Climate Prediction of Tropical Cyclones Activity in the Vicinity of Taiwan Using the multivariate least absolute deviation regression method

Pao-Shin Chu[1], Xin Zhao[2], Mong-Ming Lu[3] and Ching-Teng Lee[3]
1.Department of Meteorology, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa
2.Department of Information and Computer Science, University of Hawaii at Manoa, U.S.A.
3.Research and Development Center, Central Weather Bureau, Taipei, Taiwan, Republic of China

Abstract

In this study, a multivariate linear regression model is applied to predict the annual tropical cyclone (TC) counts in the vicinity of Taiwan using large-scale climate variables available in May. The model is based on the least absolute deviation (LAD) so that regression estimates are more resistant (i.e., not unduly influenced by outliers) than those derived from the ordinary least square method. Through correlation analysis, five variables, including sea surface temperature (SST), sea level pressure (SLP), precipitable water (PW), relative vorticity in key locations of the tropical western North Pacific and CLIPER are identified as predictor data sets. Results from cross-validation suggest that the statistical model is skillful in predicting TC activity, with a correlation coefficient of 0.75 for the period of 1970-2003 (34 years).

1. Introduction

W. Gray pioneered the seasonal hurricane prediction enterprise using a regression-based statistical model called the least absolute deviation method (Gray et al., 1992). They showed that nearly half of the interannual variability of hurricane activity in the North Atlantic could be predicted in advance. This is amazing because a hurricane is a small system, and physical mechanisms governing its formation are complicated. Along the same line of statistical modeling, Chan et al. (1998) developed a model to predict seasonal typhoon activity over the western North Pacific and the South China Sea. When tested against a 30-yr sample through the jackknife method, skillful forecasts are noted for a suite of predictands (e.g., the annual number of typhoons and the annual number of tropical storms and typhoons). The informative results of Chan et al. (1998) are pertinent to the vast western North Pacific basin and the South China Sea. For a smaller geographic domain such as in the vicinity of Taiwan, the frequency of typhoon occurrences may be very different from that of the basin-wide numbers. For example, tropical cyclone counts over the entire western North Pacific during the developing year of the El Niño do not differ much from the long-term climatology, yet the genesis location during the peak and late season of the El Niño developing year is dramatically shifted east- and southward so fewer storms are found to the west of 150°E (Wang and Chan, 2002; Chu, 2004). Therefore, it has yet to be shown that new predictive models would work for a smaller area.

In this study, we attempt to predict the number of tropical cyclones (TCs) in a season in the vicinity of Taiwan area on the basis of the Least Absolute Deviation (LAD) regression method. This method has been tested for many years by Gray and his associates and is quite mature. Section 2 discusses the dataset, and section 3 outlines the LAD model. In section 4, procedures for selecting appropriate predictor variables are described. Section 5 discusses the forecast results. A summary is found in section 6.

2. Data and data processing

The annual tropical cyclone (tropical storms and typhoons) series in the vicinity of Taiwan from 1970 to 2003 is obtained from the Central Weather Bureau. This series covers an area between 21°N-26°N and 119°E-125°E. Monthly mean sea level pressure, wind data at 850- and 200-hPa levels, relative vorticity data at the 850 hPa level, and total precipitable water over the western North Pacific (0°-30°N) are derived from the NCEP/NCAR reanalysis dataset (Kalnay et al., 1997; 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 (Clark and Chu, 2002). The monthly mean sea surface temperatures, at 2° horizontal resolution, are taken from the NOAA Climate Diagnostic Center in Boulder, Colorado. As our interest is to develop a predictive model, only the May data prior to the peak typhoon season are derived.

3. Least Absolute Deviations regression

A linear regression model can be generally written as

$$y(t)=\sum_{j=1}^{K}c_j x_j(t)+N(t) \qquad (1)$$

where $y(t)$ is the desired predictive variable or predictand, $x_i(t)$ for $i=1,\ldots,K$ represent the predictors and c_i for $i=1,\ldots,K$ are the corresponding regression parameters, while $N(t)$ is a random variable and represents the regression deviation (residual). The least square error (LSE) is probably the best known method for fitting linear regression models and by far the most widely used due to its simplicity in computation. However, the LSE is not necessarily the optimum fitting method if the deviation $N(t)$ is not of the Gaussian distribution. Moreover, the residuals in the LAD are computed from the median, whereas in the LSE they are derived from the mean. Because the median is a much more robust estimator of the location than the mean, LAD regression estimates are less sensitive to large outliers (e.g., extreme values) than the LSE method. In particular, if the deviation is double exponentially distributed, the optimum method for linear regression will be LAD.

