocessing steps are typically performed on the cyclone tracks prior to clustering. For example, Blender et al. (1997) subtract the coordinates of the initial points of each extratropical cyclone track so that they all begin at the latitude–longitude position of 0°, 0°. In addition they also normalize the latitude and longitude measurements to have the same variance. In our experiments below we did not use any such preprocessing—clustering the tracks directly produced results that were easier to interpret and more meaningful than the clustering of preprocessed tracks.

c. Number of clusters

To select the most appropriate number of clusters, we looked at both the in-sample and out-of-sample log-likelihood values. The log-likelihood is defined as the log-probability of the observed data under the model, which can be seen as a goodness-of-fit metric for probabilistic models. Used as an objective measure, one se-

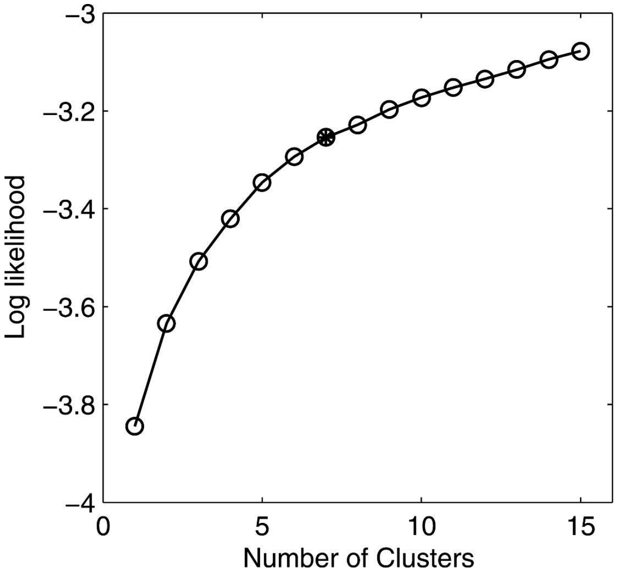

FIG. 1. Log-likelihood values for different number of TC track clusters. The log-likehood values shown are the maximum of 16 runs, obtained by a random permutation of the tropical cyclones given to the cluster model.

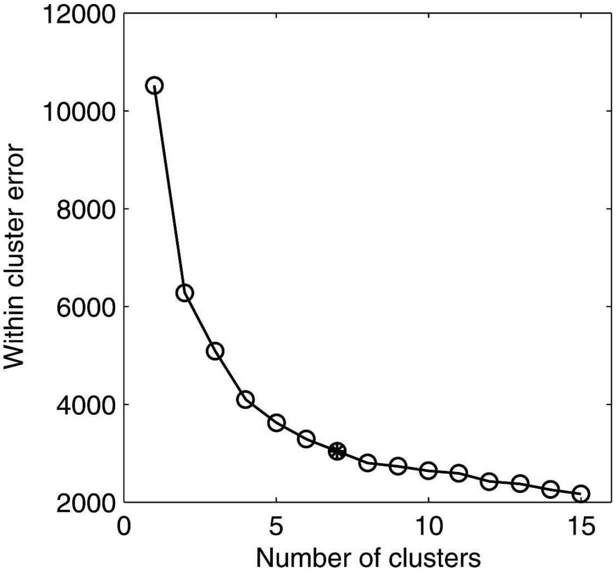

FIG. 2. Within cluster error for different number of TC track clusters. The cluster error values shown are the minimum of 16 runs, obtained by a random permutation of the tropical cyclones given to the cluster model.

lects the number of clusters for which the log-likelihood is largest across a candidate set of values. Our resulting in-sample score curve is shown in Fig. 1 (the out-of-sample curve is similar and is not shown). The observed log-likelihood values increased in direct relation to the number of clusters, and thus did not directly provide an optimal number of chosen clusters. In addition the within-cluster spread is plotted in Fig. 2 and can be used as an additional measure for goodness of fit. The curves in Figs. 1 and 2 mirror each other, showing obvious diminishing returns of improvement in fit beyond $K =$ 6–8, suggesting a reasonable stopping point somewhere in-between.

To evaluate the values $K = 6$–8 as candidates for the number of clusters, we also carried out a qualitative analysis based on how much the track types differ from one cluster to another as the number of clusters increases. Preliminary results carried out with six clusters (Camargo et al. 2004) are very similar to those presented here. The main difference is that one of the $K =$ 6 track types splits in two when K is set to 7, with slightly different characteristics. Most of the results presented here and in Camargo et al. (2007) are not sensitive to the choices between $K = 6$–8. As described by Camargo et al. (2007), the choice of $K = 7$ is found to produce particularly interpretable results with respect to ENSO and was thus taken to be our final choice.

Figure 3 illustrates how the choice for the number of clusters from $K = 2$–9 affects the final regression curves. To emphasize differences in shape, the mean regression trajectories are plotted with their initial positions collocated at the origin. The two main types of TC behavior found in previous studies (Harr and Elsberry 1991, 1995) are evident in these plots, namely, “straight movers” and “recurvers.” The differentiation between the two types is achieved for $K \geq 3$. For each of these two broad types, additional clusters yield differences in compass bearing for the straight movers and differences in the recurving portion for the recurvers. This remark is particularly valid for odd values of K (Figs. 3b,c,f,h). Although some of the regression curves look very similar in Fig. 3, their initial positions differ in several cases and there are also differences in trajectory length. Since the regression curves are plotted with the same number of points, the distances between plotted points are smaller or larger based on average speed over such a period. It is interesting to note that along the recurving trajectories, the points are very close to each other within the recurving portion, showing that TCs slow down before changing direction. The recurving usually occurs when the storms move from a region of easterlies to a region of westerlies, with the wind speed decreasing near the recurve point. It is important to note, however, that the clustering technique has no access to the wind fields.

The regression trajectories for the six, seven, and eight clusters are shown in Fig. 4; in this case, the initial positions were retained. Note that the odd (even) clusters share greater similarity than adjacent values of K. For the chosen number of clusters (K = 7), shown in

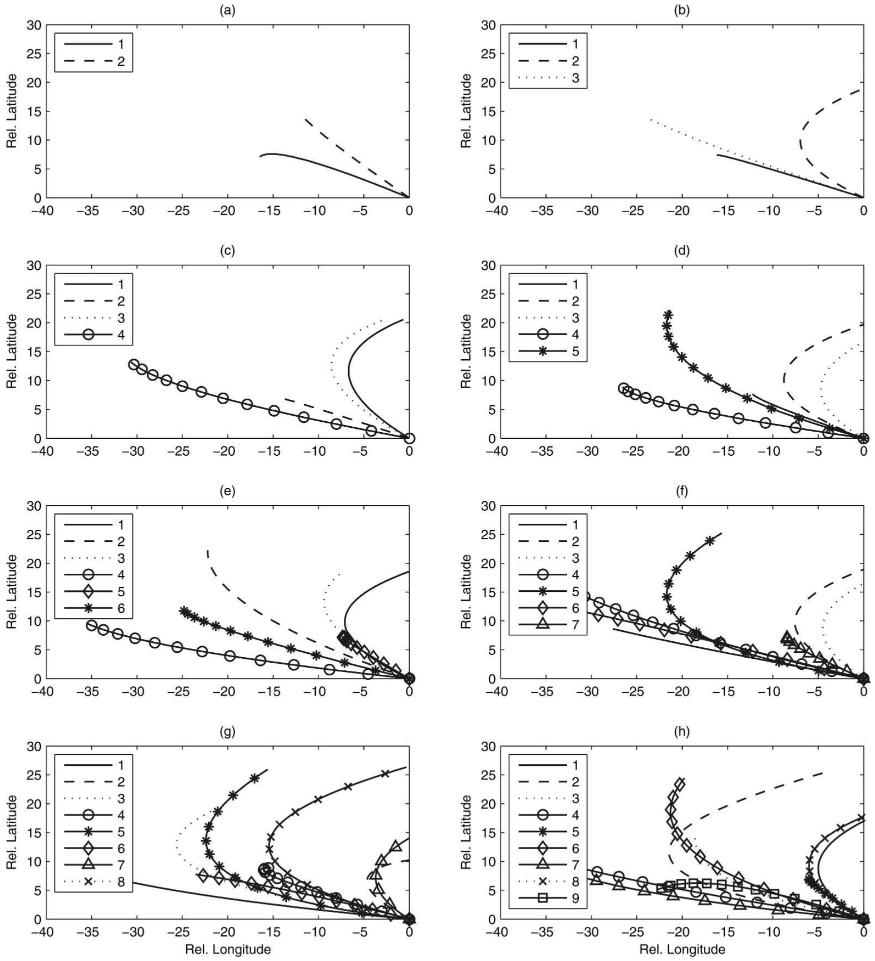

FIG. 3. Mean regression trajectories of the western North Pacific TCs with (a) two, (b) three, (c) four, (d) five, (e) six, (f) seven, (g) eight, and (h) nine clusters. The mean trajectories start at 0° lat and lon, for plotting purposes only.

Fig. 4c there are four clusters of straight movers and three of recurvers. Notice the strong separation between the clusters in terms of their genesis location: five clusters have genesis positions near 10°N in latitude, but spread in longitude from near the Philippines to just west of the date line. The other two clusters (both recurvers) start near 20°N.

