AMERICAN
METEOROLOGICAL
SOCIETY

Journal of Climate

The DOI for this manuscript is doi: 10.1175/2010JCLI3710.1

Bayesian Forecasting of Seasonal Typhoon Activity:

A Track-Pattern-Oriented Categorization Approach [1]

Pao-Shin Chu, Xin Zhao[#], Chang-Hoi Ho*, Hyeong-Seog Kim*,

Mong-Ming Lu[@], and Joo-Hong Kim[+]

Department of Meteorology

School of Ocean and Earth Science and Technology

University of Hawaii

Honolulu, Hawaii 96822

Sanjole Inc., Honolulu, Hawaii

* School of Earth and Environmental Sciences, Seoul National University, Seoul, Korea

@ Central Weather Bureau, Taipei, Taiwan

+ Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan

August 10, 2010

[1] Corresponding Author: Pao-Shin Chu, Department of Meteorology, 2525 Correa Road, School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, chu@hawaii.edu

Abstract

A new approach to forecasting regional and seasonal tropical cyclone (TC) frequency in the western North Pacific using the antecedent large-scale environmental conditions is proposed. This approach, based on TC track types, yields probabilistic forecasts and its utility to a smaller region in the western Pacific is demonstrated. Environmental variables used include the monthly mean of sea surface temperatures, sea-level pressures, low-level relative vorticity, vertical wind shear, and precipitable water of the preceding May. The region considered is the vicinity of Taiwan and typhoon season runs from June through October. Specifically, historical TC tracks are categorized through a fuzzy clustering method into seven distinct types. For each cluster, a Poisson or probit regression model cast in the Bayesian framework is applied individually to forecast the seasonal TC activity. With a noninformative prior assumption for the model parameters, and following Chu and Zhao (2007) for the Poisson regression model, a Bayesian inference for the probit regression model is derived. A Gibbs sampler based on the Markov Chain Monte Carlo method is designed to integrate the posterior predictive distribution. Because the cluster 5 is the most dominant type affecting Taiwan, a leave-one-out cross-validation procedure is applied to predict seasonal TC frequency for this type for the period of 1979–2006 and the correlation skill is found to be 0.76.

1. Introduction

Typhoon is one of the most destructive natural catastrophes that cause loss of lives and enormous property damage on the coasts of East Asia - western North Pacific (WNP). To mitigate the potential destruction caused by the passing of typhoons, understanding climate factors that are instrumental for the year-to-year typhoon variability in this area and developing a consistent and innovative method for predicting seasonal typhoon counts have become increasingly important.

To this purpose, numerous efforts have been made to improve the capability of typhoon or tropical cyclone (TC) activity forecasting. William Gray and his team pioneered the seasonal hurricane prediction enterprise using regression-based linear statistical models (Gray et al., 1992, 1993, 1994). They showed that nearly half of the interannual variability of hurricane activity in the North Atlantic could be predicted in advance. Klotzbach and Gray (2004, 2008) have continued to revise their forecasts as peak seasons approach, and they operationally issue seasonal forecasts for the Atlantic basin (http://hurricane.atmos.colostate.edu/Forecasts). Chan et al. (1998) used a different kind of deterministic regression model called the projection pursuit method to predict typhoon activity over the western North Pacific and the South China Sea for the period 1965–1994. Skillful forecasts are noted for some basin-wide predictands such as the number of annual typhoons.

Elsner and Schmertmann (1993) considered a different approach to predicting intense annual Atlantic hurricane counts. Specifically, the annual hurricane occurrence is modeled as a Poisson process, which is governed by a single parameter, the Poisson intensity. The intensity of the process is then linked to a set of covariates such as the

stratospheric zonal winds and the west Sahel rainfall via a multiple regression equation. Elsner and Jagger (2004) introduced a Bayesian approach to this Poisson linear regression model so that the predicted annual hurricane numbers could be cast in terms of probability distributions. This is an advantage over the deterministic forecasts because the uncertainty inherent in forecasts can be quantitatively expressed in the probability statements. They especially addressed the issue regarding the unreliable records by introducing an informative prior for the coefficient parameters of the model via a bootstrap procedure. With a similar Bayesian regression model, Elsner and Jagger (2006) attempted to predict annual U.S. hurricane counts. The model includes predictors representing the North Atlantic Oscillation (NAO), the Southern Oscillation (SO), the Atlantic multidecadal oscillation, as well as an indicator variable which is either 0 or 1 depending on the time period specified.

Apart from the Atlantic, Bayesian analysis has been applied to analyze TC variability in the North Pacific. For example, Chu and Zhao (2004) applied a hierarchical Bayesian change-point analysis to detect abrupt shifts in the TC time series over the central North Pacific (CNP). Following this research line, they (Zhao and Chu (2006, 2010)) further developed more advanced methods for detecting multiple change-points in hurricane time series for the eastern North Pacific and for the western North Pacific, respectively. Extending from the change-point analysis to forecasting, Chu and Zhao (2007) developed a generalized Poisson regression Bayesian model to predict seasonal TC counts over the CNP prior to the peak hurricane season so the forecasts are expressed in probabilistic distributions. In particular, the "critical region" concept is introduced. A critical region is defined as an area over the tropical North Pacific where the linear

correlation between the TC counts in the peak season and the preseason, large-scale environmental parameters is statistically significant at a standard test level. This "critical region" identification approach is further applied to forecast the typhoon activity in the vicinity of Taiwan area (Chu et al. 2007; Lu et al., 2010) and in the East China Sea (Kim et al. 2010a) , and satisfactory forecasting skill was achieved as well.

In the methods aforementioned, attempts have been made to either forecast TC activity for an entire ocean basin or for a specific region within a basin. In this regard, seasonal forecasts for an area are categorized by their geographic locations without considering the nature and variability of typhoon tracks. This spatial TC classification approach has been proved effective. However, even for a limited region, such as the vicinity of the Taiwan area, the origin of each typhoon and its tracks within a season are not the same. Some typhoons are straight movers and others are prone to recurve from the Philippines Sea or even from the South China Sea. Therefore, a categorization of the historical typhoon tracks and forecasting of each individual track types may result in a better physical understanding of the overall forecast skills.

Motivated by this fact, in this study, we extend the probabilistic Bayesian framework suggested in the prior works from CNP (Chu and Zhao 2007), the east China Sea (Ho et al., 2009), and the Fiji region (Chand et al., 2010) to WNP with a particular focus towards the vicinity of the Taiwan area. Different from prior studies, we adopt a feature classification approach based on the fuzzy clustering analysis of TC tracks in this study. Then we analyze the time series of each cluster type respectively. The structure of this paper is as follows. Section 2 discusses the data used and section 3 outlines the fuzzy clustering approach. The mathematical model of the TC counts, Bayesian inference and

Gibbs sampler for our proposed probabilistic models are described in section 4. Section 5 describes the procedure to select the appropriate predictors for each type of the TC count series. Results are presented in section 6. The conclusion is found in section 7.

2. Data

The present study used TC data obtained from the Regional Specialized Meteorological Center–Tokyo Typhoon Center. The data contain information on the name, date, position (in latitude and longitude), minimum surface pressure, and maximum wind speed of TCs in the WNP and the South China Sea for every 6-h interval. A TC is categorized as one of three types depending on its 10-min maximum sustained wind speed (w_{max}). These are tropical depression ($w_{max} < 17$ m s^{-1}), tropical storm (17 m s^{-1} $\leq w_{max} <$ 34 m s^{-1}), and typhoon ($w_{max} \geq 34$ m s^{-1}). In this study, we consider only tropical storms and typhoons for the period from 1979 to 2006.

Monthly mean sea level pressure, wind data at 850- and 200-hPa levels, relative vorticity at the 850 hPa level, and total precipitable water over the WNP and the South China Sea are derived from the NCEP/NCAR reanalysis dataset (Kalnay et al., 1996; Kistler et al., 2001). The horizontal resolution of the reanalysis dataset is 2.5° latitude-longitude. Tropospheric vertical wind shear is computed as the square root of the sum of the square of the difference in zonal wind component between 850- and 200-hPa levels and the square of the difference in meridional wind component between 850- and 200-hPa levels (Chu, 2002). The monthly mean sea surface temperatures, at 2° horizontal resolution, are taken from the NOAA Climate Diagnostic Center in Boulder, Colorado

(Smith et al., 1996). Monthly circulation indices such as NAO, Arctic Oscillation (AO), and Niño 3.4 are downloaded from the NOAA's Climate Prediction Center.

3. Fuzzy clustering of typhoon tracks

The basic structure of large-scale circulation variability or TC tracks have been grouped into several distinct types by many researchers (Harr and Elsberry, 1995; Elsner 2003; Camargo et al. 2007). Through the use of a vector empirical orthogonal function analysis and fuzzy clustering technique, Harr and Elsberry (1995) defined six recurrent circulation patterns that represent the monsoon trough and subtropical ridge characteristics over the tropical western North Pacific. Elsner (2003) used a K-means cluster analysis for the North Atlantic hurricanes. Based on a regression mixture model, Camargo et al. (2007) classified historical typhoon tracks from 1950–2002 into seven types although they claimed that the optimum types would range from six to eight types in the WNP.

A fuzzy clustering method (FCM) was applied to the TC tracks in this study. Because the FCM requires equal data length for all target objects, all TC tracks are interpolated into same data points with equal length by leaving out time information. The mean TC lifetime in the WNP is about five days, so we simply choose 20 segments (i.e., four times daily times five days) as the points of interpolated TC tracks, which retains the shape, length, and geographical path information covering the TC tracks (Kim et al., 2010b). The dissimilarity between two tracks is defined as the Euclidean norm of the difference of two vectors which contain the interpolated latitudes and longitudes for each TC track. With the defined dissimilarity, the fuzzy *c*-means algorithm was applied to

each of the tracks (Bezdek, 1981). The fuzzy clustering is in essence an extension of the soft k-means clustering method. This algorithm allows objects to belong to several clusters simultaneously, with different degrees of membership. Fuzzy clustering algorithm is more natural than hard clustering algorithm as objects on the boundaries among several clusters are not forced to fully belong to one of the classes, which means that partial membership in a fuzzy set is possible.

Based on this fuzzy clustering method, we analyze a total of 557 TCs over the entire WNP basin during the typhoon season (June to October) from 1979 to 2006 and categorize them into seven major groups. The TC tracks and its mean path for each of the seven types over the WNP are depicted in Fig. 1. The overall TC tracks are shown at the right bottom panel in Fig. 1. It is apparent that each type of TC has its own active region and distinct track patterns. For example, the cluster 1 represents the TC track pattern mainly striking Japan and Korea and eastern China coast. Most TCs in this cluster type develop over the Philippine Sea, move northwestward then turn northeastward toward Korea or Japan. For the cluster 2, most TCs develop in the subtropics farther away from the East Asian continent and move northward or northeastward over the open ocean; they have the least number of occurrences among all seven clusters (56). The cluster 3 represents the TCs which tend to develop to the east of Taiwan and move northward to the east of Japan. Its mean track is shorter than that in type 1 and the genesis location is more poleward than type 1. For the cluster 4, most TCs develop over the South China Sea and are confined in the same region. The cluster 5 is particularly of our interest in this study since this type represents the TCs which develop over the core of the Philippine Sea and move northwestward through Taiwan and

southeast China coast. Among all seven clusters, the clusters 4 (90) and 5 (92) have the largest numbers. For the cluster 6, most TCs are straight movers from the Philippine Sea through the South China Sea to south China and Vietnam. The cluster 7 TCs tend to form near 15°N and between 140°E and 180°E; they pass through the east of Japan after recurving poleward over mainly the open ocean. Overall, the mean cluster tracks identified in this study are similar to those of Camargo et al. (2007).

