Indian J. Met. Hydrol. Geophys. (1978), **29**, 3, 487-500

551.509.313:551.515.2(267)

Statistical prediction of tropical storm motion over the Bay of Bengal and Arabian Sea

CHARLES J. NEUMANN

NOAA, National Hurricane Center, Miami, Florida, U.S.A.

and

G. S. MANDAL*

Meteorological Office, New Delhi

(*Received on 13 September 1976*)

ABSTRACT. Statistical analog models for the prediction of tropical cyclone motion have been developed for all of the world's tropical cyclone basins. The popularity and success of these models is related to their ability to optimize the variance reducing potential of climatology and persistence. At the same time, they effectively recognize certain repetitious synoptic patterns without having to use fields of sometimes unavailable and often unreliable synoptic data over the tropics.

The analog method can be simulated by the use of continuous functions derived through the use of multivariate regression analysis. This method simulates all aspects of the analog process including the derivation of probability ellipses while avoiding certain operational problems inherent to the purely analog approach.

This paper describes the derivation of such a model for the north Indian Ocean tropical cyclone basin. It is patterned after an earlier paper (Neumann and Randrianarison 1976) which describes the development of a similar model for the southwest Indian Ocean tropical cyclone basin.

1. Introduction

Operational models for the statistical prediction of tropical cyclone motion fall into one of two categories, *i.e.*, analog models or multivariate regression equation models. The latter, in order of increasing sophistication, can be further subdivided into, (*a*) those models which exclude predictors derived from synoptic data, (*b*) those models which include current and 24-hour old synoptic data and (*c*) those models which include predictors derived from both current and numerically forecast synoptic data (statistical-dynamical models). Recent experience in the tropical cyclone belt of the North Atlantic Ocean shows that increased sophistication is not always a guarantee of better operational performance. A priori reasoning suggests otherwise. Such a situation is quite discouraging to those engaged in the development of these models.

The enigma lies in the fact that statistical models which use synoptic data fields are geared to older developmental data sets extending back in time some 30 years. These typically consist of gridded geopotential heights over vast data-void areas. In many instances, the data are no better than pure climatology. Current objective analysis methods are de-emphasizing height values and are more concerned with obtaining direct wind analysis as augmented with thousands of satellite derived winds. The statistical models, being tied to a rather archaic set of development data, will have to be re-evaluated.

Similar difficulties will likely be encountered in other tropical regions. Accordingly, it is important that statistical models being developed in these regions take maximum advantage of the variance reducing potential offered by climatology and persistence. This study outlines the development of such a model for the north Indian Ocean tropical cyclone basin.

2. Analog and simulated analog models

The HURRAN (Hurricane Analog) model developed in 1968 by Hope and Neumann (1970) was initially put into operational use at the National Hurricane Center for the 1969 season (Simpson *et al.* 1970). Gupta and Datta (1971) adapted the method to the Bay of Bengal. Having improved on the HURRAN and Gupta and Datta model, the U. S. Navy has developed analog models for virtually all of the world's tropical cyclone basins. Their model for the north Indian area is described by Brand *et al.* (1974).

Sikka and Suryanarayana (1968) developed a model based on climatology and persistence for forecasting the movement of tropical storm in Indian

*Work performed while detailed as a World Meteorological Organization Fellow at the National Hurricane Center

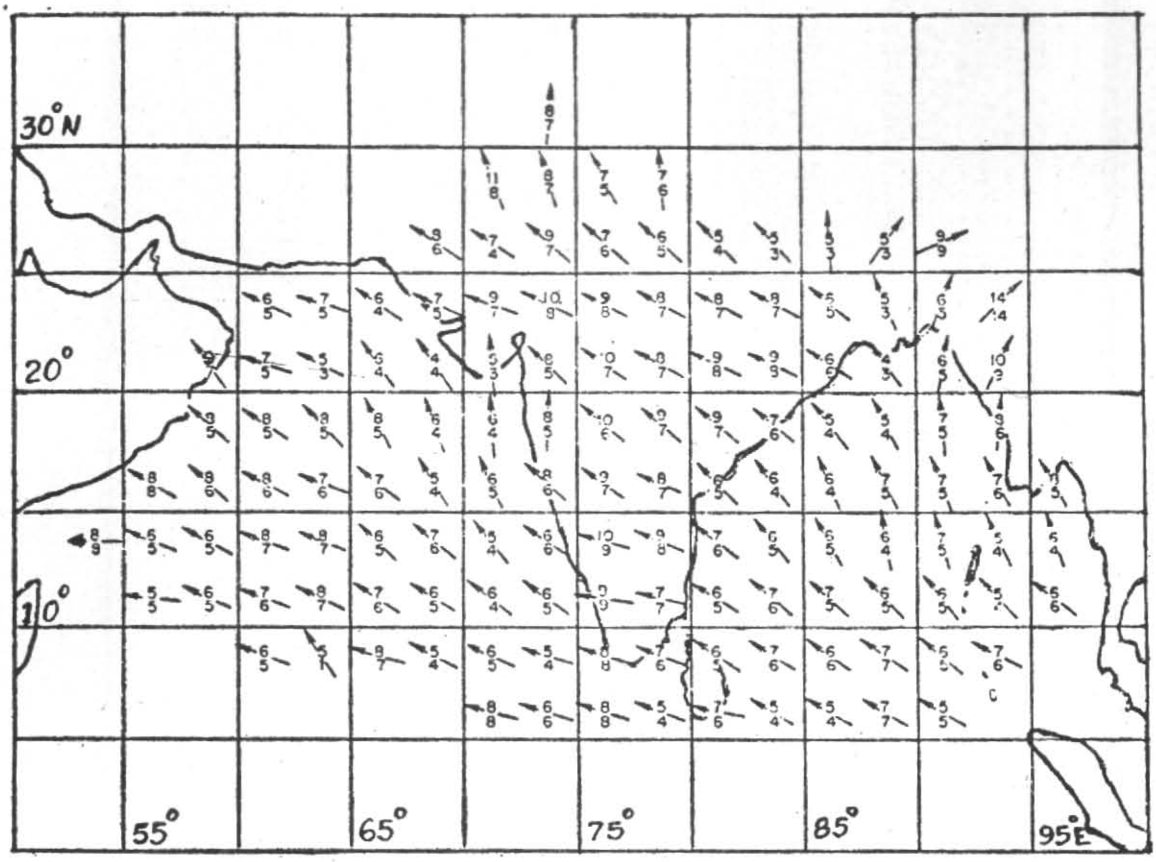

Fig. 1. Motion climatology of all recorded north Indian Ocean storms 1877-1974 by $2\frac{1}{2}°$ Lat./Long. squares
Arrows—Resultant direction, Upper Nos.—Scaler speeds (kt), Lower Nos.—Vector speeds (kt).
No entry when storms < 5

seas for 24 hr period. Later work includes a brief review of earlier works done in India on forecasting storm track. In addition, the Malagasy Republic, the People's Republic of China and possibly others have developed localised analog programmes for their areas of interest. The popularity of such method rests in the fact that they do not require current synoptic data but nevertheless, senses a synoptic pattern in the analog selection process.

One difficulty with analog models is their inability to arrive at a forecast under temporal or spatial anomalies. To offset this difficulty, Neumann (1972) developed a simulated analog model. Neumann and Randrianarison (1976) extended the concept to the southwest Indian Ocean. To complete the analog simulation process, the authors describe a method of constructing probability ellipses from residual (error) data.

The analog model and its simulated counterpart have been in operational use over the Atlantic for a number of years. Differences in performance based on average displacement error statistics are insignificant. However, the fact that the simulated model always produces a forecast is a definite operational advantage. Another advantage is its computational simplicity. The entire forecast takes but a few seconds of computer clock time. Analog models, on the other hand, can require as much as 45 minutes of clock time for older generation computers to scan and process the historical storm track file.

