Cluster Analysis of Typhoon Tracks.

Part I: General Properties

Suzana J. Camargo[1,*], Andrew W. Robertson[1], Scott J. Gaffney[2],

Padhraic Smyth[2], and Michael Ghil[3,†]

[1]International Research Institute for Climate and Society,

The Earth Institute at Columbia University,

Lamont Campus, Palisades, NY

[2]Department of Computer Science,

University of California, Irvine, CA

[3]Department of Atmospheric and Oceanic Sciences

and Institute for Geophysics and Planetary Physics,

University of California, Los Angeles, CA

[*]Corresponding author: Dr. Suzana J. Camargo, IRI - Monell 225, 61 Route 9W, Palisades, NY 10964-8000. Phone: +1 845 680 4416, Fax: +1 845 680 4865, E-mail: suzana@iri.columbia.edu.

[†]Additional affiliation: Département Terre-Atmosphère-Océan and Laboratoire de Météorologie Dynamique du CNRS/IPSL, Ecole Normale Supérieure, Paris, France.

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 data set 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 socio-economic 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 dateline, between about 8°N 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," i.e. slow down, turn northward, and then accelerate eastward as they enter the midlatitude westerlies (e.g. Harr and Elsberry 1995). A further fraction track northward over the ocean, posing no threat to land.

The large-scale circulation of the atmosphere has a 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 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 timescales 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 data set 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, 2003; Gaffney 2004). Part II examines 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 therms 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, 3 clusters were chosen to describe the trajectories. The *K*-means approach has also been used to cluster North Atlantic extra-tropical 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; providing 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.

The clustering methodology is briefly described in Section 2, and applied to the JTWC best track data set 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 the companion paper, Camargo et al. (2005b), 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 data set available at 6-hourly sampling frequency over the time interval 1950–2002 (JTWC 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°–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 (TS), typhoons (TY; categories 1 and 2 in the Saffir-Simpson scale (Saffir 1977; Simpson and Riehl 1981), and intense typhoons (ITY; categories 3–5); tropical depressions are not included in the analysis.

The observed data quality is thought to be considerably poorer during pre-satellite years (pre-1970). We assume that although some of the TCs may be missing in the JTWC (2005) database for the pre-satellite data, especially those that remain over the ocean, the tracks for those that do appear in the data set 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. A more complete discussion is given by Gaffney (2004); with an application of the clustering method to extratropical cyclones over the North Atlantic.[1]

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 densities, including multi-modal ones by 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 functions model a cyclone's longitudinal position versus time using a quadratic polynomial regression, as discussed in Gaffney (2004). 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 interpretation and goodness-of-fit. Separately (and simultaneously), a similar regression is also used for the latitudinal positions. The

[1] A Matlab toolbox with the clustering algorithms described in this paper is available at http://www.ics.uci.edu/software/CCT/ .

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.

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 (a) learn the parameters of all K models given the TC tracks; and (b) infer which of the K models are most likely to have generated each TC track. Each track can be assigned to the mixture component (or cluster) that is most likely (under the model) to have generated that track, that is the cluster which 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 Peel 2000). The resulting EM algorithm is straightforward to implement and use and its computational complexity is linear in the number of observations.

Certain pre-processing 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 track so that they all begin at the latitude-longitude position of $(0, 0)$; these authors also normalize the latitude and longitude measurements to have the same variance. In our experiments below we did not use any such pre-processing—clustering the tracks directly produced results that were easier to interpret and more meaningful than the clustering of pre-processed tracks.

c. Number of clusters

In selecting the number of clusters, we used first the in-sample log-likelihood values (Fig. 1), defined as the log-probability of the observed data under the model, akin to a goodness of fit metric for probabilistic models. As is often the case, however, the log-likelihood values keep increasing with the number of clusters. Out-of-sample log-likelihood values result in a very similar curve (not shown). As additional measure of goodness of fit, the within-cluster-spread, is plotted in Fig. 2. The curves in Figs. 1 and 2 mirror each other and both show diminishing returns in terms of improvement in fit beyond $K = 6 - 8$, suggestion this as a reasonable stopping point.

To evaluate $K = 6-8$ 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 6 clusters (Camargo et al. 2004b) are very similar to those presented here. The main difference is that one of the track-types for $K = 6$ splits in two for $K = 7$, with slightly different characteristics. Most of the results presented here and in Camargo et al. (2005b) are not sensitive to choices between $K = 6 - 8$. As described by Camargo et al. (2005b), 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 of number of clusters, for $K = 2-7$, affects the final regression curves. To emphasize differences in shape, the mean regression trajectories are plotted with their initial positions co-located 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 (panels *b*, *c* and *f* in Fig. 3). 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 plotted have the same number of points, the points are nearer or farther from each other in the shorter vs. longer trajectories. 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.

For the chosen number $K = 7$ of clusters, we show the regression trajectories in Fig. 4; in this case, the initial positions were retained. There are 4 clusters of straight-movers and 3 of recurvers. Notice the strong separation between the two set of 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 dateline; the other two (both recurvers) start near 20°N.

Looking at the population of each cluster in Table 1, we see there are 3 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 years, 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 clusters 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 (TC) 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 the same time interval are also shown (Fig. 5(h)). 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. 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 leads to a TC activity that 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° further 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 they stay closer to the equator in angle as well as position. 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 travelling 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 clusters contains a fair number of cyclones that recurve. The TC activity in cluster G extends, therefore, much wider region than in the case of 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, before is that, before reaching more northern latitudes, these tracks follow stay quite close to the mean regression track.

