Jour. Korean Earth Science Society, v. 31, no. 5, p. 454–464, September 2010

Development of a Probability Prediction Model for Tropical Cyclone Genesis in the Northwestern Pacific using the Logistic Regression Method

Ki-Seon Choi[1], KiRyong Kang[2,*], Do-Woo Kim[3], and Tae-Ryong Kim[2]

[1]Korea Meteorological Administration, Seoul 156-720, Korea
[2]National Typhoon Center, Korea Meteorological Administration, Jeju 699-942, Korea
[3] Pukyong National University, Busan 608-737, Korea

Abstract: A probability prediction model for tropical cyclone (TC) genesis in the Northwestern Pacific area was developed using the logistic regression method. Total five predictors were used in this model: the lower-level relative vorticity, vertical wind shear, mid-level relative humidity, upper-level equivalent potential temperature, and sea surface temperature (SST). The values for four predictors except for SST were obtained from difference of spatial-averaged value between May and January, and the time average of Niño-3.4 index from February to April was used to see the SST effect. As a result of prediction for the TC genesis frequency from June to December during 1951 to 2007, the model was capable of predicting that 21 (22) years had higher (lower) frequency than the normal year. The analysis of real data indicated that the number of year with the higher (lower) frequency of TC genesis was 28 (29). The overall predictability was about 75%, and the model reliability was also verified statistically through the cross validation analysis method.

Keywords: Tropical cyclone genesis, Logistic regression model, Cross validation

Introduction

There are many researches on applying the statistical models behind the seasonal development of tropical cyclones (TCs). Leading researches over the Atlantic ocean include Gray et al. (1992, 1993, 1994), Elsner and Schmertmann (1993), Hess et al. (1995), and Mestre and Hallegatte (2008). In addition, researches over the Southern Pacific ocean include Nicholls (1992), McDonnell and Holbrook (2004), and Leroy and Wheeler (2008). In contrast, the researches covering the western North Pacific (WNP) is low compared to other oceanic regions with relatively short history (Chan, 1998; Lee et al., Kwon et al., 2007; Choi et al., 2010). One major commonality among these different researches is the fact that they implied climatological factor of various different synoptic scale including ENSO and QBO index as the predictor in their statistical models. However, the results from this methodology are limited to the particular oceanic region in terms of the statistical analysis. In order to overcome this limitation, we adopted the basic and general parameter revised by Gray (1968) as the predictors in the statistical model. These parameters were composed of the following six elements: 1) the sea surface temperature (SST), 2) conditional instability, 3) cyclonic absolute vorticity in the lower level, 4) high relative humidity in mid-level, 5) anticyclonic absolute vorticity in the upper level, and 6) vertical wind shear. Ryan et al. (1992), Watterson et al. (1995) and Royer et al. (1998) have claimed of optimum predictors in the statistical and dynamic model for the TC forecasting by combining the listed parameters. These researchers incorporated this idea by defining “Seasonal Genesis Parameter (SGP).” In addition, Ward (1995), Lehmiller et al. (1997), Elsner et al. (2000), DeMaria et al. (2001), Jagger et al. (2001, 2002), and McDonnell and Holbrook (2004) utilized factors of a pre-TC activity season as the predictor in their statistical models and have produced a successful prediction results.

On the other hand, majority of the aforementioned researches consists of the multiple linear regression model aiming at predicting the TC genesis frequency

*Corresponding author: krkang@kma.go.kr
Tel: 82-64-801-0224
Fax: 82-64-805-0366

in quantitative for the short-term period. However, for long-term prediction case, the qualitative prediction could complement the inaccuracy of the quantitative predictions, i.e. predicting the probability in the relative abundance of summer precipitation of the specified year. One of most frequently used technique for the probability prediction is the logistic regression model (Wilks, 1995). This model can predict and determine the increase or decrease pattern of the actual event calculating the probability form. Therefore, the larger the predicted value, the higher the development probability of the concerned event. It has been widely used to forecast the precipitation probability (Vislocky and Young, 1989; Vislocky and Fritsch, 1995b; Frei and Schär, 2001).

In this study, we are trying to provide a simplified statistical model for the qualitative prediction from a long-term perspective, like determining the higher or lower genesis frequency of seasonal TC activity. We used the predictors suggested by Gray (1968), and also adopted the concept by Ward (1995). In addition, the predictability of the model was tested to compare the frequency of TC genesis in the WNP between June and December and to verify the accuracy of this model.

The data and methodology used in this study were described in section 2, and the climatological characteristics of TC genesis frequency was analyzed in section 3. The more detail about the logistic regression model used in this study is described in section 4. The model validation results were shown in section 5, and summary was finally given in section 6.