The basic idea of LAD regression problem is generally stated as below.

Given a sample size of n points $\{\underline{\mathbf{x}}_{\mathrm{i}}, y_i\}$, where $\underline{\mathbf{x}}_{\mathrm{i}} \in R^k$ for $i=1,\ldots,n$, the LAD fitting problem is to find a minimizer, $\hat{\underline{\mathbf{c}}} \in R^K$, of the distance function (absolute deviation). That it,

$$f(\underline{\mathbf{c}})=\sum_{i=1}^{n}\left|y_i-\sum_{j=1}^{K}c_j x_{ij}\right|=\sum_{i=1}^{n}\left|y_i-\langle\underline{\mathbf{c}},\underline{\mathbf{x}}_{\mathrm{i}}\rangle\right|=\left\|\underline{\mathbf{y}}-\mathbf{x}\underline{\mathbf{c}}\right\|_1 \qquad (2)$$

where $\underline{\mathbf{y}}=[y_1,y_2,\ldots,y_n]'$, $\mathbf{x}=[\underline{\mathbf{x}}_1,\underline{\mathbf{x}}_2,\ldots,\underline{\mathbf{x}}_n]'$. $\underline{\mathbf{c}}=[c_1,c_2,\ldots,c_K]'$

such that, $f(\hat{\underline{\mathbf{c}}})=\min(f(\underline{\mathbf{c}}))$.

This problem is solvable because function $f(c)$ is continuous and convex. Due to the nonlinearity of absolute operation, solving LAD model is no longer a linear problem. Since LAD and LP are similar in their very basic nature, an abundance of algorithms was developed based on the well-studied linear programming (LP) problem in the past. The LP problem in standard form is to find $\hat{\underline{x}}$ that maximizes $f(\underline{x})=\langle\underline{c},\underline{x}\rangle$ subject to linear constraint $A\underline{x}\le\underline{b}$ and $\underline{x}\ge 0$ with given vector $\underline{c}$, $\underline{b}$ and matrix A. With some straightforward but tedious derivations, it can be shown that any LAD curve-fitting can be expressed as an equivalent bounded feasible LP problem. We choose the Bloomfield-Steiger algorithm to find the minimizer (Bloomfield and Steiger, 1980). The basic idea of this algorithm is to find the normalized steepest direction in each iteration of the algorithm. Suppose the current fit is $\underline{c}$, and $\underline{\delta_1},\underline{\delta_2},\ldots,\underline{\delta_K}$ is a set of directions along which the next iteration could move, the optimum descent direction being $\underline{\delta_p}$ along which

$$\min(f(\underline{c}+t\underline{\delta_p}),t\in R)=\min[\min f(\underline{c}+t\underline{\delta_i})),i\le K] \quad (3)$$

the inner minimization over t in R. To find this direction, the K -weighted median calculations would need to be done (one for each i in the right hand side of the equation (3)). The pseudo code for the Bloomfield-Steiger (BS) algorithm is listed by (A1) in Appendix A.

For the sake of stability of the algorithm, before applying the raw data to the BS algorithm, it's better to normalize each predictor beforehand. That is, for each sample value of any target predictor, one subtracts it by the associated sample mean, followed by dividing the centered value by the sample standard deviation. This ensures that the normalized sample values of each predictor have zero sample mean and unit sample variance.

4. Procedures for selecting predictor variables

We first calculate the correlation coefficient between the annual TC count series in the vicinity of Taiwan and the relative data for each climate variable (including SST, SLP, PW, VWS and vorticity) on each available point on the grid (all within the same rectangular plane bounded by 0N, 30N in N-S direction and 120E, 180E in E-W direction). This procedure will end with a matrix of correlation coefficients. Then, for simplicity, we choose the point associated with the maximum correlation coefficient (in absolute value) as the key location and apply a two-tailed t-test to this maximum value. If the resulted t-score is bigger than the critical value under some significance level (such as 0.05), we will choose the data series on this key location as a predictor. The t-score can be calculated as below:

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (4)$$

where, t is distributed as Student's t with $n-2$ degrees of freedom, r is the tested correlation coefficient and n is the sample size (Bevington 2002). Specifically, for this study, $n=34$, leading the critical value under significance level 0.05 is 0.34.

4.1 Sea surface temperatures (SSTs)

SSTs are known to be important for TC formation and intensification. Warmer SSTs are expected to fuel the overlying atmosphere with additional warmth and moisture, thereby reducing atmospheric stability and increasing the likelihood of deep tropical convection. In this context, we calculate the correlation between the number of TCs in the vicinity of Taiwan area and the preseason SST (i.e., May) over the western North Pacific. The contour plot for the correlation is shown in Fig. 1a where a maximum (0.58) is found in the core of the warm pool (2°N, 146°E). This value is statistically significant; thus the SST series at this point is chosen as a predictor.