In order to quantify the differences in track-direction between the clusters, we plot kernel distributions of tangential direction in Fig. 6; the distributions are based on all neighboring pairs of points along each trajectory and were produced using a 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 hr later, when the angle this vector makes with the east lies between $157.5^\circ - 202.5^\circ$ (i.e., $180^\circ \pm 22.5^\circ$).

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 classfication 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 number of cyclones originating in each 2° latitude $\times 2^\circ$ 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 further northwestward. The individual clusters partition genesis position even more than in track density, resulting in very little overlap between clusters. Clearly, the genesis position plays a large role in defining the clusters, although is not given any particular weight in the clustering algorithm itself.

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 3 clusters are the most highly populated ones and together comprise 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 to the east compared to cluster D, and to the south compared to cluster C.

The two least populated clusters are F and G, whose 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 clima-

tological basis, about one-third of 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 clusters have very different genesis locations, with cluster E forming much further south (see Fig. 7). Cluster G also contains a large number of intense typhoons (59%); like cluster E it has a a genesis region near 10°N and 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 long 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 largest in this cluster. This relationship of intensity and life-time 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.

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 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.

Examining the regional landfall distributions in Fig. 10, 61% of the tropical cyclones in cluster A (recurving) make landfall in the Philippines, China, Taiwan, 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, eastern coast of China, Korean Peninsula, and Japan. Only 15% of the cyclones in cluster G (with genesis positions furthest east) make landfall. 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 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 basin-wide 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 ae almost absent in JFM, 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. 2005a), 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 latitutde 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 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. Due to the presence of the western North Pacific's warm pool the climatological SSTs in the Philippines Sea are above $26^{\circ}C$ 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. Cluster 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 the regions of the Pacific warm pool and low latitudes; since the SSTs are too cold elsewhere.

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 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-transitions occurrences in all clusters, with 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, the South China Sea cluster, occurrence seems to be independent of any of the other clusters and even 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 cluster (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 data set of tropical cyclones (TCs) over the western North Pacific (JTWC 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. We compare these results with previous ones in the next subsection and draw conclusions in the following one.

Previous work on western North Pacific TC tracks used manual classification into straight-moving and recurving types (Miller et al. 1988; Harr and Elsberry 1991), 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 3 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 in our analysis includes more TCs; our including tropical storms, too, probably expanding the landfall region. 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, 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 4*a*) were less clear.

The regression-model methodology used here results in a finer differentiation between track types. 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 characterisitcs — 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 1979–1987 years, 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, 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. 8(a)) are very similar to those in our cluster D. The tracks associated with

two of their other clusters (their Fig. 8(b) and (c)), 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 7 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; 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 previous analyses. The tightness of the three straight moving clusters B, D and F is of interest, because it suggests the possibility of fairly sharp landfall predictons 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 seasonality between the clusters. Since neither information was available to the clustering algorithm, this is good evidence that the resulting partition is meaningful.

We have argued that the 7 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, 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.

Acknowledgements SJC 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 (SJC and AWR), Department of Energy grant DE-FG02ER6413 (MG and AWR), and the National Science Foundation under grants No. SCI-0225642 and IIS-0431085 (SJG and PS). MG also wishes to acknowledge the European Commission's Project No. 12975 (NEST) "Extreme events: Causes and consequences (E2-C2)".

References

Adem, J., 1956: A series solutions 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. Royal 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. G. Barnston, and S. E. Zebiak, 2005a: A statistical assessment of tropical cyclones in atmospheric general circulation models. *Tellus*, **57A**, 589–604.

—, A. W. Robertson, S. J. Gaffney, and P. Smyth, 2004: Cluster analysis of western North Pacific tropical cyclone tracks, in Proc. of the 26th Conf. on Hurricanes and Tropical Meteorology, pp. 250–251, American Meteorological Society, Miami, FL.

—, —, —, —, and M. Ghil, 2005b: Cluster analysis of typhoon tracks, Part II: Large scale circulation and ENSO, submitted to *J. Climate*.

Chan, J. C. L., 1985: Tropical cyclone activity in the Northwest Pacific in relation to 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. Classification*, **5**, 249–282.

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

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

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

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

Frank, W. M., 1982: Large-scale characteristics of tropical cyclones. *J. Atmos. Sci.*, **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, Departament of Information and Computer Science, University of California, Irvine, CA (available on line 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 International Conf. Knowledge Discovery and Data Mining*, S. Chaudhuri and D. Madigan, eds., ACM Press, New York, NY, 63–72.

—, and —, 2003: Curve clustering with random effects regression mixtures. *Proc. of the Ninth International Workshop Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, eds., Key West, FL, ISBN 0-9727358-0-1.

Ghil, M., and S. Childress, 1987: *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics*. Springer-Verlag, New York/Berlin/London/Paris/Tokyo, 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.*, **99 (Suppl. 1)**, 2493–2500.

Gray, W. M., 1979: *Meteorology over the Tropical Oceans*, Roy. Meteor. Soc., chapter Hurricanes: Their formation, structure and likely role in the tropical circulation. 155–218.

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.

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.

JTWC, 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. *Current Science*, **88**, 602–612.

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. om Mathematical Statistics and Probability*, University of California, Berkeley, CA, 281–297.

McLachlan, G. J., and K. E. Basford, 1988: *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York, NY.

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

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 93943, 511 pp.

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

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

Molteni, F., S. Tibaldi, and T. N. Palmer, 1990: Regimes in the wintertime circulation over northern extratropics I: Observational evidence. *Quart. J. Royal 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. Technical Report Preprint No. 2830, 20 pp., ASCE, 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.