Data and methodology

TC activity information from 1951 to 2007 was obtained from the best track archives of the Regional Specialized Meteorological Centers-Tokyo (RSMC-Tokyo). The dataset is composed of names, the migration path (longitude and latitude positions), minimum surface pressures and maximum sustained wind speed of the TC for every six hour interval. In addition, atmospheric data was based on the reanalysis dataset provided by the National Centers for Environmental Prediction-National Center for Atmospheric Research of the United States (NCEP-NCAR) (Kalnay et al., 1996; Kistler et al., 2001) on zonal/meridional current (ms^{-1}), relative humidity (%), air temperature (^{o}C) and specific humidity (gkg^{-1}). The dataset is composed of space-time resolution of longitude/latitude of $2.5^{o}\times2.5^{o}$, 17 vertical levels, and of the monthly average. We also implemented the Niño 3.4 index averaged for February to April provided by the Climate Prediction Center (CPC), an affiliated organization of the NOAA. This index represents the moving average from December of the previous year on a tri-monthly basis.

We first investigate the climatological characteristics of the TC genesis over the WNP and, construct the logistic regression model to predict the TC genesis. This logistic regression model will be analyzed by using the SPSS (Statistical package for the Social Sciences Ver.10.0.7), a PC type statistical analysis package. The validity of this regression model will be conducted through the cross validation method using the hindcast.

Climatological characteristics of TC genesis

Figure 1a represents the climatological monthly mean TC genesis frequency over the WNP for the past 57 years. The annual average number of TC genesis for this period was 26.6. About 70% of this number is occurred between July and October. It was only 0.6 for January and May, and 3.4 for June and December. During the period between January and May, it was only 3, however during the period between June and December, almost 90% of annual average frequency of TC genesis (23.7) was occurred. We can safely assert that majority of the TC is formed during the latter half of the year. So, in this study we focused on the total genesis frequency of TC between June and December. Figure 1b represents the time series of total TC genesis frequency between June and December for the past 57 years. From this time series

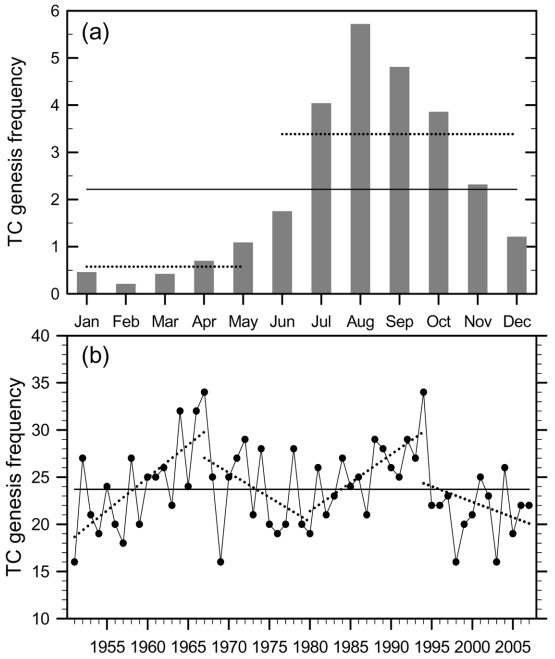

Fig. 1. (a) Climatological mean monthly tropical cyclone (TC) genesis frequency and (b) time series of total TC genesis frequency for June-December for 57 years in the western North Pacific (WNP). Left and right dotted and solid lines in the upper figure denote averages for January-May, for June-December, and for January-December, respectively. Solid and dotted lines in the lower figure indicate a trend and a normal year (1971-2000) mean, respectively.

we can witness the distinct inter-decadal variation, change of approximately 10-year period. When simply eye-watching the pattern of this periodic variation, we can classify them into groups of early 1950s to late 1960s and early 1980s to mid 1990s, representing the increasing stage; late 1960s to early 1980s and mid 1990s to late 2000 as decreasing stage. We can also see that two periods with approximately 20 years cycle is formed from the 1950s to 1970s and from 1980s to 2000. However, since post mid-1990s, there exists only two recorded years, 2001 (24 cases) and 2004 (26 cases), which show a higher frequency level than the average number for the period from 1971 to 2000. This fact, although the analysis result is limited to TC genesis between June and December, corresponds to the results of Webster et al. (2005) stating that the TC genesis frequency in the WNP is on a decrease.

