TRENDS IN TRACK FORECASTING FOR TROPICAL CYCLONES THREATENING THE UNITED STATES, 1970–2001

BY JAMES L. FRANKLIN, COLIN J. MCADIE, AND MILES B. LAWRENCE

An examination of long-term trends in forecasts for tropical cyclones threatening the United States shows statistically significant improvements in forecast accuracy.

Trends in the accuracy of tropical cyclone track forecasts issued by the National Hurricane Center (NHC) have been the subject of recent studies by McAdie and Lawrence (2000) and Powell and Aberson (2001). McAdie and Lawrence found that official NHC track forecasts over the period 1970–98 improved at an annual average rate of 1.0%, 1.7%, and 1.9% for the 24-, 48-, and 72-h forecast periods, respectively, for the Atlantic basin as a whole. They also found that each of these trends was statistically significant at the 95% confidence level. Powell and Aberson, noting that "although these trends (were) promising, neither forecast landfall position nor time error trends (had) been quantified," examined inferred landfall forecasts at time periods roughly 12, 24, 36, 48, and 60 h prior to landfall. They showed that none of the NHC landfall location error trends, and only the 24-h landfall timing error trend, showed a statistically significant improvement. Powell and Aberson attributed the overall apparent lack of improvement in landfall forecast errors to a "conservative least-regret" forecast philosophy for storms threatening to make landfall, or to deficiencies in numerical models or the observing network in the Caribbean and Central America.

It would be tempting to conclude from the results of Powell and Aberson that the accuracy of NHC forecasts close to the United States has not followed the basinwide trends reported by McAdie and Lawrence. However, differences in verification methodology between the two studies make such a conclusion problematic. The present study attempts to bridge the gap between these two verification methodologies by posing two questions: what are the long-term trends of NHC forecast errors for storms threatening the coastline, and are these forecast trends detectably different from basinwide trends?

AFFILIATIONS: FRANKLIN, MCADIE, AND LAWRENCE—NOAA/NWS/Tropical Prediction Center, Miami, Florida
CORRESPONDING AUTHOR: Mr. James L. Franklin, NOAA/NWS/Tropical Prediction Center, 11691 SW 17th St., Miami, FL 33165-2149
E-mail: James.Franklin@noaa.gov
DOI: 10.1175/BAMS-84-9-1197

In final form 30 January 2003

Although Powell and Aberson restricted their analysis to forecast tracks making landfall or passing within 75 km of the coastline, we prefer to take a broader view and consider tropical cyclones *threatening* land, whether or not a landfall is specifically forecast. This is motivated by the recent example of Hurricane Michelle, an extremely dangerous 120-kt hurricane in the northwestern Caribbean during November 2001. NHC official forecasts for Michelle called for the hurricane to turn away after approaching to within 140 km of the Florida Keys. Due to the anticipated close approach, and the extent of hurricane force winds from the center, a hurricane warning was issued for the Keys. Interest on the part of emergency managers and the general public was extremely high for this event, and evacuations were ordered, even though no (U.S.) landfall was ever forecast.

To identify tropical cyclone threats, one could use a "distance to land/direction of motion" threshold. This would be extremely complex computationally, especially if one wanted to include the effects of storm size. A simpler method that implicitly includes these effects is to consider those periods when watches or warnings (either hurricane or tropical storm) were in effect.[1] Although inadequate for some purposes, the watch/warning process is one of the primary mechanisms through which the meteorology of the official forecast triggers actions on the part of the general public. For the remainder of this discussion then, we will define a "land-threatening" tropical cyclone as one for which watches or warnings are in effect.