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., 1999: Typhoons affecting Taiwan: Current understanding and fugure challenges. *Bull. Amer. Meteor. Soc.*, **80**, 67–80.

—, 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.

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

List of Figures

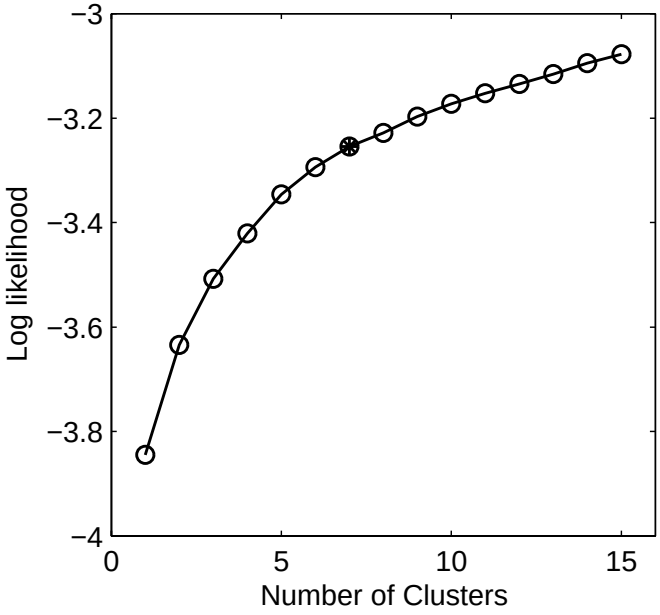

Figure 1: Log-likelihood values for different number of tropical cyclone (TC) track clusters. The log-likehood avlues shown are the maximum of 16 runs, obtained by a random permutation of the tropical cyclones given to the cluster model.

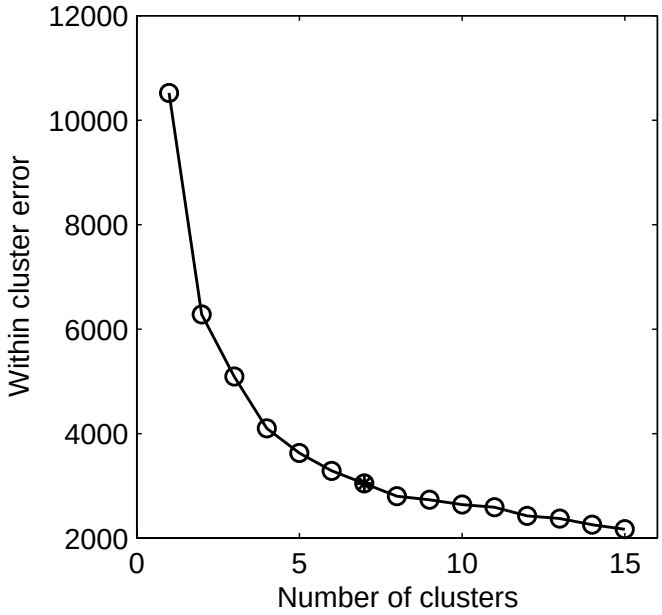

Figure 2: Within cluster error for different number of tropical cyclone (TC) track clusters. The cluster error values shown are the minimum of 16 runs, obtained by a random permutation of the tropical cyclones given to the cluster model.