Logistic regression model

Predictors

We, first, try to define the predictor for developing a statistical forecast model of TC genesis in the WNP region (Table 1). As indicated in Table 1, the predictors are composed of the dynamic parameters (relative vorticity in the lower-level and vertical wind shear) and thermal parameters (relative humidity in the mid-level, equivalent potential temperature in the upper-level and sea surface Temperature (SST)) defined as necessary elements for the TC genesis by Gray (1968). 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 TC. Strong vertical wind shear (VWS) also disrupts the organized deep convection that inhibits the TC intensification. In addition, SST is known as an important for TC formation and

Table 1. List of the logistic regression model variables used in the development of models of TC genesis in the WNP region. Area-averaged value of 110-180E, 5-25N was used for the independent variables except for SST

	Description of logistic regression model variable	
	Predictor	Name
Independent variables	Difference of 925 hPa relative vorticity (10^{-5} s^{-1}) between May and January	RVOR
	Difference of 200-850 hPa vertical wind shear (m s^{-1}) between May and January	VWS
	Difference of 700 hPa relative humidity (%) between May and January	RHUM
	Difference of 300 hPa equivalent potential temperature (^{o}C) between May and January	EPT
	Niño-3.4 index averaged for February-April	SST
	Predictand	
Dependent variable	Entire TC number from June to December	TC

Table 2. Correlation coefficients between variables. Bold values are significant at 95% level

Variable	TC	RVOR	VWS	RHUM	EPT	SST
TC	1	-0.13	-0.09	-0.19	0.07	-0.10
RVOR	-0.13	1	0.03	0.14	0.11	0.13
VWS	-0.19	0.14	1	0.25	0.17	0.38
RHUM	-0.09	0.03	0.25	1	-0.17	0.18
EPT	0.07	0.11	-0.17	0.17	1	0.31
SST	-0.10	0.13	0.18	0.38	0.31	1
Multicollinearity		0.98	0.75	0.75	0.73	0.75

intensification. Warmer SST is expected to fuel the overlying atmosphere with additional heating and moisture supply and the warm and moisture atmosphere can play an important role as energy for the convection, reducing atmospheric stability and increasing the likelihood of deep tropical convection. These five predictors were implemented as independent variables to predict the TC genesis frequency (dependent variable) between the months of June and December. Among the predictors which are going to be input into the model, the value of all variables except for the SST was calculated through the difference of area-averaged values in the WNP TC genesis region (5-25°N, 110-180°E) between May and January. This is because the environment conditions in the tropics or subtropics of the WNP during the spring can have an effect on the genesis of TC in the region for the rest of the year. Concerning this phenomenon, Chan (2008) claims when the environment fails to establish favorable conditions for the TC genesis during spring and early summer, there is a high possibility that the concerned year shows a low level of TC genesis compared to normal year. Concerning the predictor of SST, there are many published research results showing the dominant effect of ENSO on the TC genesis frequency and location (Chan, 1985; Dong, 1988; Wu and Lau, 1992; Chen et al., 1998; Wang and Chan, 2002; Camargo and Sobel, 2005).

There should be no strong correlation among the independent variables comprising the multiple linear regression model. If there is high correlation, the characteristics of one independent variable should be overlap with the characteristics of another independent variable, thus lowering the validity of the regression coefficient of the model. In other words, symptom of multicollinearity may be observed in situations: (1) small changes in the data produce wide swings in the parameter estimates; (2) coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high; (3) coefficients may have the "wrong" sign or implausible magnitude. Therefore, in the current study we examined multicollinearity among the independent variables and relationship between dependent and independent variables (Table 2). Here as the value of multicollinearity approaches closer to 1, the lower the correlative relationship, signifying its adaptability as a regression coefficient in the regression model. More detailed information regarding the multicollinearity can be found from the study of Greene (2000). The result of this study showed that the correlation between the independent variables was below ±0.4, and the multicollinearity was high values above 0.7, thus indicating the statistical validity of the independent variables used in this study. For example, even though the relative humidity of the mid-level and SST showed negative relationship with the TC genesis, the correlation coefficients are not statistically significant because of above reason (Table 2).

Forecasting model and its results

The simple linear or multiple linear regression analysis method is used when the dependent variable need to be quantitatively determined. However, there are many cases that the dependent variables are determined by not quantitatively but qualitatively. For

example, estimating the possibility of success and failure in the cooperation or the sales of products is some kind of a good case. This kind of concept could be implemented in the field of meteorology: When we need to answer if it is going to be rain during the weekend from a short-term perspective, or if there is a drought during the summer from a long-term perspective. When dependent variables are comprised of bisect variables, we can analyze the relationship between the dependent and independent variables using the logistics regression model. This methodology is similar to the simple and multiple linear regression analysis to explain the relationship between the dependent variables and independent variable through their linear combination, however it does not directly predict the outcome for a certain event, because it predicts the possibility of the genesis of a certain event. Therefore, the dependent variable has either a value of 0 or 1, and does not require any additional hypothesis on normal distribution, thus has the merit of applying to many different cases.