3. The data set

A magnetic tape containing the tracks of tropical cylones over the north Indian Ocean basin was obtained from the U. S. Department of Commerce, NOAA, National Climatic Center, Asheville, N. C. These data consisted of twice daily positions of 1282 tropical storms for the 98 year period 1877 through 1974.

As dictated by the requirements of the regression analysis, a minimum of four consecutive 12-hourly storm positions (from -24 to $+12$ hours) were required to develop prediction equations for the 12-hour forecast period. This necessitated deleting shorter duration storms. The few storms occurring during the January through April (*see* Fig. 3) off-season were also deleted. The final restructured data set consisted of 5,526 cases on 1,076 storms.

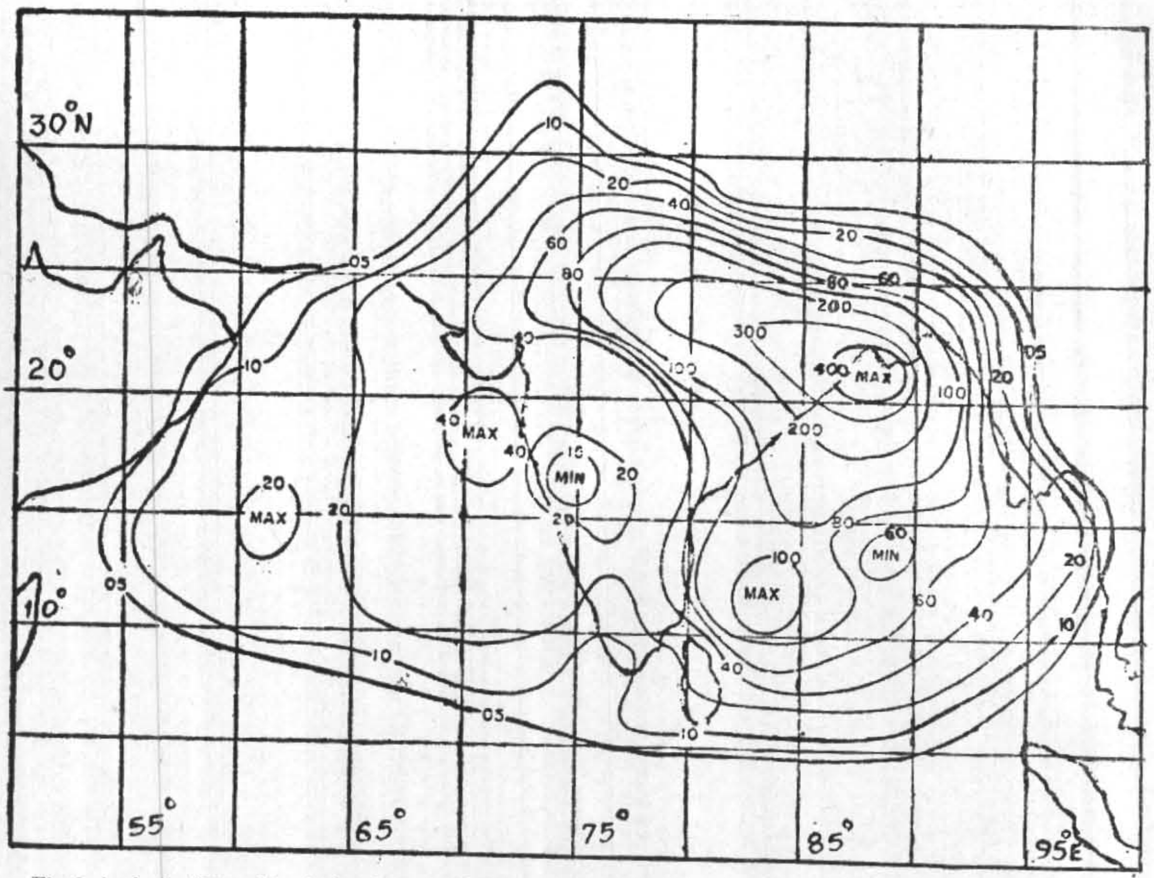

Fig. 2. Analysis of No. of tropical cyclones (including depressions) passing through $2\frac{1}{2}$° Lat./Long. squares, 1877-1974

Additional deletions were required when dealing with the longer-range forecasts. Development of prediction equations for the 72 hr forecast period, for example, require nine consecutive 12-hourly (from -24 to $+72$ hours) storm positions. The average durations of storms over the area is but slightly over 4 days. Consequently, the requirement for 9 consecutive positions reduced the number of cases to 1,778 on 553 storms.

This method of processing the data renders the number of cases a function of the forecast projection, the 12-hour forecast having over three times the number of developmental cases than the 72-hour forecast. While this requires the formulation of separate covariance matrices for each forecast projection rather than a single matrix had a uniform number of cases been used, it avoids a forecast bias toward the longer duration storms.

The decision to include all classes of tropical storms in the data set was made after careful analyses of the motion characteristics of these storms. Computer plots similar to those shown in Figs. 1 and 2 were prepared for the storms including and excluding depressions and also for different seasons. These figures (to be the subject of a separate publication) disclose only minor differences in motion climatology depending on the storm stage. Initial latitude and longitude being the predictors some weightage have been given to the seasonal changes. Further, more rigid, Mote Carlo tests as described in section 8 showed that the discriminatory information offered by a storm satge stratification was only slightly better than that offered by a random stratification.

On the other hand deletion of depressions from the data set would decrease the number of cases to 50 per cent which might lead to proportionate deterioration of the performance of the technique. Tracking of depressions and deep depressions in the Bay of Bengal and in Arabian Sea was also another important consideration to include all classes of tropical storm in the data set. However, this is to be considered as one of the limitations of the technique.

4. Climatological considerations

The data set just described include all classes of tropical cyclones. The annual occurrence rate averages about 13 storms, two to three being classified as severe cyclonic storms ($\geqslant$48 kt), three