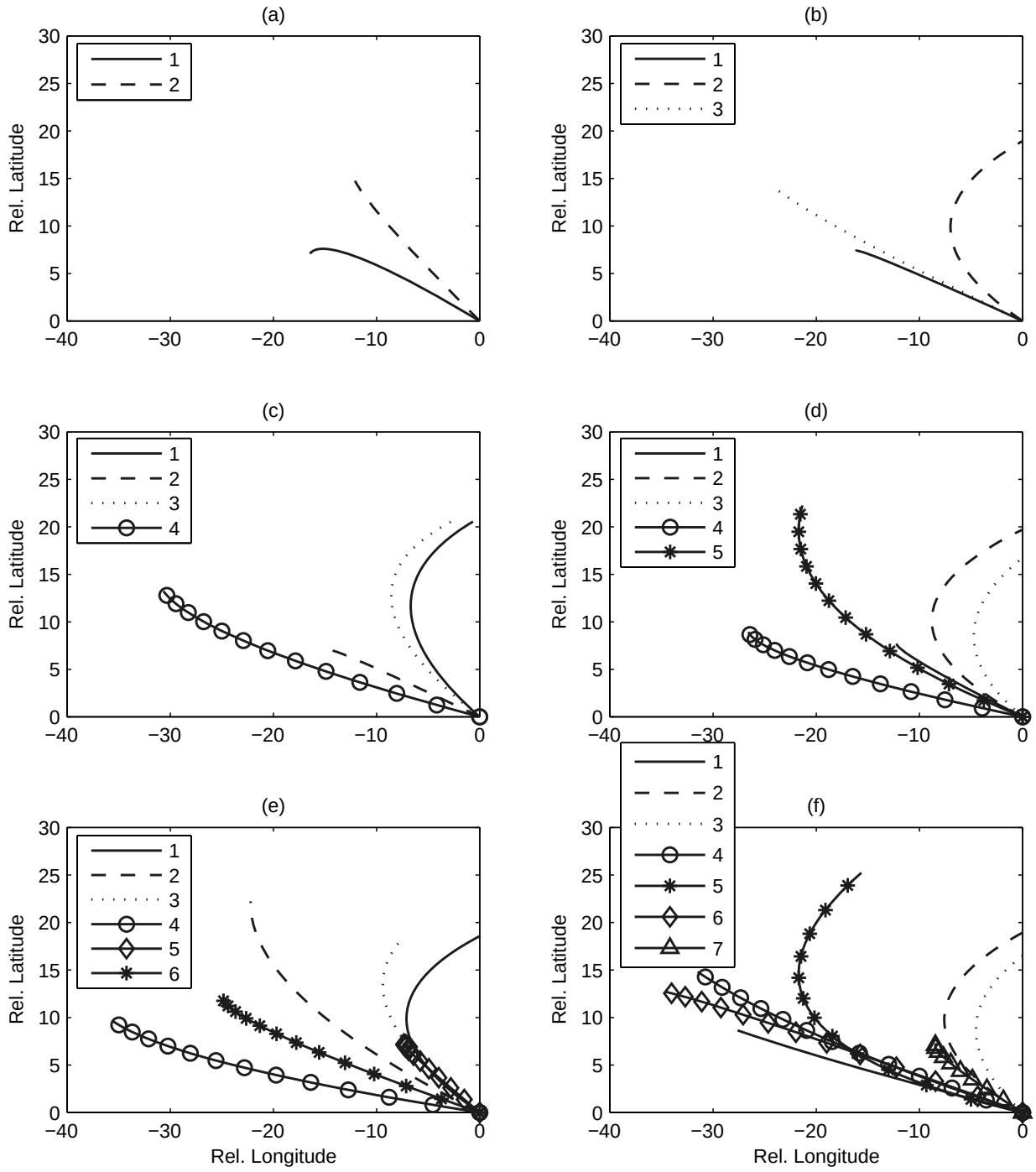

Figure 3: Mean regression trajectories of the western North Pacific TCs with 2,3,4,5,6 and 7 clusters. The mean trajectories start at zero latitude and longitude, for plotting purposes only.

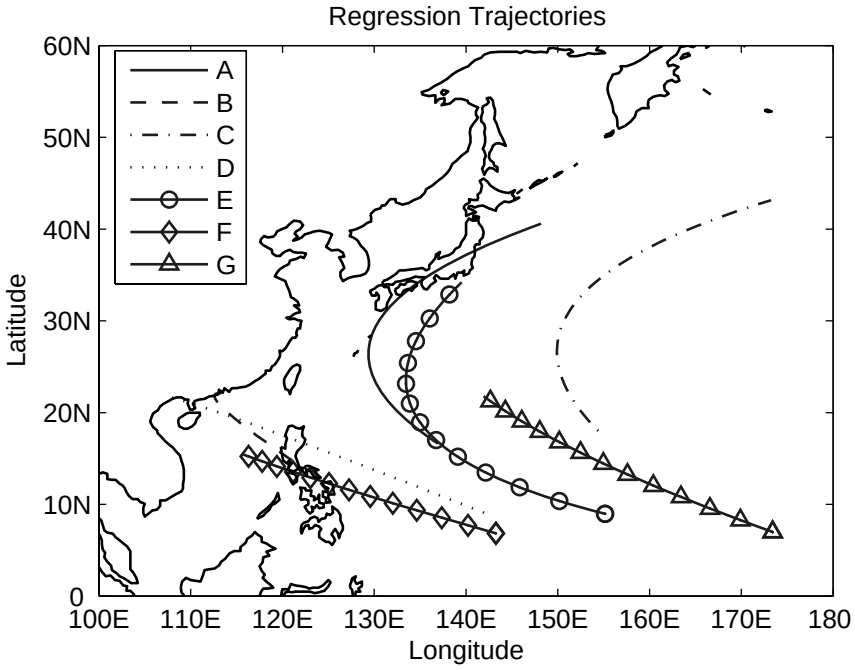

Figure 4: Mean regression trajectories of the TCs over the western North Pacific, with $K = 7$ clusters.

