Joint Hurricane Testbed Final Report for Year 1
September 1, 2005-May 1, 2006
Updated July 10, 2006 (See section 5 below)

Project: *Continued Development of Tropical Cyclone Wind Probability Products*
Principal Investigator: John Knaff and Mark DeMaria
Affiliation: Knaff (CIRA) and DeMaria (NESDIS)
Project dates: September 2005-May 2007
TPC Points of Contact: Rick Knabb, Michelle Mainelli, Chris Sisko and Jim Gross

Note: This final report was submitted 3 months early to re-align the project with the NOAA grant cycle.

Background

This project is continue the development of the Monte Carlo wind probability program and assist with the implementation of new products that are derived from the output. A verification system for the probabilities will also be developed. At the request of TPC, a new task involving the evaluation of the probabilities associated with hurricane watches and warnings from the 2004 and 2005 hurricane landfalls was added. The timeline and deliverables for this project are listed below in the Appendix.

1. Accomplishments

M. DeMaria coordinated with Rick Knabb of TPC to provide feedback on a training session that was developed to help explain the new probabilities to NWS forecasters and other users of the new products. In addition, several cases from the 2004 and 2005 seasons were re-run using the most current version of the program for Pablo Santos from the Miami WFO, for the development of an experimental algorithm that utilizes the probability output. A web site was created at CIRA displaying the complete set of probabilities from all of the cases for Hurricanes Charley, Frances, Jeanne and Ivan from 2004 and Katrina and Rita from 2005
(see http://rammb.cira.colostate.edu/projects/tc_wind_prob). A short description of the MC program is also provided on the web site to assist with training.

Work has begun on the evaluation of the probabilities associated with hurricane warnings from the 2004 and 2005 seasons. Table 1 lists all the storms that had a warning issued for at least on time period. The probability program was adapted so that it provides probabilities directly at the same set of coastal breakpoints that are used to issue warnings. This set includes 195 points along the U.S. coastline from Brownsville, Texas to Eastport, Maine. The distance between these points is fairly irregular with spacing ranging from about 5 to 50 nmi. To provide more even coverage, the official breakpoints were supplemented by additional coastal points, so that the difference between points is no more than 15 nmi. The final set includes 342 coastal points. The MC model runs at the supplemented breakpoint set for all 14 storms in Table 1 were completed.

A program to match the points with a hurricane warning with the probability output has also been developed. Results show that for all the coastal points for which a warning was issued for these 14 storms, the average 5-day cumulative probability was 28%. This is consistent with previous analysis of the warning regions which suggests that when a warning is issued there is actually only about a 1 in 4 chance of the point experiencing hurricane winds. This data has been further analyzed to determine the distribution of probabilities and the values at the end points of the warning areas. It is anticipated that this work may lead to a new application of the MC probability program, which would provide objective guidance for issuing and revising hurricane watches and warnings.

Table 1. Atlantic Storms with at Least One Hurricane Warning

Storm Name	Year
Alex	2004
Charley	2004
Frances	2004
Gaston	2004
Ivan	2004
Jeanne	2004
Arlene	2005
Cindy	2005
Dennis	2005
Emily	2005
Katrina	2005
Ophelia	2005
Rita	2005
Wilma	2005

The 14 storm cases in Table 1 are also being used as a test dataset for the development of the verification program. There were 375 times when a warning was issued or hurricane winds were observed along the coast for at least one breakpoint for these 14 storms, which provides 128250 points (375 x 342) for development of the verification program.

The development of a verification program involves data handling and the development of probabilistic verification methods. Much progress has been made in the data handling aspects of this project. Routines have been written to access the GRIB files containing the probabilities, access the necessary files from the ATCF (b-deck, a-deck) and methods to combine these information to create a deterministic forecast, keeping in mind that wind radii forecasts are issued out to only 72 hours. While the individual methods have been developed the integration of the pieces is not completed at this time.