Looking at the population of each cluster in Table 1, we see that there are three dominant clusters (A, B, and C), each accounting for approximately 20% of the tracks. Clusters D and E occur less often (13%), while clusters F and G (each containing about 100 cyclones) are relatively rare (8%). When only considering the last 33 yr, 1970–2002, the number and characteristics of the clusters did not change (see section 2a), but their relative sizes did change somewhat (not shown), with the dominant clusters (such as A and C) decreasing and the least populated ones (E, F, and G) increasing. This significant change in relative cluster sizes could be due to either a decadal shift in the occurrence of tracks (Ho et al. 2004), or to data issues, with fewer TCs being detected over open waters before the satellite era.

3. Tropical cyclone clusters

a. Trajectories

The TC tracks in clusters A–G from the time interval 1983–2002 are shown in Fig. 5, along with the mean regression curves for each cluster. For comparison, the tracks of all TCs in the same time interval are also shown (Fig. 5h). The figure illustrates the high degree of geographic localization achieved by the cluster analysis, mainly due to the fact that the tracks were not reduced to a common origin before performing the clustering. The spread about the mean track for the straight-moving clusters B, D, and F is particularly small. Although the mean regression trajectories of

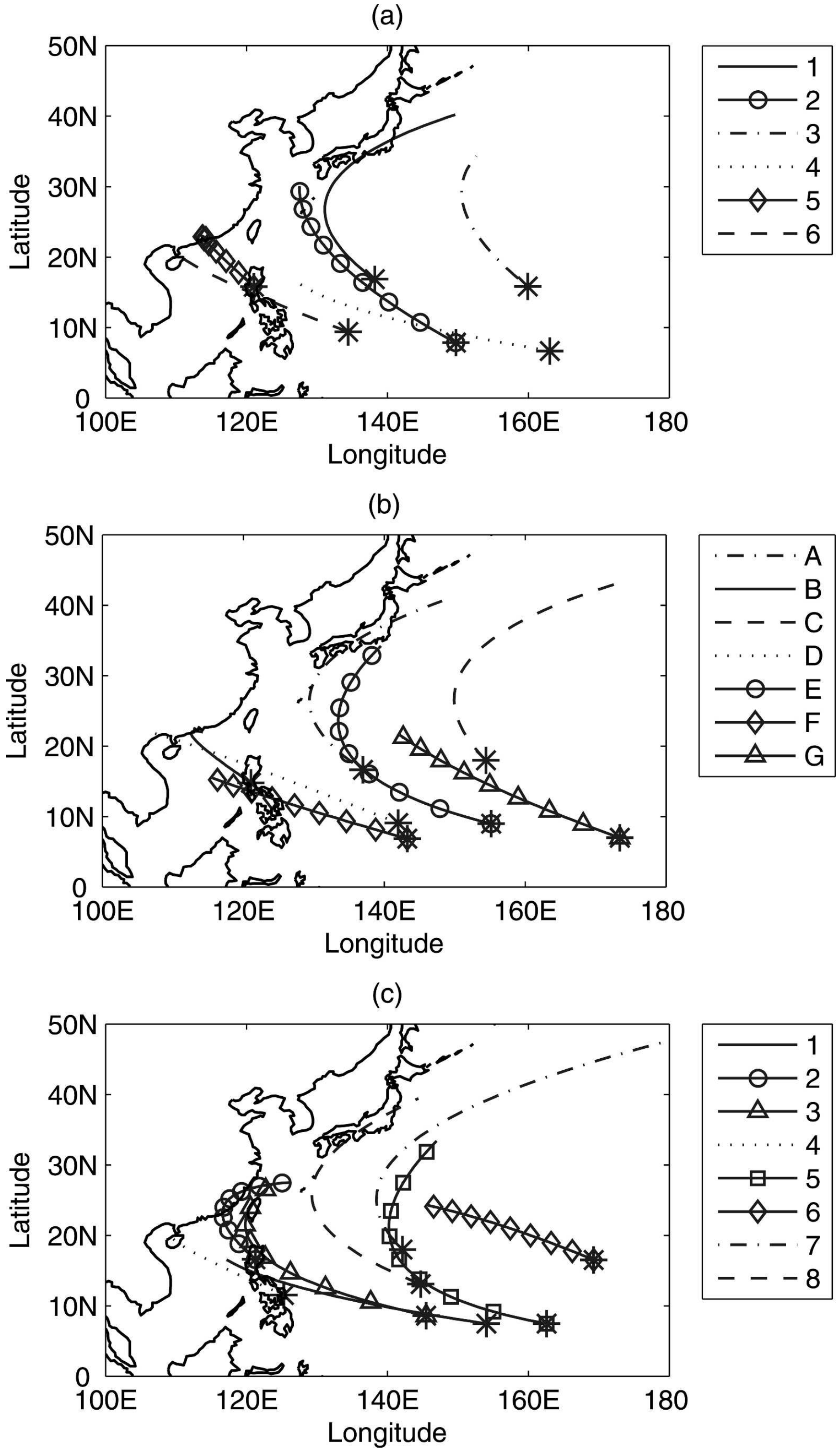

FIG. 4. Mean regression trajectories of the TCs over the western North Pacific, with (a) six, (b) seven, and (c) eight clusters. The origin of the trajectories are marked with an asterisk (*).

TABLE 1. Main TC statistics. The seven clusters are labeled from A to G, in decreasing order of NTCs in each. The subsequent columns indicate percentage of TCs (PTC), number of landfalls (NLF), and percentage of landfall (PLF) in each cluster. The last row (ALL) summarizes the data for the seven clusters.

Cluster	NTC	PTC	NLF	PLF
A	306	22%	188	61%
B	280	20%	238	85%
C	235	17%	17	7%
D	178	13%	129	72%
E	176	13%	56	32%
F	112	8%	71	63%
G	106	8%	16	15%
ALL	1393	100%	715	51%

clusters D and F are very similar, many of their characteristics are very different, as will be further explored. The recurving clusters A, C, and E are more diffuse, though A and C are still quite geographically limited.

The typical track in cluster A is a recurved trajectory, as shown by the mean regression curve of the cluster. Most of the TC activity in cluster A occurs between Japan and the Philippines. The cyclones in cluster B typically follow straight tracks across the Philippines and the South China Sea. The typical recurving track in cluster C stays mostly over the ocean.

The cyclone tracks in cluster D are typically straight and usually cross the Philippines. The TC activity in cluster D is restricted to a long narrow region that extends from east of the Philippines to the south China coast. The first-position patterns of clusters D and E are somewhat similar, with the mean track of the latter originating about 10° farther east (see Fig. 4). The typical track of cluster E, though, is recurving, and leads to a much larger area of TC activity over the Pacific Ocean and land.

The typical tracks in cluster F are very similar to those in cluster D, but the former are typically longer and localized farther to the south. The TC activity for cluster F is also limited to a long narrow strip, as for cluster D. The cyclones that cross from the western North Pacific to the Indian Ocean belong to cluster F. Frequently the remnants of these TCs traveling westward into the Indian Ocean are responsible for monsoon disturbances and cyclones in the Bay of Bengal (e.g., Krishnamurti et al. 1977; Saha et al. 1981; Kumar and Krishnan 2005). The typical track in cluster G is straight, with a more northward trending compass angle than in the case of clusters D and F; still, this cluster contains a fair number of cyclones that recurve. The TC activity in cluster G extends, therefore, over a much wider region than for clusters D and F, with a maximum over the southeastern corner of the basin. The most plausible reason for some recurving tracks to belong to cluster G is that before reaching more northern latitudes, these tracks follow the mean regression track.

To quantify the differences in track direction between the clusters, we plot kernel distributions of tangential direction in Fig. 6. These distributions are based on all neighboring pairs of points along each trajectory and were produced using an 11.25° resolution. Each compass point marked on the plot—E, NE, N, NW, W, SW, S, and SE—corresponds to a range of 45°; thus a westward direction is associated with a vector from one point on a track to the next one, 6 h later, when the angle this vector makes with the east lies between 157.5° and 202.5° (i.e., 180° ± 22.5°).

As expected, when all TCs are considered, the peak direction occurs for western and northwestern trends, with a second peak smaller to the north and northeast associated with recurving trajectories. The TC movement in each cluster is in agreement with the mean regression trajectory. Thus, the recurving clusters A and C exhibit flatter distributions with small peaks to the NW and NE. The straight-moving clusters B, D, and F exhibit a single, fairly sharp peak between NW and W. Clusters E and G have two peaks: a larger one at NW and W, and a second smaller one at NE. The tracks in both of these clusters are a mix of straight moving and recurving, more similar to the distribution of all TCs, with more (fewer) recurving tracks occurring in cluster E (G), and therefore justifying their classification as recurving (straight). Note that the kernel distributions for clusters B, D, and F are much sharper than that of all TCs.

b. Genesis position

The tracks presented in the previous subsection are strongly localized. Here we examine the extent to which this geographic localization can be accounted for by differences in genesis position. The numbers of cyclones originating in each 2° latitude × 2° longitude "square," for each cluster, and for all TCs are given in Fig. 7. The median latitude and longitude of first positions for all TCs are given in Table 2, grouped by cluster.