4.2 Sea level pressures (SLPs)

The contour plot for the correlation between the TC frequency in the vicinity of Taiwan and the May SLP is shown in Fig. 1b. The highest (in absolute value sense) negative correlation (-0.48) is found at 15°N, 130°E, very close to Taiwan. This result is physically reasonable as lower SLPs to the southeast of Taiwan in May correspond to higher TC frequency near Taiwan, and vice versa. Dynamically, the juxtaposition of the maximum correlations found in Figs. 1a and 1b suggests a Rossby-wave type response of atmosphere to equatorial heating as demonstrated in Gill's model. Note that the maximum correlation in Fig. 1b is also statistically significant.

4.3 Precipitable water (PW)

The entrainment of drier air in the midtroposphere results in less buoyancy for the tropical convection systems as well as diminish the upper-level warming due to decreased release of latent heat (Knaff, 1997). Consequently, drier atmosphere tends to suppress deep convection and inhibits TC activity. Positive and strong correlations between PW and TC frequency are found in the core of the tropical western North Pacific where the correlation coefficient reaches more than 0.6 and is statistically significant (Fig. 1c). Thus, more moisture in the atmosphere in the tropical western North Pacific in May is conducive to more TC activity near Taiwan and vice versa. Of particular note is a smaller area of high correlation near 10°N, 137.5°E, and this maximum center lies between the points with maximum correlations found in Figs. 1a and 1b.

4.4 Relative vorticity

The monsoon trough in the western North Pacific is characterized by the strong relative cyclonic vorticity in the lower troposphere and is known to be the birthplace of typhoons. Figure 1d displays a positive and high correlation in the Philippine Sea (17.5°N, 132.5°E) with a value of 0.47, which is statistically significant. This result is internally consistent with those in Figs. 1b and 1c, in that lower SLPs in the Philippine Sea induce stronger cyclonic vorticity near the surface and higher moisture in the atmosphere, leading to more TC frequency in the vicinity of Taiwan.

4.5 Vertical wind shear (VWS)

Strong VWS disrupts the organized deep convection (the so-called ventilation effect) which inhibits intensification of the TCs. Negative but weak correlations exist in the low latitudes with a center near 5°N, 155°E (Fig. 1e). This correlation (-0.3) is, however, statistically insignificant. A positive and strong correlation is noted near 5°N, 120°E but that location is too close to the western boundary of the domain and there is a lack of physical explanation for the VWS and TC frequency. As a result, VWS is not used in the subsequent forecasting simulation.

4.6 CLIPER

The analysis of variance (ANOVA) method is a statistical technique to test the existence of hidden periods in a time series. After the hidden periods for the series are given, we can use persistence (CLIPER) to find the periodical oscillation for this series.

The basic idea for CLIPER prediction is very simple. We first find the most significant period (with the maximum score), say p . If this score is less than the critical value under the desired confidence level, we stop and don't use CLIPER as a predictor. Otherwise, we re-group the series with respect to the period p . Then we can calculate each group's mean, which will be the prediction for each year of this group. To avoid the overfitting problem, we only use the first significant hidden period to reconstruct the CLIPER predictor for this study. Fig. 2a displays the result of applying ANOVA to the annual TC series in the vicinity of Taiwan for each possible period, where we can see, a 16-year hidden period has the F-score way above the 95% confidence line. By using this hidden period, we construct the CLIPER predictor and the resulted series is plotted in Fig. 2b. Actually, the correlation coefficient between the resulting series and the original TC series is 0.51 The CLIPER shown in Fig. 2b is thus considered as a predictor.

5. Prediction results

With the various predictor variables selected through correlation analysis, including SST, SLP, PW, Vorticity and CLIPER, we then use a cross-validation method to establish the overall forecasting ability of LAD model. The approach is as follows (Yu et al., 1997). The predictor and predictand data set of T time points are divided into L segments. A model is then developed using the data of L-1 segments. This model is then used to predict TC frequency in the remaining segment. This process is repeated by changing the segment that has been excluded from the model development. In this study, we remove only one observation at a time for each case. By doing this, we obtain N predictions. These predicted values are then correlated with N observations and the overall forecast skill can thus be determined.