Several verification methods were proposed for evaluating the probabilities, including a bias check and conditional distributions, discrimination distances, a Brier Skill Score, and a Relative Operating Characteristic (ROC) score. In addition we were tasked with examining the distribution of probabilities at the break points when warnings were in place. The progress thus far in the verification procedures using the test dataset discussed previously are presented below.

a) Bias

The bias program has already been developed, and compares the integrated probability to the total number of points with observed hurricane winds. Results for the 128250 point verification set showed that there were 5025 points with observed winds, and 4489 were expected from the summed probabilities. Thus, the bias is about 0.89. The slight low bias is probably due to the size and the atypical nature of the sample listed in Table 1.

In response to feedback from TPC during a visit in December, 2005, the bias verification was expanded to include stratification by probability categories or conditional distributions. For this calculation, the points are divided into 20 categories according to the probabilities estimated by the MC program (0-5, 5-10, …, 95-100%). The average estimated probability in each group is calculated, and then compared to the percentage of points in each group that actually experienced hurricane winds. If the probability program was perfect, the estimated and observed probabilities would be the same. If the observed frequencies are above (below) the forecast probabilities under forecasting (over forecasting) is indicated. If on the other hand, the observed frequencies cross the forecast probabilities a calibration issue is indicated.

Figure 1 shows the forecast probabilities and the observed frequencies for each group. The correspondence is quite reasonable, given the limited sample size of the test verification set. The observed frequency and estimated probably agree to within ~20% in every category. The under forecasting of events, indicated by the black dashed line, is also in agreement with the overall bias (i.e. ~10% less than observed), but gives more details as to where the bias is most prominent. In this dataset, much of the bias was associated with Hurricane Wilma, whose wind field grew unexpectedly as it approached landfall in Florida.