Taking all TCs together, there is a large spread in genesis position, with highest density west of 160°E and south of 20°N. This corresponds roughly to the density of the tracks themselves, though the latter extends farther northwestward. The individual clusters partition genesis position even more clearly than in track density, resulting in relatively little overlap between clusters. Clearly, the genesis position plays a large role in defining the clusters, although it is not given any particular weight in the clustering algorithm itself.

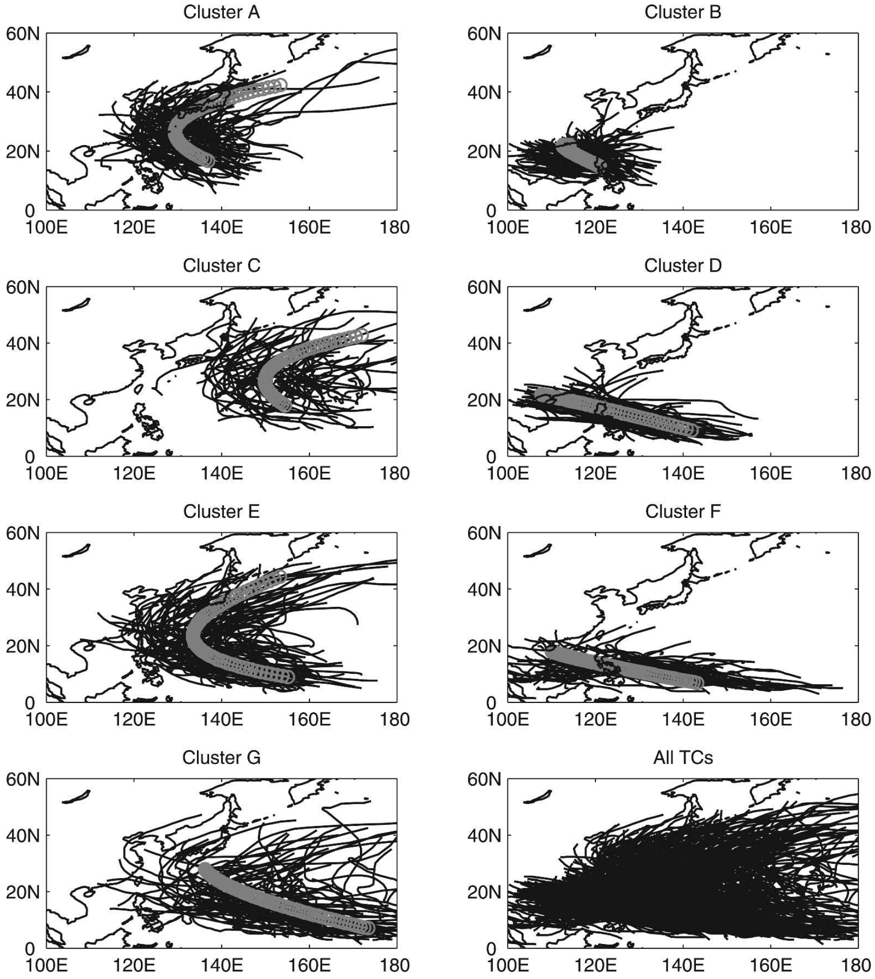

FIG. 5. TC tracks (black) over the western North Pacific, during the period 1983–2002 in each of the seven clusters and for all TCs; the mean regression curve of each cluster is shown in gray open circles.

Typically, cluster A formation occurs east of the Philippines and Taiwan. Cluster B first positions lie mainly in the South China Sea and to the east and northeast of the Philippines, but staying south of Taiwan. The genesis of the cyclones in cluster C occurs more diffusely in the central part of the western North Pacific basin. These three clusters are the most highly populated ones and together make up 59% of the TCs in the basin (see Table 1). Cyclones in cluster D form east of the Philippines, while cluster E genesis also occurs east of the Philippines, but shifted farther to the east compared to cluster D, and to the south of cluster C.

The two least populated clusters are F and G, and their genesis occurs around 10°N. The cyclones in cluster F form within a narrow, long strip that extends from the Gulf of Thailand to the east of the Philippines, until near the date line. The TCs in G originate the farthest east of all the clusters, in close proximity to the date line, and closer to the equator than the other clusters (except F).

c. Intensity and lifetime

The percentage of cyclones with tropical storm, typhoon (categories 1–2), and intense typhoon (categories 3–5) strength is given in Fig. 8, cluster by cluster, as well as for all TCs. On a climatological basis, about one-third of all TCs fall into each intensity category. In contrast, all the clusters except D and F—the long-track, straight movers crossing the Philippines—show large deviations in intensity percentage from the all-TC climatology. The two recurving clusters with large density between Japan and the Philippines (A and E) exhibit highly skewed intensity distributions: cluster E toward intense typhoons (60%), and cluster A toward tropical storms (43%) while only 22% of TCs are intense typhoons in this cluster. These two recurving clus-

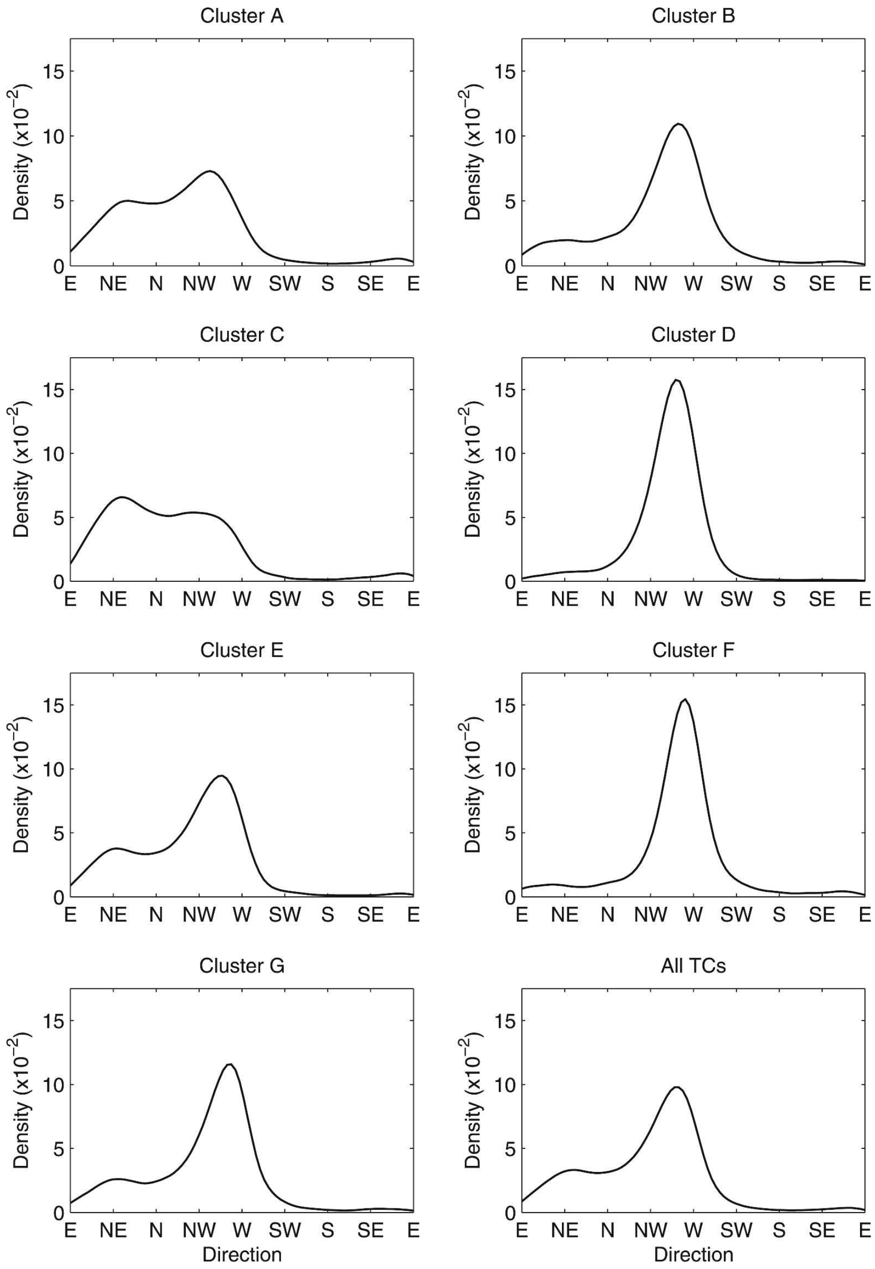

FIG. 6. Distribution of TC angles of movement (as function of compass points E, NE, N, NW, W, SW, W, and SE) for each cluster and for all TCs.

ters have very different genesis locations, with cluster E forming much farther south (see Fig. 7). Cluster G also contains a large number of intense typhoons (59%), and like cluster E it has a genesis region near 10°N and contains many long tracks. By contrast, the numerous short tracks of cluster B (see Fig. 5) rarely develop into intense typhoons (10%). Like cluster A, cluster C has a genesis region near 20°N, and it too has relatively few intense typhoons (21%).