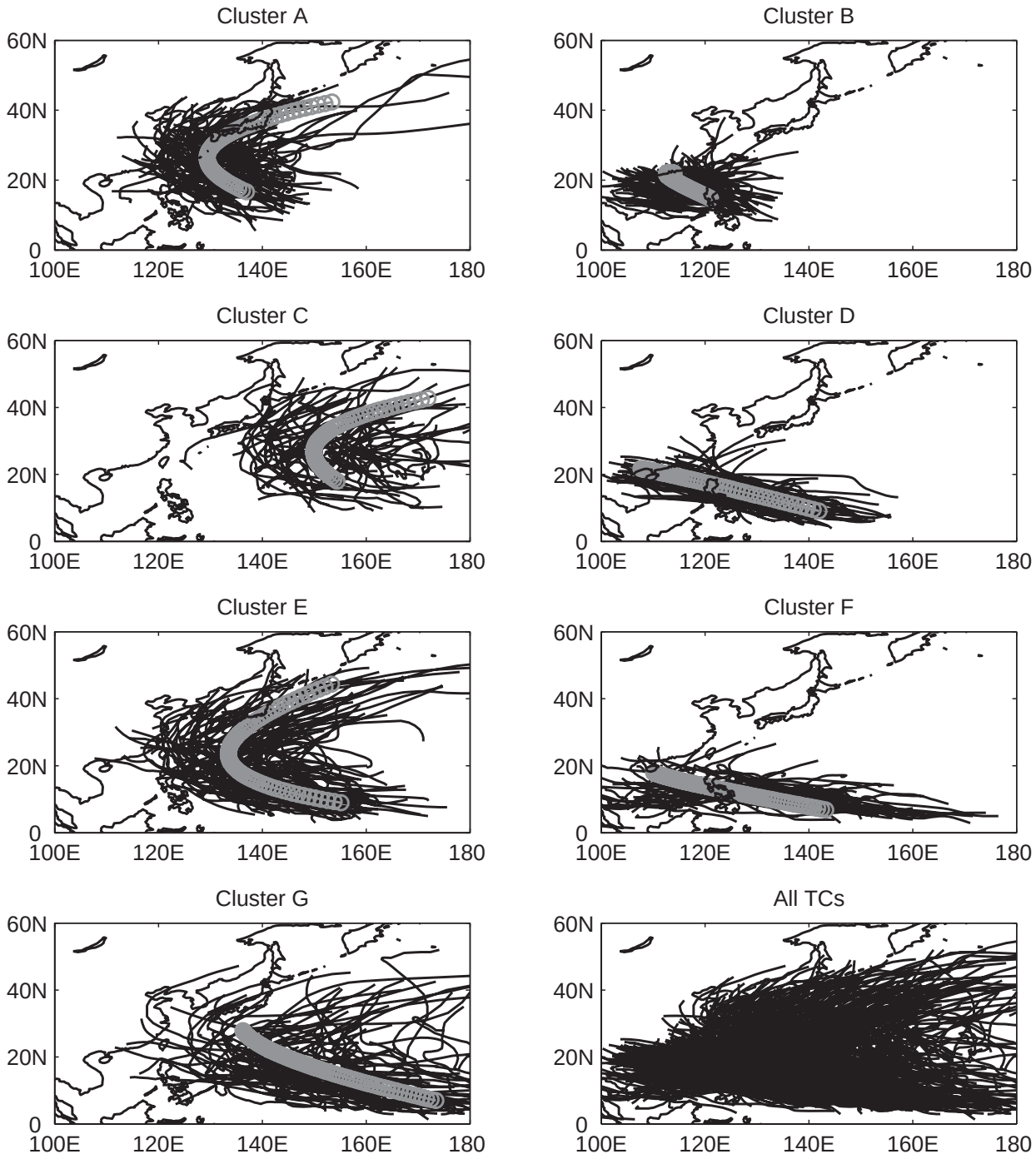

Figure 5: TC tracks (black) over the Western North Pacific, during the period 1983–2002 in each of the 7 clusters and for all TCs; the mean regression curve of each cluster is shown in gray open circles.

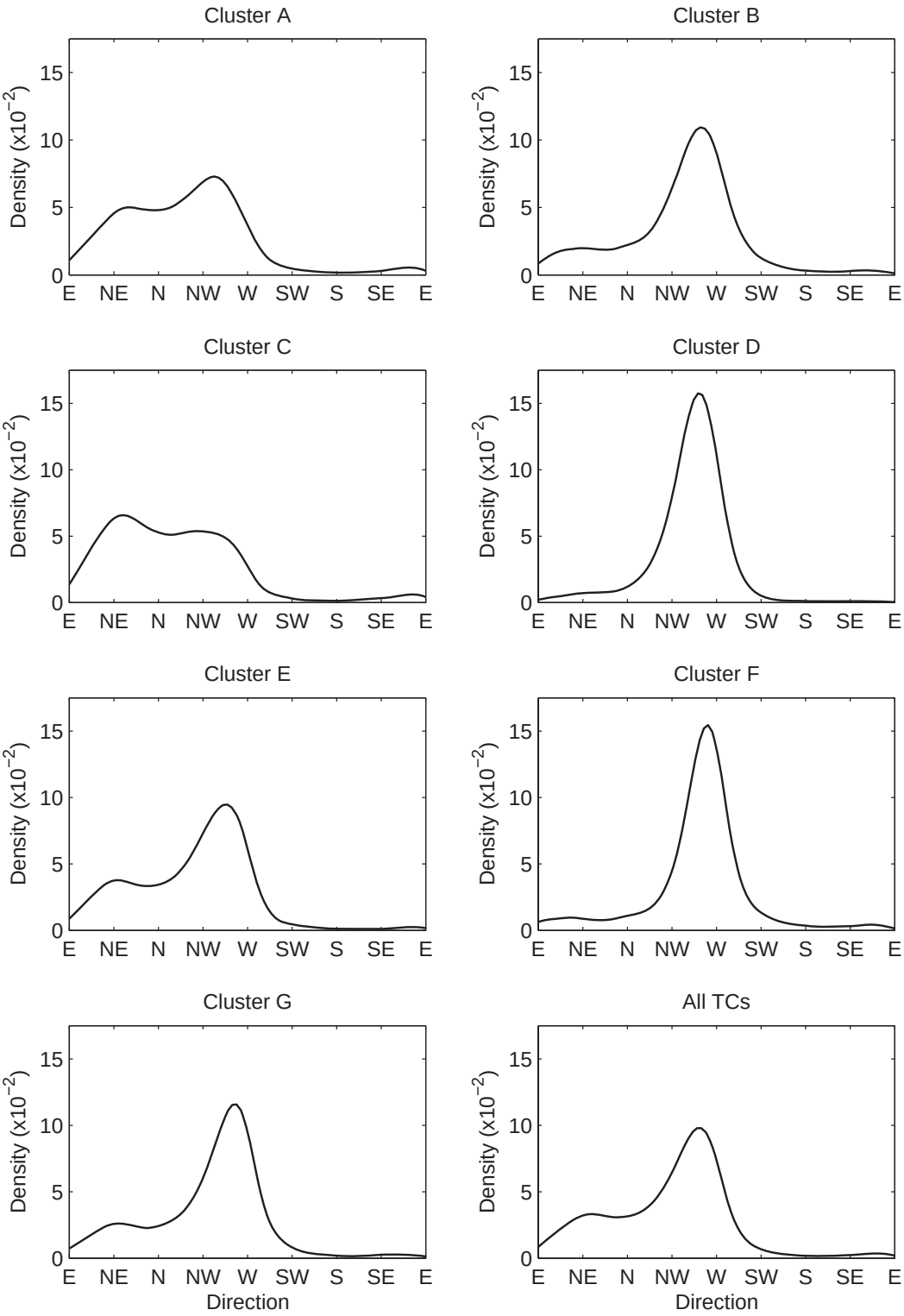

Figure 6: Distribution of TC angles of movement (as function of compass points – E, NE, N, NW, W, SW, W, SE –) for each cluster and for all TCs.