Albeit the method developed in this paper is applicable for the entire East Asian coast and the WNP, only a case study is presented for the vicinity of Taiwan which is defined as a region bordered between 21°N–26°N and 119°E–125°E. This is justified because of the relatively high annual number of TCs observed there and the huge damage typhoons inflicted (Tu et al., 2009). Table 1 lists the seasonal typhoon counts affecting Taiwan, as stratified by the seven cluster types, from 1979 to 2006. We notice that about 63% TCs that have affected Taiwan are classified as cluster 5. This is followed, in descending order of historical occurrence, by clusters 1, 6, 3 and 4. Not surprisingly, because of their distant geographic locations, clusters 2 and 7 have no effects on Taiwan.

4. Prediction Methods and Bayesian inference

Once historical TC tracks are classified into distinct clusters, the next goal is to develop a modern methodology for predicting seasonal TC counts for a target region (Taiwan) influenced by various track types. In this section, we will first describe the two statistical models used and the Bayesian inference for each model. We will then discuss the predictor selection method followed by the overall forecast scheme.

4.1 Model description

4.1.1 The generalized Poisson regression model

Poisson distribution is a proper probability model for describing independent (memory-less), rare event counts. Given the Poisson intensity parameter λ, the probability mass function (PMF) of h counts occurring in a unit of observation time, say one season, is (Epstein, 1985)

$$P(h \mid \lambda) = \exp(-\lambda)\frac{\lambda^h}{h!}, \text{ where } h = 0,1,2,... \text{ and } \lambda > 0. \quad (1)$$

The Poisson mean is simply λ, so is its variance. In many applications, Poisson rate λ is not treated as a fixed constant but rather as a random variable.

Through a regression model, the relationship between the target response variable, seasonal typhoon counts, and the selected predictors can be mathematically built. In this study, we adopt the Poisson linear regression model. Assume there are N observations that are conditional on K predictors. We define a latent random N-vector $\mathbf{Z}$, such that for each observation h_i, $i = 1,2,...,N$, $Z_i = \log \lambda_i$, where λ_i is the Poisson rate for the i-th observation. The link function between the latent variable and its associated predictors is expressed as $Z_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$ is a random vector; noise ε_i is assumed to be identical and independently distributed (IID) and normally distributed with zero mean and σ^2 variance; $\mathbf{X}_i = [1, X_{i1}, X_{i2}, ..., X_{iK}]$ denotes the predictor vector. In vector form, the general Poisson linear regression model is formulated as below:

$$P(\mathbf{h} \mid \mathbf{Z}) = \prod_{i=1}^{N} P(h_i \mid Z_i), \text{ where } h_i \mid Z_i \sim Poisson(h_i \mid e^{Z_i}),$$

$$\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim Normal(\mathbf{Z} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N), \text{ where, specifically}$$

$\mathbf{X}' = [\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_N']$, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and

$\mathbf{X}_i = [1, X_{i1}, X_{i2}, ..., X_{iK}]$ is the predictor vector for h_i, $i = 1,2,..., N$,

$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$. (2)

Here, *Normal* and *Poisson* stand for the normal distribution and the Poisson distribution, respectively. In model (2), β_0 is referred to intercept.

It is worth noting that Poisson rate λ is a real value while the TC counts h is only an integer. Accordingly, λ contains more information relative to h. Furthermore, because h is conditional on λ, λ is subject to less variations than h is. Taken together, λ should be preferred as the forecast quantity of the TC activity than h for decision making. We also notice the fact that this hierarchical structure essentially fits well for Bayesian inference.

4.1.2 The probit regression model for a binary classification problem

The Poisson regression model detailed in section 4.1.1 has been approved very effective for most rare event count series. However, if the underlying rate is significantly below one, this model may introduce significant bias. In this study, for a given typhoon type, if the mean of its historical seasonal occurrence is less than 0.5, we shall instead adopt a binary classification model. That is, the response variable here is a binary class label, which is termed by "Y". For each observation period, we define a class "Y = 1" if one or more TC is observed and "Y = 0" otherwise.

As below we formulate the probit regression model (Albert and Chib 1993; Zhao and Cheung 2007). Again, we assume there are N observations conditional on K selected predictors. We define a latent random N-vector $\mathbf{Z}$, such that for each

observation y_i, $i = 1,2,...,N$, $y_i = 1$ if $Z_i \geq 0$ and $y_i = 0$ otherwise. The link function between the latent variable $\mathbf{Z}$ and its associated predictors is also linear, $Z_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$ is a random vector; noise ε_i is assumed to be identical and independently distributed (IID) and normally distributed with zero mean and σ^2 variance; $\mathbf{X}_i = [1, X_{i1}, X_{i2}, ..., X_{iK}]$ denotes the predictor vector. In vector form, the probit regression model is described by:

$$P(\mathbf{y} \mid \mathbf{Z}) = \prod_{i=1}^{N} P(y_i \mid Z_i), \text{ where } y_i = \begin{cases} 1 & if\ Z_i \geq 0 \\ 0 & if\ Z_i < 0 \end{cases} \text{ and}$$

$\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim Normal(\mathbf{Z} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$, where, specifically

$\mathbf{X}' = [\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_N']$, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and

$\mathbf{X}_i = [1, X_{i1}, X_{i2}, ..., X_{iK}]$ is the predictor vector for h_i, $i = 1,2,...,N$,

$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$. (3)

Classification model (3) is very similar to the Poisson regression model (2). Actually the probability of class "Y = 1" can be viewed as the rate of the TC counts.

4.2 Bayesian inference for the constructed models

With the built models provided in section 4.1, we shall derive the posterior distribution for the model given by (2) and (3) separately in this section.

4.2.1 Bayesian inference of the Poisson regression model

Since we do not have any credible prior information for the coefficient vector $\boldsymbol{\beta}$ and the variance σ^2, it is reasonable to choose the non-informative prior. In formula, it is (Gelman et al., 2004, p. 355)

$$P(\boldsymbol{\beta},\sigma^2) \propto \sigma^{-2}. \tag{4}$$

This is not a proper probability distribution function; however, it leads to a proper posterior distribution.

The posterior distribution of a hidden variable Z, which is conditionally independent from each other given the model parameters $\boldsymbol{\beta}$ and σ^2, is derived in Chu and Zhao (2007) and given in (A2). With the newly observed predictor set $\widetilde{\mathbf{X}} = [1, \widetilde{X}_{i1}, \widetilde{X}_{i2}, ..., \widetilde{X}_{iK}]$, the predictive distribution for the new latent variable $\widetilde{Z}$ and TC counts $\widetilde{h}$ will be

$$P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \iint_{\boldsymbol{\beta},\sigma^2} P(\widetilde{Z} \mid \widetilde{X}, \boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h}) d\boldsymbol{\beta} d\sigma^2 \tag{5a}$$

$$P(\widetilde{h} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \int_{\widetilde{Z}} \frac{\exp(-e^{\widetilde{Z}} + \widetilde{Z}\widetilde{h})}{\widetilde{h}!} P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) d\widetilde{Z}. \tag{5b}$$

Here the "new" variables refer to those not involved in the model construction and only used for prediction. With the non-informative prior, the posterior distribution for the model parameter set $(\boldsymbol{\beta}, \sigma^2)$ in (5) is not standard and directly sampling from it is difficult. In this section, we design a Gibbs sampler, which has $P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h})$ as its stationary distribution, and then we can use an alternative approach, Monte Carlo method, to integrate (5) by

$$P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L} \sum_{i=1}^{L} P(\widetilde{Z} \mid \widetilde{X}, (\boldsymbol{\beta}, \sigma^2)^{[i]}) \tag{6a}$$

$$P(\widetilde{h} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L}\sum_{i=1}^{L} \frac{\exp(-e^{\widetilde{Z}^{[i]}} + \widetilde{Z}^{[i]}\widetilde{h})}{\widetilde{h}!}, \tag{6b}$$

where $(\boldsymbol{\beta}, \sigma^2)^{[i]}$ is the i-th sampling from the proposed Gibbs sampler after the burn-in period, $\widetilde{Z}^{[i]}$ is sampled from $\widetilde{Z}^{[i]} \mid \widetilde{X}, (\boldsymbol{\beta}, \sigma^2)^{[i]} \sim Normal(\widetilde{Z}^{[i]} \mid \widetilde{X}\boldsymbol{\beta}^{[i]}, \sigma^{2[i]})$ subsequently, and L is a sufficiently large number (for example, throughout this study, we use L = 10000). The concept and detailed algorithms of a Gibbs sampler can be found in many literatures such as Gelman et al. (2004).

Based on the inference analysis derived in appendix, the Gibbs sampler yields the following algorithm:

1. Select proper initial value for $\mathbf{Z}^{[0]}, \boldsymbol{\beta}^{[0]}, \sigma^{2[0]}$ and set $i = 1$.

2. Draw $Z_j^{[i]}$ from $Z_j^{[i]} \mid \mathbf{h}, \boldsymbol{\beta}^{[i-1]}, \sigma^{2[i-1]}$ for $j = 1,2,..., N$ via (A3), a conditional distribution for the hidden variable *Z* from the Poisson regression model.

3. Draw $\boldsymbol{\beta}^{[i]}$ from $\boldsymbol{\beta}^{[i]} \mid \mathbf{h}, \mathbf{Z}^{[i]}, \sigma^{2[i-1]}$ via (A6), a multivariate Gaussian distribution.

4. Draw $\sigma^{2[i]}$ from $\sigma^{2[i]} \mid \mathbf{h}, \mathbf{Z}^{[i]}, \boldsymbol{\beta}^{[i]}$ via (A7), an inverse χ^2 distribution.

5. Set $i = i + 1$ then go back to step 2 until meeting the required number of iterations.

(7)

With the observation data $\mathbf{h}$ and following (7), after a burn-in period, one can sample set $\mathbf{Z}, \boldsymbol{\beta}, \sigma^2$ within each iteration, which will have the desired posterior distribution that facilitates the numerical computation of (6a) and (6b).

A practical issue in the step 2 of algorithm (7) is that the distribution governed by Eq. (A3) is not standard. We resort to the Metropolis-Hasting algorithm in this study,

which is relatively computationally expensive. Some other approaches can be considered here. For example, based on our simulation results, the estimated posterior probability density functions for the hidden variables are all Gaussian like, which theoretically also can be proven log-concave. Therefore, using Laplace approximation in this context should be a sound choice as well.