It is clear from Figs. 5 and 8 that track length plays an important role in intensity. This is also consistent with the distribution of the cyclone's lifetime shown in Fig. 9. The median lifetime of the TCs in cluster A is 6.75 days, compared to 11.25 days for cluster E. The typical lifetime of cluster B is the smallest of all (5.25 days), with very few cyclones in this cluster reaching extended lifetimes and intense typhoon status (10%). The median lifetime in cluster G (whose TCs are often intense; see Fig. 8) is the largest of all clusters (13.1 days); the standard deviation is also the largest in this cluster. This

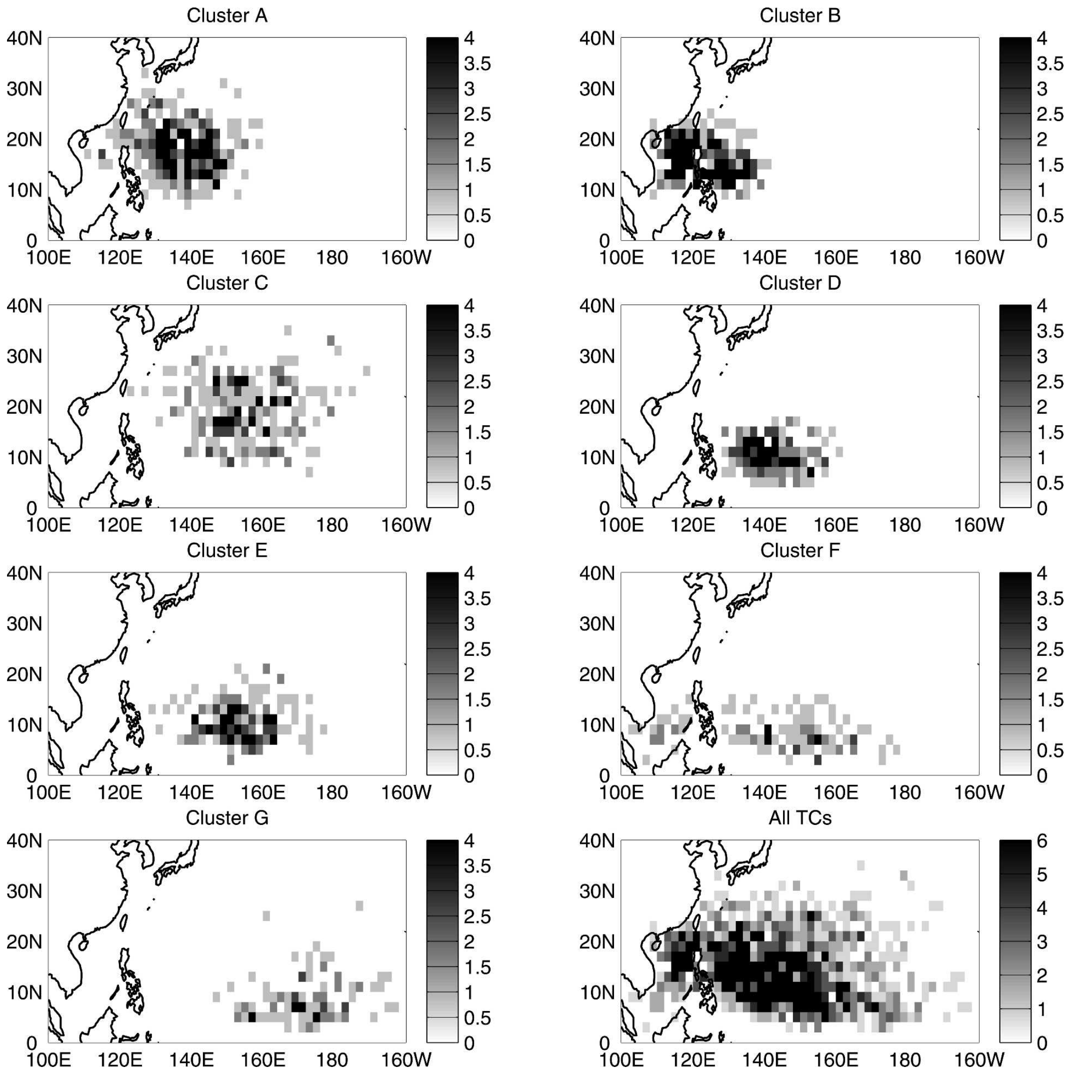

FIG. 7. Number of cyclones with genesis position in each 2° lat × 2° lon square for each cluster and all TCs.

positive relationship of intensity and lifetime of cyclones is consistent with the results of Camargo and Sobel (2005), with longer-living and more intense typhoons in El Niño years and the opposite in La Niña years.

TABLE 2. The median genesis location (latitude and longitude) for TC first positions, in each cluster and for all TCs.

Cluster	Lat (N)	Lon (E)
A	17.0°	136.0°
B	14.9°	120.8°
C	18.3°	153.0°
D	9.2°	140.7°
E	8.9°	152.2°
F	6.9°	146.6°
G	6.6°	171.7°
All	13.0°	141.7°

The straight-moving cyclones in clusters D and F are quite equally divided among tropical storm, typhoon, and intense typhoon strengths. However, the median lifetime in cluster F is the second largest of all clusters (11.5 days). Despite the length of its tracks, their trajectory over land, when crossing the Philippines, could be responsible for the relative lack of intense typhoons in this cluster.

d. Landfall

Landfall risk is the paramount parameter of concern for societies in SE and East Asia. Figure 10 shows the location and percentage of landfall in each cluster, defining landfall where the center of the TC intersects the coast (which could be an island). Each asterisk (*) shown in Fig. 10 represents the landfall of one TC in

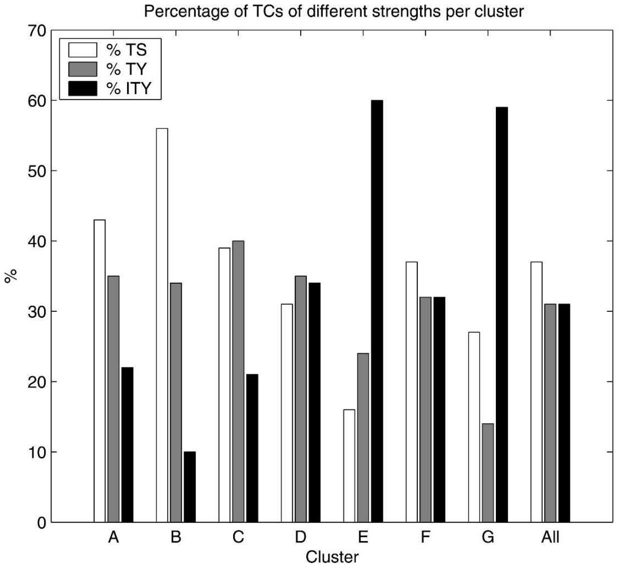

FIG. 8. Percentage of cyclones that have TS, TY (categories 1–2), and ITY (categories 3–5) strength in each cluster and for the whole basin in the period 1970–2002.

that location. In clusters with a high percentage of landfall, these asterisks cover continuous segments of coast.

Some 51% of all cyclones over the basin make landfall, but the percentage varies from 7% to 85% between the clusters. There is thus considerable landfall discrimination between the clusters in landfall likelihood: the landfall percentages in each cluster are significantly different from the landfall percentage of the whole JTWC dataset at the 95% significance level, using a binomial significance test. The percentages of landfall are also significantly distinct from cluster to cluster, with the exception of cluster F when compared with either A or D. In the case of the pair (A, F), it is the percentages (61% and 63%) that are very close to each other, while for the pair (D, F) the percentages (72% and 63%) are not that close, but the population of the two clusters is too small for statistical significance at the 95% level.

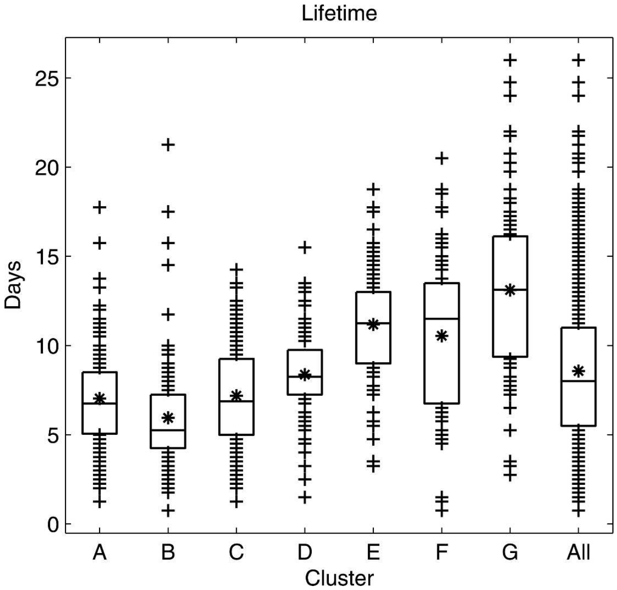

FIG. 9. Distribution of lifetime per cyclone, in each cluster and for all TCs in the period 1970–2002. The boxes show the 25th and 75th percentiles, the lines in the boxes mark the median, the asterisks (*) the mean, and the crosses (+) the values below (above) the 25th (75th) percentiles of the distributions.

Examining the regional landfall distributions in Fig. 10, 61% of the tropical cyclones in cluster A (recurving) make landfall in the Philippines, China, Taiwan, the Korean Peninsula, and Japan. Cluster B (short straight tracks over the South China Sea) has the highest percentage (85%) of landfall of all clusters, with landfall from northern Vietnam to the south China coast, including the Philippines and Taiwan. In contrast, cluster C has the lowest landfall rate among all clusters (7%), as its main TC activity occurs over the ocean. The few landfall cases of cluster C occur over Japan.