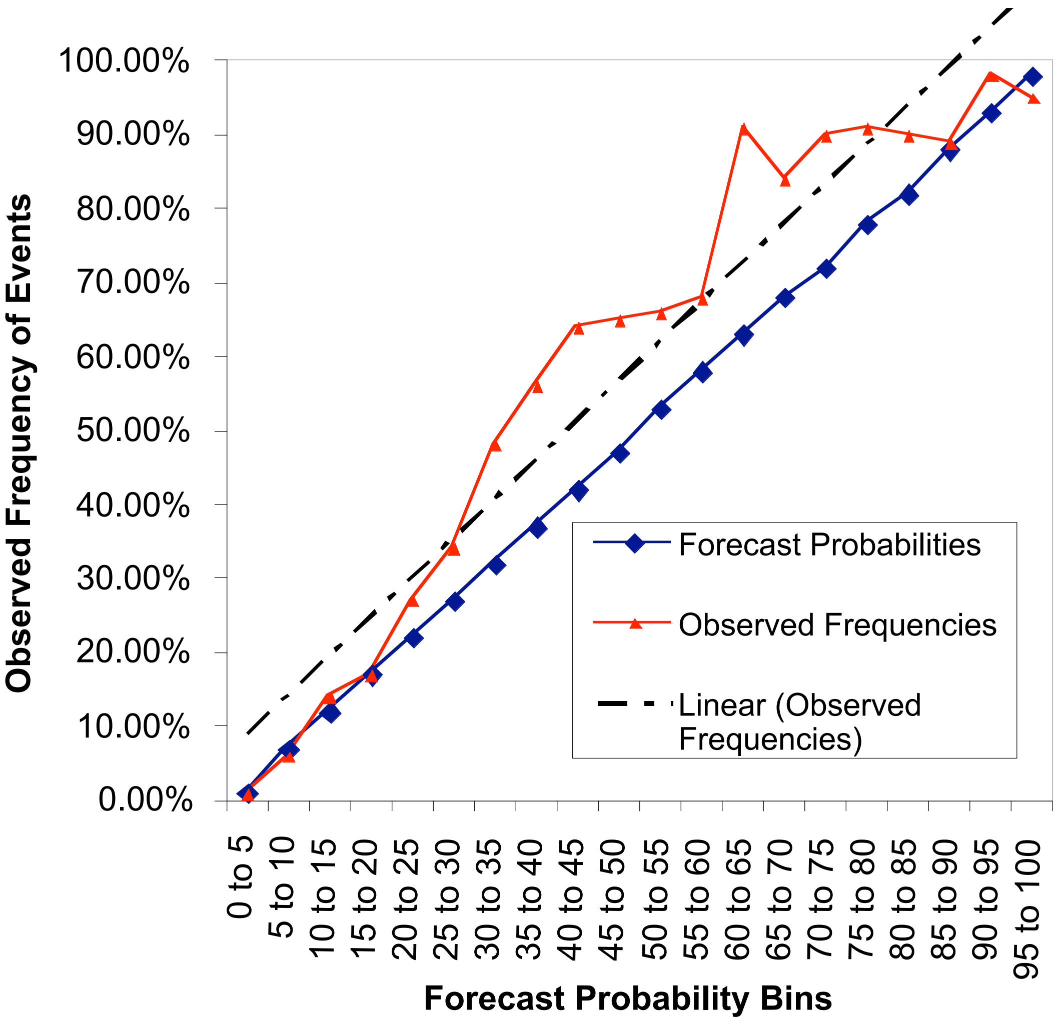

Figure 1. A comparison of the average forecast probability from the MC model and the observed frequency of occurrence of hurricane winds for 5% probability intervals of the estimated probability. The black dashed line is the best linear fit to the observed frequency and indicates approximately 10% under forecasting in this sample.

b) Discrimination distance

The discrimination distance measures the separation between two likelihood distributions created from the forecast probabilities when there is an event and when no event occurs. This involves the examination of the forecast probabilities at the points along the coast that experienced hurricane force winds and compare them with the forecast probabilities associated with the points not experiencing hurricane force winds. The differences between the means of these two likelihood distributions is the discrimination distance and measures the skill to discriminate between events and non-events. Figure 2 shows the distributions of the forecast probabilities at the landfalling points and non-landfalling points. The means and standard deviations of the landfalling points are 43% and 32%, respectively. At the non-landfalling points the mean likelihood is 2% with a standard deviation of 7% giving a discrimination distance of 41%.