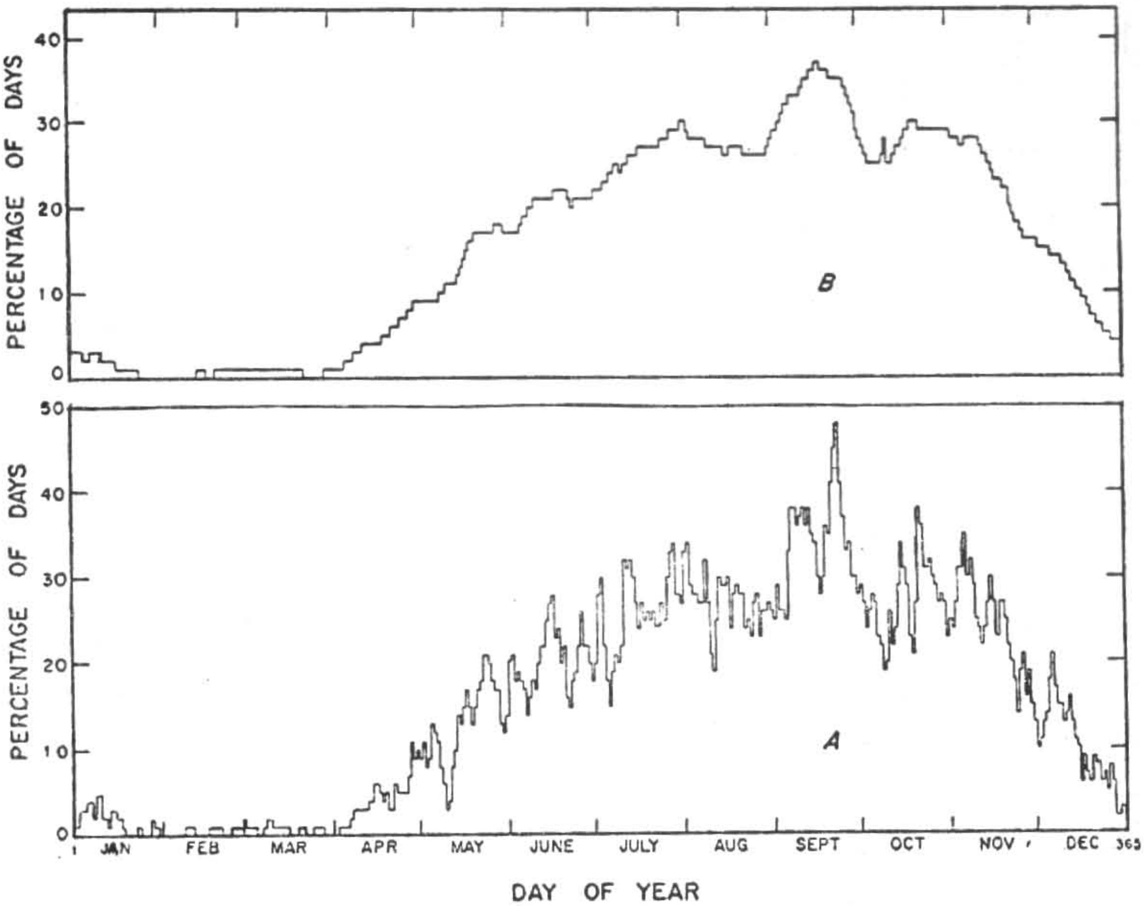

Fig. 3. Percentage of days, 1877-1974 with tropical cyclones (including depressions) over Bay of Bengal or Arabian Sea. Upper panel shows smoothing the 15-day moving average. Lower panel gives unsmoothed data.

to four as cyclonic storms ($\geqslant$34 to 47 kt) and the remainder as depressions.

Initial processing of the data led to an overview of north Indian Ocean tropical cyclone climatology as described in Figs. 1, 2 and 3. The computer routine which produced Figs. 1 and 2 required hourly storm positions. Hourly positions were estimated from the 12 hourly positions using curve fitting method described by Akima (1970) to avoid computational error by using less precise linear method. Interpolation of the short period positions of the storms were essential for finding accurate distribution of the storms in $2\frac{1}{2}$-degree latitude longitude box. If at least one of these hourly positions fell within the areal bound of the zone, the storm was considered to have passed through the zone.

The fifteen day moving average smoothing function used in Fig. 3 was determined from trial and error and represents a trade-off between the desire to smooth out obvious 'noise' and preserve recognized cyclical variations in the annual frequency of tropical cyclones. Although further discussions of these data could be made, this must be considered as beyond the scope of the present paper. The main purpose of including Figs. 1, 2 and 3 as discussed in a later section, relates to an optimization of the prediction algorithm.

There are many references dealing with various aspects of the tropical cyclone climatology. In this connection earlier climatological studies of the storms and depressions in Bay of Bengal by Rai Sircar (1956) and Rao and Jayaraman (1958) may be mentioned. A comprehensive work by Crutcher and Quayle (1974) is significant in that it gives homogeneous statistics for all tropical cyclone basins. Other works include Crutcher and Nicodemus (1973) and Crutcher and Hoxit (1973). These latter publications use the bivariate normal distribution to describe the tropical cyclone motion climatology and the strike probabilities at given sites.

5. Derivation of prediction equations

Multivariate regression analysis has been widely used in the meteorological profession for a numbe of years, Gradually, the older classical concepts of such analysis as described by Mills (1955) have been replaced by the "perfect-prog" and Model Output Statistics (MOS) approach as described by Klein and Glahn (1974). These modern concepts, however, require developmental data fields not generally available over data-sparse areas of the oceanic tropical cyclone belts. Accordingly, the older concepts still have applicability in tropical regions.

More advanced methods of the classical concept involve the introduction of non-linear effects and the stepwise screening regression process. These are described by Efroymson (1964). Regardless of the method employed, the computational procedures of multivariate analysis are similar. All require a considerable amount of matrix manipulation involving products (variance), cross-products (co-variance) and sums of variables. An important but often overlooked aspect of multivariate analysis relates to the question of statistical significance. This subject is discussed in section 8.

Predictands — Tropical cyclone forecasts are normally made for periods out through 72 hours. Prediction algorithms require displacement forecasts in both the meridional ($\triangle Y$) and the zonal ($\triangle X$) directions. Through vectorial addition, the $\triangle Y$ and $\triangle X$ components are used to position the latitude and longitude of the storm for any of the forecast projections. Accordingly, 12 predictands are defined. These, listed along with their means and standard deviations are given in Table 1. As is typical of other tropical cyclone regions, the standard deviations of zonal motion are seen to be considerably greater than meridional motion. As will be shown later, this has significance in the variance analysis.

The seven primary predictors — Analog models make use of storm selection criteria based on such factors as initial and past storm motion, initial storm position and time of year. To simulate these criteria by functional methods, seven primary predictors are introduced. These, together with their means and standard deviations are identified in Table 2. The 12 predictands given in Table 1 are assumed to be functions of these seven primary or basic predictors,

$$\triangle Y_j = f_j\,(P_1, P_2, P_3, P_4, P_5, P_6, P_7) \quad j = 1, 6 \quad (1)$$

$$\triangle X_j = g_j\,(P_1, P_2, P_3, P_4, P_5, P_6, P_7) \quad j = 1, 6 \quad (2)$$

Polynomial functions — The functions defined in (1) and (2) are usually taken as simple first-order polynomials. These have the advantage of simplicity and minimal degree of freedom loss. Non-linear effects, if persent, can be handled by stratifying the data into strategic periods or geographical zones. An example of this is shown by the data in Table 3 as partially illustrated in Fig. 4. These non linear data could be fitted by two linear equations, one for the period prior to 15 July and the other thereafter. With the availability of computers, it is often more convenient to use higher order polynomials to handle non-linearities. Such functions have the advantage of being everywhere continuous, whereas, the pair of linear equations would most likely give dual solutions near mid-July.

TABLE 1

Means and standard deviations of the 12 predictands (nautical miles). Southward and westward motions are negative

	Symbol	Mean	Standard deviation
	Meridional diplacement		
12 hr	$\triangle Y_1$	43·3	37·3
24 hr	$\triangle Y_2$	84·7	67·7
36 hr	$\triangle Y_3$	123·4	91·3
48 hr	$\triangle Y_4$	160·3	111·6
60 hr	$\triangle Y_5$	190·0	126·2
72 hr	$\triangle Y_6$	225·6	141·4
	Zonal displacement		
12 hr	$\triangle X_1$	—40·9	67·7
24 hr	$\triangle X_2$	—83·1	126·3
36 hr	$\triangle X_3$	—125·8	174·4
48 hr	$\triangle X_4$	—167·4	218·3
60 hr	$\triangle X_5$	—204·3	252·4
72 hr	$\triangle X_6$	—239·7	285·7

One restriction in using higher order polynomials relates to the large number of terms generated by these functions. As given in Neumann and Randrianarison (1976), the number of terms (T) in an nth order polynomial having m primary (basic) predictors is given by,

$$T = (m+n)!/(m!n!) \quad (3)$$

Thus, a polynomial containing 7 primary predictors will expand to 8, 36 and 120 terms, respectively for order one, two and three. The large number of terms required by the latter seldom justify the added complexities and degree of freedom loss. Accordingly, the second-order polynomial was selected as not having an excessive number of terms but still being able to approximate non-linear trends. The additional predictors P_8 through P_{35} generated by this expansion are defined in Table 4. They can be identified by

TABLE 2

Means and standard deviations of the seven primary predictors. Westward and southward motions are negative