4.2.2 Bayesian inference of the probit regression model

The probit regression model for a binary classification problem is detailed in (3), which implies that the posterior distribution of any hidden variable Z is conditionally independent from each other given the model parameters $\boldsymbol{\beta}$ and σ^2. Therefore, similar to the Poisson regression model, with the newly observed predictor set $\widetilde{\mathbf{X}} = [1, \widetilde{X}_{i1}, \widetilde{X}_{i2}, ..., \widetilde{X}_{iK}]$, the predictive distribution for the latent variable $\widetilde{Z}$ and TC counts $\widetilde{h}$ will be

$$P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \iint_{\boldsymbol{\beta},\sigma^2} Normal(\widetilde{Z} \mid \widetilde{X}\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h}) d\boldsymbol{\beta} d\sigma^2 \tag{8a}$$

$$P(\widetilde{h} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \int_{\widetilde{Z} \geq 0} P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) d\widetilde{Z} . \tag{8b}$$

The posterior distribution for the model parameter set ($\boldsymbol{\beta}, \sigma^2$) in (8a) is not standard with a non-informative prior. We hence design a Gibbs sampler, which has $P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h})$ as its stationary distribution, and thereby we can use Monte Carlo method to integrate (8) by

$$P(\widetilde{Z} \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L} \sum_{i=1}^{L} Normal(\widetilde{Z} \mid \widetilde{X}\boldsymbol{\beta}^{[i]}, \sigma^{2[i]}) \tag{9a}$$

$$P(\widetilde{y} = 1 \mid \widetilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L} \sum_{i=1}^{L} \Phi(\widetilde{X}\boldsymbol{\beta}^{[i]} / \sqrt{\sigma^{2[i]}})$$

$$P(\tilde{y}=0 \mid \tilde{X}, \mathbf{X}, \mathbf{h}) = 1 - P(\tilde{y}=1 \mid \tilde{X}, \mathbf{X}, \mathbf{h}), \tag{9b}$$

where $\boldsymbol{\beta}^{[i]}$ and $\sigma^{2[i]}$ is the i-th sampling from the proposed Gibbs sampler after the burn-in period; $\Phi(.)$ denotes the probability cumulative function of the standard normal distribution; and L is a sufficiently large number.

Since the hierarchical probit regression model in (3) is very similar to the Poisson regression model in (2), we adapt most of the formulae provided in Appendix for the Bayesian inference. The only major difference is the conditional posterior distribution of the latent variable, which is from a truncated Gaussian distribution based on the definition in (3). In formula, it is:

$$\begin{aligned}
&Z_i \mid \mathbf{X}_i, \beta_0, \sigma^2, y_i = 1 \propto N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2) \quad \text{truncated at the left by 0,} \\
&Z_i \mid \mathbf{X}_i, \beta_0, \sigma^2, y_i = 0 \propto N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2) \quad \text{truncated at the right by 0.} \\
&i = 1,2,..., N, \ \mathbf{X}_i \text{ represents the } i\text{-th row the predictor matrix } \mathbf{X}.
\end{aligned} \tag{10}$$

The overall Gibbs sample for the probit regression model (3) is executed as:

1. Select proper initial value for $\mathbf{Z}^{[0]}, \boldsymbol{\beta}^{[0]}, \sigma^{2[0]}$ and set $i = 1$.

2. Draw $Z_j^{[i]}$ from $Z_j^{[i]} \mid y_j, \boldsymbol{\beta}^{[i-1]}, \sigma^{2[i-1]}$ for $j = 1,2,..., N$ via (10), a truncated Gaussian distribution.

3. Draw $\boldsymbol{\beta}^{[i]}$ from $\boldsymbol{\beta}^{[i]} \mid \mathbf{Z}^{[i]}, \sigma^{2[i-1]}$ via (A6), a multivariate Gaussian distribution.

4. Draw $\sigma^{2[i]}$ from $\sigma^{2[i]} \mid \mathbf{Z}^{[i]}, \boldsymbol{\beta}^{[i]}$ via (A7), an inverse χ^2 distribution.

5. Set $i = i + 1$ then go back to step 2 until meeting the required number of iterations.

(11)

In step 2 of (11), we choose the Robert's (1995) fast algorithm to draw sample from a truncated Gaussian distribution.

5. Predictor selection procedure

In model (2) or (3), we assume the predictors are given a priori. In real applications, however, choosing the appropriate environmental parameters which are physically related to the formation and typhoon tracks are crucial for the success of the final forecast scheme. In Chu and Zhao (2007) and Chu et al. (2007), environmental parameters such as sea surface temperatures, sea level pressures, low-level relative vorticity, vertical wind shear (VWS), and precipitable water were chosen.

5.1 Critical region determination

In this study, we apply the same procedure suggested in Chu and Zhao (2007) and Chu et al. (2007) to determine the critical region for each candidate environmental parameter. We calculate the Pearson correlation between the count series of each type of typhoon track and the preseason environmental parameters. If the Pearson correlation between the predictor and the target count series is statistically significant, it is deemed as critical. Based on the linear regression theory, for a sample size of 28, the critical value for a correlation coefficient with two tails is 0.374 at the 99% confidence level (Bevington and Robinson, 2002). Hence, a correlation coefficient with its absolute value greater than 0.374 at a grid point is deemed locally significant and this point is then selected as a critical region. To avoid the large dimensionality of the predictor matrix which would easily lead to overfitting the model, a simple average over the critical regions is chosen to serve as a final predictor. We also examined the lagged correlations between the circulation index (e.g., NAO, AO) and the TC counts for each of seven

clusters listed in Table 1. However, none of those correlations are statistically significant at the 95% confidence level. Therefore, circulation indices are not chosen as predictors.

5.2 Large-scale circulations and track types

Because track type 5 accounts for almost three-fourth of the overall TC activity near Taiwan, we will focus on the selection of predictors for this type and the interim results are illustrated in Fig. 3. The isocorrelate map of seasonal TC track type 5 and SSTs over the WNP during the antecedent May is displayed in Fig. 3a; positive and significant correlations are found near Taiwan and the equatorial western Pacific. Warmer SSTs are expected to fuel the overlying atmosphere with additional warmth and moisture, possibly reducing atmospheric stability and increasing the likelihood of deep tropical convection. The occurrence of deep convection is important for typhoon formation because it provides a vertical coupling between the upper level outflow and lower tropospheric inflow circulations. For SLPs (Fig. 3b), negative correlations are observed in the eastern half of the western Pacific, suggesting that type 5 TCs are more abundant if the antecedent SLP in the western Pacific subtropical high is anomalously low. In Fig. 3c, a dipole structure of correlation patterns is seen with the negative correlation region to the north and positive region to the south. Accordingly, more atmospheric moisture, or an increase of the depth of the moisture layer, in low latitudes is attributable to more type 5 TCs.

Fig. 3d shows the correlations between type 5 TCs and the low-level relative vorticity in the preceding May. An elongated band of positive correlations in the subtropics is noted with a critical region between 140°-170°E. Local concentration of cyclonic vorticity in the critical region would enhance spin-up process by strengthening

boundary layer convergence. An increase in cyclonic vorticity near Taiwan may also reflect southward shift of the Mei-Yu front in May. Such a possibility is supported by the positive SLP correlation near Taiwan (Fig. 3b), and the negative PW (Precipitable Water) correlation over the subtropical WNP (Fig. 3c). Mei-Yu front is a prominent feature during the developing stage of East Asian summer monsoon. However, to the best of our knowledge, the relationship between Mei-Yu front and the subsequent TC activity over the WNP has not been well studied. This relationship can be particularly important for the type 5 TCs.

It is possible that as easterly waves in the subtropics approach the monsoon confluence region (say, near 140°E), they will interact with monsoon westerlies to the west to increase cyclogenesis potential. Together with moist convergent flow at low levels, TC may form on the cyclonic shear side of the monsoon circulation. This pattern, known as the cyclogenesis in the monsoon confluence region, is one of the distinctive flow fields for TC formation in the WNP during the peak typhoon season (Ritchie and Holland, 1999).

The isocorrelate map of seasonal typhoon frequency of type 5 and VWSs over the WNP during the antecedent May is displayed in Fig. 3e. Negative correlations are generally found in the subtropics and the mid-latitudes. It is well known that strong VWS disrupts the organized deep convection (the so-called ventilation effect) that inhibits intensification of the typhoons. Conversely, weak vertical shear allows TC development. A small critical region is found near Taiwan (Fig. 3e).

5.3 Overall model

With the predictors selected through correlation analysis, the regression model (2) and classification model (3) are set. Through the algorithms provided in section 4.2, the analysis and forecast procedure deliberated in Fig. 2 can be executed.

The selected predictor vector $\mathbf{X}_i$ and the associated coefficient parameter vector $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$ in model (2) and (3) can both be explicitly formulated by

$$\mathbf{X}_i = [1, SST_i, SLP_i, VWS_i, RV_i, PW_i],\ i = 1,2,..., N \text{ and}$$

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'. \quad (12)$$

In (12), if an environmental variable is not selected, its associated parameters and coefficients are set as null. In case of a variable with two predictors (one with positive correlation and the other with negative), its associated parameters and coefficients represent two vectors. In practical applications, it is desirable to normalize each predictor series before further analysis to avoid the scaling problem among the different predictors.

6. Prediction results

As briefly discussed earlier, we have a total of 28 years (1979 – 2006) of tropical cyclone track records and environmental variable data in the WNP. Following the flow chart in Fig. 2, we apply the method detailed in section 4 to this dataset.

In detail, we first categorize each of the typhoons observed in WNP into seven classes based on the fuzzy clustering algorithm. Thereafter, we tabulate each type of typhoons that occurred in the Taiwan area. Historically, types 2 and 7 never had any impact on Taiwan (Table 1). For type 3, type 4 and type 6, the average typhoon rate in peak season is 0.143, 0.286 and 0.357, respectively, all of which are well below 0.5. Hence, we apply the probit regression model to analyze type 3, type 4 and type 6 typhoon

tracks, and use the Poisson regression model for the type 1, and 5 (with average typhoon rate 0.714 and 2.50, respectively). Again, since type-5 typhoon has been the dominant typhoon type for the typhoon activity in Taiwan, we shall provide the detail analysis results for this type.

A general way to verify the effectiveness of a regression or classification method is to apply a strict cross-validation test for the relevant dataset. Considering the fact that the typhoon activity variation is approximately independent from year to year, it is proper to apply a leave-one-out cross-validation (LOOCV) in this context (Elsner and Schmertmann, 1994; Chu et al., 2007). That is, a target year is chosen and a model is developed using the remaining 27-year data as the training set. The observations of the selected predictors for the target year are then used as inputs to forecast the missing year. This process is repeated successively until all 28 forecasts are made. We shall apply this LOOCV process for each typhoon type and thereby the overall activity.