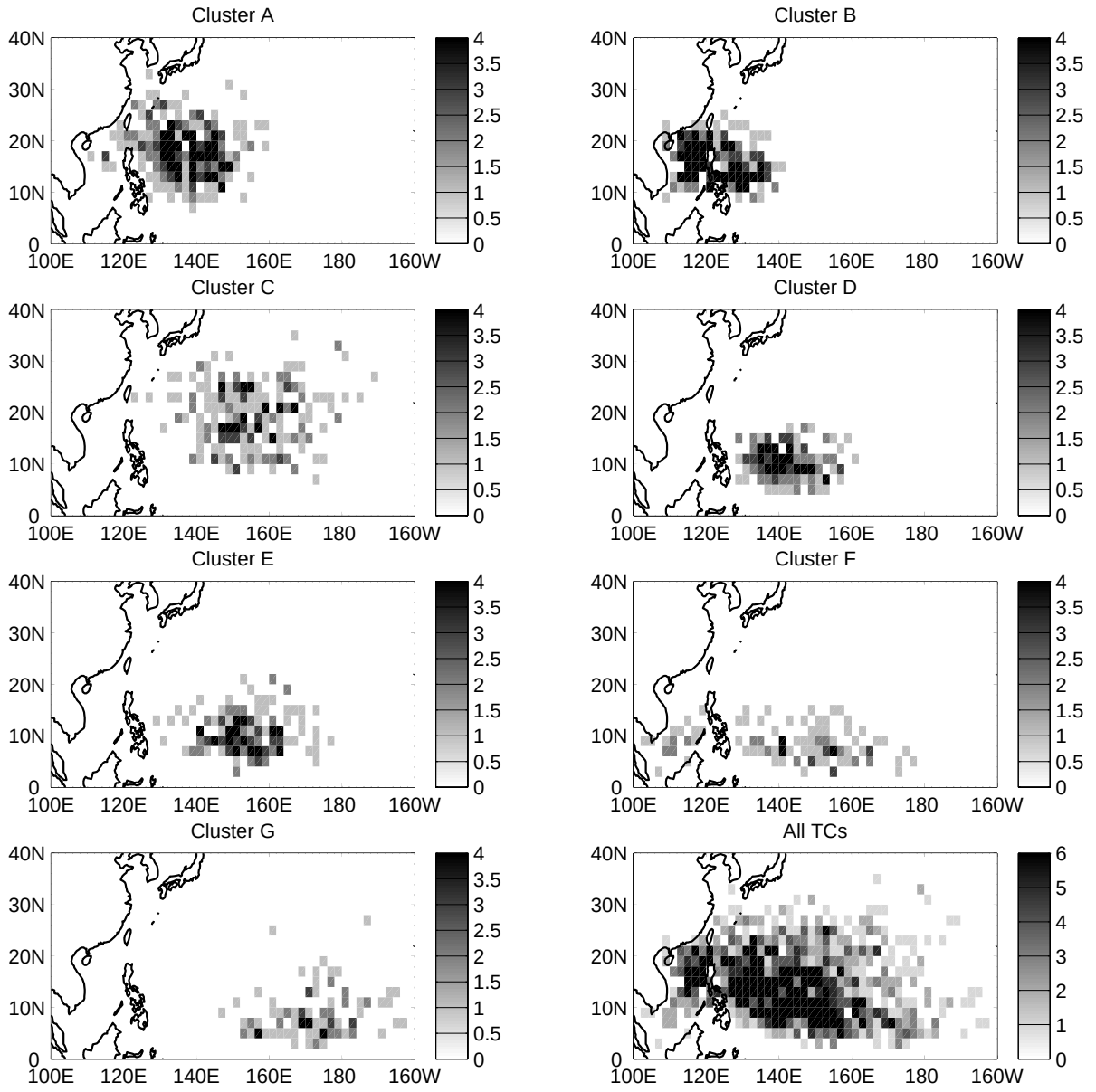

Figure 7: Number of cyclones with genesis position in each 2° latitude $\times 2^\circ$ longitude square for each cluster and all TCs.