Predictor	Symbol	Means (Forecast period (hr)) 12	24	36	48	60	72	Standard deviations (Forecast period (hr)) 12	24	36	48	60	72
Day number	P_1	242	242	243	243	244	243	58	57	57	56	56	55
Initial latitude (°N)	P_2	19·8	19·5	19·2	18·9	18·7	18·6	4·6	4·6	4·5	4·6	4·6	4·6
Initial longitude (°E)	P_3	83·3	83·5	83·7	84·0	84·1	84·3	5·7	5·5	5·4	5·2	5·2	5·2
Average meridional speed, past 12 hr (kt)	P_4	3·4	3·2	3·0	2·9	2·7	2·6	3·0	2·8	2·7	2·6	2·6	2·5
Average meridional speed, past 24 hr (kt)	P_5	−3·7	−3·8	−4·0	−3·9	−3·9	−3·8	5·2	4·9	4·7	4·5	4·3	4·2
Average zonal speed, past 12 hr (kt)	P_6	3·3	3·1	2·9	2·8	2·7	2·5	2·6	2·6	2·4	2·3	2·3	2·3
Average zonal speed, past 24 hr (kt)	P_7	−3·7	−3·9	−3·9	−3·9	−3·9	−3·8	4·7	4·5	4·2	4·0	3·9	3·8
Number of cases		5526	4572	3656	2944	2270	1778	5526	4572	3656	2944	2270	1778

Note : Day number 243 is 30 Aug.

TABLE 3

Average zonal, meridional and vector tropical cyclone motion by months, April through December. North and east are positive

	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Zonal	−0·7	−0·1	−3·0	−6·0	−5·3	−4·5	−2·0	−2·0	−2·4
Meridional	4·8	4·9	3·1	2·3	2·5	3·3	4·5	4·0	3·7
Vector	352/5	358/5	315/4	291/6	296/6	306/6	336/5	333/4	327/4
Number of cases	99	309	523	867	922	1194	782	629	261

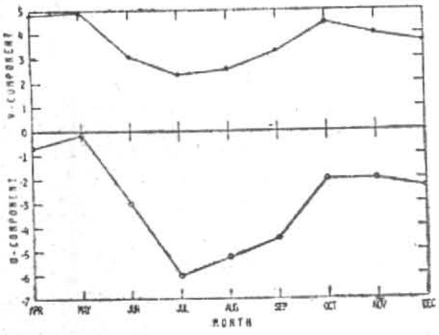

Fig. 4. Average V and U components (kt) of north Indian Ocean tropical cyclone motion April through December. V components are towards north. U components are toward west

TABLE 4

Additional predictors P_8 through P_{35} generated by a second-order polynomial with seven predictors. The meaning of P_1 through P_7 is given in Table 2

$P_8 = P_1^2$	$P_{15} = P_4P_2$	$P_{22} = P_5^2$	$P_{29} = P_7P_1$
$P_9 = P_2P_1$	$P_{16} = P_4P_3$	$P_{23} = P_6P_1$	$P_{30} = P_7P_2$
$P_{10} = P_2^2$	$P_{17} = P_4^2$	$P_{24} = P_6P_2$	$P_{31} = P_7P_3$
$P_{11} = P_3P_1$	$P_{18} = P_5P_1$	$P_{25} = P_6P_2$	$P_{32} = P_7P_4$
$P_{12} = P_3P_2$	$P_{19} = P_5P_2$	$P_{26} = P_6P_4$	$P_{33} = P_7P_5$
$P_{13} = P_3^2$	$P_{20} = P_5P_3$	$P_{27} = P_6P_5$	$P_{34} = P_7P_6$
$P_{14} = P_4P_1$	$P_{21} = P_5P_4$	$P_{28} = P_6^2$	$P_{35} = P_7^2$

TABLE 5

Regression coefficients $C(I, J)$ for use with meridional motion prediction equation (4). Index I refers to predictor number. Intercept is given by element $C(36, J)$. Index J refers to forecast projection at 12 hourly intervals, 12 through 72 hr

	$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$
$I=1$	0·26343	0·86020	1·76512	3·13866	4·46446	5·88013
$I=2$	9·84851	30·19603	47·76279	57·15314	59·90614	58·88515
$I=3$	3·75820	9·65990	15·96362	22·76817	30·56889	35·37262
$I=4$	1·12522	24·51108	43·53468	11·20905	4·09838	−17·20822
$I=5$	−3·61793	−18·65553	−24·40282	−99·59052	−111·99460	−117·87840
$I=6$	7·73040	−7·03645	−19·27299	38·73376	85·76770	157·23630
$I=7$	−0·82257	9·24858	9·60812	86·79787	106·77991	120·53040
$I=8$	−0·00042	−0·00098	−0·00240	−0·00420	−0·00646	−0·00917
$I=9$	0·00610	0·01145	0·01358	0·01626	0·01572	0·01262
$I=10$	−0·09127	−0·32478	−0·54548	−0·75188	−0·85670	−0·96841
$I=11$	−0.00260	−0.00793	−0.01227	−0.01901	−0.02146	−0.02254
$I=12$	−0·11212	−0·29239	−0·44988	−0·51582	−0·52865	−0·49694
$I=13$	−0·00559	−0·01504	−0·03231	−0·05930	−0·10266	−0·13467
$I=14$	0·01820	0·02266	0·02466	0·05921	0·06028	0·08416
$I=15$	0·07368	−0·25259	−0·69545	−0·16633	0·08658	−1·16188
$I=16$	0·03167	−0·14435	−0·24725	−0·11992	−0·10848	0·44803
$I=17$	−0·03414	−0·33868	−0·62900	−1·23145	−0·85828	−0·70693
$I=18$	−0·00881	−0·01404	−0·01641	0·03348	0·06663	0·07851
$I=19$	−0·01286	−0·02265	−0·08121	0·27562	0·28160	0·02217
$I=20$	0·05627	0·24933	0·33646	0·96378	1·04387	1·13563
$I=21$	−0·19797	−0·33273	−0·41875	−0·22339	0·23348	0·74009
$I=22$	−0·13362	−0·20868	−0·17159	−0·32005	−0·97610	0·00936
$I=23$	−0·01313	−0·01752	−0·01112	−0·06157	−0·12289	−0·16544
$I=24$	−0·05192	0·11509	0·31428	−0·68794	−1·48933	−0·73082
$I=25$	−0·04674	0·13802	0·26803	0·05138	−0·26374	−1·09210
$I=26$	−0·14163	0·42044	0·73193	1·93822	−0·09352	−2·09380
$I=27$	0·61084	1·35296	1·82994	2·01700	1·20465	0·25589
$I=28$	0·02417	−0·43208	−0·87565	−2·19423	−1·07125	0·94595
$I=29$	0·00856	0·01436	0·01424	−0·03300	−0·07075	−0·08751
$I=30$	0·00152	−0·03858	−0·07703	−0·48631	−0·45572	−0·17719
$I=31$	−0·00027	−0·12134	−0·11430	−0·75438	−0·94554	−1·14832
$I=32$	0·03470	−0·12756	−0·26753	−1·17138	−1·82244	−2·74871
$I=33$	0·18409	0·32670	0·30504	0·34155	0·11906	−0·22017
$I=34$	−0·33537	−0·89613	−0·94523	−0·58281	0·70283	2·29618
$I=35$	−0·08368	−0·12993	−0·10074	0·06027	0·02197	0·17952
$I=36$	−245·04770	−680·72949	−1113·79810	−1553·41895	−2000·48999	−2280·93311

considering all the combinatorial products and cross-products of P_1 through P_7. Eqns. (1) and (2) can thus be defined,

$$\triangle Y_j = C_{36,j} + \sum_{\substack{i=1,35 \\ j=1,7}} C_{i,j}\, P_i \qquad (4)$$

$$\triangle X_j = Q_{36,j} + \sum_{\substack{i=1,35 \\ j=1,7}} Q_{i,j}\, P_i \qquad (5)$$

where the arrays C and Q are constants. The array elements $C_{36,j}$ and $Q_{36,j}$ are defined as the intercept values. Determination of these constants require the formulation of 36 "normal" equations using methods described in Neumann and Hope (1972). However, the scientific subroutine packages available through most large computer facilities normally contain statistical programs for this purpose. In the present application, the IBM multivariate analysis program was used. The list of constants so determined are given in Tables 5 and 6, the former being for meridional and the latter for zonal motion in n. miles. For computer application, these constants can conveniently reside in a disk or tape file, permitting the solution of Eqns. (4) and (5) in generally less than 1 second of computer time. This offers a distinct