Landfall percentages for clusters D and F (straight movers, crossing the Philippines Sea) are high (72% and 63%, respectively), both with large landfall rates over the Philippines. Their landfall distributions, however, are quite different over mainland Asia: cluster D impacts southern China to northern Vietnam, including Taiwan, while cluster F affects Vietnam, Malaysia, and Thailand.

Cluster E is recurving and has a relatively low number of landfalls (32%), mostly over the northern Philippines, Taiwan, the eastern coast of China, the Korean Peninsula, and Japan. Only 15% of the cyclones in cluster G (with genesis positions farthest east) make landfall. These landfall locations are relatively widespread, occurring in the Philippines, Taiwan, China, Korea, and Japan.

4. Temporal evolution

a. Seasonality

Based on track shape and geographic location, the regression-curve mixture model identifies clusters of TC tracks with quite distinct properties, as described in the previous subsections. A key motivation for this decomposition is to obtain a probabilistic description of the temporal behavior of TC activity. We begin here with the mean seasonal evolution and transition probabilities between clusters. In Part II of this paper, we extend the analysis to interannual variability.

Figure 11 shows the mean number of cyclones per

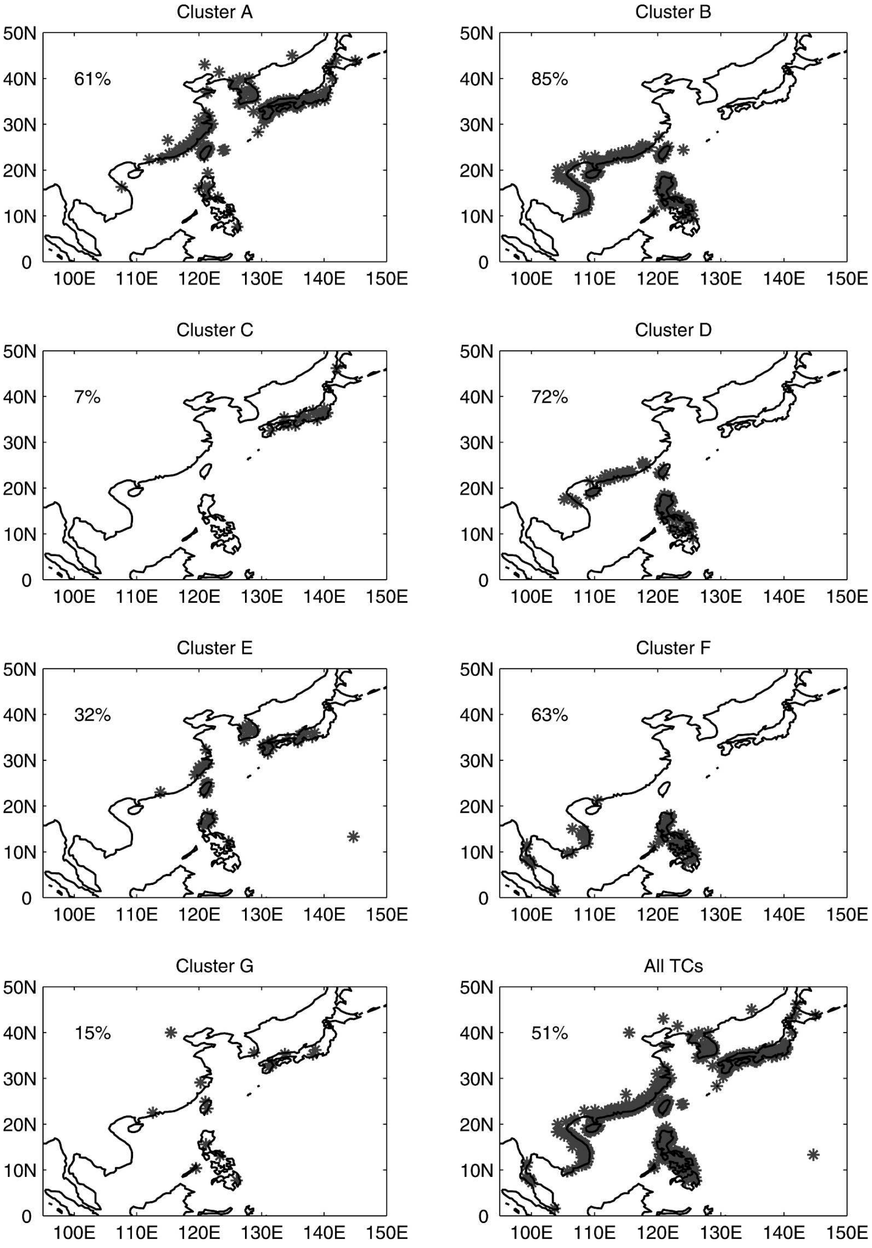

FIG. 10. Landfall location and percentage of cyclones making landfall, in each cluster and for all TCs.

calendar month, for each cluster and for all TCs, based on genesis date. The seasonal evolution of TC activity is remarkably different from cluster to cluster, facilitating therewith a description of the basinwide seasonal cycle. Certain clusters are highly season specific. Among the recurving clusters, cluster A (peak activity in JAS) precedes cluster C (peak in ASO), while cluster E has a smoother evolution. Of the straight movers, cluster B cyclones are almost absent in JFM, and there is little cluster D activity in FMA, and even less cluster F activity in JJAS.

The mean genesis location of the TCs in the western North Pacific has a well-defined annual cycle (Chia and Ropelewski 2002; Camargo et al. 2005), with the average latitude reaching its northernmost position in August and its most equatorward position around February. This cycle is consistent with the seasonal occurrence of the clusters, as the genesis locations vary in latitude among the clusters (see Table 2). Clusters A, C, and to a lesser extent B, have the northernmost genesis positions, and this is reflected in their prevalence during summer and fall, with almost no activity during

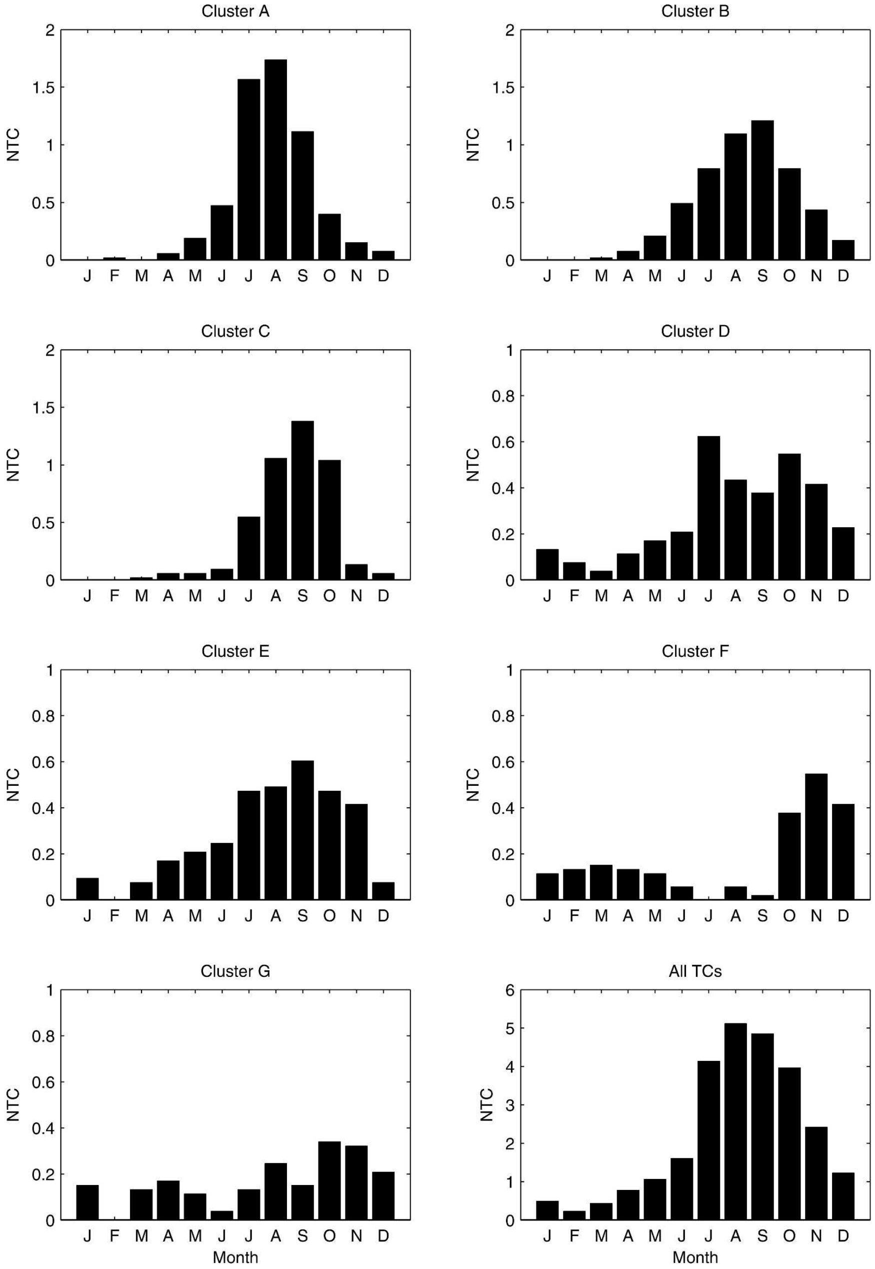

FIG. 11. Mean number of cyclones per month, in each cluster and all TCs.