With all the samples drawn, we can estimate any statistic deemed as important. To demonstrate this, we illustrate the analysis results for type 5 typhoon seasonal series in the following. We first apply the Poisson regression algorithm (7) to the data and the output median, the upper and lower quartiles (the upper 75% and lower 25%) of the predicted rates, through a LOOCV, are plotted together with the actual observation for each year in Fig. 4a. The distance between the upper quartile and lower quartile locates the central 50% of the predicted TC variations. The Pearson correlation between the median of predictive rate and independent observations is as high as 0.76, which implies that about 58% of the variation of this type near Taiwan can be predicted. In Fig. 4b, the median, and the upper and lower quartiles of the predicted typhoon counts are plotted

together with the actual observation for each year. Out of a total of 28 years, there are only 1 year in which the actual TC counts lie outside the predictive central 50% boundaries, achieving 96% accuracy.

The similar Poisson regression procedure is applied to the type-1 typhoon series as well. The correlation coefficient between the LOOCV median rate and true observation series is 0.63 for type 1. Note that the actual typhoon occurrence rate of this type near Taiwan is actually very small (0.71 per year).

We also apply the probit regression algorithm in (11) to the seasonal typhoon series of type 3, 4 and 6, respectively. The labeling of the year with more than one TC as belonging to one group (i.e., class 1) or the other is arbitrary. With an LOOCV procedure, we obtain the median, upper quartile and lower quartile of the probability of class "1" (equivalently, with a typhoon) for each season. Specifically, the correlations between the median probability of class "1" and the observation series of type 3, type 4 and type 6 are 0.65, 0.72 and 0.74. Based on the median (or quartile) probability, we can make a class decision.

From each individual simulation, we summarize the relative probability outputs, and then obtain the marginal forecast for the typhoon frequency in Taiwan (Fig. 5). For simplicity, for all the simulations in this study, we take the first 2,000 samples as burn-in and use the following 10,000 samples as the output of the Gibbs sampler. Fig. 5a displays the median, the upper and lower quartiles of the predicted (LOOCV) overall seasonal typhoon rates in the vicinity of Taiwan. The correlation between the median of predictive rate and observations is 0.71. In comparison to the correlation skill of 0.63 from a different regression model and without clustering TC tracks (Chu et al. 2007), the

current result is noticeably an improvement. In Fig. 5b, the median, and the upper and lower quartiles of the summed predicted typhoon counts are plotted together with the actual observation for each year. Out of a total of 28 years, only 2 years (93% accuracy) falls outside the interquartile range. This result further supports the efficiency of the proposed feature-oriented regional typhoon frequency forecast framework.

As Poisson or probit regression model provides probability forecasts, it is also of interest to evaluate the model performance using the Brier Skill Score (BSS), which provides a measure of improvement percentage of model forecast over a climatology model (BSS = 0 indicates no skill relative to the climatological forecast and BSS = 1 means perfect prediction. A detailed definition, for example, can be found in Jagger et al. (2002)). If we treat the seasonal typhoon activity occurred in the Taiwan area as a binary-class problem (seasonal count either above normal or below normal), the BSS score of the proposed track-pattern based forecast model is 0.32.

7. Summary and conclusions

The importance of typhoon prediction research cannot be overemphasized. Heavy rain, destructive winds, and coastal storm surges associated with typhoons cause flood and landslide disasters, often resulting in loss of life and enormous property damage. Improving forecast skill of seasonal TC counts before the peak season has become increasingly important for society and economy. Traditionally, seasonal TC forecasting has been attempted for an entire ocean basin or for a specific region within a basin. That is, seasonal forecasts for a basin (or an area) are categorized by its geographic location without considering the nature and variability of typhoon tracks. However, even

for a limited region, the formation point of each typhoon and its subsequent track within a season are not the same. Therefore, a categorization of the historical typhoon tracks and forecasting of each individual track types may result in a better physical understanding of large-scale circulation characteristics and an improvement in overall forecast skills. Motivated by this, based on a TC track oriented categorization approach, we apply a marginal mix of Poisson regression and probit regression model to predict the seasonal TC activity in the Taiwan area which has been repeatedly ravaged by typhoons and typhoon activity there has undergone a significant upward shift since 2000 (Tu et al., 2009).

Following a fuzzy clustering algorithm, we first projected all the recorded TC tracks from 1979–2006 into seven distinct groups featured by their genesis locations and pathways. Then for each type of cluster, we apply a Poisson regression model or probit regression model to construct the relationship between the large-scale circulations and the seasonal TC frequency. As an example, for the case of Taiwan which is mainly affected by track type 5, we resort to the Poisson regression model (Fig. 2, Table 1). For other types with less than 0.5 average seasonal typhoon rate such as types 3, 4 and 6, we adopt the probit regression to solve a binary classification problem. Because Taiwan is not affected by track types 2 and 7, no analysis is applied to these two types.

In the analysis of each type of TC, we choose the prediction selection procedure suggested in Chu and Zhao (2007) and Chu et al. (2007). That is, for each target TC cluster, we identify the associated critical regions for each considered environmental variable via a simple correlation analysis, forming the relative predictors. The variables include SST, SLP, PW, relative vorticity, and VWS. Subsequently, we derive Bayesian

inference for both the Poisson regression model and the probit regression model by assuming a non-informative prior. In this study, Markov Chain Monte Carlo (MCMC) method is adopted to numerically analyze the data because it is difficult to analytically evaluate complex integral quantities of the posterior distribution as it is not a standard probability density function. The MCMC is based on drawing values of the parameters of interest from probability distributions and then correcting these draws to better approximate the posterior distribution. For details on the MCMC, see Zhao and Chu (2006). The designed Gibbs samplers for both regression models are very similar, through which we are able to forecast the probabilistic distribution of TC activity of each type prior to the peak season. When tested for the period 1979–2006, the leave-one-out cross-validation correlation test delivers satisfactory results as described in section 5. Especially for type 5 TC, the correlation between the leave-one-out forecasts and actual observations is as high as 0.76, highlighting the efficiency our proposed feature-oriented approach (Fig. 3a). By summarizing the marginal distributions of the forecasts for all five track types (1, 3, 4, 5 and 6), the overall correlation skill is 0.71, an improvement over the geographic based categorization approach (Chu et al., 2007). The proposed forecast model also provides a 0.32 Brier score skill, showing significant enhancement over a simple climatology prediction.

The TC forecast framework developed in this study is valuable. First, it is physically based on the TC origin and track path feature. Hence, it should be easier to interpret and forecast the TC activity in terms of the mean genesis location, mean tracks, and the preferred landfall location for a give type (Fig. 1). Second, forecasts of seasonal TC counts are presented in probabilistic format, which is preferred for decision-makers,

since it provides the uncertainty of the prediction. In addition, the proposed hierarchical probabilistic structure for both regression models can serve as the perfect platform for further researches because any probabilistic model can be treated as an independent modulo and seamlessly plugged into a unified Bayesian framework. For example, albeit in this study we assume that the link function between the TC rate and the predictors is linear (or generalized linear), which is not necessarily the best approximation for the true underlying physical model. In principle, these models can be extended to non-linear link function via proper non-linear probabilistic model such as kernel-based Gaussian processes. Obviously, the predictor selection procedures are also needed to be revised accordingly in this regard. This promising approach is however beyond the scope of this study.

Acknowledgments

We acknowledge the kind financial support from the Pacific Disaster Center, in particular, Ray Shirkhodai. Partial support for this study came from the Central Weather Bureau, Taiwan (MOTC-CWB-98-3M-01), from the Korea Meteorological Administration Research and Development Program under grant CATER 2006-4202, and from the National Science Council of the Republic of China under grant NSC98-2625-M-052-009 to Central Weather Bureau.

Appendix: Conditional posterior distribution for a Poisson regression model

For the sake of simplicity, in the following derivation, we will drop the notation of the predictor matrix $\mathbf{X}$, which is always given by default.

Based on formula (3) and model (2), it follows that

$$P(\mathbf{Z} \mid \mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto P(\mathbf{h} \mid \mathbf{Z}, \boldsymbol{\beta}, \sigma^2) P(\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2) = P(\mathbf{h} \mid \mathbf{Z}) P(\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2). \quad \text{(A1)}$$

Substituting the probability model (2) into (A1) and ignoring the constant part yields

$$P(\mathbf{Z} \mid \mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^N} \prod_{i=1}^{N} \exp\left\{ -e^{Z_i} + Z_i h_i - \frac{1}{2\sigma^2} (Z_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right\}. \quad \text{(A2)}$$

This is not a standard density distribution, but we can design a Gibbs sampler through which the output of each of its iteration will be of the distribution given by (A2).

From (A2), one can see that Z_i is conditionally independent from each other for $i = 1,2,...N$ given $\boldsymbol{\beta}$ and σ^2, therefore sampling from $Z_i \mid \mathbf{h}, \boldsymbol{\beta}, \mathbf{Z}_{-i}, \sigma^2$, where $\mathbf{Z}_{-i} = [Z_1, .., Z_{i-1}, Z_{i+1}, ..., Z_N]'$, is equivalently sampling from $Z_i \mid \mathbf{h}, \boldsymbol{\beta}, \sigma^2$. We ignore the constant part and obtain

$$P(Z_i \mid \mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto \exp\left\{ -e^{Z_i} + Z_i h_i - \frac{1}{2\sigma^2} (Z_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right\}, \; i = 1,2,...N. \quad \text{(A3)}$$

To sample Z_i from (A3), in this paper we apply the Metropolis-Hasting algorithm. One can refer to Ripley (1987), Gelman et al. (2004), or originally Hastings (1970) for the details of this algorithm.

After the latent vector $\mathbf{Z}$ is obtained, the model is exactly the same as the so-called ordinary linear regression and its Bayesian inference derivation is straightforward. The joint posterior distribution for $(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)$ can be expressed as

$$P(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{h}) \propto P(\mathbf{h} \mid \mathbf{Z}, \boldsymbol{\beta}, \sigma^2) P(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2) = P(\mathbf{h} \mid \mathbf{Z}) P(\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2). \quad \text{(A4)}$$

With (A4) and under the non-informative prior for the parameter given by Eq. (6), we have

$$P(\boldsymbol{\beta},\sigma^2 \mid \mathbf{Z},\mathbf{h}) \propto P(\mathbf{Z},\boldsymbol{\beta},\sigma^2 \mid \mathbf{h}) \propto P(\mathbf{Z} \mid \boldsymbol{\beta},\sigma^2)P(\boldsymbol{\beta},\sigma^2)$$
$$\propto (\sigma^2)^{-(N/2+1)} \exp(-\frac{(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}) \quad . \tag{A5}$$

From (A5), if σ^2 is given, the conditional posterior distribution for $\boldsymbol{\beta}$ obviously is multivariate Gaussian:

$$\boldsymbol{\beta} \mid \mathbf{Z},\mathbf{h},\sigma^2 \sim \mathit{Normal}(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}},(\mathbf{X}'\mathbf{X})^{-1}\sigma^2),$$

$$\text{where } \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}. \tag{A6}$$

Alternatively, if $\boldsymbol{\beta}$ is given, the conditional posterior distribution for σ^2 is scaled-inverse-χ^2 distribution. That is

$$\sigma^2 \mid \mathbf{Z},\mathbf{h},\boldsymbol{\beta} \sim \text{Inv}-\chi^2(\sigma^2 \mid N,s^2) \ ,$$

$$\text{where } s^2 = \frac{1}{N}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}). \tag{A7}$$

In (A7), $\text{Inv}-\chi^2$ refers to the scaled-inverse-χ^2 distribution. With (A3) (A6) and (A7), we have completed the proposed Gibbs sampler, and its stationary output within each iteration will be equivalently sampled from the joint posterior distribution of set $(\mathbf{Z},\boldsymbol{\beta},\sigma^2)$ from the model given by formula (2).