The cross-validation results are shown in Fig. 3 and a reasonably skillful forecast is seen. In some years, forecast values are smaller than actual observations (e.g., 1982) but in other years they are larger than observations (e.g., 1996). In fact, the mean of the real observation is 3.85 while the mean of leave-one-out cross-validation is 3.82. That is, there is no systematic bias revealed in the prediction scheme. The correlation coefficient between the cross-validation result and the raw TC data is 0.75, which means that almost 57% variation of the TC activity in the vicinity of Taiwan area can be predicted based on the large-scale climate information one month prior to the peak season and the past history of TC records.

6. Summary

Climate prediction of tropical cyclone activity has been carried out for the North Atlantic and the western North Pacific by various research teams. Because of the vast expanse of ocean basins and pronounced interannual climate variations in the tropics, there is no guarantee that such basin-wide prediction is also applicable to smaller regions within a basin. In this study, a multivariate least-absolute-deviation regression method is adopted to predict the annual tropical cyclone frequency in the vicinity of Taiwan using large-scale climate information available in May. Through correlation analysis between TC frequency and each individual climate variables (e.g., SST, SLP) over the western North Pacific, we identified key locations to be used as the predictor data sets. We then used the leave-one-out cross-validation technique to test the predictability of TC frequency. The cross validation provides a nearly unbiased estimate of true forecast skill. The linear correlation between the cross-validation predictions and the corresponding actual observations for the test period 1970-2003 is 0.75. This result implies that it would be possible to predict the annual TC counts for a small area with reasonable skill using a physically based regression model. In the future, it would be of interest to determine the predictability of TC frequency when climate variables chosen are for months prior to May. Apart from pure scientific inquiry, if good skills can be obtained, say, when the April predictors are chosen, it would allow decision makers more lead time to response.

Acknowledgments

This project has been funded by the Central Weather Bureau, CWB 95-3M-06.

References:

Bevington, P. R. and Robinson, D. K., 2002: Data Reduction and Error Analysis for the Physical Sciences, 3rd Edition. New York: McGraw-Hill.

Bloomfield, P., and W. L. Steiger, 1980: Least absolute deviations curve-fitting.
J. Sci. Stat. Comput., 1, 290-301.

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.

Chu, P.-S., 2004: ENSO and tropical cyclone activity. Hurricanes and Typhoons: Past, Present, and Future, R.J. Murnane and K.-B Liu, Eds., Columbia University Press, 297-332.

Clark, J.D., and P.-S. Chu, 2002: Interannual variation of tropical cyclone activity over the central North Pacific. J. Meteor. Soc. Japan, 80, 403-418.

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.

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

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

Knaff, J.A., 1997: Implications of summertime sea level pressure anomalies in the tropical Atlantic region. J. Climate, 10, 789-804.

Wang, B., and J.C.L. Chan, 2002: How strong ENSO affect tropical storm activity over the western North Pacific. J. Climate, 15, 1643-1658.

Figure caption

Fig. 1. (a) Correlation map between tropical cyclone count series in the vicinity of Taiwan and the May sea surface temperatures over the tropical western North Pacific (0°N - 30°N in latitude and 120°E – 180°E in longitude).

(b) Same as (a) except for the May sea level pressures.

(c) Same as (a), except for the May precipitable water.

(d) Same as (a), except for the May relative vorticity at 850 hPa.

(e) Same as (a), except for the May vertical wind shear

Fig. 2. (a) F-scores for each possible hidden period by applying the ANOVA to the annual TC series in the vicinity of Taiwan.

(b) Constructed CLIPER predictor for the TC series with a 16-year hidden period. The Solid line presents the CLIPER prediction (in leave-one-out cross-validation) and dotted line shows the original series.

Fig. 3. Time series of observed and cross-validated forecasts of tropical cyclone counts. The solid line denotes the leave-one-out cross-validation results by using LAD method, while the dotted line displays the original observation.