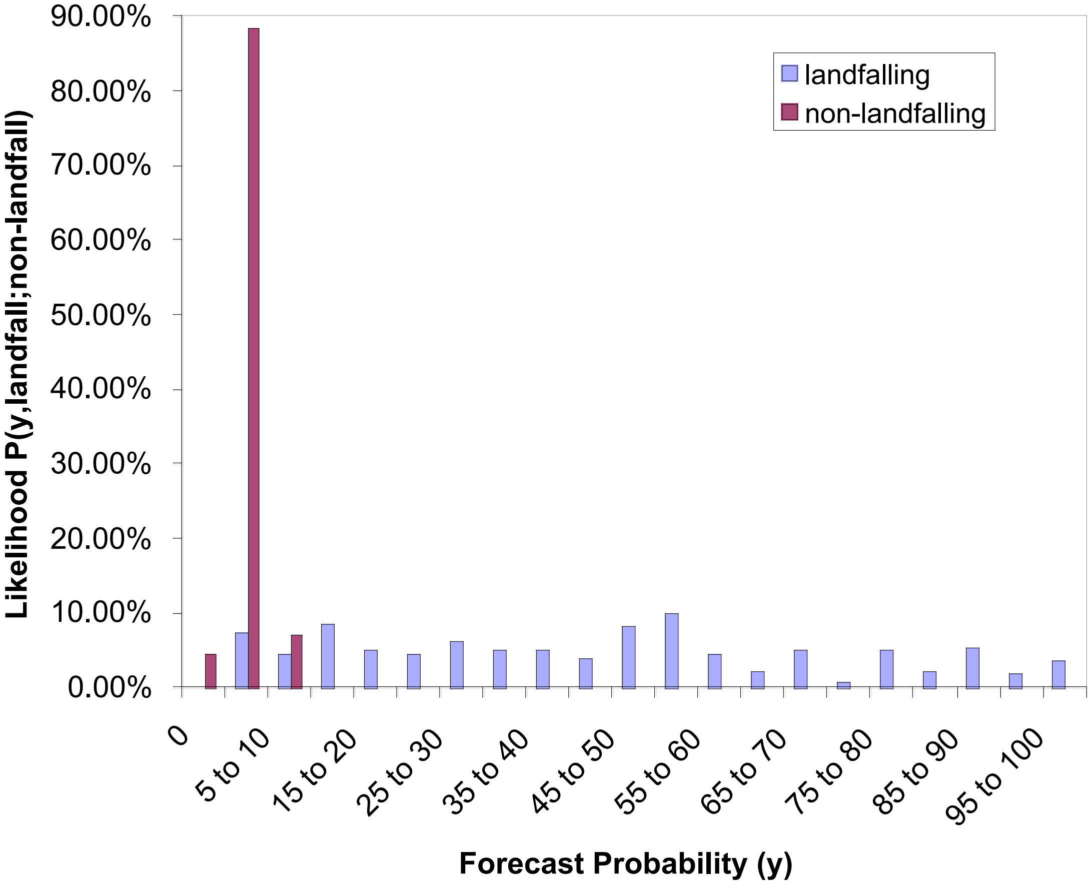

Figure 2. The likelihood distributions associated with landfalling and non-landfalling events in the warning breakpoint dataset.

c) Brier skill score

The program for the Brier Skill Score (BSS) first requires the development of a program for the Brier score. The Brier Score is the mean square error of a probabilistic forecast and measures the accuracy of a given forecast with values ranging from 0 (perfect) to 1 (worst). The BSS is the improvement of the Brier Score, relative to some baseline. The baseline that will be used is what would be available without the probability, which is just the deterministic forecast. This will be converted to a probability by assigning 100% for all points that fall within the 64 kt radii along the forecast track, and 0% for points that fall outside the radius. Results from the warning break point dataset indicate that the probabilities are skillful with respect to the OFCL forecast (i.e. BSS=28%). Since the number of verifying (OFCL) breakpoints is small, it was also instructive to create a Brier Skill Score based on a no-warning-all-the-time forecast. Here again skill is indicate; BSS=37%. As part of this process we also verified the OFCL deterministic forecast for this dataset, which indicates in this dataset that the OFCL verifies 51% of the time.

d) Relative Operating Characteristic (ROC) score

Contingency tables can be constructed from the outcomes of events and warnings (i.e., a binary forecast system), where the issuance/non issuance of a warning is contingent on a threshold probability (e.g., one may put up warnings when the probabilities along the coast greater than or equal to 10%). Table 2 shows the organization of the contingency table, where *h* is the number of hits, *f* is the number of false alarms, *m* is the number of misses, and *c* is the number of correct rejections. For our project we are interested in the trade-off between two quantities that can be estimated from the values compiled in the contingency tables; the hit rate (***hr***) = *h/(h+m)* and the false-alarm rate (***far***)=*f/(f+c)*. As the threshold probability for issuing warning varies (i.e., 10%, 20% …100%), ***hr*** and ***far*** also vary. By calculating ***hr*** and ***far*** over a range of different threshold probabilities for issuance of a warning for the sample of 128250 points, a curve (e.g. the ROC curve) can be constructed as shown in Fig. 3. The area under this curve is related to the skill of the probabilistic forecasts. If the curve is above the line where far is equal to hr, there is an indication of skill for that portion of the curve. The overall skill of a probabilistic forecast scheme can be estimated from the area above the ***hr***=***far*** line; ***skill*** = *2 X (A - 0.5)*, where *A* is the area under the curve. The ***skill*** will vary from 1.00 for perfect forecast to -1.0 for a perfectly bad forecast. For our test dataset, ***skill*** = 0.885 and the probability threshold where ***hr*** is maximized relative to ***far*** (i.e., the maximum likelihood ratio) is 60%. With the hurricane warning problem, however, it is likely more important to determine the threshold that maximize the ***hr*** suggesting a threshold closer to 10%. Such issues will be investigated during the rest of the project.