Here we try to answer whether the total number of TC formed from June to December is higher or lower than normal year based on a logistic regression model. The value of normal year is defined as the average number of TC generated between June and December in the WNP for the past 30 years (1971-2000). First, we calculate the value of the independent variables for each concerned year, and at the same time, allocate '0' if the total TC genesis frequency is lower than the value of normal year's, and '1' if it is higher according to the equitation (1). The average number of TC genesis (=23.7) for 30 years is used as a standard classifying a binary value (0 or 1). The value of the independent variables and the observed binary value of the TC genesis frequency are indicated in Table 3.

$$OBSTC = \begin{cases} 0, TC < 23.7 \\ 1, TC > 23.7 \end{cases} \quad (1)$$

The finalized logistic regression model to predict the possibility of higher or lower TC genesis frequency compared to normal year can be expressed through equation (2), and the multiple linear regression of the independent variables in relations to this model can be expressed through equation (3).

$$P(TC) = \frac{e^f}{1+e^f} \quad (2)$$

$$f = 0.53 - 0.02(RVOR) - 0.11(RHUM) + 0.02(VWS) + 0.85(EPT) - 0.69(SST) \quad (3)$$

The binary value of the hindcasted TC genesis frequency (HCST TC) from this model and the forecast probability (P) is expressed in Table 3. Here the division line for P is 0.5, which is corresponding to the TC genesis frequency of normal year between June and December. If the P is greater (smaller) than the division value, 0.5, it means the probability of the total TC genesis frequency between June and December in certain year is higher (lower) than normal year.

In order to determine the hindcasting performance of the hindcasted results through the logistic regression model, the analysis of probability of detection (POD; Wilks, 1995) must be preceded (Table 4): Closer the POD results to 1, more accurate the predictability. Among the 28 years that showed the higher TC genesis frequency than normal year, the actual number of years which were hindcasted to have been higher, were 21, and the POD was 0.75. Therefore the average POD for these two incidences were analyzed to be at 0.75, which signifies that about 75 cases would be hindcasted accurately out of 100 cases. We also calculated the mean value for the actual TC genesis frequency for every 0.1 forecast probability (P) (Fig. 2). If the probability model is inaccurate, then there is a high possibility with low P (i.e. P< 0.5). So, based on the standard P value, for the group P<0.5, it was lower frequency (21.1) than the normal year (23.7), and for the group P>0.5, it showed higher value (25.3) than normal year. Therefore, it could conclude that the annual forecast probability from this model in Table 3 is in general valid.

Table 3. Annual values of the logistic regression model variables and observation (OBS) and hindcasted (HCST) TCs and FCST probability (P). Shaded areas mean that OBS TC is not consistent with FCST TC

Year	RVOR	VWS	RHUM	EPT	SST	OBS TC	HCST TC	P
1951	-2.33	-17.37	9.08	1.68	-0.7	0	1	0.76
1952	6.27	1.25	17.53	0.69	0.1	1	0	0.28
1953	6.55	-13.78	12.03	1.12	0.4	0	0	0.37
1954	19.61	0.42	9.94	0.42	-0.2	0	0	0.39
1955	-5.16	18.61	13.79	1.33	-0.9	1	1	0.79
1956	14.36	1.95	14.46	0.67	-0.6	0	0	0.43
1957	-9.71	-48.68	7.61	2.06	0.3	0	1	0.60
1958	3.84	-15.39	8.67	2.34	1.1	1	1	0.60
1959	9.34	-16.81	18.64	1.54	0.4	0	0	0.27
1960	9.25	-3.41	12.79	1.52	-0.3	1	1	0.60
1961	8.35	10.34	5.30	1.31	-0.2	1	1	0.78
1962	16.25	6.72	4.55	0.28	-0.4	1	1	0.59
1963	39.84	-8.53	3.32	1.97	0.0	0	1	0.68
1964	18.68	-16.72	17.87	1.42	0.0	1	0	0.28
1965	-0.10	-1.08	1.85	1.97	-0.2	1	1	0.89
1966	18.44	-5.21	15.51	1.83	0.8	1	0	0.34
1967	-8.19	-4.12	5.75	0.60	-0.6	1	1	0.72
1968	-3.31	-10.6	11.99	-0.27	-0.8	1	0	0.36
1969	-1.83	-17.42	20.24	2.95	0.9	0	0	0.48
1970	-2.63	-7.36	12.04	1.76	0.2	1	1	0.62
1971	3.81	-4.42	11.90	0.52	-1.2	1	1	0.59
1972	-3.67	-29.46	9.98	2.52	0.0	1	1	0.74
1973	-5.15	16.22	14.42	0.22	0.5	0	0	0.34
1974	16.79	-7.30	7.25	0.90	-1.2	1	1	0.69
1975	-17.79	2.89	11.65	0.43	-0.7	0	1	0.65
1976	6.97	3.20	10.94	1.29	-0.9	0	1	0.73
1977	-2.27	1.08	14.96	0.35	0.3	0	0	0.29
1978	-0.74	41.64	15.29	1.18	0.0	1	1	0.70
1979	-5.49	-24.07	18.03	1.35	0.1	0	0	0.33
1980	39.24	-3.51	11.69	1.37	0.3	0	0	0.33
1981	-11.13	-12.59	-2.22	1.01	-0.4	1	1	0.87
1982	16.01	-17.89	5.65	1.04	0.2	0	0	0.48
1983	-3.47	35.56	12.61	0.86	1.6	0	0	0.42
1984	-12.7	6.11	8.42	1.61	-0.2	1	1	0.83
1985	-7.49	-16.73	10.51	1.74	-0.8	1	1	0.78
1986	18.13	-2.27	6.28	0.56	-0.3	1	1	0.52
1987	5.21	15.81	6.73	0.28	1.2	0	0	0.37
1988	-24.74	20.87	10.55	0.71	0.1	1	1	0.73
1989	-11.99	-2.72	4.64	-0.33	-1.2	1	1	0.69
1990	-1.07	12.75	12.32	0.86	0.3	1	1	0.51
1991	18.45	-30.77	8.80	0.60	0.3	1	0	0.23
1992	-16.72	-29.22	9.43	1.67	1.5	1	0	0.41
1993	-14.62	-10.79	5.22	0.99	0.5	1	1	0.64
1994	-0.40	-7.56	12.49	1.56	0.2	1	1	0.56
1995	11.20	16.59	13.80	0.50	0.6	0	0	0.31
1996	-15.68	-15.04	6.39	0.93	-0.5	0	1	0.73
1997	14.86	-32.51	6.49	0.96	-0.1	0	0	0.42
1998	13.43	75.3	27.27	1.48	1.4	0	0	0.33
1999	-3.53	3.87	3.31	-1.71	-0.9	0	0	0.38
2000	-23.22	-21.34	6.92	0.81	-1.0	0	1	0.77
2001	-3.16	4.57	10.08	0.70	-0.4	1	1	0.62
2002	9.83	-16.35	17.79	0.97	0.2	0	0	0.22
2003	20.72	9.11	14.58	0.88	0.5	0	0	0.30
2004	17.57	-31.58	17.72	0.94	0.2	1	0	0.21
2005	-9.92	15.37	9.59	0.24	0.4	0	0	0.42
2006	7.58	-13.56	7.62	-0.97	-0.3	0	0	0.19
2007	-11.28	-27.43	16.56	-0.70	0.1	0	0	0.11