TABLE 6

Regression coefficients $Q(I, J)$ for use with zonal motion prediction equation (5). Index I refers to predictor number. Intercept is given by element Q (36, J). Index J refers to forecast projection at 12 hourly intervals, 12 through 72 hr

	$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$
$I=1$	−1·10296	−3·09466	−5·24750	−6·92603	−8·49573	−10·22943
$I=2$	5·43039	25·14336	52·52481	94·45428	136·02850	167·55330
$I=3$	1·53067	6·04916	27·08588	56·66257	111·69189	142·78349
$I=4$	−4·52247	−28·83974	−39·62119	−73·76862	23·63422	173·47290
$I=5$	14·23036	38·33446	54·77888	131·37550	192·35159	242·80920
$I=6$	-4·27768	4·70108	8·84996	55·37897	8·54507	−92·01617
$I=7$	−4·44658	−27·63277	−39·65843	−104·97130	−159·30000	−212·17780
$I=8$	0·00257	0·00744	0·01275	0·01857	0·02475	0·03054
$I=9$	0·01013	0·03439	0·05705	0·08295	0·09935	0·12169
$I=10$	−0·05199	−0·23297	−0·50918	−0·91327	−1·31348	−1·64977
$I=11$	−0·00254	−0·00943	−0·01552	−02966	−0·04666	−0·08020
$I=12$	−0·05409	−0·26156	−0·51603	−0·90094	−1·24737	−1·51671
$I=13$	−0·00821	−0·01616	−0·12723	−0·26697	−0·55568	−0·71614
$I=14$	0·01249	0·04970	0·08779	0·06202	0·00647	−0·04444
$I=15$	−0·08880	−0·05907	−0·33889	−0·61917	−1·94409	-3·39371
$I=16$	0·03594	0·19867	0·31511	0·79579	0·14692	−1·02959
$I=17$	0·58372	1·14957	1·15357	0·95541	−0·25967	0·64892
$I=18$	0·00694	0·00064	0·01290	−0·00215	−0·05066	−0·04777
$I=19$	0·18284	0·00756	0·10647	−1·12353	−1·69527	−2·41801
$I=20$	−0·11998	−0·26603	−0·48601	−1·00544	−1·44736	−1·81309
$I=21$	0·63447	1·43872	1·79772	3·22583	2·34810	4·49679
$I=22$	−0·11857	−0·06451	−0·07240	−0·11740	−0·44239	−0·26581
$I=23$	−0·01307	−0·04482	−0·08696	−0·08348	−0·06393	−0·08429
$I=24$	0·02298	0·08225	0·23463	0·26485	1·23972	2·15842
$I=25$	0·11578	0·15161	0·19891	−0·25313	−0·01799	0·84943
$I=26$	−1·12787	−1·67738	−2·85826	−1·20330	−0·32176	−3·76540
$I=27$	0·08467	0·34092	0·75917	0·41948	2·05382	−0·09522
$I=28$	0·41543	0·31625	1·56578	0·16003	0·02787	2·72731
$I=29$	−0·01178	−0·01384	−0·03403	−0·03641	−0·02490	−0·05247
$I=30$	−0·21978	−0·15730	−0·42805	0·58023	1·02093	1·85390
$I=31$	0·15627	6·44282	0·75963	1·32899	1·87509	2·35547
$I=32$	−1·15828	−2·64025	−4·14273	−6·15051	−6·98076	−9·13587
$I=33$	0·93456	1·21241	1·49898	2·66552	3·39079	3·62971
$I=34$	0·25920	0·54853	1·10935	1·58943	1·22430	3·92556
$I=35$	−0·29870	−0·00088	−1·03990	−1·99002	−2·38217	−2·39664
$I=36$	27·58488	−86·58549	−890·91772	−2225·71899	−4638·21484	−5947·91016

advantage over the analog approach which requires scanning and processing the entire storm history file each time the program is activated.

6. Stepwise screening regression methods

An alternate approach in developing prediction equations could have been stepwise screening regression. This method eliminates those predictors which, for one reason or another, fail to explain incremental reductions of variance beyond some set minimum value. To see if this offered any advantages, a separate set of prediction equations was developed in this way. The resultant dependent data variance analysis is given in Table 7.

The data given in this table are quite thought provoking. As would be expected, the principal reductions of variance for the short period forecasts of meridional motion are given by P_4 (average meridional speed over past 12 hr). However the relationships become more complex with increasing forecast interval. At 72 hours, P_2 (initial latitude) works in combination with P_{10} (square of initial latitude) to provide the main variance reduction. It is interesting to note that in the case of zonal motion, P_{20} (the product of average zonal speed over the past 12 hr and initial longitude) provides, by far, the maximum reduction of variance. Further, in depth discussions of these results could be made but will have to be considered beyond the scope of the present paper.

TABLE 7

Predictor selection and reduction of variance (per cent) using a stepwise screening procedure. Missing entries and missing predictors contributed less than ½% incremental variance reduction

Predictors	Predictands											
	$\triangle Y_1$	$\triangle Y_2$	$\triangle Y_3$	$\triangle Y_4$	$\triangle Y_5$	$\triangle Y_6$	$\triangle X_1$	$\triangle X_2$	$\triangle X_3$	$\triangle X_4$	$\triangle X_5$	$\triangle X_6$
P_1	—	—	—	0·6	1·0	1·3	—	0·8	1·3	0·8	3·2	3·6
P_2	—	1·0	1·4	2·1	2·1	2·4	—	—	—	0·8	1·0	1·1
P_3	0·5	0·8	0·9	0·6	—	—	—	1·0	3·0	4·8	—	0·6
P_4	42·7	30·0	25·3	18·4	15·6	9·2	2·9	—	—	—	—	—
P_7	—	—	—	—	—	0·6	—	—	—	—	—	—
P_8	—	—	—	—	—	—	—	—	—	2·6	—	—
P_9	—	—	—	—	—	—	0·5	1·1	1·4	—	0·7	0·6
P_{10}	0·7	2·1	3·8	6·2	7·7	12·2	—	—	—	—	—	0·5
P_{11}	—	—	—	—	—	—	—	—	—	0·5	—	—
P_{12}	—	—	0·6	0·7	0·8	0·9	—	—	0·6	—	0·7	1·0
P_{13}	—	—	—	—	—	—	—	—	—	—	6·9	12·0
P_{14}	—	—	—	—	—	—	—	5·6	7·4	—	8·4	7·0
P_{15}	—	—	0·5	0·5	0·7	1·1	—	—	—	—	—	—
P_{16}	—	—	—	—	—	—	—	—	—	0·6	—	—
P_{17}	—	—	0·6	1·1	1·4	1·4	—	—	—	—	—	—
P_{20}	—	—	—	—	—	—	60·8	45·2	37·3	26·5	22·1	14·2
P_{21}	—	—	—	—	—	—	—	—	—	0·6	—	—
P_{23}	—	—	—	—	—	—	—	—	—	8·7	—	—
P_{27}	—	0·6	—	—	—	—	—	—	0·5	—	0·9	—
P_{33}	—	—	—	—	—	—	—	1·5	1·3	1·3	1·0	1·2
P_{35}	—	—	—	—	—	—	0·6	—	—	—	—	—
Total reduction	43·9	34·5	33·1	30·2	29·3	29·1	64·8	55.2	52·8	47·7	44·9	41·8
No. of cases	5526	4572	3656	2944	2270	1778	5526	4572	3656	2944	2270	1778

It is customary and often desirable to exclude those predictors which fail to explain minimum incremental reduction of variance. Thus, the prediction equation for 12 hr meridional motion would contain but three predictors, P_3, P_4 and P_{10}, while the prediction equation for 72 hr meridional motion would contain 8 predictors. However, in the case of tropical cyclone forecasts, the use of equations with differing numbers of predictors has the very undesirable side-effect of producing discontinuous forecast tracks. Such a meandering, unrealistic track is immediately dismissed by the operational forecaster as being totally unrealistic.