Table 2. Two-by-two contingency table for verification of a binary forecast system

	Forecasts		
Observation	**Warning (*W*)**	**No Warning (*W'*)**	**Total**
Event (*E*)	*H*	*m*	*E*
Nonevent (*E'*)	*F*	*c*	e'
Total	*W*	w'	*N*

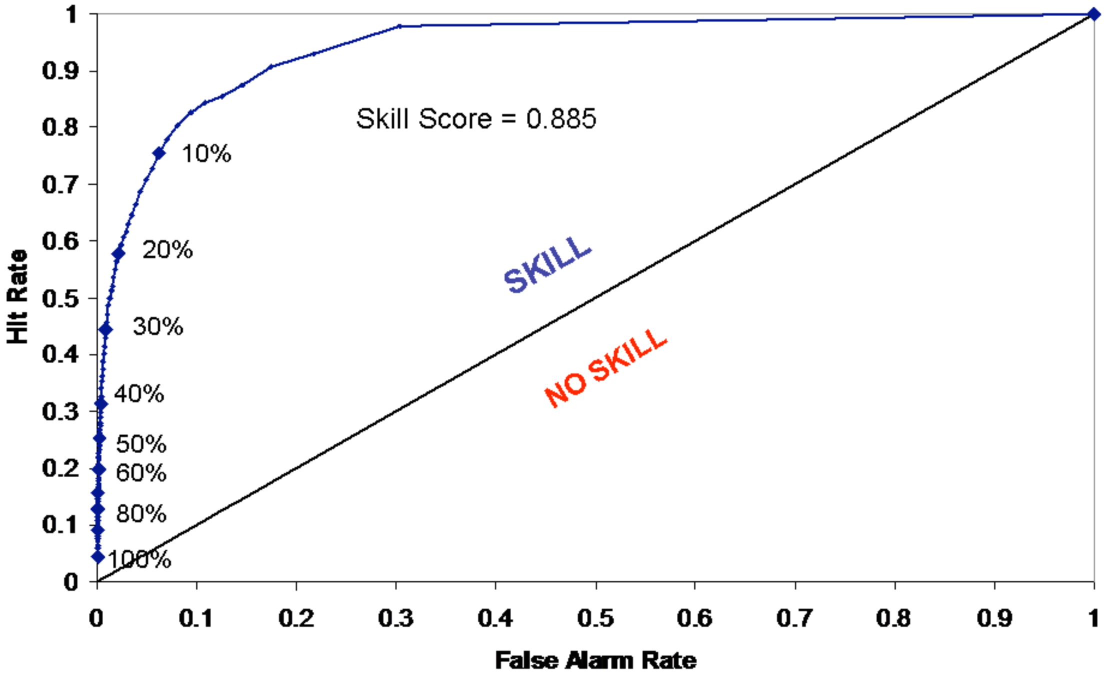

Figure 3. Hit rates vs. false-alarm rates associated with varying the threshold probabilities given by the 5-day cumulative probabilities from the MC model when warning were issued for the storms listed in Table 1.

e) Distribution of probabilities at the warning breakpoints

It is desirable for the probabilities to provide guidance as to where and when warnings should be issued. To begin to address this issue, the 2004-2005 warning dataset was utilized to examine the probabilities associated with the issued warnings, as requested by TPC. Figure 4 shows the distribution of forecast probabilities associated with the warnings, which has a mean of 28%. Also of interest was the distribution of probabilities in the un-warned areas also shown in Fig. 4, which has a mean of 1%. This figure is much like Fig. 2 showing a relatively broad range of probabilities observed in the warning areas and a relatively narrow range of probabilities in the unwarned areas. Unlike the likelihood shown in Fig.2, Fig. 4 indicates that there are quite a few instances when the probabilities are zero or near zero and the warnings are still in place.