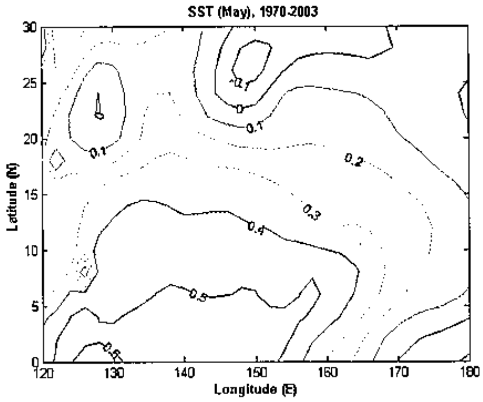

Fig. 1a

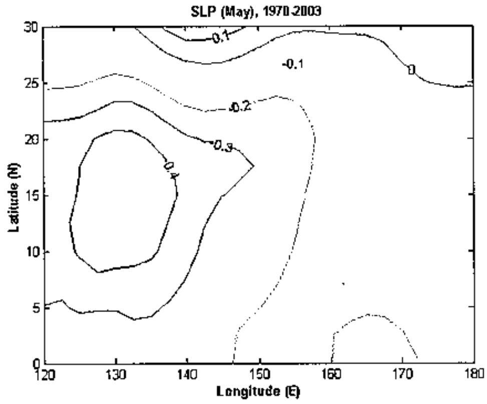

Fig. 1b

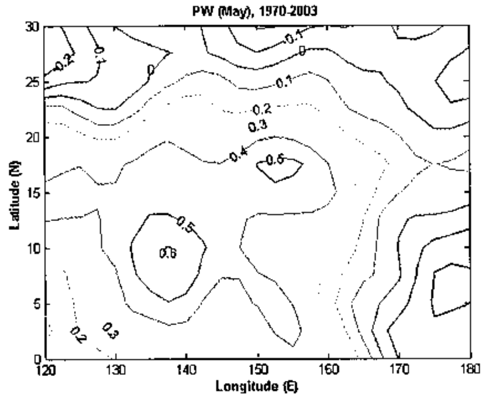

Fig. 1c

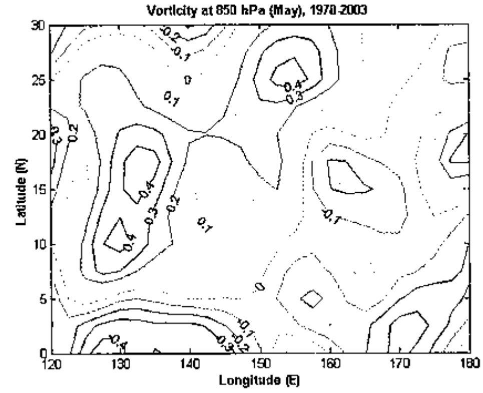

Fig. 1d

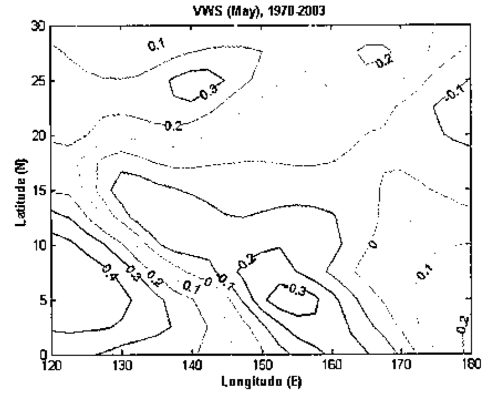

Fig. 1e

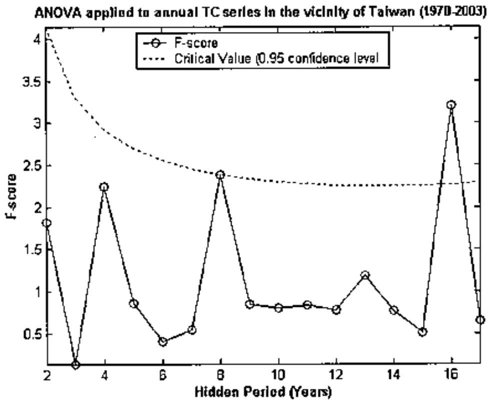

Fig. 2a

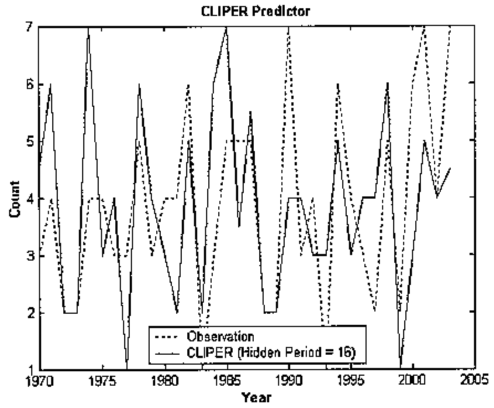

Fig. 2b

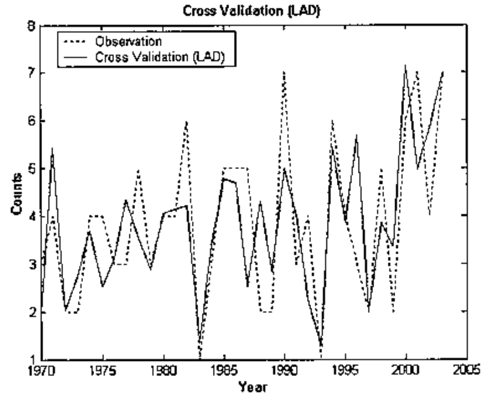

Fig. 3