Table 4. Statistics on TC genesis frequency predicted from the logistic regression model

		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	21	7	28	21/28=0.75
	TC>23.7	7	22	29	22/29=0.76
Total		28	29	57	(21+22)/57=0.75

* OBS - Observation, FCST - Forecast, and POD - Probability of detection

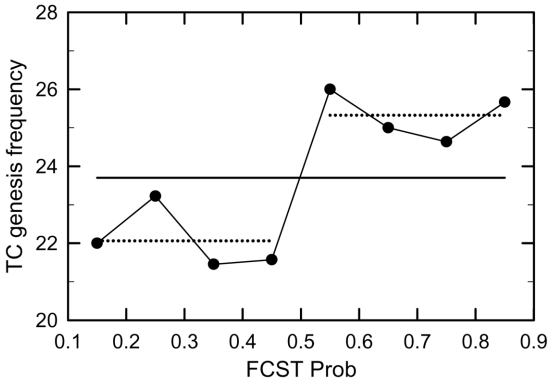

Fig. 2. Mean TC genesis frequency according to a forecast probability yielded from a logistic regression model. Solid line denotes the average TC genesis frequency (=23.7) during the period from 1971 to 2000. Left and right dotted lines indicate TC genesis frequencies averaged between 0.1-0.5 FCST probability and between 0.6-0.9 FCST probability, respectively.

Validation of the Model Predictability

Testing the validity of the logistic regression model developed in this study through the simple analysis is not sufficient to determine the overall validity of the model. In fact, when the data period for the model development is reduced or expanded, it is possible to become a different regression model and forecast results from the model. In this section, we examine the validity of the developed model through the cross validation analysis. The cross validation analysis is conducted through the following stages utilizing the hind cast method.

1) To develop the logistic regression model from the independent variables of the years excluding the 1950s (1961-2007: 47 years) in order to forecast the TC genesis frequency in the 1950s (1951-1960: 10 years).
2) To perform the forecasting TC genesis frequency for 1950s using the model.
3) To conduct POD analysis to investigate the forecast performance.
4) To repeat the above process for every 10-year interval till 2000s (7 years for 2000s).
5) To collect the POD for each 10-year from 2000s to calculate the total POD for 57 years, and compare this POD with the POD from the developed model based on the independent variables for the past 57 years, and determine the validity for the latter model.