Although we focus later on the accuracy of NHC forecasts *issued* during the watch/warning period, one can just as well evaluate forecasts that *verify* during the watch/warning period. Both sets of forecasts are reasonably described as representing land-threatening storms, but the two datasets are not equivalent. Forecasts issued during the watch/warning phase are typically made under intense media scrutiny and pressure, when the landfall threat is imminent and psychological factors that could potentially influence the forecast process would most come into play. However, the longer-range portions of forecasts issued during the watch/warning phase will often not be relevant to coastal areas. This is because the 48- and 72-h datasets consist only of forecasts for which a storm was expected to threaten land at an earlier projection, and these forecasts may or may not represent a longer-range threat. Therefore it is necessary to also consider those forecasts *verifying* during the watch/warning phase. This strategy has the advantage of capturing the landfall threats, regardless of whether the threat is short or long range. The disadvantage is that many of these forecasts will have been issued for storms still far from shore, up to 4–5 days from landfall, when threat levels, media attention, and public interest, that is, those factors that might induce a conservative forecast philosophy, are relatively low. Because of this, our interest lies primarily with forecasts *issued* during the watch/warning period; however, we will show that the conclusions to be drawn are similar regardless of which set of forecasts are considered.

Our analysis follows that of McAdie and Lawrence, except that we have updated their period of study (1970–98) to include the subsequent years through 2001. NHC official track forecast errors for the 24-, 48-, and 72-h forecast periods are compiled into annual averages, and then adjusted for forecast difficulty by comparing the average annual official forecast errors to forecast errors from a climatology/persistence model, CLIPER (Neumann 1972). The adjustment process follows both McAdie and Lawrence and Powell and Aberson. As in the aforementioned studies, forecasts are not included in the verification if the initial or verifying intensity was below tropical storm strength. The adjusted forecast errors for each sample are given in Table 1.

A linear trend of the adjusted errors was computed, in which each annual average error was weighted by the number of forecasts from that particular year. Analysis of long-term trends was first performed for the full sample of forecasts (essentially repeating the analysis of McAdie and Lawrence). A second analysis was then performed for just the land-threatening storms, that is, for those forecasts issued when U.S. mainland watches or warnings were in effect. The resulting trend lines are shown in Fig. 1 and the results are summarized in Table 2.

For the sample of all Atlantic basin forecasts, trend lines at each forecast period indicate improvements, with annual average percentage improvements of 1.3, 1.9%, and 2.0% at 24, 48, and 72 h, respectively. These trend lines explain roughly 60%–70% of the variance in annual adjusted forecast error and are significant beyond the 99% level. Due to relatively low forecast errors during 2000 and 2001, these improvement rates are slightly larger than those reported by McAdie and Lawrence. Interestingly, these improvement rates for

[1] A hurricane (tropical storm) watch means that hurricane (tropical storm) conditions are possible within the watch area within 36 h. A hurricane (tropical storm) warning means that hurricane (tropical storm) conditions are likely within 24 h.

TABLE 1. NHC average annual official forecast errors, adjusted for forecast difficulty (Err) and number of cases (*N*) at 24, 48, and 72 h, for the period 1970–2001. Units are nautical miles (n mi). Errors are given for all forecasts (All), and separately for those forecasts issued when watches or warnings were in effect for the mainland United States (W/W).