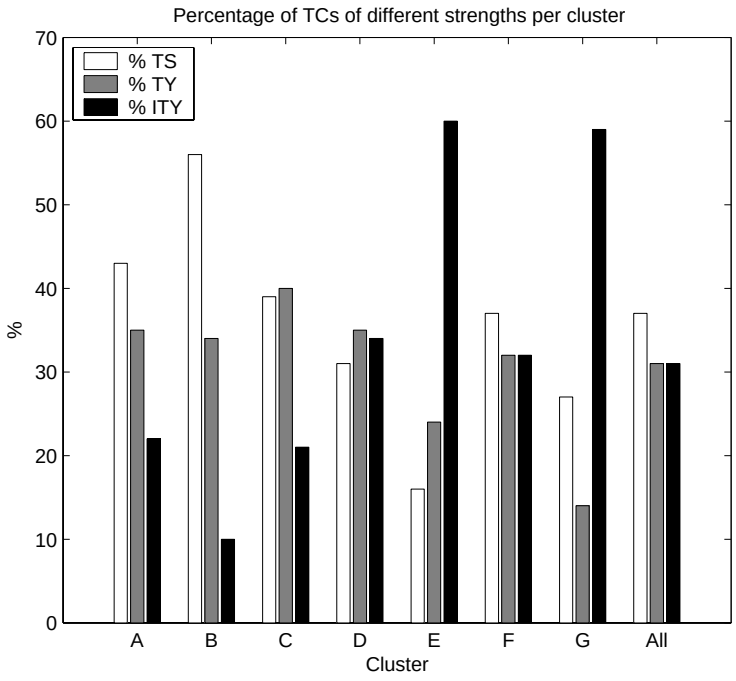

Figure 8: Percentage of cyclones that have tropical storm (TS), typhoon (TY, categories 1–2) and intense typhoon (ITY, categories 3–5) strength in each cluster and for the whole basin in the period 1970–2002.

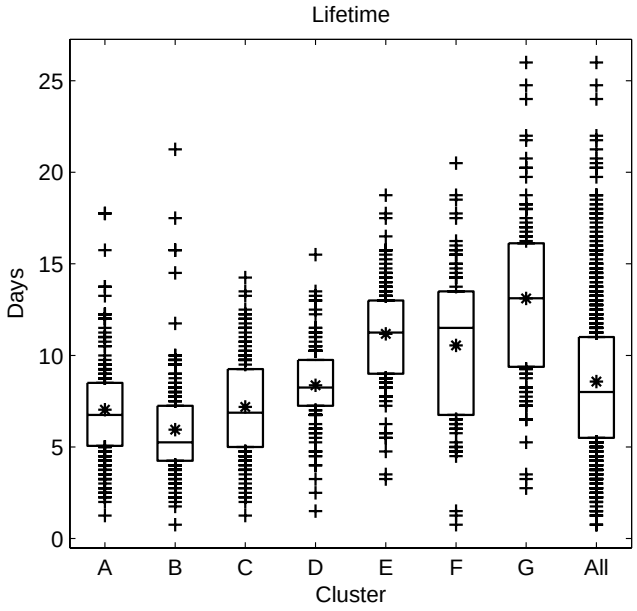

Figure 9: Distribution of lifetime per cyclone, in each cluster and for all TCs in the period 1970–2002. The boxes show the 25th and 75th percentiles, the lines in the boxes mark the median, the asterisks ($*$) the mean, and the crosses ($+$) the values below (above) the 25th (75th) percentiles of the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