Table 5 represents the difference between the regression coefficient calculated through the pre-stated methods and the regression coefficient calculated through the independent variables from the past 57 years. The sign of majority of the regression coefficient calculated from the cross validation analysis correlates to the sign of regression coefficient calculated from the independent variables from the past 57 years. In addition, the difference between the two regression coefficients is below the decimal point range. In particular the Mean Absolute Error (MAE) of the regression coefficient were all below 0.3 thus signifying no major significant difference with the regression model derived from the independent variables from the past 57 years. The MAE is defined as equation 4.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (4)$$

Here, n denotes the sum of the numbers of correlation coefficient and constant. And, f and y represent regression coefficients in a logistic

Table 5. Regression coefficients in each logistic regression model constructed by using independent variables of the remaining decades except for each decade used to forecast. For example, "1950s" means that the logistic regression model was constructed by using independent variables of the remaining 47-year except for 1950s. The regression coefficients of the second column (A) are used in formula (3). Mean absolute error (MAE) indicates the average of absolute values yielded by subtraction of the coefficient (A) from the former coefficients

RC	All	1950s		1960s		1970s		1980s		1990s		2000s	
	A	B	B-A	C	C-A	D	D-A	E	E-A	F	F-A	G	G-A
Constant	0.532	0.182	-0.350	-0.107	-0.639	0.061	-0.471	0.031	-0.501	-0.121	-0.653	0.436	-0.096
RVOR	-0.023	-0.018	0.005	-0.015	0.008	-0.022	0.001	-0.005	0.018	-0.022	0.001	-0.021	0.002
VWS	0.022	0.003	-0.019	0.018	-0.004	0.012	-0.010	0.010	-0.012	0.032	0.010	0.020	-0.002
RHUM	-0.106	-0.082	0.024	-0.108	-0.002	-0.049	0.057	-0.076	0.030	-0.053	0.053	-0.096	0.010
EPT	0.846	0.992	0.146	1.259	0.413	0.784	-0.062	0.773	-0.073	0.978	0.132	0.809	-0.037
SST	-0.692	-0.883	-0.191	-0.656	0.036	-0.800	-0.108	-0.363	0.329	-1.556	-0.864	-0.654	0.038
MAE			0.123		0.184		0.118		0.161		0.286		0.031

Table 6. As in Table 4., but for decadal prediction. Here, a logistic regression model is constructed using the remaining decades except for each decade used to forecast

(a) 1950s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	4	1	5	0.80
	TC>23.7	2	3	5	0.60
Total		6	4	10	0.70

(b) 1960s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	2	2	4	0.50
	TC>23.7	1	5	6	0.83
Total		3	7	10	0.70

(c) 1970s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	4	0	4	1.00
	TC>23.7	2	4	6	0.67
Total		6	4	10	0.80

(d) 1980s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	4	1	5	0.80
	TC>23.7	0	5	5	1.00
Total		4	6	10	0.90

(e) 1990s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	4	3	7	0.57
	TC>23.7	2	1	3	0.67
Total		6	4	10	0.50

(f) 2000s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	5	2	7	0.71
	TC>23.7	0	0	0	0
Total		5	2	7	0.71

(g) 50s-00s		OBS		Total	POD
		TC<23.7	TC>23.7		
FCST	TC<23.7	23	9	32	0.72
	TC>23.7	7	18	25	0.72
Total		30	27	57	0.72

regression model developed from the independent variables excluding each decade and in a logistic regression model derived from the independent variables from the past 57 years, respectively. In addition, although Table 5 does not directly indicate this, the majority of POD for the forecasting of 47 years (if excluding 2000s, total 50 years) was faired within the 0.7 range (1950s: 0.68, 1960s: 0.72, 1970s: 0.70, 1980s: 0.68, 1990s: 0.77, 2000s: 0.74), thus providing that the high accuracy and validity of the regression model derived from the independent variables for the past 57 years was not a coincidence.

Table 6 represents the forecast results from the developed logistic regression model excluding each decade. The highest forecast performance was shown during the 1980s (POD: 0.90) and the lowest forecast performance was shown during the 1990s (POD: 0.50). The lowest forecast performance could be cause of the fact that the trend of TC genesis frequency during this decade is significantly steep comparing to that during other decades, as described in Figure 2b. All Other decades showed higher POD than 0.7.

When we exclude the 1990s, the logistic regression models excluded each decade accurately predicted for 8 years out of any given 10 years. This, when we exclude the 1990s, is much more accurate result than the forecast performance of the regression model developed from the independent variables of the past 57 years in Table 4. However, the average POD for the entire year was 0.72, thus it shows no significant difference from the POD resulting from the regression model based on the independent variables of the past 57 years.