References

Albert, J. and Chib, S., 1993: Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, **88**, 669–679.

Bevington, P. R., and D. K. Robinson, 2002: *Data reduction and error analysis for the physical sciences,* 3rd edition, New York, McGraw-Hill.

Bezdek, J.C., 1981: Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, 256 pp.

Camargo, S.J., A.W. Robertson, S.J. Gaffney, P. Smyth, and M. Ghil, 2007: Cluster analysis of typhoon tracks. Part I: General properties. *J. Climate*, 20, 3635–3653.

Chan, J.C.L., J.S. Shi, and C.M. Lam, 1998: Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea. *Wea. Forecasting*, **13**, 997–1004.

Chand, S.S., K.J.E. Walsh, and J.C.L. Chan, 2010: A Bayesian regression approach to seasonal prediction of tropical cyclones affecting the Fiji region. *J. Climate*, **23**, 3425-3445.

Chu, P.-S., 2002: Large-scale circulation features associated with decadal variations of tropical cyclone activity over the central North Pacific. *J. Climate*, **15**, 2678–2689.

Chu, P.-S., and X. Zhao, 2004: Bayesian change-point analysis of tropical cyclone activity: The Central North Pacific case. *J. Climate*, **17**, 4893–4901.

Chu, P.-S., and Zhao, X., 2007: A Bayesian regression approach for predicting seasonal tropical cyclone activity over the Central North Pacific, *J. Climate*, **15**, 4002–4013.

Chu, P-S, Zhao, X., Lee, C-T and Lu, M-M, 2007: Climate Prediction of Tropical Cyclone Activity in the Vicinity of Taiwan Using the Multivariate Least Absolute Deviation Regression Method. *Terr. Atmos. Ocean. Sci*., **18**, 805–825.

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

Elsner, J.B., and C.P. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting*, **8**, 345–351.

Elsner, J.B., and T.H. Jagger, 2004: A hierarchical Bayesian approach to seasonal hurricane modeling. *J. Climate*, **17**, 2813–2827.

Elsner, J.B., and T.H. Jagger, 2006: Prediction models for annual U.S. hurricane counts. *J. Climate*, **19**, 2935–2952.

Epstein, E.S., 1985: *Statistical Inference and Prediction in Climatology: A Bayesian approach*. Meteor. Monogr., No. 42, Amer. Meteor. Soc., 199 pp.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin, 2004: *Bayesian Data Analysis*. 2nd edition, Chapman & Hall/CRC, 668 pp.

Gray, W.M., C.W. Landsea, P.W. Mielke, and K.J. Berry, 1992: Predicting Atlantic seasonal hurricane activity 6-11 months in advance. *Wea. Forecasting*, **7**, 440–455.

Gray, W.M., C.W. Landsea, P.W. Mielke, and K.J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, **8**, 73–86.

Gray, W.M., C.W. Landsea, P.W. Mielke, and K.J. Berry, 1994: Predicting Atlantic basin seasonal tropical cyclone activity by 1 June. *Wea. Forecasting*, **9**, 103–115.

Hastings, W.K., 1970: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Harr, P.A., and R.L. Elsberry, 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., H.-S. Kim, and P.-S. Chu, 2009: Seasonal prediction of tropical cyclone frequency over the East China Sea through a Bayesian Poisson-regression method. *Asia-Pacific J. Atmos. Sci*., **45**, 45-54.

Jagger, T. H., X. Niu, and J. B. Elsner, 2002: A space-time model for seasonal hurricane prediction. *Int. J. Climatol.*, **22**, 451-465.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc*., **77,** 437–471.

Kim, H.-S., C.-H. Ho, P.-S. Chu, and J.-H. Kim, 2010a: Seasonal prediction of summertime tropical cyclone activity over the East China Sea using the least absolute deviation regression and the Poisson regression. *Int. J. Climatol.*, **30**, 210- 219.

Kim, H.-S., J.-H. Kim, C.-H. Ho, and P.-S. Chu, 2010b: Pattern classification of typhoon tracks using the fuzzy *c*-means clustering method. *J. Climate*, accepted.

Kistler, R., and Coauthors, 2001: The NCEP-NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc*., **82**, 247–267.

Klotzbach, P., and W.M. Gray, 2004: Updated 6-11 month prediction of Atlantic basin seasonal hurricane activity. *Wea. Forecasting*, **19**, 917–934.

Klotzbach, P.J., and W.M. Gray, 2008: Multidecadal variability in North Atlantic tropical cyclone activity. *J. Climate*, **21**, 3929-3935.

Lu, M.-M., P.-S. Chu, and Y.-C. Chen, 2010: Seasonal prediction of tropical cyclone

activity in the vicinity of Taiwan using the Bayesian multivariate regression method. *Wea. Forecasting*, in press.

Ripley, B.D., 1987: *Stochastic Simulation*. John Wiley, New York, 237 pp.

Ritchie, E.A., and G.J. Holland, 1999: Large-scale patterns associated with tropical cyclogenesis in the western Pacific. *Mon. Wea. Rev*., **127**, 2027-2043.

Robert, C., 1995: Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121-125.

Smith, T.M., R.W. Reynolds, R.E. Livezey, and D.C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9**, 1403–1420.

Tu, J.-Y., C. Chou, and P.-S. Chu, 2009: The abrupt shift of typhoon activity in the vicinity of Taiwan and its association with western North Pacific-East Asian climate change. *J. Climate*, **22**, 3617-3628.

Zhao, X., and L.W.K. Cheung, 2007: Kernel-Imbedded Gaussian Processes for Disease Classification using Microarray Gene Expression Data, *BMC Bioinformatics*, 8:67

Zhao, X., and P.-S. Chu, 2006: Bayesian multiple change-point analysis of hurricane activity in the eastern North Pacific: A Markov Chain Monte Carlo approach. *J. Climate*, **19**, 564–578.

Zhao, X., and P.-S. Chu, 2010: "Bayesian Change-Point Analysis for Extreme Events (Typhoons, Heavy Rainfall, and Heat Waves): A RJMCMC approach", *J. Climate*, **23**, 1034-1046.

Year	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Total
1979	1	0	0	0	2	0	0	3
1980	0	0	0	0	3	1	0	4
1981	1	0	0	1	1	0	0	3
1982	2	0	0	1	2	1	0	6
1983	0	0	0	0	1	0	0	1
1984	0	0	0	1	2	0	0	3
1985	0	0	2	0	4	0	0	6
1986	0	0	0	1	2	0	0	3
1987	1	0	0	0	4	0	0	5
1988	1	0	0	0	1	0	0	2
1989	0	0	1	0	1	0	0	2
1990	1	0	0	0	3	1	0	5
1991	0	0	1	1	1	1	0	4
1992	2	0	0	0	2	0	0	4
1993	0	0	0	1	0	0	0	1
1994	0	0	0	0	6	1	0	7
1995	1	0	0	0	2	0	0	3
1996	0	0	0	0	1	1	0	2
1997	0	0	0	0	2	0	0	2
1998	1	0	0	1	2	0	0	4
1999	0	0	0	0	1	1	0	2
2000	1	0	0	0	4	1	0	6
2001	1	0	0	0	4	1	0	6
2002	3	0	0	0	1	0	0	4
2003	1	0	0	0	4	1	0	6
2004	2	0	0	1	5	0	0	8
2005	0	0	0	0	5	0	0	5
2006	1	0	0	0	4	0	0	5

Table 1: Seasonal (JJASO) tropical cyclone counts in the vicinity of Taiwan, stratified by seven cluster types, from 1979 to 2006. The last column refers to the total number of tropical cyclones for each year.

Figure captions

Fig. 1: Track pattern of each type of tropical cyclones in the western North Pacific.

Fig. 2a: Flow chart of analysis procedure for predicting seasonal typhoon activity in the vicinity of Taiwan.

Fig. 2b: Flow chart of forecast procedure for seasonal tropical cyclone activity in the vicinity of Taiwan.

Fig. 3: Predictor selection for type 5. (a) Isocorrelates of seasonal (JJASO) tropical cyclone frequency in the vicinity of Taiwan (the box) with the antecedent May SSTs. (b) Same as in (a), but for SLPs. (c) Same as in (a) but for PW. (d) Same as in (a), but for low-level relative vorticity. (e) Same as in (a) but for vertical wind shear. The hatching denotes the critical region for which the local correlation is statistically significant at the 99% confidence level.

Fig. 4: Simulation results for the seasonal tropical cyclone activity near Taiwan based on track type 5. (a) The median (solid), upper, and lower quartiles (broken) of the predicted TC rate are plotted together with the actual observed TC rate (dotted) during 1979-2006. (b) Same as in (a), but for the predicted and observed tropical cyclone counts.

Fig. 5: Simulation results for the seasonal tropical cyclone activity near Taiwan based on a mix of track types. (a) The median (solid), upper, and lower quartiles (broken) of the LOOCV-predicted TC rate are plotted together with the actual observed tropical cyclone rate (dotted) during 1979-2006. (b) Same as in (a), but for the predicted and observed tropical cyclone counts.