There are other reasons that argue against using the stepwise screening approach. One such reason relates to convenience. It is much easier to program (4) nad (5) for a digital computer if the arrays C and Q are of fixed dimension.

A third reason, discussed at some length in Neumann and Randrianarison (1976), relates to uncertainties in the operational specification of P_4 and P_5. A final and perhaps most important reason is simply that Eqns. (4) and (5) were tested alongside the stepwise regression equations on a homogeneous independent data set and found to be superior. It was concluded, therefore, that prediction Eqns. (4) and (5) were preferred over similar equations developed from the stepwise screening process.

7. System performance

Dependent data — Data relating to the performance of (4) and (5) on dependent (development) data are given in Table 8. Since system performance on either dependent or independent data is generally based on displacement error (item 13), subsequent comparisons will principally concern this item. The displacement error is defined as the great circle distance between a forecast (X_f, Y_f) and an observed (X_0, Y_0) storm position. The expression for the great circle distance error (E) is derivable from the law of cosines for oblique spherical triangles and is given by,

$$E = \cos^{-1} [\sin Y_0 \sin Y_f + \cos Y_0 \cos Y_f \cos(X_0 - X_f)] \quad (6)$$

The variance analysis given in Table 8 reveals what appears to be (compare, for example, item 3 and item 8) better performance of the model for

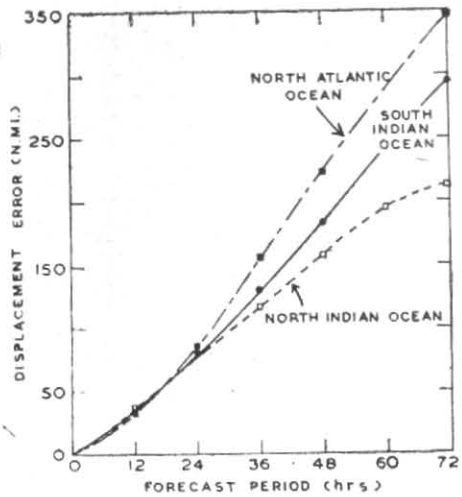

Fig. 5. Independent data performance of the model on 3 different tropical cyclone basins

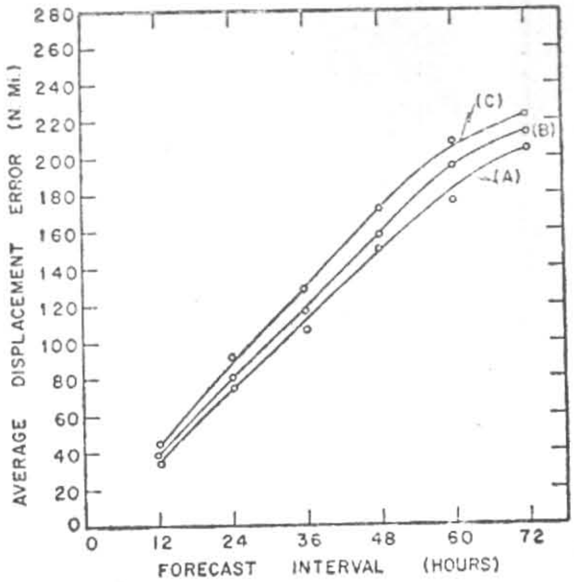

Fig. 6. Performance of the model on dependent data (A) and independent data (B). Curve (C) gives estimated performance on operational data

TABLE 8

System performance on development data. Errors are given in nautical miles

	Forecast periods (hr) 12	24	36	48	60	72
1. Number of cases	5526	4572	3656	2944	2270	1778
2. Number of storms	1071	1053	815	798	576	553
3. Meridional motion, reduction of variance (%)	46	37	35	33	32	32
4. Meridional motion, multiple correlation coefficient	0·68	0·61	0·60	0·57	0·57	0·56
5. Meridional motion, standard error	27	54	73	92	104	117
6. Mean absolute meridional motion error	19	40	56	70	80	90
7. Mean meridional motion error (*y*-bias)	0	0	0	0	0	0
8. Zonal motion, reduction of variance (%)	67	57	55	50	49	45
9. Zonal motion, multiple correlation coefficient	0·82	0·76	0·74	0·71	0·70	0·67
10. Zonal motion, standard error	39	83	118	155	181	211
11. Mean absolute zonal motion error	27	61	90	120	143	166
12. Mean zonal motion error (*x*-bias)	0	0	0	0	0	0
13. Mean displacement error	37	80	116	151	177	205

TABLE 9

System performance on independent data. Errors are given in nautical miles. Numbered items correspond to similarly numbered items given in Table 8

	Forecast period (hr) 12	24	36	48	60	72
1. Number of cases	141	130	98	87	52	46
2. Number of storms	141	130	98	87	52	43
5. Meridional motion, standard error	30	57	73	96	97	109
6. Mean absolute meridional motion error	22	44	55	73	73	86
7. Mean meridional motion error (*y*-bias)	3	4	5	10	23	23
10. Zonal motion, standard error	39	84	123	162	212	227
11. Mean absolute zonal motion error	27	58	91	125	164	174
12. Mean zonal motion error (*x*-bias)	0	0	–8	1	–14	–18
13. Mean displacement error	37	81	117	158	195	212

zonal motion than for meridional motion. Before interpreting these results, it is well to consider the relationship between the reduction of variance (R_V), the multiple correlation coefficient (R_m) the standard error (S_E) and the standard deviation of the motion (S_D). These relationships are

$$S_E = S_D\,(1-R_m^2)^{\frac{1}{2}} \qquad (7)$$

$$R_V = R_m^2. \qquad (8)$$

Thus, even though Table 8 shows substantially greater variance reduction for zonal motion the greater standard deviations of zonal motion (*see* Table 1) provide, according to (7), greater zonal standard errors. Similar results are noted in other tropical cyclone basins.

Independent data — Consider now the performance of the model on independent data. For this purpose, every 40th case, or the equivalent of approximately 2½ years of data had been withheld from the master data set. This sample provide an independent test of Eqns. (4) and(5), the results of which are given in Table 9. The differences in performance between dependent and independent data can be obtained by comparison with Table 8. For convenience, a common numbering system has been used in both tables. Items 4, 5, 8 and 9 were intentionally omitted from Table 9 since these are generally associated only with dependent (development) data.

Comparison between the two tables reveals the usual deterioration in performance between dependent and independent data. These differences are quite small, however. The x- and y-biases always zero in the dependent data, are also seen to be quite small. These biases are typical and result from dependent and independent data having somewhat different statistical properties.