Summary and Conclusion

While the general simple linear regression or the multiple linear regression model is utilized and used to forecast based on quantitative data, the logistic regression model is different because it can provides qualitative forecast value. In this study, we tried to develop a simple statistical model in order to forecast a probability of the higher or lower level of genesis frequency of tropical cyclone (TC) between June and December comprising climatologically 90% of the TC genesis in the western North Pacific (WNP). Total five independent variables (predictors) were utilized in the forecast model including the dynamic parameters (relative vorticity of the lower-level, vertical wind shear) and the thermal parameters (relative humidity in the mid-level, equivalent potential temperature in the upper-level and Niño 3.4 index). Four predictors, excluding the Niño 3.4 index, were used to get the difference between the area-averaged values in the WNP TC genesis region (5-25^{o}N, 110-180^{o}E) between May and January. Niño 3.4 index is averaged for the months from February to April. The binary value 1 and 0 were used to indicate the status of the total TC genesis frequency between June and December each year: For 1 (0), it means that the total TC genesis frequency is higher (lower) than normal value of TC genesis frequency.

Using the newly developed logistic regression model, the total TC genesis between June and December of each concerned year was predicted for the past 57 years (from 1951 to 2007). The model produced a forecast probability (P) between 0 and 1. If P is greater (less) than 0.5, it shows high probability that the TC genesis frequency is higher (lower) than normal year. The predicted results showed that 21 (22) years were higher (lower) frequency than the normal year's. In the real data, the number of year that showed the higher (lower) frequency of TC genesis was 28(29).

In order to determine the validity of this model, the cross validation analysis method was applied to the predictability of the TC genesis frequency for excluded each decade, using the predictors excluding each decade from 1950s to 2000s to construct the regression model. The results of the regression coefficient and forecast by the model showed no significant difference to the model results derived from the utilization of independent variables of the past 57 years, thus suggesting the validity of the newly developed model as an accurate statistical model. Therefore, the forecast probability (P) from this model could be effectively utilized as a probability forecast value in predicting the status of TC genesis frequency in one particular year (from June to December) in lieu with normal year.

In general, the logistic regression model has beneficial point that it can provide fast forecasting information, while the multiple linear regression model to predict the TC genesis (i.e. Choi et al. 2010) has somewhat complicate procedures which take more time to produce the final information. And, the logistic regression model developed in this research requires many independent variables to predict the non-linear characteristics of dependent variable, thus it still has some limitations. It is crucial for this model to minimize the number of the independent variables with maximum predictability. In this study when comparing the frequency of TC genesis to normal year, we focused on only two ranges of TC genesis status: higher or lower than normal. In the future, we will focus on more precise prediction model with more categories and shorter time scale like monthly in predicting a probability of the TC genesis frequency.

Acknowledgments

This research was supported by the National Typhoon Center of Korea Meteorological Administration as a part of major project named 'Operation of the National Typhoon Center (1131-301-210-13). And also the authors deeply thank two anonymous reviewers for giving us great comments for this article to have less error in the logical development.

References

Camargo, S.J. and Sobel, A.H., 2005, Western North Pacific tropical cyclone intensity and ENSO. Journal of Climate, 18, 2996-3006.

Chan, J.C.L., 1985, Tropical cyclone activity in the northwest Pacific in relation to the El Niño/Southern Oscillation phenomenon. Monthly Weather Review, 113, 599-606.

Chan, J.C.L., 1998, Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea. Weather and Forecasting, 13, 997-1004.

Chan, J.C.L., 2008, A simple seasonal forecast update of tropical cyclone activity. Weather and Forecasting, 23, 1016-1021.

Chen, T.C., Weng, S.P., Yamazaki, N., and Kiehne, S., 1998, Interannual variation in the tropical cyclone formation over the western North Pacific. Monthly Weather Review, 126, 1080-1090.

Chia, H.H. and Ropelewski, C.F., 2002, The interannual variability in the genesis location of tropical cyclones in the Northwest Pacific. Journal of Climate, 15, 2934-2944.

Choi, K.-S., Moon, J.-Y., Chu, P.-S., and Kim, D.-W., 2010, Seasonal prediction of tropical cyclone genesis frequency over the western North Pacific using teleconnection patterns. Theoretical and Applied Climatology, 100, 191-206, doi: 10.1007/s00704-009-0182-1.

Crosby, S.C. and Ferraro, R.R., 1995, Estimating the probability of rain in an SSM/I FOV using logistic regression. Journal of Applied Meteorology, 34, 2476-2480.

DeMaria, M., Knaff, J.A., and Connell, B.H., 2001, A tropical cyclone genesis parameter for the tropical Atlantic. Weather and Forecasting, 16, 219-233.

Dong, K.Q., 1988, El Niño and tropical cyclone frequency in the Australian region and the northwest Pacific. Australian Meteorological Magazine, 28, 219-225.

Elsner, J.B. and Schmertmann, C.P., 1993, Improving extendedrange seasonal predictions of intense Atlantic hurricane activity. Weather and Forecasting, 8, 345-351.