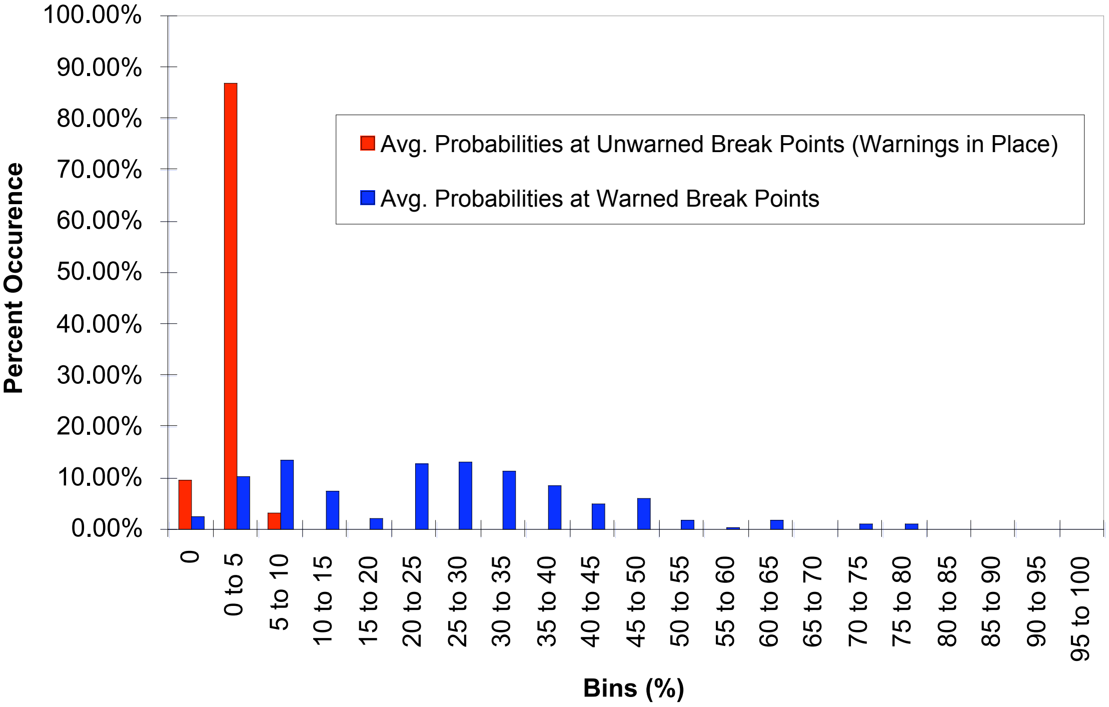

Figure 4. Probability density distribution of forecast probabilities at the warned and un-warned break points.

To more thoroughly examine the forecast probabilities at the warning break points, the distribution of the forecast probabilities was examined at just the starting and ending break points (Fig. 5). Here again there is an indication that the warnings are up when the probabilities in some instances are zero. This result inspired us to examine individual warning cases, and it was found that the warning are often left up too long despite the probabilities of 64-kt winds is less than 1 %. Thus changes in the operational procedures to revise previously issued warning based on the wind speed probabilities may result in a decrease of average area warned during each event.