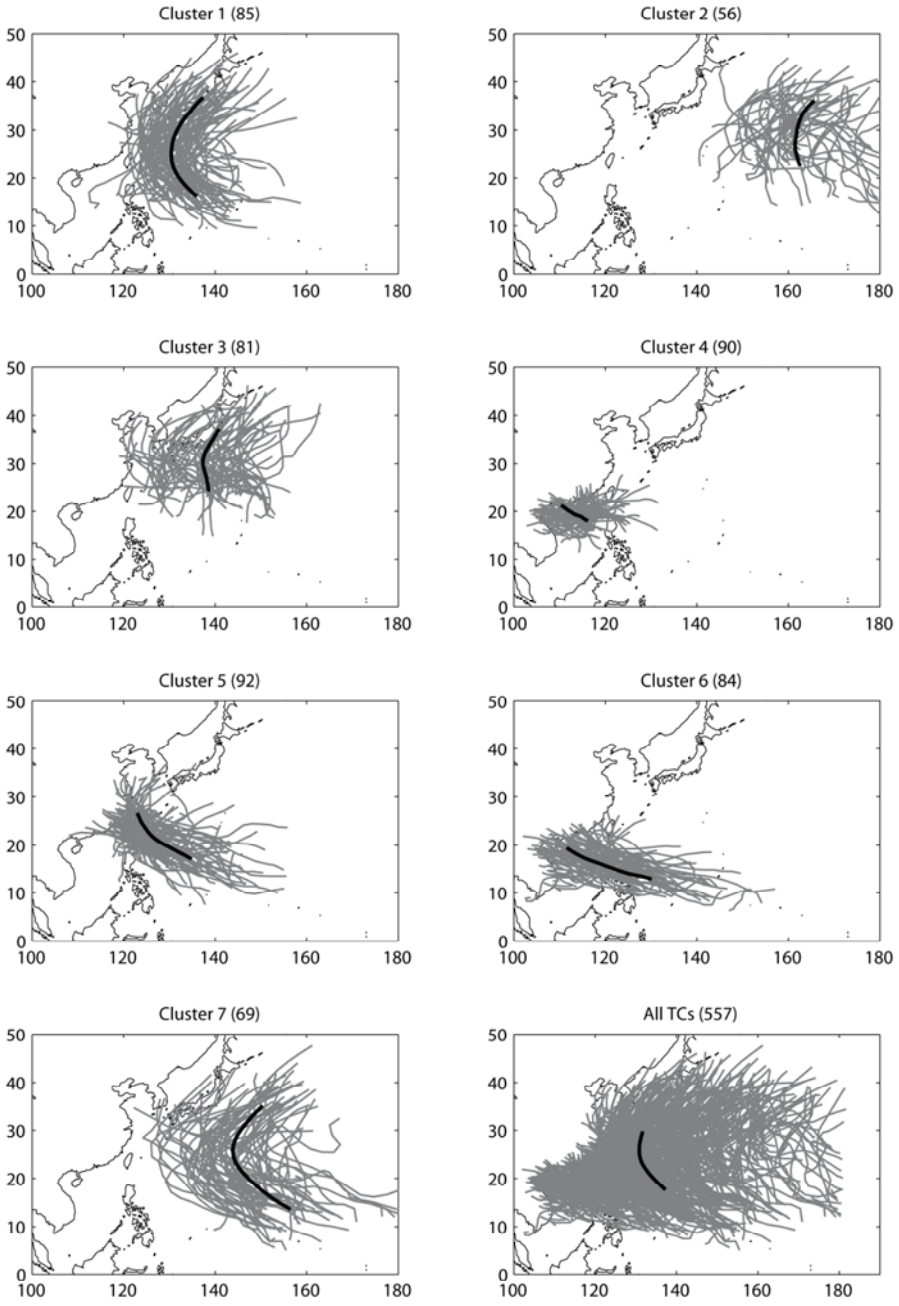


Fig. 1: Track pattern of each type of tropical cyclones in the western North Pacific.

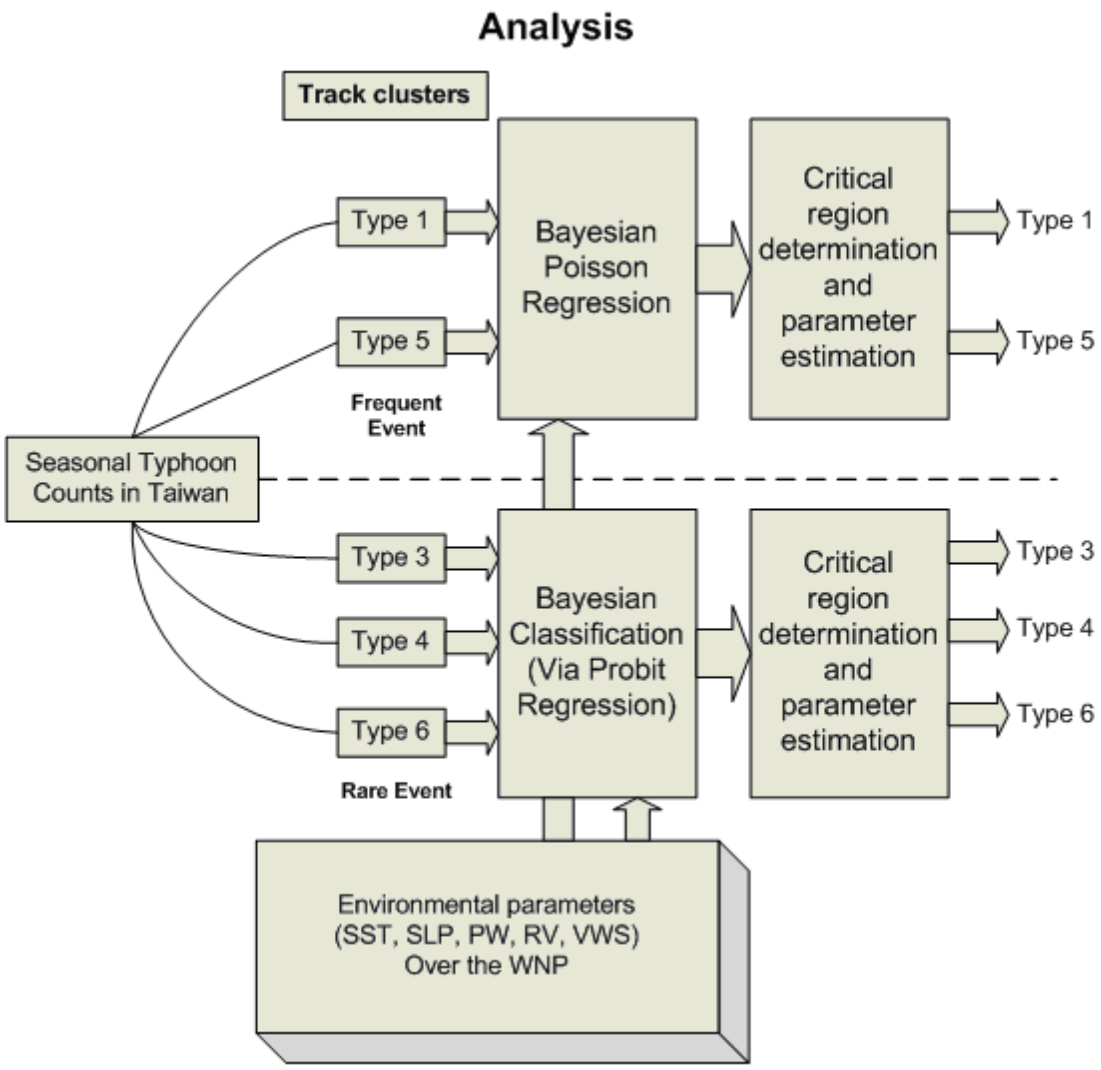


Fig. 2a: Flow chart of analysis procedure for predicting seasonal typhoon activity in the vicinity of Taiwan

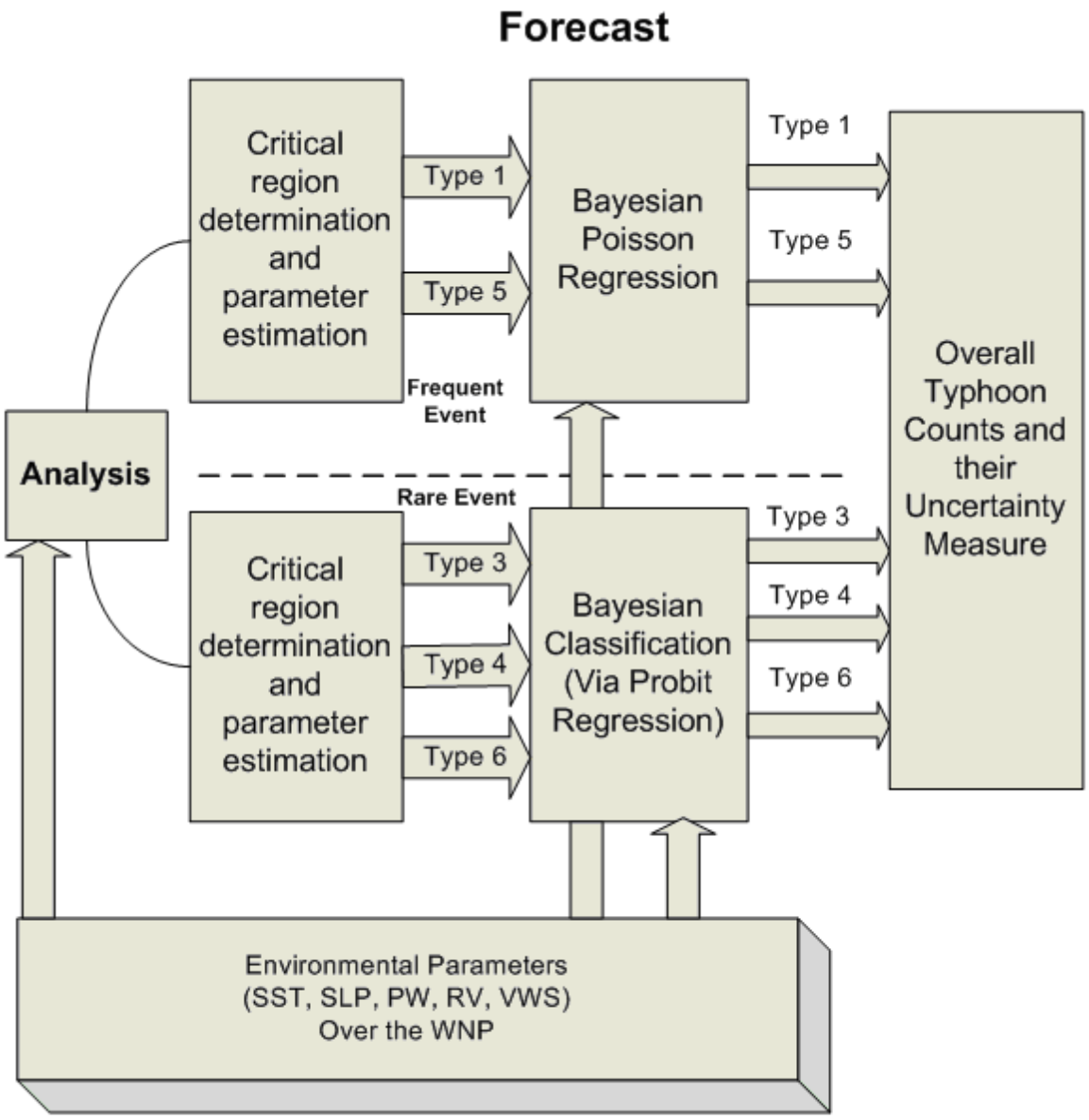


Fig. 2b: Flow chart of forecast procedure for seasonal tropical cyclone activity in the vicinity of Taiwan