January–April. The larger spread in cluster B could be explained by more persistent warm sea surface temperatures (SSTs) over its genesis region, namely the South China and Philippines Seas. Because of the presence of the western North Pacific's warm pool, the climatological SSTs in the Philippines Sea are above 26°C year round, and this holds for the South China Sea from May to November too.

The activity in clusters with low-latitude genesis points (D, E, F, and G) is distributed more irregularly throughout much of the year. Cyclones in clusters D and E occur mainly from July to November, with maxima in July and October, or in September, respectively. Most of the TCs in cluster F occur after the peak typhoon season, from October to December, with a peak in November. Clusters F and G have the latest median genesis day among the clusters (not shown), along with the most equatorward genesis latitude among the clusters (see Table 2). The formation of TCs in the late boreal fall and early winter is restricted to

TABLE 3. Transition matrix between cyclone clusters: number of occurrences of a TC in a specific cluster (column) given that a TC in a given cluster occurred (row). Transitions between clusters that are more (less) likely than pure chance at the 95% significance level are in bold (underlined). Significance at the 99% levels is denoted with an asterisk (*). The statistical significance was determined following Vautard et al. (1990).

From/to	A	B	C	D	E	F	G	Sum	Mean	Std dev
A	**103***	70	**64**	26*	27	5*	11*	306	43.7	36.0
B	71	52	**60**	44	24	15	14	280	40.0	22.6
C	59	53	**54***	21	29	5*	14	235	33.6	21.7
D	35	29	17*	**38***	**33**	16	10	178	25.4	10.9
E	22*	36	25	19	**31**	**23***	**20**	176	25.1	6.2
F	6*	20	6*	17	16	**30***	**17***	112	16.0	8.3
G	10*	20	9*	13	15	**18***	**20***	105	15.0	4.5
Sum	306	280	235	178	175	112	106	1392		

the Pacific warm pool and low latitudes. This is due primarily to the presence of high values of vertical wind shear in other parts of the basin.

b. Transitions between clusters

The simplest temporal dependency given by a cluster analysis is in terms of transition probabilities between clusters (Fraedrich and Klauss 1983; Mo and Ghil 1987, 1988). Table 3 is a transition matrix of TC number of occurrences in each cluster (column), given a previous occurrence in another cluster (row). The bottom row gives the TC occurrences in each cluster, as in the number of tropical cyclones (NTC) column of Table 1. The statistical significance that the occurrence of a transition between clusters is more or less likely than pure chance was determined following Vautard et al. (1990).

The main diagonal of Table 3 contains the self-transition occurrences; that is, the number of times two consecutive TCs occur in the same cluster. The self-transition occurrences in all clusters, with the exception of B, are more likely to occur than pure chance at the 95% significance level. In clusters A, D, F, and G the highest values of occurrence are for self-transition, indicating that the same type of cyclone is likeliest to recur, once the conditions are right. Even in the other three clusters, the self-transition occurrences are among the highest. This is probably related to the observed fact that around 5 days after a cyclone has formed, the environmental conditions are favorable to the formation of other TCs (Frank 1982; Holland 1995). If the environmental conditions persist, there is a large chance that the next cyclone track will be in the same cluster as the previous one. A possible mechanism for the formation of TCs at short intervals from each other is wave accumulation (Sobel and Bretherton 1999).

In the case of A (the largest cluster), besides persistence, there is a statistically significant chance that a transition to cluster C occurs, while transitions to clusters D, E, F, and G are less likely (see Table 3). On the other hand, cluster B occurrence (the South China Sea cluster) seems to be independent of both the other clusters and persistence, as none of the transitions are statistically significant. Clusters F and G, though rare, seem to be persistent and occur often following one another, as well as following cluster E.

By examining the sequence of cyclone occurrences within different clusters (not shown), we find that the most populated clusters, especially A and C, seem to exhibit groups of TCs, with several successive cyclones belonging to the same cluster. In contrast, the less populated clusters (F, G) have a clearer interannual variability, with no cyclones at all in some years and several in others. The interannual variability will be further explored in Part II.

5. Concluding remarks

a. Discussion

We have applied a novel clustering methodology to the best-track dataset of tropical cyclones (TCs) over the western North Pacific (Joint Typhoon Warning Center 2005). The analysis included tropical storms as well as typhoons but not tropical depressions. The methodology combined regression modeling of tracks of arbitrary length with mixture modeling of quadratic track shapes obtained by the regression. Our classification resulted in seven clusters labeled A–G. Here, we discuss our results and compare these to previous work, followed by our conclusions in section 5b.

Our main categories of cluster track types are consistent with the two principal track types identified in previous studies (e.g., Sandgathe 1987; Harr and Elsberry 1995; Lander 1996), namely recurving and straight-moving track types. Previous work on western North Pacific TC tracks used manual classification into straight-moving and recurving types (Sandgathe 1987;

Miller et al. 1988; Harr and Elsberry 1991; Lander 1996), and more recently, *K*-means cluster analysis (Elsner and Liu 2003). The latter study analyzed typhoon tracks (excluding tropical storms) based on the typhoon's position at maximum intensity and its final intensity. The three clusters obtained were associated with straight-moving, recurving, and north-oriented tracks. Their straight-moving cluster includes not only South China Sea typhoons (our cluster B), but also typhoons that are classified in our other straight-moving clusters, D, F, and G. The landfall region for their straight-moving cluster includes southern China, the Philippines, and Vietnam, similar to the result of pooling all our straight-moving clusters. Their second cluster is defined as recurving, with landfall predominantly in Japan and Korea. This region falls within the landfall area of clusters A and E, keeping in mind that our study includes more TCs. While our analysis includes tropical storms, only typhoons are included in Elsner and Liu (2003). By including tropical storms in our analysis, our landfall region is probably larger than if we had considered typhoons only, similarly to Elsner and Liu (2003). Finally, their third, north-oriented cluster has very few landfalls and could be associated with our cluster C.

We have also repeated our analysis with tracks of fixed length, using the middle part of the tracks (5 middle days), and excluding TCs that lasted less than 5 days. The clusters so obtained did not have the same characteristics as the ones presented here. For instance, no clear track types—straight or recurving—could be identified for each of the clusters obtained.

Our model allows all tracks to be included in the classification, independent of shape, while in Harr and Elsberry (1991), tropical cyclones classified as "odd" were excluded from their analysis. Our analysis includes TCs forming throughout the year, while Harr and Elsberry (1991) considered only the peak season. We have repeated our analysis with June–November TCs only, and although the clusters are qualitatively similar, the differences in seasonal timing among them (section 4a) were less clear.

Our model did not identify the north-oriented and "S"-shaped tracks identified subjectively in Lander (1996) and found to be associated with a reverse-oriented monsoon trough. This synoptic configuration is relatively rare, occurring on average about once per typhoon season, usually between mid-July and mid-October. Lander (1996) identified only 35 cases of S tracks out of 508 TCs (6.9% of the cases). The cluster methodology is not expected to represent such relatively rare events in terms of mean trajectories, but rather as stochastic excursions from them. The typhoons cited as north oriented or with an S track in Lander (1996) are classified using our cluster algorithm into one of the recurving clusters (A, C, E, and G), with the exception of typhoon Yunia (1994), which was classified in cluster B.

Comparison with the synoptic track types identified subjectively by Lander (1996) serves to emphasize the probabilistic nature of our model. By identifying relatively broad-brush categories, the intent is to isolate types of TC behavior that may contain predictability at the seasonal scale in a probabilistic sense. This aspect is pursued further in Part II. On this time scale, individual synoptic events will not be predictable and will thus need to be treated stochastically.

The regression-model methodology used here nonetheless results in a finer differentiation between track types compared to other previous studies. With respect to genesis location, Harr and Elsberry (1991) showed that TCs with first positions situated north of 20°N, or east of 150°E and north of 10°N, tend to have recurving tracks. This is consistent with our recurving clusters A, C, and E. These three clusters, however, do have distinguishing characteristics that are useful–such as landfall characteristics—but cannot be identified a priori. Harr and Elsberry (1991) also identified the region south of 20°N and west of 135°E as more likely to give rise to straight-moving storms. This region includes our cluster B genesis locations, with tracks that we also classify as straight moving. Our straight-moving clusters D, F, and G, though, have distinct genesis regions. Therefore, the clustering presented here tends to provide greater initial-position separation than Harr and Elsberry's (1991). Another method that has been used to classify TCs is based on clustering of anomalous large-scale circulation patterns and makes use of relationships between track type and these patterns. Harr and Elsberry (1995) used a fuzzy cluster analysis (Mo and Ghil 1987, 1988) in the subspace of leading empirical orthogonal functions of the 700-hPa wind field. A posteriori, TCs were then classified according to these large-scale circulation patterns. Tracks were classified into four types—straight mover, recurve-south, recurve-north, and South China Sea—and tracks not fitting into these four types were discarded. These authors considered only the June–October seasons of nine yr (1979–87), resulting in a much smaller number of tracks—172 as compared to 1393 here. Two of their large-scale clusters are clearly dominated by straight movers and recurving-south tracks (see Table 4 in Harr and Elsberry 1995), two others by recurving-south and recurving-north TCs, and another by recurve-south tracks, while the two remaining clusters are not associated with a dominant track type. The straight tracks associated with one of the clusters in Harr and Elsberry (1995; see their Fig. 8a) are very similar to those in our cluster D. The tracks associated with two of their other clusters (their Figs. 8b and 8c), however, could not be related to our analysis. We will turn to connections between our TC clusters and large-scale circulation anomalies in Part II.