Year	24 h				48 h				72 h			
	All		W/W		All		W/W		All		W/W	
	Err	*N*	Err	*N*	Err	*N*	Err	*N*	Err	*N*	Err	*N*
1970	101.8	34	104.7	17	157.1	13	194.2	3	62.4	3		0
1971	125.7	183	97.9	17	268.7	137	172.4	6	383.1	118	124.0	2
1972	102.8	57	56.1	6	213.7	38	144.5	2	322.5	25	349.6	1
1973	100.1	84	124.5	7	203.9	54	271.5	3	320.4	29	648.5	1
1974	113.6	89	88.7	6	268.1	64	282.1	2	458.8	42		0
1975	109.6	121	113.9	5	253.4	92	424.7	1	417.6	68		0
1976	109.2	143	75.6	10	223.5	113	181.9	4	350.7	85		0
1977	98.5	30	77.9	11	205.5	14	174.3	6	242.4	5	222.1	2
1978	120.6	101	128.6	8	266.7	59	320.9	8	396.7	33	589.0	7
1979	107.3	138	69.9	23	224.4	98	181.8	17	326.0	83	191.9	8
1980	110.0	188	102.2	9	254.0	140	232.1	5	378.4	109	175.2	1
1981	120.8	190	70.1	6	233.2	146	116.3	6	360.9	106	247.0	6
1982	118.6	45	171.7	2	235.8	29		0	335.2	21		0
1983	92.8	34	97.9	8	187.7	18	131.6	3	337.2	10		0
1984	117.4	157	87.1	16	207.8	122	129.4	16	307.0	89	267.0	10
1985	90.4	151	85.0	50	181.1	106	198.4	33	314.3	69	348.3	16
1986	112.5	66	111.7	8	254.5	42	338.7	6	391.5	27	437.3	3
1987	103.2	119	83.1	4	190.2	95		0	261.5	67		0
1988	88.4	133	75.7	13	193.1	109	182.8	5	293.6	90		0
1989	107.5	215	56.5	9	221.5	166	191.5	1	329.4	129		0
1990	110.7	209	85.8	8	214.6	157	65.5	4	319.9	114		0
1991	88.2	55	66.4	10	152.3	31	60.4	5	238.4	17	79.8	1
1992	91.7	124	82.4	16	152.0	99	121.4	9	186.9	76	132.5	5
1993	91.4	82	64.2	7	150.1	60	123.2	7	224.8	41	186.0	7
1994	60.3	83	115.6	17	125.0	62	275.2	12	291.7	50	409.2	11
1995	94.7	408	103.7	34	177.0	341	172.0	22	270.5	278	240.7	17
1996	87.7	260	88.3	36	157.1	217	155.6	22	203.5	183	198.5	12
1997	92.7	75	66.8	9	179.0	51	132.1	5	237.0	38	161.3	1
1998	89.7	286	92.7	44	163.1	233	173.8	34	248.6	191	313.6	25
1999	80.1	267	68.2	74	165.4	223	119.2	55	247.5	182	172.4	39
2000	79.1	202	64.5	8	140.4	164	83.6	6	224.8	136	79.6	4
2001	63.5	183	54.2	21	116.1	134	136.8	15	174.8	100	225.4	10