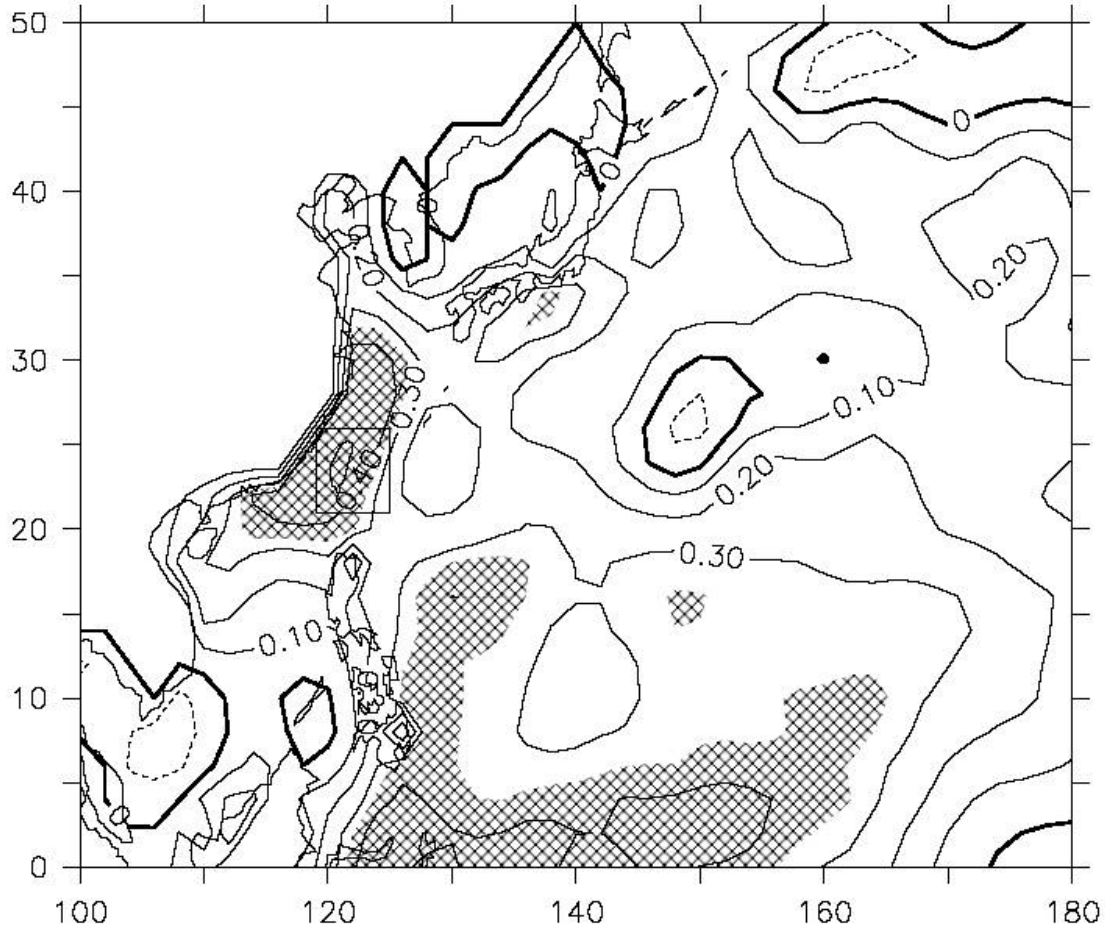
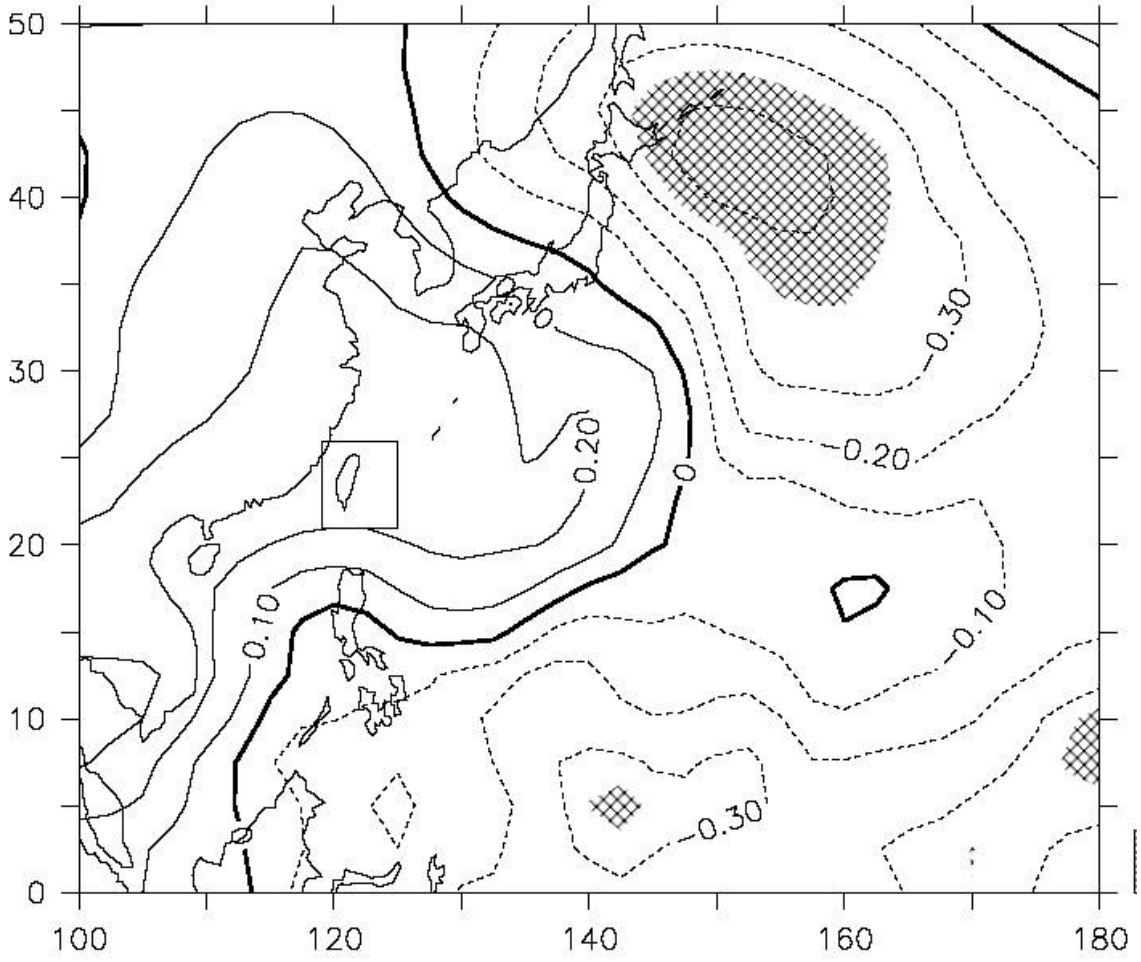
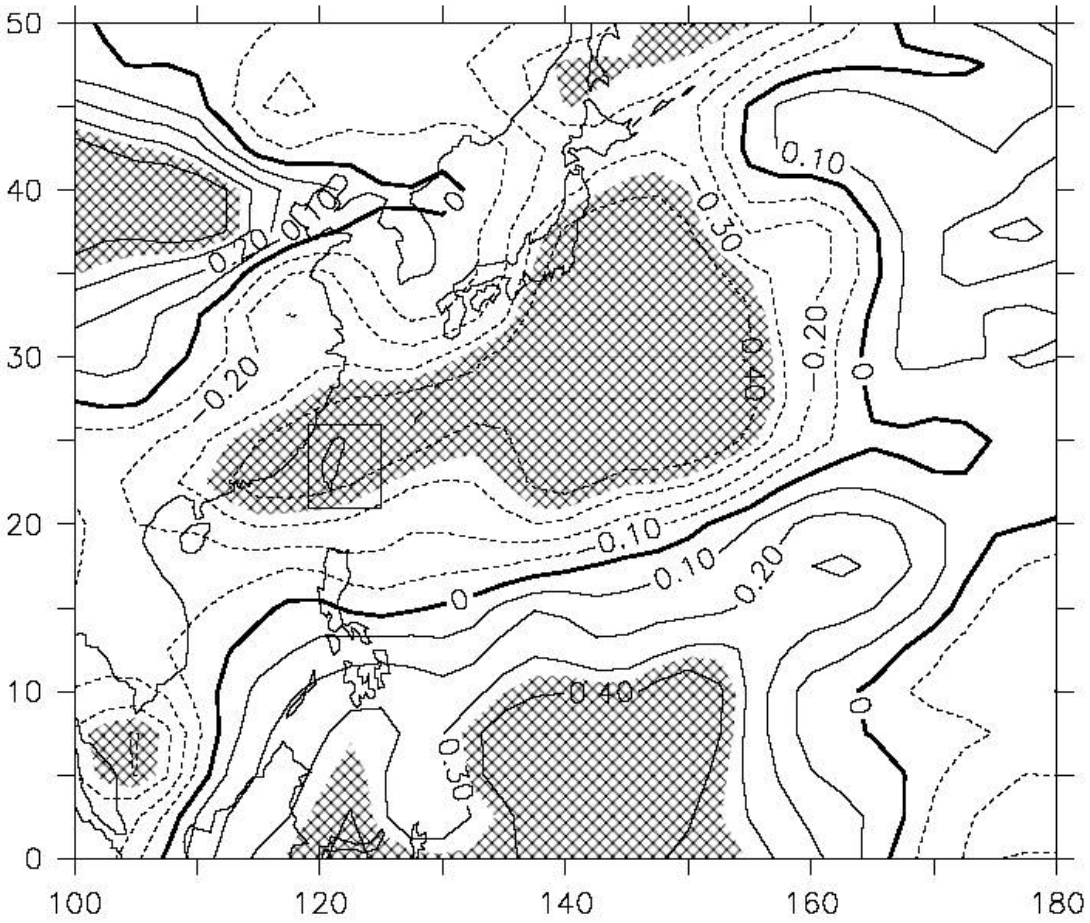