In Fig. 5, a comparison is made with the displacement errors obtained from the similar set of prediction equations developed by Neumann and Randrianarison (1976) for the south Indian Ocean and by Neumann (1972) for the North Atlantic. The figure implies what appears to be better performance of the model beyond 36 hr over the north Indian basin. However, this is merely a reflection of different standard deviations of tropical cyclone motion in these three basins. The 72 hr meridional and zonal standard deviations, for example, are 372 and 669 n. miles respectively for the north Atlantic and 141 and 286 n. miles for the north Indian area. If one accounts for these differences; the performance characteristics of the model is similar for all three basins. Fig. 5 well illustrates the pitfalls of unqualified comparison between different tropical cyclone basins.

Operational data — Some further deterioration in performance is expected when running the program in an operational mode. The reason relates to uncertainties in specifying the current position of a storm. Later information or a post-analysis sometimes indicates that the original estimate was somewhat in error. The problem is discussed in detail by Neumann (1975 a and 1975 b). Some estimate of the additional deterioration can be obtained by considering the operational degradation of a similar set of prediction equations for the Atlantic. Here, the degradation over independent data was found to be 15, 13, 10, 9 and 5 per cent respectively, for the 12, 24, 36, 48 and 72 hr forecast periods. Applying hese same percentages to the north Indian independent data gives the results shown in Fig. 6. However, the limited aircraft reconnaissance and satellite coverage over the north Indian area may cause additional deterioration.

8. The question of statistical significance

Establishing the statistical significance of a regression equation is tantamount to stating a hypothesis that the difference in the two variances, one about the mean of the development data and the other about the fitted regression hyperplane, are greater than one would reasonably expect from pure chance. The independent data test, described in the preceding section, establishes, at least qualitatively, in that this hypothesis is true.

More rigid tests can be used. One such test described in Neumann *et al.* (1976) uses a Monte Carlo approach. The method concerns itself with replacing the ordered tropical cyclone displacements (predictands) with another randomly ordered set. Prediction equations are then developed from this latter data set. Repeated application of this test gave reductions of variance of only about 1 per cent. Since the actual variance reductions were much greater, significance at well above the 99 per cent level is indicated.

Serial correlations between individual predictor/predictand sets preclude direct use of the classical F-test named after R. A. Fisher by Snedecor (1946). The presence of these correlations effectively reduces the degrees of freedom in the denominator of the F-variance ratio. Thus, the value of F one obtains by not allowing for this reduction is seriously inflated. To counter this effect, a smaller "effective degrees of freedom" can be introduced. Crutcher (1976) suggests reducing the original degree of freedom by a factor of three. Applying the F-test in this manner still leads to acceptance of the regression equations (rejection of a null-hypothesis).

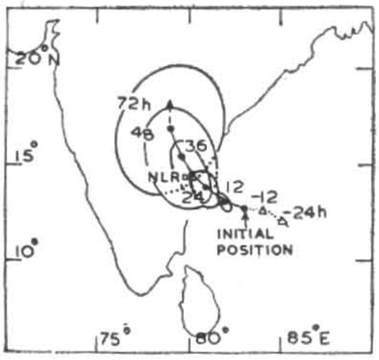

Fig. 7. Simulated example showing 12, 24, 36, 48 and 72 hr forecasts with associated 50 per cent probability ellipses. Storm initially located at 12·7°N, 87·3°E on 5 Nov. Positions at — 12 hr and — 24 hr are at 12·5°N, 84·0°E and 12·0°N, 85·0°E

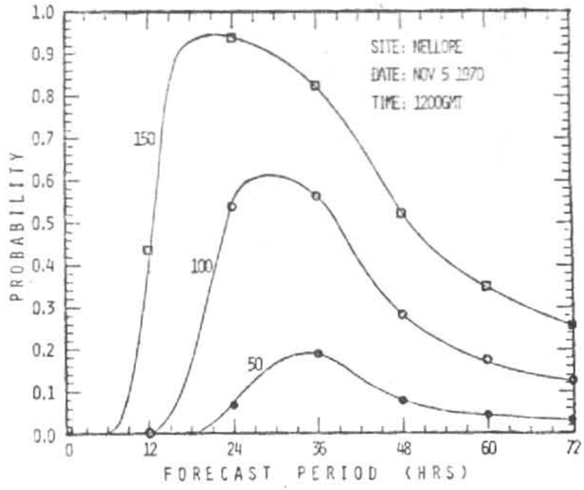

Fig. 8. Probability (0<P<1) of the storm shown in Fig. 7 being within 50, 100 and 150 n. miles of Nellore at any time during the 72 hr forecast period

9. Probability ellipses

The computer output packages produced by analog models typically contain supplemental data for the construction of probability ellipses. If an x, y plotter is available, the complete forecast package might appear as in Fig. 7. The forecast track shown in the figure was derived from Eqns. (4) and (5) according to the input data as specified in the legend. A fictitious example was intentionally chosen since the purpose of the figure is merely to illustrate the method and not to depict a single, perhaps not representative, verification. Ample documentation on this latter subject is given elsewhere in the study.

The complete forecast package, as given in Fig. 7, afford the forecaster effective decision making capabilities (Simpson 1971). He may, for example, use the intersection of say, the 48 hr 50 per cent probability ellipse with the coast-line as objective guidance in fixing the extent of coastal warnings. Or, he may want to confine his forecast track to fall within some elliptical bounds.

In addition to the Mercator map background, computer plotting of Fig. 7 requires, (*a*) the forecast positions of the storm, (*b*) a curve-fitting routine as, for example, given by Akima (1970) for plotting the track and (*c*) the rotation angle and lengths of the major and minor ellipse axes. These latter quantities are determined by fitting selected meridional and zonal dependent data residuals to a bivariate normal distribution. Crutcher (1971) provides statistical justification for the use of this particular density function.

The bivariate normal density function in terms of the five population parameters, x, y, S_x, S_y and r_{xy} is expressed as

$$f(x,y) = [2\pi S_x S_y \,(1-r_{xy}^2)\,]^{-1} \exp(-G/2) \qquad (9)$$

where,

$$G = (1-r_{xy}^2)^{-1} \; [(x-\bar{x})^2 S_x^{-2} - -2r_{xy}\,(x-\bar{x})\,(y-\bar{y})\,(S_x S_y)^{-1} + (y-\bar{y})^2 S_y^{-2}]. \qquad (10)$$

In the above expressions, x, y, $\bar{x}$, $\bar{y}$, S_x and S_y are respectively, the zonal and meridional forecast displacements, the mean of the zonal and meridional forecasts, the standard deviation of the zonal and meridional forecasts and the linear correlation coefficient between the individual components. For further mathematical treatment, the readar is referred to the above Crutcher reference or to Hope and Neumann (1970).

In analog models, the five population parameters are determined from the clusters of analog storm positions. In the present simulated case, the "clusters" are obtained from the residuals (errors) one obtains in applying Eqns. (4) and (5) to the dependent data set. The two methods are entirely analogous.

The algorithm for performing the necessary calculations is computer dependent and is thus beyond the scope of the paper. Briefly, however, a magnetic tape or disk file is structured to contain ordered sets of predictors P_1 through P_5 along with the associated meridional and zonal residuals. Cases are then systematically eliminated using selection criteria based on day number (P_1
)

latitude (P_2) and longitude (P_3). The remaining cases are ordered depending on the magnitude of the vector distance (Z) between current storm motion (V_c, U_c) and the "analog" storm motion (V_a, U_a), the latter being given by P_4 and P_5. This distance is given by

$$Z = [(U_c - U_a)^2 + (V_c - V_a)^2]^{\frac{1}{2}} \qquad (11)$$

The upper portion of the sorted Z-array (25 cases were found to be optimal in the Atlantic) are then paired with the associated residuals and it is from these data that the five population parameters needed to define the distribution are computed. The ellipse centroid (mean of the meridional and zonal residuals) may not necessarily equal zero. That is to say, it may not correspond exactly to the forecast storm position given by Eqns. (4) and (5). However, these differences are normally so small that they can be ignored.