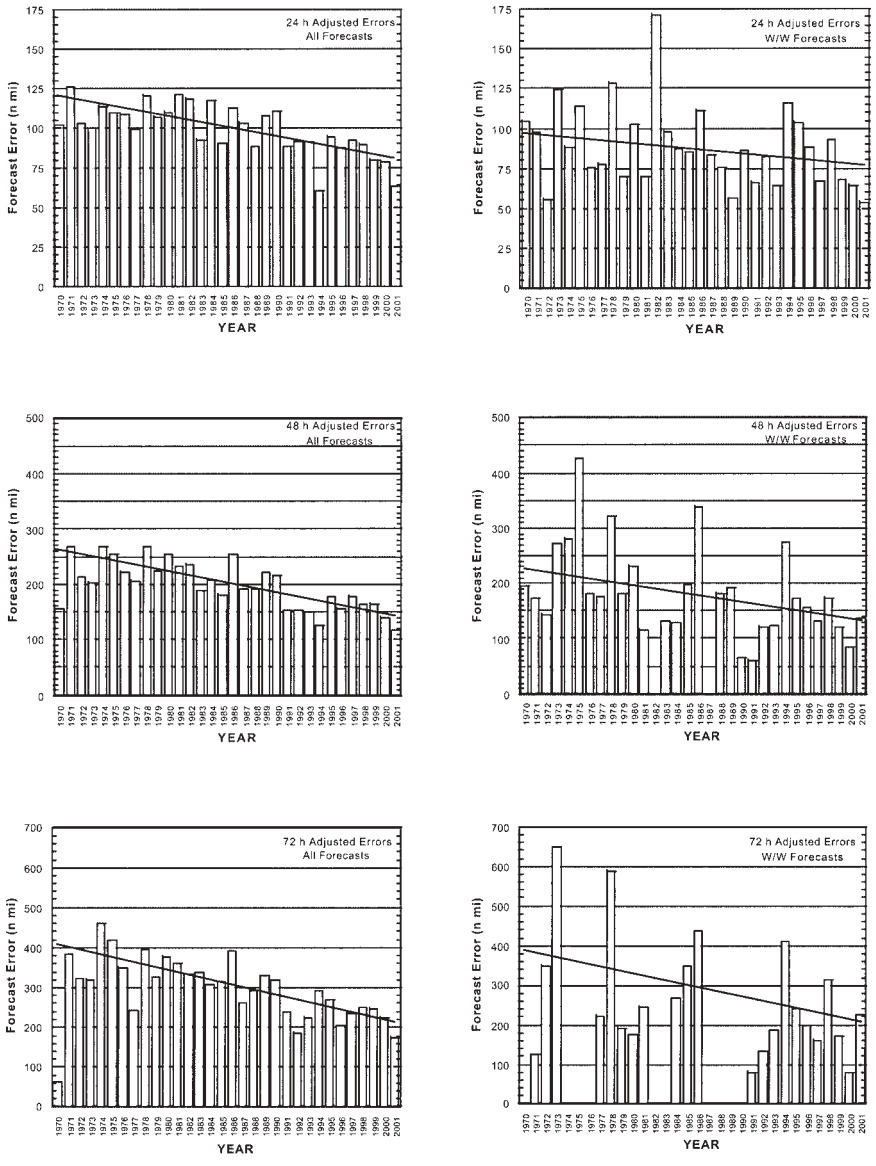

Fig. 1. Annual average NHC official track forecast errors, adjusted for forecast difficulty, over the period 1970–2001 for the Atlantic basin. Errors are given for the (top) 24-, (middle) 48-, and (bottom) 72-h forecast periods (left column) for all forecasts and (right column) those forecasts issued when either hurricane or tropical storm watches or warnings were in effect.

NHC forecasts are very close to improvement rates for the ensemble of operational model guidance found by Aberson (2001) for the period 1976–2000 (1.3%, 1.8%, and 2.4% for 24, 48, and 72 h, respectively.)

For the sample of tropical cyclones threatening land, trend lines at each forecast period indicate that NHC official forecasts also have been improving, with annual average percentage improvements of 0.7%, 1.6%, and 1.9% at 24, 48, and 72 h, respectively. Improvement trends at 48 and 72 h are about as large as those for the Atlantic basin as a whole, while the 24-h trend is about half as large as for the basin as a whole. Note that there is considerably more scatter among the annual average errors for this relatively small sample of forecasts issued under watches or warnings; the variance explained by the trend lines is roughly 10%–20%. Nevertheless, the 48-h trend line is significant at the 95% level, while the 24- and 72-h trend lines exceed the 90% level of significance but fall just below the 95%

Table 2. Average annual percentage improvement (Imp), variance explained (Var), and statistical significance (Sig) of trend lines shown in Fig 1. Level of statistical significance is indicated by (*) for the 90% level, (**) for the 95% level, and (***) for the 99% level.

Forecast period	All forecasts				Issued under watches/warnings			
	N	Imp (%)	Var (%)	Sig	*N*	Imp (%)	Var (%)	Sig
24 h	4512	1.3	61	***	519	0.7	11	*
48 h	3427	1.9	72	***	323	1.6	19	**
72 h	2614	2.0	70	***	189	1.9	17	*

level. One should recall that failure of the 24-h trend line to reach the 95% significance level does *not* mean that there has been no long-term improvement in official 24-h NHC forecasts for storms threatening land. What the test tells us, in fact, is that the likelihood of finding a relationship this strong by chance where none exists is rather low (between 5% and 10%). The more reasonable conclusion to be drawn collectively from these three regressions is that NHC forecasts for land-threatening storms *do* show a long-term improvement trend.

The second question noted above was whether forecast trends near land are different from those for the basin as a whole. The Chow (1960) test can be used to evaluate the equivalence of regression parameters (slope and intercept) of a sample split into subsamples—for example, the sample of all forecasts split into those made near land and those made away from land. However, this test cannot be applied to the annual average error trend lines computed above; in order to use the Chow test we must perform regressions on the individual forecasts themselves.