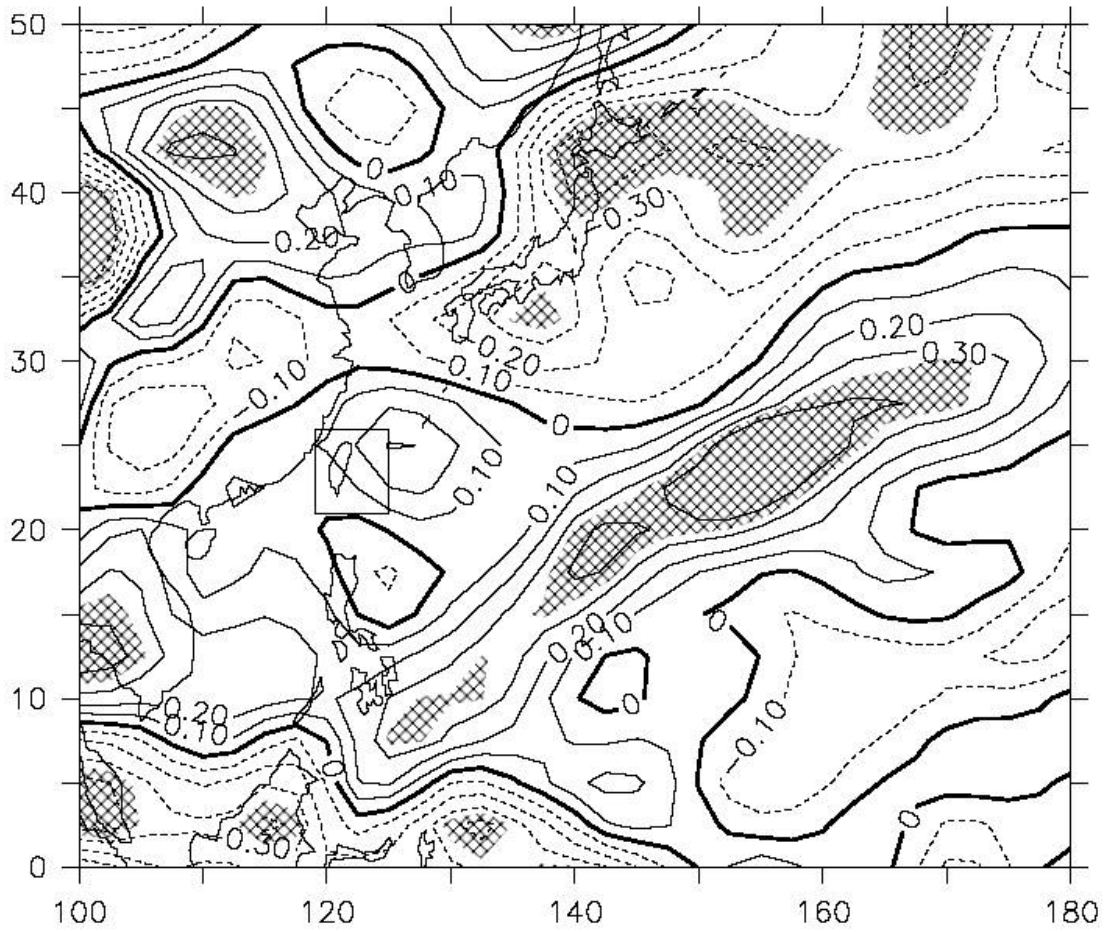
Fig. 3: Predictor selection for type 5. (a) Isocorrelates of seasonal (JJASO) tropical cyclone frequency in the vicinity of Taiwan (the box) with the antecedent May SSTs. The hatching denotes the critical region for which the local correlation is statistically significant at the 99% confidence level.



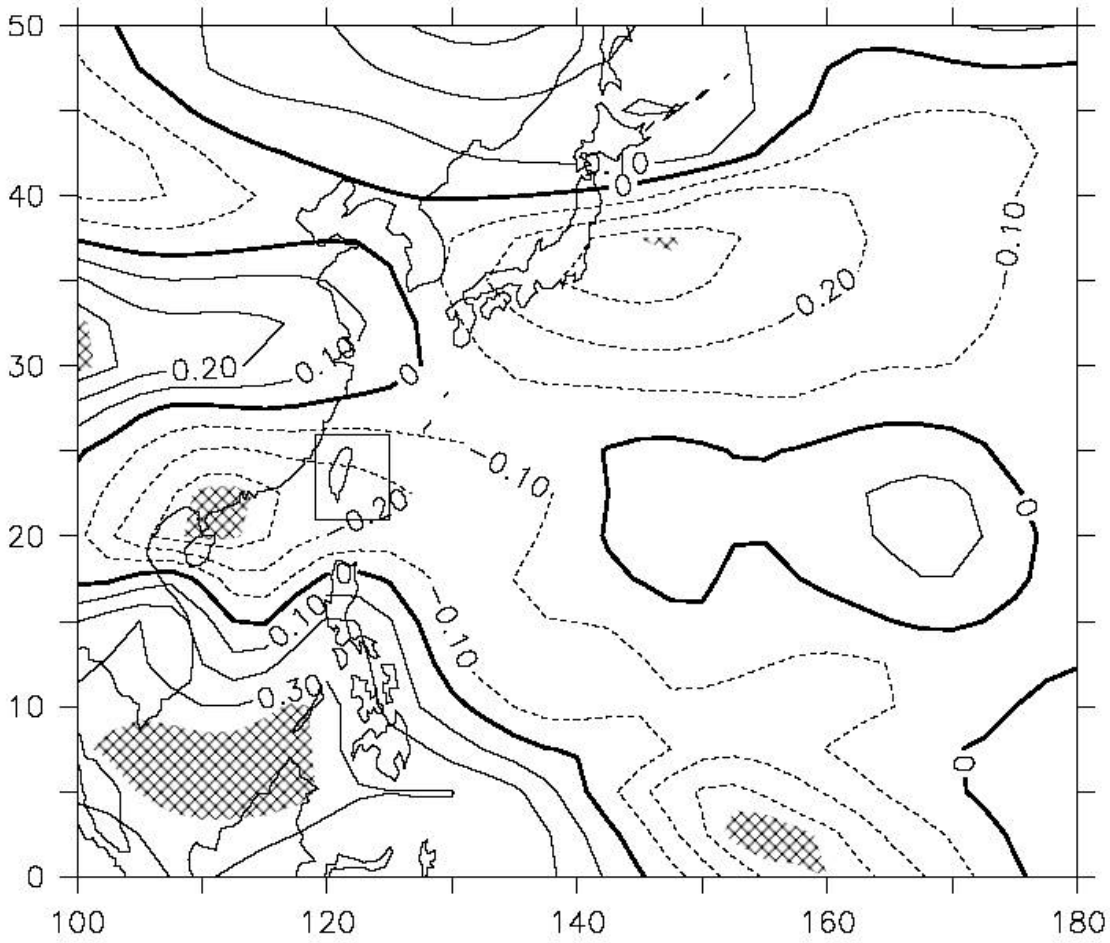
(b) Same as in (a), but for SLPs.



(c) Same as in (a) but for PW.



(d) Same as in (a), but for low-level relative vorticity.



(e) Same as in (a) but for vertical wind shear.

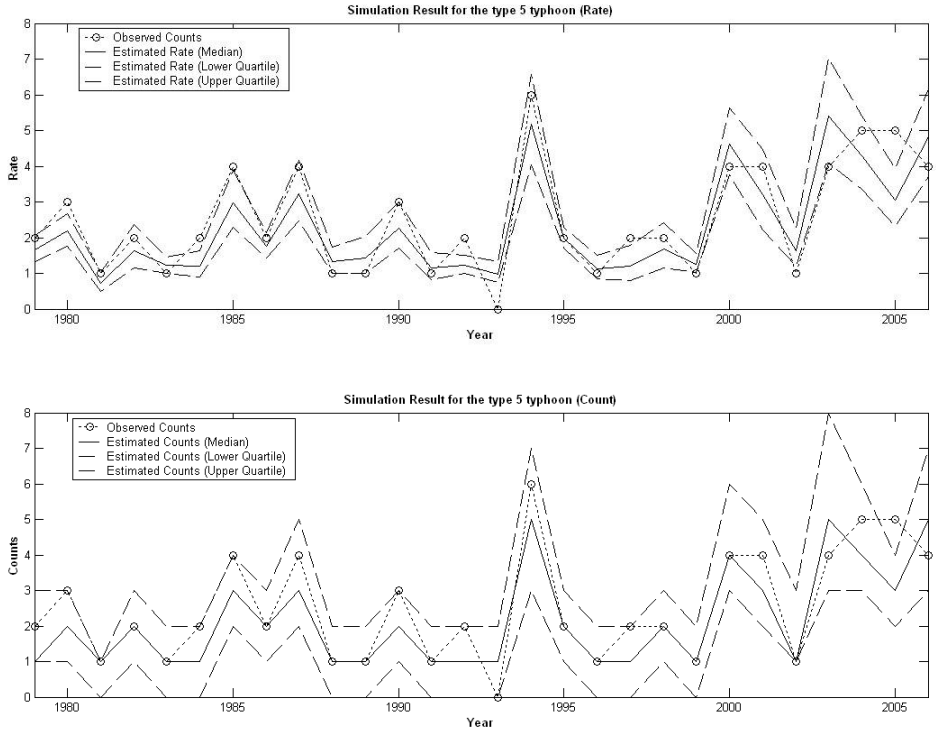


Fig. 4: Simulation results for the seasonal tropical cyclone activity near Taiwan based on track type 5. (a) The median (solid), upper, and lower quartiles (broken) of the LOOCV-predicted TC rate are plotted together with the actual observed tropical cyclone rate (dotted circle) during 1979-2006. (b) Same as in (a), but for the predicted and observed tropical cyclone counts.

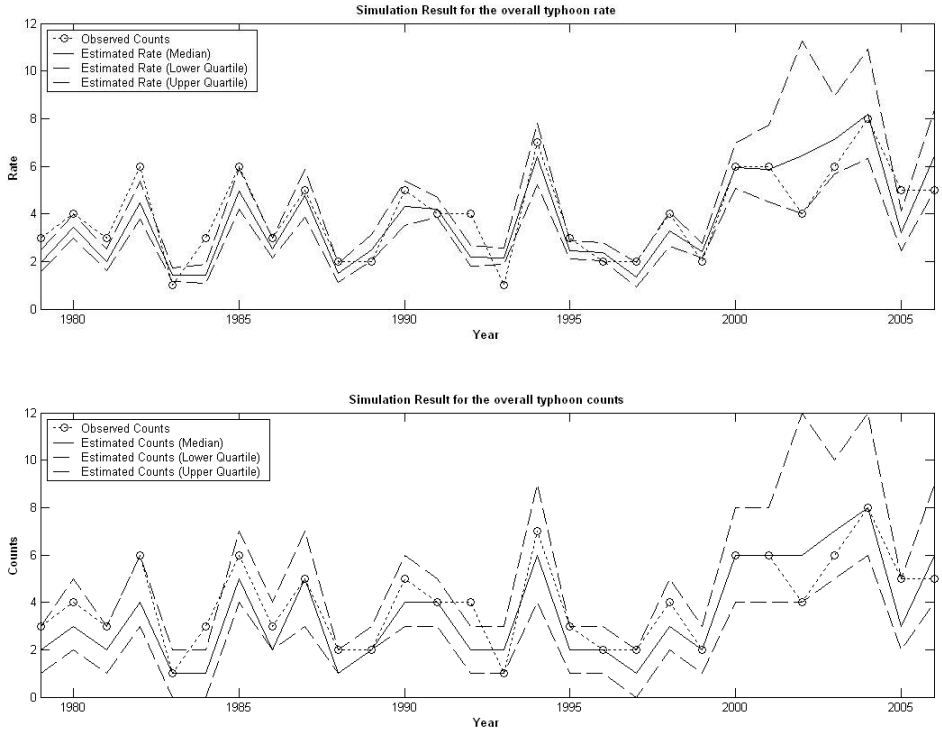


Fig. 5: Simulation results for the overall seasonal tropical cyclone activity near Taiwan area. (a) The median (solid), upper, and lower quartiles (broken) of the LOOCV-predicted TC rate are plotted together with the actual observed tropical cyclone rate (dotted circle) during 1979-2006. (b) Same as in (a), but for the predicted and observed tropical cyclone counts.