b. Conclusions

In Part I of this two-part study, we have examined the seven clusters obtained by our curve-based mixture model in terms of distributions of tracks, genesis positions, intensity and lifetimes, landfalls, seasonality, and cluster transitions. Any cluster analysis will produce a set of clusters according to the chosen metric, and an important task is to demonstrate that the resulting partition is meaningful. Our metric combines track shape and position, using all recorded 6-hourly latitude–longitude positions of the TC. Additional properties, such as intensity, could have been included in the metric, but were left out on purpose, for verification of the results. By using the entire trajectory, we are able to include the path between the important positions of genesis and landfall. Genesis position clearly plays an important role in the resulting partition. However, genesis location alone cannot be used to determine the landfalling locations, as there is some overlap of the genesis location among the clusters. We found that using only the fixed-length, middle portion of the tracks yielded poorly defined clusters. Trajectory shape is also important, as small differences in the shapes lead to different landfall and impact regions.

Several distinct straight-moving and recurving TC clusters were identified, refining analyses from previous studies. The latitudinal tightness of the three straight-moving clusters B, D, and F is of interest because it suggests the possibility of fairly sharp landfall predictions for these clusters. The recurving clusters clearly exhibit more spread. This is physically plausible because of the much larger intrinsic variability of the midlatitude atmosphere into which recurvers penetrate.

The model partition according to position and shape lead also to high differentiation in storm intensity and seasonally between the clusters. Since neither information was available to the clustering algorithm, this is good evidence that the resulting partition is meaningful. The statistics of intensity of different clusters could be potentially useful in forecasts. Once a tropical cyclone is identified as belonging to one of the clusters, the historical information on the probabilities of certain intensity (i.e., major tropical cyclone) can be used as a guidance to forecasts.

We have argued that the seven clusters are supported by the data, and more evidence of this is presented in Part II, where the accompanying large-scale circulation patterns and the relationship with ENSO are investigated. The analysis yields a differentiated picture of landfall probabilities in terms of distinct trajectory types. If this picture is robust, it could yield new predictors for landfall, several days in advance. The potential advantage of this approach is that we have not keyed the analysis to landfall at a particular location, instead we analyzed the whole western North Pacific simultaneously, so that cyclones with highly variable trajectories can be treated consistently. Since the individual clusters have distinct regional landfall probability distributions, the methodology could form the basis for improved landfall risk maps and probabilistic seasonal forecasts of TC risk.

Acknowledgments. S. J. Camargo would like to thank Anthony G. Barnston and Adam H. Sobel for valuable discussions. This work was support in part by NOAA through a block grant to the International Research Institute for Climate and Society (S. J. Camargo and A. W. Robertson), Department of Energy Grant DE-FG02ER6413 (M. Ghil and A. W. Robertson), and the National Science Foundation under Grants SCI-0225642 and IIS-0431085 (S. J. Gaffney and P. Smyth). M. Ghil also wishes to acknowledge the European Commission's Project No. 12975 (NEST) "Extreme Events: Causes and Consequences (E2-C2)."

APPENDIX

Clustering Trajectories Using Mixtures of Regression Models

Let $\mathbf{z}_i$ be an $n_i \times 2$ matrix of latitude and longitude measurements for TC track i and let $\mathbf{t}_i$ be an $n_i \times 1$ vector of corresponding discrete time indices $\{0, 1, \ldots, n_i - 1\}$.

We model both longitude and latitude with a polynomial regression model of order p (where $p = 2$ for results in this paper), with time $\mathbf{t}_i$ as the independent variable. Under the assumption that TC track i was generated by cluster k we have

$$\mathbf{z}_i = \mathbf{T}_i\beta_k + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sum\nolimits_k\right). \tag{A1}$$

Here $\mathbf{T}_i$ is the $n_i \times (p + 1)$ Vandermonde regression matrix associated with the vector $\mathbf{t}_i$, defined as $(p + 1)$ columns corresponding to $\mathbf{t}_i$ such that the components of $\mathbf{t}_i$ in the mth column are taken to the power of m for $0 \leq m \leq p$. Here, β_k is a $(p + 1) \times 2$ matrix of regression coefficients for cluster k, containing 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 Gaussian noise, with zero mean and a 2×2 covariance matrix Σ_k. The covariance matrix Σ_k contains diagonal elements σ_{k1}^2 and σ_{k2}^2, which are the noise variances for each longitude and latitude observation, respectively. The cross covariance is set to 0 for simplicity.

The conditional density for the ith cyclone, conditioned on membership in the kth cluster, is thus defined as

$$p(\mathbf{z}_i|\mathbf{t}_i, \theta_k) = f(\mathbf{z}_i|\mathbf{T}_i\beta_k, \Sigma_k) = (2\pi)^{-n_i}|\Sigma_k|^{-n_i/2} \exp\left\{-\frac{1}{2}\mathrm{tr}[(\mathbf{z}_i - \mathbf{T}_i\beta_k)\Sigma_k^{-1}(\mathbf{z}_i - \mathbf{T}_i\beta_k)']\right\}, \tag{A2}$$

where $\theta_k = \{\beta_k, \Sigma_k\}$.

This results in the following unconditional (unconditional on k) regression mixture model with K clusters:

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

where α_k is the probability of a randomly selected TC track belonging to cluster k (and $\Sigma_k\ \alpha_k = 1$) and ϕ represents the overall set of mixture parameters (β_k, Σ_k, and α_k; $1 \le k \le K$).

If we let $\mathbf{Z} = \{z_1, \ldots, z_n\}$ be the complete set of n cyclone trajectories and $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ be the set of associated measurement times, then the full probability density of Z given $\mathbf{T}$, also known as the conditional likelihood, is

$$p(\mathbf{Z}|\mathbf{T}, \phi) = \prod_i^n \sum_k^K \alpha_k f_k(\mathbf{z}_i|\mathbf{T}_i\beta_k, \Sigma_k). \tag{A4}$$

Clustering is performed by maximizing this likelihood expression to find estimates of the parameters ϕ given data. The EM algorithm (described below) provides an iterative algorithm for finding local (not necessarily global) maxima of the likelihood. Given a learned model one can then infer for each TC track which of the K models it is most likely to be associated with.

Let $N = \Sigma_i^n n_i$, let $\mathbf{Y}$ be the $N \times 2$ concatenated matrix $(\mathbf{y}_1', \ldots, \mathbf{y}_n')'$, and let $\mathbf{X}$ be the $N \times (p + 1)$ concatenated matrix $(\mathbf{X}_1', \ldots, \mathbf{X}_n')'$ of regression matrices. In the E step, we calculate the *membership* probability,

$$w_{ik} = \frac{\alpha_k f_k(\mathbf{z}_i|\mathbf{T}_i\beta_k, \Sigma_k)}{\Sigma_j^K \alpha_j f_j(\mathbf{z}_i|\mathbf{T}_i\beta_j, \Sigma_j)}, \tag{A5}$$

that trajectory i was generated from cluster k. Note that w_{ik} is equal to the ratio of the likelihood of trajectory i under cluster k, to the total likelihood of trajectory i under all clusters. Let $\mathbf{w}_{ik} = w_{ik}\ \mathbf{I}_{ni}$, where $\mathbf{I}_{ni}$ is an n_i vector of ones. Let $\mathbf{W}_k = \mathrm{diag}(\mathbf{w}_{1k}', \ldots, \mathbf{w}_{nk}')$ be an $N \times N$ diagonal matrix. In the **M** step we use $\mathbf{W}_k$ to calculate the mixture parameters

$$\hat{\beta}_k = (\mathbf{X}'\mathbf{W}_k\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_k\mathbf{Y}, \tag{A6}$$

$$\hat{\Sigma}_k = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}_k)'\mathbf{W}_k(\mathbf{Y} - \mathbf{X}\hat{\beta}_k)}{\sum_i^n w_{ik}}, \tag{A7}$$

and

$$\hat{\alpha}_k = \frac{1}{n}\sum_i^n w_{ik}.$$

These updated equations are equivalent to the well-known weighted least squares solution in regression (Draper and Smith 1981). The diagonal elements of $\mathbf{W}_k$ represent the weights to be applied to $\mathbf{Y}$ and $\mathbf{X}$ during the weighted regression.

The EM algorithm iterates through pairs of E and M steps until convergence. For the results in this paper the algorithm was initialized by randomly selecting a set of membership weights $\mathbf{W}_k$ and then executing an M step. Convergence was detected when the ratio of the incremental change in log-likelihood of the current iteration to the change from the second iteration drops below a threshold of 10^{-8}. To avoid poor local maxima in parameter space the highest likelihood solution obtained from 10 starts of EM was chosen, where for each run the algorithm was started with a different randomly selected set of membership weights $\mathbf{W}_k$.