We have repeated the analysis of long-term forecast trends for the period 1970–2001, this time considering each forecast individually. Three regressions are performed: one for those forecasts issued when watches and warnings were in effect, one for those forecasts issued when watches and warnings were *not* in effect, and one for the entire sample of forecasts. As before, forecast errors are adjusted for difficulty, except that the adjustment is made to each individual forecast. It is no longer necessary to do a weighted regression to determine trend lines from individual forecasts; however, it is necessary to consider serial correlation in the significance tests, since Neumann et al. (1977) and Aberson and DeMaria (1994) suggested that independence may not be fully attained between forecasts separated by less than 24–30 h. Following Franklin and DeMaria (1992), an effective sample size is calculated (using the more conservative 30-h criterion) and used to compute the F and Chow statistics as well as the degrees of freedom.

Results of these regressions are summarized in Table 3. Not surprisingly, improvement rates for the watch/warning forecast errors are nearly the same as those computed previously from the annual average errors (Table 2). However, because of the larger sample size associated with the individual forecasts, the level of significance of the trend lines has increased: the 24-h trend line is now significant at the 95% level and the 48-h trend line is significant at the 99% level. Repeating this analysis for forecasts verifying during the watch/warning phase gives similar results (Table 4). This increases confidence in our earlier conclusion that NHC forecasts for land-threatening storms are improving.

These calculations do support the notion, implied by Powell and Aberson and shown earlier in Fig. 1, that forecasts for nonland-threatening storms improved more rapidly than those for storms threatening land, at least at 24 and 48 h. Examination of the trend lines in Fig. 1 shows that forecast accuracy is currently comparable for the land-threatening and nonland-threatening samples, but that this was not the case in the 1970s. In our view, changes in observing systems over time, in particular the increasing use of satellite observations in numerical models over data-sparse oceanic regions (which would preferentially improve the analysis of the hurricane environment in these regions), could account for this more plausibly than a "conservative" forecast philosophy on the part of the NHC.

TABLE 3. Results of trend analysis based on regressions of all individual forecasts during the period 1970–2001. Average annual percentage improvement (Imp), variance explained (Var), and level of significance (Sig) for the trend line are given for those forecasts issued when watches or warnings were in effect and for those forecasts issued when no watches or warnings were in effect. The "*N" represents the effective sample size after the serial correlation of individual forecasts is taken into account; actual sample sizes are roughly 3–5 times larger. Results of the Chow test for equivalence between the two regressions are given in the column labeled "Diff." Level of statistical significance is indicated by (*) for the 90% level, (**) for the 95% level, and (***) for the 99% level.**

Forecast period	Issued under watches/warnings				Not issued under watches/warnings				Diff
	*N**	Imp (%)	Var (%)	Sig	*N**	Imp (%)	Var (%)	Sig	Sig
24 h	170	0.8	3	**	1048	1.6	7	***	***
48 h	111	1.7	6	***	821	2.2	13	***	***
72 h	66	1.9	5	*	645	2.3	14	***	

Table 4. As in Table 3, except that forecasts are stratified into the "Watch/warning" or "No-watch/warning" samples based on when the forecasts verify, rather than on when the forecasts are issued.

Forecast period	Verified under watches/warnings				Not verifed under watches/warnings				Diff
	*N**	Imp (%)	Var (%)	Sig	*N**	Imp (%)	Var (%)	Sig	Sig
24 h	178	0.9	3	**	1043	1.5	7	***	***
48 h	134	1.5	5	***	810	2.2	14	***	***
72 h	100	1.5	4	**	628	2.3	14	***	***

The different improvement rates found for the land-threatening versus nonthreatening forecasts is consistent with Powell and Aberson. However, our finding of improvement trends for land-threatening storms appears to contradict their assessment of a lack of improvement in landfall forecasts; indeed it is hard to accept the notion that NHC forecasts near land are improving but the specific landfall points contained within these forecasts are not. While the sample of forecasts that contain a landfall is clearly distinct from those forecasts that are issued when watches and warnings are in effect, the two samples should be quite similar in terms of forecaster philosophy, data availability, and model performance. The apparent contradiction in results could simply reflect the different methodologies and samples between our study and Powell and Aberson; however, another possibility is that there have simply been too few landfalls (an average of 5 yr^{-1} at 24 h and only 3 yr^{-1} at 48 h over the period 1976–2000, or only about 3% of the total number of forecasts issued) to extract meaningful and robust long-term trends in the accuracy of landfall forecasts.