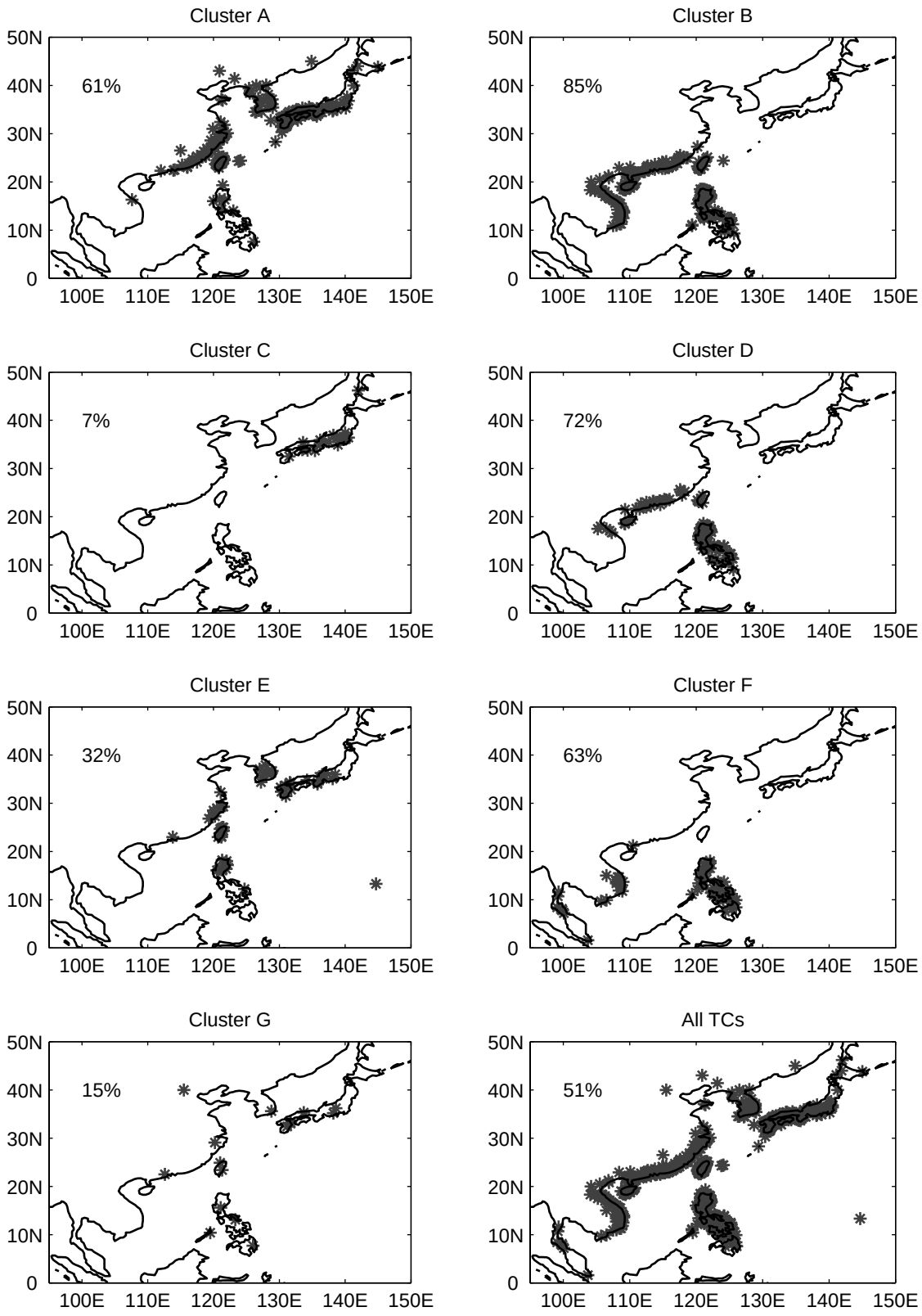

Figure 10: Landfall location and percentage of cyclones making landfall, in each cluster and for all TCs.

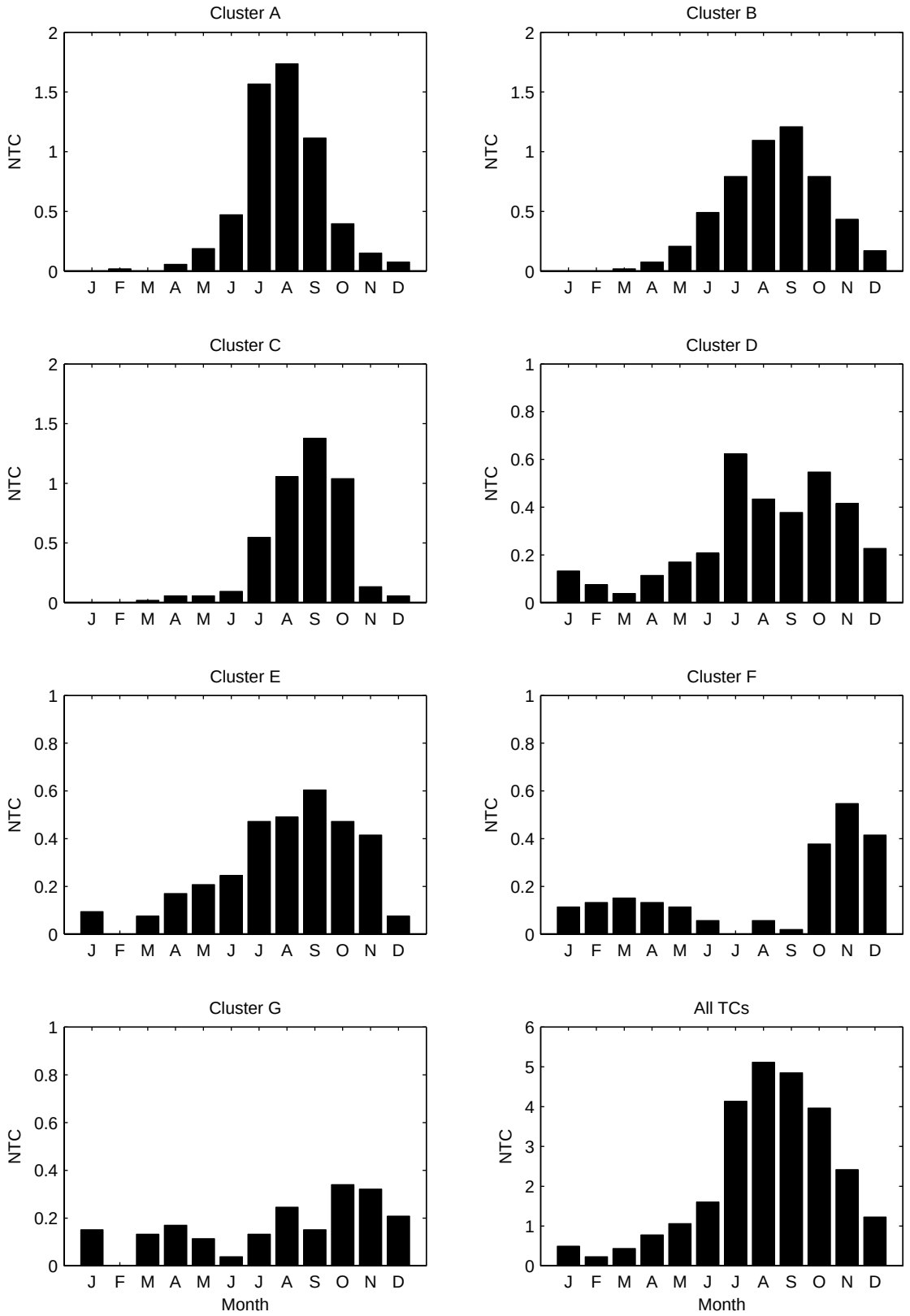

Figure 11: Mean number of cyclones per month, in each cluster and all TCs.

List of Tables

Table 1: Main tropical cyclone (TC) statistics. The seven clusters are labeled from A to G, in decreasing order of number NTC of TCs in each. The subsequent columns indicate percentage PTC of TCs, 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%

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°

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), and at the 99% significance level with an asterisk (∗). The statistical significance was determined following Vautard et al. (1990).

From/To	A	B	C	D	E	F	G	Sum	Mean	SD
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		