Elsner, J.B., Liu, K.B., and Kocher, B., 2000, Spatial variations in major U.S. hurricane activity: Statistics and a physical mechanism. Journal of Climate, 13, 2293-2305.

Frei, C. and Schär, C., 2001, Detection Probability of Trends in Rare Events: Theory and Application to Heavy Precipitation in the Alpine Region. Journal of Climate, 14, 1568-1583.

Gray, W.M., 1968, Global view of the origin of tropical disturbances and storms. Monthly Weather Review, 96, 669-700.

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

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

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

Greene, W.H., 2000, Econometric Analysis (Fourth edition). Prentice-Hall, NJ, USA, 256 p.

Hess, J.C., Elsner, J.B., and LaSeur, N.E., 1995, Improving seasonal hurricane predictions for the Atlantic basin. Weather and Forecasting, 10, 425-432.

Jagger, T.H., Elsner, J.B., and Niu, X., 2001, A dynamic probability model of hurricane winds in coastal counties of the United States. Journal of Applied Meteorology, 40, 853-863.

Jagger, T.H., Niu, X., and Elsner, J.B., 2002, A space-time model for seasonal hurricane prediction. International Journal of Climatology, 22, 451-465.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D., 1996, The NCEP/NCAR 40-Year reanalysis project. Bulletin of American Meteorological Society, 77, 437-471.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., Dool, H., Jenne, R., and Fiorino, M., 2001, The NCEP/NCAR 50-year reanalysis. Bulletin of the American Meteorological Society, 82, 247-267.

Kwon, H.-J., Lee, W.-J., Won, S.-H., and Cha, E.-J., 2007, Statistical ensemble prediction of the tropical cyclone activity over the western North Pacific. Geophysical Research Letters, 34, L24805, doi: 10.1029/2007GL032308.

Lee, W.-J., Park, J.-S., and Kwon, H.-J., 2007, A statistical model for prediction of the tropical cyclone activity

over the western North Pacific. Journal of the Korean Meteorological Society, 43, 175-183.

Lehmiller, G.S., Kimberlain, T.B., and Elsner, J.B., 1997, Seasonal prediction models for North Atlantic basin hurricane location. Monthly Weather Review, 125, 1780-1791.

Leroy, A. and Wheeler, M.C., 2008, Statistical prediction of weekly tropical cyclone activity in the Southern Hemisphere. Monthly Weather Review, 136, 3637-3654.

McDonnell, K.A. and Holbrook, N.J., 2004, A Poisson regression model of tropical cyclogenesis for the Australian-southwest Pacific Ocean region. Weather and Forecasting, 19, 440-454.

Mestre, O. and Hallegatte, S., 2008, Predictors of tropical cyclone numbers and extreme hurricane intensities over the North Atlantic using generalized additive and linear model. Journal of Climate, 22, 633-648.

Nicholls, N., 1992, Recent performance of a method for forecasting Australian seasonal tropical cyclone activity. Australian Meteorological Magazine, 40, 105-110.

Royer, J.F., Chauvin, F., Timbal, B., Araspin, P., and Grimal, D., 1998, A GCM study of the impact of greenhouse gas increase on the frequency of occurrence of tropical cyclones. Climatic Change, 38, 307-343.

Ryan, B.F., Watterson, I.G., and Evans, J.L., 1992, Tropical cyclone frequencies inferred from Gray's yearly genesis parameter: Validation of GCM tropical climates. Geophysical Research Letter, 19, 1831-1834.

Vislocky, R.L. and Young, G.Y., 1989, The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. Weather and Forecasting, 4, 202-209.

Vislocky, R.L. and Fritsch, J.M., 1995b, Improved model output statistics forecasts through model consensus. Bulletin of American Meteorological Society, 76, 1157-1164.

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

Ward, G.F.A., 1995, Prediction of tropical cyclone formation in terms of sea-surface temperature, vorticity and vertical wind shear. Australian Meteorological Magazine, 44, 61-70.

Watterson, I.G., Evans, J.L., and Ryan, B.F., 1995, Seasonal and interannual variability of tropical cyclogenesis: Diagnostics from large-scale fields. Journal of Climate, 8, 3052-3066.

Webster, P.J., Holland, G.J., Curry, J.A., and Chang, H.R., 2005, Changes in tropical cyclone number, duration, and intensity in a warming environment. Science, 309, 1844-1846.

Wilks, D.S., 1995, Statistical Methods in the Atmospheric Sciences. Academic Press, UK, 240 p.

Wu, G. and Lau, N.C., 1992, A GCM simulation of the relationship between tropical storm formation and ENSO. Monthly Weather Review, 120, 958-977.

Manuscript received: February 3, 2010
Revised manuscript received: March 23, 2010
Manuscript accepted: May 10, 2010