We would like to close with some thoughts on the attention that is often attached to the precise location and timing of forecast landfall. The average official 24-h track error over the period 1992–2001 was 81 n mi (149 km). Precisely because of the errors associated with tropical cyclone forecasts, the NHC tries to focus attention *away* from the precise forecast track of the center. Coastal residents under a hurricane warning are risking their property and lives if they fail to respond adequately because they see an official forecast indicating landfall in some community other than their own. The "strike" probabilities, in combination with the forecast storm size and peak intensity, can be used to help those not directly in the forecast path of a hurricane assess their particular level of risk. While the timing and location of landfall of the center are important, particularly to the news media, it is not clear that given the current state of the art, these "forecast" parameters are of any more practical importance than, say, the 24–36-h forecast positions, which help determine the coastal warning zones, or the 36–48-h forecast positions, which help determine the location of the watch zones. Adding to the importance of the prelandfall portion of the forecast is that many preparedness activities cease with the arrival of tropical storm force winds along the coast. For a storm with a 150 n mi (278 km) radius of tropical storm force winds traveling at 5 kt (2.6 m s^{-1}), this could occur 30 h in advance of landfall.

While the most severe hazards generally occur fairly close to the center, dangerous conditions associated with tropical cyclones cover a large area and may last for a day or more. In fact, a majority of the 600 U.S. deaths directly associated with tropical cyclones or their remnants for the period 1970–99 were associated with inland flooding (Rappaport 2000). This distribution of hazards is difficult to assess from the official forecast track alone. The current level of uncertainty in tropical cyclone track forecasts, and the distribution of hazards, dictate that actions to protect life and property should be more closely tied to threats defined by the tropical storm or hurricane watches and warnings.

REFERENCES

Aberson, S. D., 2001: The ensemble of tropical cyclone track forecasting models in the North Atlantic basin (1976–2000). *Bull. Amer. Meteor. Soc.,* **82,** 1895–1904.

——, and M. DeMaria, 1994: Verification of a nested barotropic hurricane track forecast model (VICBAR). *Mon. Wea. Rev.,* **122,** 2804–2815.

Chow, G. C., 1960: Tests of equality between sets of coefficients in two linear regressions. *Econometrica,* **28,** 591–605.

Franklin, J. L., and M. DeMaria, 1992: The impact of Omega dropwindsonde observations on barotropic hurricane track forecasts. *Mon. Wea. Rev.,* **120,** 381–391.

McAdie, C. M., and M. B. Lawrence, 2000: Improvements in tropical cyclone track forecasting in the Atlantic basin, 1970–98. *Bul. Amer. Meteor. Soc.,* **81,** 989–997.

Neumann, C. B., 1972: An alternative to the Hurran tropical cyclone forecasting system. NOAA Tech. Memo. NWS SR 62, 24 pp. [Available from Environmental Science Information Center, Environmental Data Service, NOAA, U.S. Department of Commerce, 3300 Whitehaven St. NW, Washington, DC 20235.]

——, M. B. Lawrence, and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *J. Appl. Meteor.,* **16,** 1165–1174.

Powell, M. D., and S. D. Aberson, 2001: Accuracy of United States tropical cyclone landfall forecasts in the Atlantic basin, 1976–2000. *Bull. Amer. Meteor. Soc.,* **82,** 2749–2767.

Rappaport, E. N., 2000: Loss of life in the United States associated with recent Atlantic tropical cyclones. *Bull. Amer. Meteor. Soc.,* **81,** 2065–2073.