REFERENCES

Adem, J., 1956: A series solution for the barotropic vorticity equation and its application in the study of atmospheric vortices. *Tellus,* **8,** 364–372.

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

Camargo, S. J., and A. H. Sobel, 2005: Western North Pacific tropical cyclone intensity and ENSO. *J. Climate,* **18,** 2996–3006.

——, A. W. Robertson, S. J. Gaffney, and P. Smyth, 2004: Cluster analysis of western North Pacific tropical cyclone tracks.

Proc. 26th Conf. on Hurricanes and Tropical Meteorology, Miami, FL, Amer. Meteor. Soc., 250–251.

——, A. G. Barnston, and S. E. Zebiak, 2005: A statistical assessment of tropical cyclones in atmospheric general circulation models. *Tellus,* **57A,** 589–604.

——, A. W. Robertson, S. J. Gaffney, P. Smyth, and M. Ghil, 2007: Cluster analysis of typhoon tracks. Part II: Large-scale circulation and ENSO. *J. Climate,* **20,** 3654–3676.

Chan, J. C. L., 1985: Tropical cyclone activity in the Northwest Pacific in relation to the El Niño/Southern Oscillation phenomenon. *Mon. Wea. Rev.,* **113,** 599–606.

——, 2005: The physics of tropical cyclone motion. *Annu. Rev. Fluid Mech.,* **37,** 99–128.

——, and W. M. Gray, 1982: Tropical cyclone movement and surrounding flow relationships. *Mon. Wea. Rev.,* **110,** 1354–1374.

Cheng, X. H., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.,* **50,** 2674–2696.

Chia, H. H., and C. F. Ropelewski, 2002: The interannual variability in the genesis location of tropical cyclones in the northwest Pacific. *J. Climate,* **15,** 2934–2944.

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

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis.* 2d ed. John Wiley and Sons, 709 pp.

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

——, and K. B. Liu, 2003: Examining the ENSO-typhoon hypothesis. *Climate Res.,* **25,** 43–54.

Everitt, B. S., and D. J. Hand, 1981: *Finite Mixture Distributions.* Chapman and Hall, 143 pp.

Fraedrich, K., and M. Klauss, 1983: On single station forecasting: Sunshine and rainfall Markov chains. *Beitr. Phys. Atmos.,* **56,** 108–134.

Frank, W. M., 1982: Large-scale characteristics of tropical cyclones. *Mon. Wea. Rev.,* **110,** 572–586.

Franklin, J. L., S. E. Feuer, J. Kaplan, and S. D. Aberson, 1996: Tropical cyclone motion and surrounding flow relationships: Searching for beta gyres in omega dropwindsond datasets. *Mon. Wea. Rev.,* **124,** 64–84.

Gaffney, S. J., 2004: Probabilistic curve-aligned clustering and prediction with regression mixture models. Ph.D. thesis, University of California, Irvine, CA, 281 pp. [Available online at http://www.ics.uci.edu/pub/sgaffney/outgoing/sgaffney_thesis.pdf.]

——, and P. Smyth, 1999: Trajectory clustering with mixture of regression models. *Proc. Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining,* San Diego, CA, Association for Computing Machinery, 63–72.

——, and ——, 2005: Joint probabilistic curve-clustering and alignment. *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference,* L. K. Saul, Y. Weiss, and L. Bottou, Eds., MIT Press, 473–580.

——, A. W. Robertson, P. Smyth, S. J. Camargo, and M. Ghil, 2007: Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dyn.,* in press.

Ghil, M., and S. Childress, 1987: *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics.* Springer-Verlag, 485 pp.

——, and A. W. Robertson, 2002: "Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Natl. Acad. Sci. USA,* **99** (Suppl. 1), 2493–2500.

Gray, W. M., 1979: Hurricanes: Their formation, structure and likely role in the tropical circulation. *Meteorology over the Tropical Oceans,* D. B. Shaw, Ed., Royal Meteorological Society, 155–218.

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

Harr, P. A., and R. L. Elsberry, 1991: Tropical cyclone track characteristics as a function of large-scale circulation anomalies. *Mon. Wea. Rev.,* **119,** 1448–1468.

——, and ——, 1995: Large-scale circulation variability over the tropical western North Pacific. Part I: Spatial patterns and tropical cyclone characteristics. *Mon. Wea. Rev.,* **123,** 1225–1246.

Ho, C.-H., J.-J. Baik, J.-H. Kim, and D.-Y. Gong, 2004: Interdecadal changes in summertime typhoon tracks. *J. Climate,* **17,** 1767–1776.

Holland, G. J., 1983: Tropical cyclone motion: Environmental interaction plus a beta effect. *J. Atmos. Sci.,* **40,** 328–342.

——, 1995: Scale interaction in the western Pacific monsoon. *Meteor. Atmos. Sci.,* **56,** 57–79.

Joint Typhoon Warning Center, cited 2005: JTWC (Joint Typhoon Warning Center) best track dataset. [Available online at https://metoc.npmoc.navy.mil/jtwc/best-tracks/.]

Kimoto, M., and M. Ghil, 1993a: Multiple flow regimes in the northern hemisphere winter. Part I: Methodology and hemispheric regimes. *J. Atmos. Sci.,* **50,** 2625–2644.

——, and ——, 1993b: Multiple flow regimes in the northern hemisphere winter. Part II: Sectorial regimes and preferred transitions. *J. Atmos. Sci.,* **50,** 2645–2673.

Krishnamurti, T. N., J. Molinari, H. L. Pan, and V. Wong, 1977: Downstream amplification and formation of monsoon disturbances. *Mon. Wea. Rev.,* **105,** 1281–1297.

Kumar, V., and R. Krishnan, 2005: On the association between the Indian summer monsoon and the tropical cyclone activity over northwest Pacific. *Curr. Sci.,* **88,** 602–612.

Lander, M. A., 1996: Specific tropical cyclone tracks and unusual tropical cyclone motions associated with a reverse-oriented monsoon trough in the western North Pacific. *Wea. Forecasting,* **11,** 170–186.

Legras, B., and M. Ghil, 1985: Persistent anomalies, blocking, and variations in atmospheric predictability. *J. Atmos. Sci.,* **42,** 433–471.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability,* Berkeley, CA, University of California, 281–297.

McLachlan, G. J., and K. E. Basford, 1988: *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, 253 pp.

——, and T. Krishnan, 1997: *The EM Algorithm and Extensions.* John Wiley and Sons, 274 pp.

——, and D. Peel, 2000: *Finite Mixture Models.* John Wiley & Sons, 419 pp.

Miller, R. J., T. L. Tsui, and A. J. Schrader, 1988: Climatology of North Pacific tropical cyclone tracks. NAVENVPREDRSCHFAC Tech. Rep. TR 88-10, Naval Oceanographic and Atmospheric Research Laboratory, Monterey, CA, 511 pp.

Mo, K. C., and M. Ghil, 1987: Statistics and dynamics of persistent anomalies. *J. Atmos. Sci.,* **44,** 877–901.

——, and M. Ghil, 1988: Cluster analysis of multiple planetary regimes. *J. Geophys. Res.,* **93D,** 10 927–10 952.

Molteni, F., S. Tibaldi, and T. N. Palmer, 1990: Regimes in the wintertime circulation over northern extratropics. I: Observational evidence. *Quart. J. Roy. Meteor. Soc.,* **116,** 31–67.

Reinhold, B. B., and R. T. Pierrehumbert, 1982: Dynamics of weather regimes: Quasi-stationary waves and blocking. *Mon. Wea. Rev.,* **110,** 1105–1145.

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

Saffir, H. S., 1977: Design and construction requirements for hurricane resistant construction. ASCE Tech. Rep. Preprint 2830, 20 pp. [Available from American Society of Civil Engineers, New York, NY 10017.]

Saha, K., F. Sanders, and J. Shukla, 1981: Westward propagating predecessors of monsoon depressions. *Mon. Wea. Rev.,* **109,** 330–343.

Sandgathe, S. A., 1987: Opportunities for tropical cyclone motion research in the northwest Pacific region. Tech. Rep. NPS-63-87-006, Naval Postgraduate School, Monterey, CA, 36 pp.

Simpson, R. H., and H. Riehl, 1981: *The Hurricane and Its Impact.* Louisiana State University Press, 398 pp.

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.

Sobel, A. H., and C. S. Bretherton, 1999: Development of synoptic-scale disturbances over summertime tropical northwest Pacific. *J. Atmos. Sci.,* **56,** 3106–3127.

Vautard, R., 1990: Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. *Mon. Wea. Rev.,* **118,** 2056–2081.

——, K. C. Mo, and M. Ghil, 1990: Statistical significance test for transition matrices in atmospheric Markov chains. *J. Atmos. Sci.,* **47,** 1926–1931.

Wu, C.-C., and K. A. Emanuel, 1995: Potential vorticity diagnostics of hurricane movement. Part I: A case study of Hurricane Bob (1991). *Mon. Wea. Rev.,* **123,** 69–92.

——, and Y.-H. Kuo, 1999: Typhoons affecting Taiwan: Current understanding and future challenges. *Bull. Amer. Meteor. Soc.,* **80,** 67–80.

Wu, L., and B. Wang, 2004: Assessing impacts of global warming on tropical cyclone tracks. *J. Climate,* **17,** 1686–1698.