In a manner similar to that employed by analog models, the analog selection criteria can be optimized depending on season and storm location. As shown by Hope and Neumann (1972), the magnitude of displacement errors is primarily a function of initial storm motion and secondarily of geographical location and time of year. Accordingly, the day number, latitude and longitude filtering described above should eliminate only about 25 per cent of the cases. The remaining 75 per cent are filtered by application of (11). Figs. 1, 2 and 3, as well as other pertinent climatology, to be published separately, are quite useful in this respect. With optimized programming techniques including the use of logical variables, the entire procedure outlined above can be accomplished very rapidly, even on older generation computer systems.

Further application — Further use can be made of the elliptical data. By integrating the density function (9) over a circular area (A),

$$\int_A \int f(x,y)\, dx\, dy \qquad (12)$$

one can obtain the probability of a storm being within the given area at a given time. The Fortran computer program to perform this numerical integration is given in the Appendix of Crutcher (1971).

As an example of the use of (12), consider the forecast track given in Fig. 7. It can be noted that the storm is forecast to pass about 20 n. miles ENE of the city of Nellore (NLR), located 14·3°N, 79·6°E. The probability of the city being within 50, 100 and 150 n. miles of the storm centre at any time throughout the 72 hr forecast period is shown in Fig. 8. Although not accomplished here, further time integration could be employed to determine the probability of the city being within a specified distance from the storm at any time during the 72-hr period. These cumulative probabilities are obviously much higher and in the case cited here, would approach 100 per cent.

The procedures, although still experimental, may well be applicable in determining the coastal extent and timing of a hurricane warning or watch zone. With suitable x, y plotting equipment, the graphical depiction, shown in Figs. 7 and 8 can be made available to the operational forecaster in but a few minutes time.

10. Discussion and Summary

In this paper, the derivation of a simulated analog model for the forecast displacement of tropical cyclones over the north Indian Ocean basin has been described. The technique is similar to that employed by Neumann and Randrianarison (1976) for the south Indian Ocean tropical cyclone basin, and by Neumann (1972) for the North Atlantic. One of the principal advantages of the simulated analog method is computational simplicity in both the developmental and operational modes.

Another advantage relates to anomalous weather situations. Analog models attempt to identify certain synoptic patterns. Occasionally, under anomalous situations, a pattern cannot be identified and a forecast is not produced. This leaves an operational void. The simulated model, on the other hand, effectively bridges these discontinuous synoptic subsets and always produces a forecast. Both the analog model and its simulated counterpart have been run simultaneously over the Atlantic for a number of years and have displayed similar error profiles.

The model makes optimum and explicit use of climatology and persistence. In this respect, it can be considered a benchmark from which to judge the performance of more sophisticated models, whether they be statistical or numerical. The output of the model may well be used as input into a model which employs synoptic data such as that described by Jagannathan and Crutcher (1968). Such an approach is used in the NHC72 (Neumann *et al.* 1972) model and the NHC73 (Neumann and Lawrence 1975) statistical-dynamical model, in use over the North Atlantic tropical cyclone basin.

Any statistical model, whether it be analog or otherwise is very sensitive to uncertainties in the operational specification of the initial motion vector. The model discussed herein is no exception. Accordingly, every effort should be made to initialize the model with present and past storm positions (from which the motion is desired which reflect the more conservative motion of the larger scale storm circulation rather than the short period, often transient, oscillations of the inner storm vortex. Only in this way can the full variance reducing potential of the system be realized.

Acknowledgements

The authors are indebted to the World Meteorological Organization for assuming most of the co-author's expenses while he was a WMO Fellow at the National Hurricane Center. The co-author wishes to record his gratefulness to the Government of India and to the India Meteorological Department for deputing him for training at the National Hurricane Center (NHC). Finally, the co-author wishes to thank officials at NHC for extending him the facilities at the Center.

REFERENCES

Akima, H.	1970	*J. Assoc. Comput. Mach.*, **17**, pp. 589-602.
Brand, S., Long, J. M. Blelloch, J. W. and Hamilton, G. D.	1974	ENVPREDRSCHFAC *Tech. Paper* No. 11-74, 20 pp.
Crutcher, H. L.	1971	Atlantic tropical cyclone statistics, NASA (Contract Rept. NASA CR-61355, NASA) George, C. Marshal Space Flight Center, Huntsville, AL, 16 pp.
Crutcher, H. L. and Hoxit, L. R.	1973	South Indian tropical cyclone strike probabilities. Commander, Naval Weather Service Command, Wash. DC, 91 pp.
Crutcher, H. L. and Nicodemus, M. L.	1973	South Indian tropical cyclone vector mean charts. Commander, Naval Weather Service Command, Wash. DC, 56 pp.
Crutcher, H. L. and Quayle, R. G.	1974	*Mariners Worldwide guide to tropical storms at sea*, NAVAIR 50-1C-61, 113 pp. plus 311 charts.
Crutcher, H. L.	1976	(Environmental Data Service, NASA, U.S. Department of Commerce, Asheville, NC), 1976 (personal communication).
Efroymson, M. A.	1964	Multiple regression analysis. *Mathematical Methods for Digital Computers* (A. Ralston and H.S. Wilf, Eds.), New York, John Wiley and Sons, Inc., pp. 191-203.
Gupta, R. N. and Datta, R. K.	1971	*Vayu Mandal*, **1**, pp. 124–127.
Hope, J. R. and Neumann, C. J.	1970	*Mon. Weath. Rev.*, **98**, pp. 925-933.
Jagannathan, P. and Crutcher, H. L.	1968	*ESSA Tech. Rep.*, EDS-4, 494 pp.
Klein, W. H. and Glahn, H. R.	1974	*Bull. Am. Met. Soc.*, **55**, pp. 1217-1227.
Mills, F. C.	1955	*Statistical Methods*, 3rd Ed., New York, Holt, Rinehart and Winston, pp. 147, 612-656, 601, 727-731.
Neumann, C. J.	1972	*NOAA Tech. Memo.*, NWS SR-62, 32 pp.
Neumann, C. J. and Hope, J. R.	1972	*Mon. Weath. Rev.*, **100**, pp. 245-255.
Neuman, C. J., Hope, J. R. and Miller, B. I.	1972	*NOAA Tech. Memo.*, NWS SR-63, 32 pp.
Neumann, C. J. and Lawrence, M. B.	1975	*Mon. Weath. Rev.*, **103**, 8, pp. 665-673.
Neumann, C. J.	1975 a	*NOAA Tech. Memo.*, NWS-SR-82, 21 pp.
	1975 b	*Ibid.*, NWS-SR-81, 16 pp.
Neumann, C. J. and Randrianarison, E. A.	1976	*Mon. Weath. Rev.*, **104**, 1, pp. 76-85.
Neumann, C. J., Lawrence M. B. and Caso, E. L.	1976	Monte-Carlo Significance testing as applied to statistical tropical cyclone prediction models, (submitted to NOAA for review).
Rai Sircar, N. C.	1956	*Indian J. Met. Geophys.*, **7**, p. 157.
Rao, K. N.	1958	*Ibid.*, **9**, p. 233.
Sikka, D. R. and Suryanarayana, R.	1968	India met. Dep. Sci. Rep. 76 (268).
Simpson, R. H. and Sugg, A. L.	1970	*Mon. Weath. Rev.*, April, 1970, pp. 293-306.
Simpson, R. H.	1971	*NOAA Tech. Memo.*, NWS SR-53, National Weather Service, 35 pp.
Snedecor, G. W.	1946	*Statistical Methods*, Ames, Iowa, The Iowa State College Press, 485 pp.