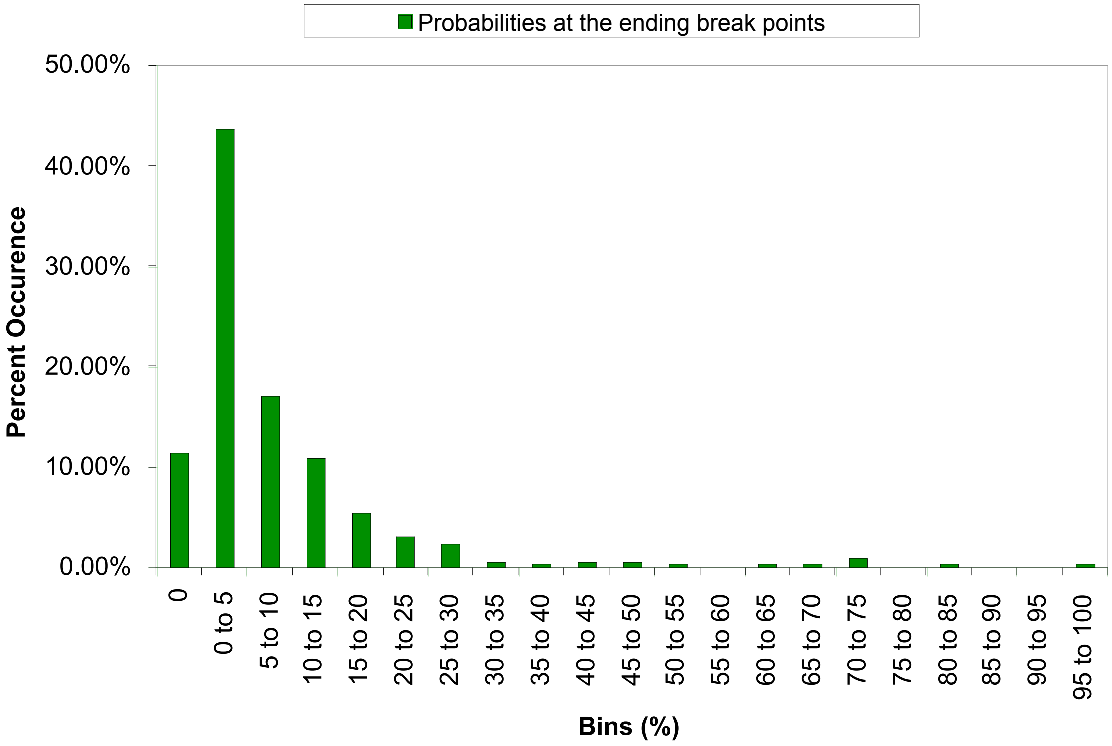

Figure 4. Distribution of the forecast 64-kt wind probabilities at the starting and ending warning break points.

In summary, the verification procedures outlined in the original proposal have been tested on a the test sample that includes the probabilities at the U.S. breakpoints for landfalling storms from the 2004 and 2005 season. The verification shows that the probabilities are skillful by the standard measures, and have a fairly small bias. This work will continue in the second year to provide TPC with a robust verification package for the probability program.

2. Things not Completed/Pending Items:

Two items are a little behind schedule. Because of the rather active 2005 hurricane season, error distributions for the MC model have not been updated because the eastern Pacific best tracks are not finalized. Because of additional requests from NHC to examine the probabilities associated with the issuance of watch/warnings during the 2004-2005 seasons and the late arrival of initial funding/early final reporting the verification code and associated training has not yet been completed, but will occur early in the second year.

3. Things that did not succeed:

None.

4. Plans for Year 2

The project will continue according to the schedule shown in the Appendix. The development of the verification program will continue using the 14 storm test set, and then run on the full 2005 and 2006 seasons. Results from the evaluation of the probabilities at the hurricane warning locations, and additional progress on the verification program will be reported at the Interdepartmental Hurricane Conference in March of 2007.

5. Updates to the Final Report for May 1-July 10, 2006

The underlying track and intensity error distributions were updated to include the 2005 cases from JTWC and NHC. The code with the new error files were implemented on the NCEP IBM and at JTWC for the 2006 season, in collaboration with Chris Lauer from TPC and Buck Sampson from JTWC. The error distributions for all basins now include a 5-year sample (2001-2005). The error distribution of the wind radii CLIPER model was not modified because that was developed from a much longer sample (1988-2004 in the Atlantic), and the error characteristics of that model should not change with time. If NHC and JTWC ever product radii forecasts out to 5 days, the wind radii CLIPER component of the method could eventually be replaced.

In May of 2006, we were notified by TPC that the basic output of the model will now be in grib2, rather than grib1 format, and there is a new requirement for cumulative and incremental probabilities at 6 hour, rather than 12 hour intervals. In order to deliver a verification code that will not immediately be archaic, the development plan was modified to take into account these new requirements. Thus, it may not be possible to run the verification code on the full 2005 sample unless all of the cases from that year are re-run with the new output format (grib2) and time frequency. Thus, the first complete season verification will be performed using the 2006 sample. Selected cases from 2005 will still be used for code development and testing. The project timeline in the Appendix below was modified slightly to account for these new developments.

Appendix

Year- two project timeline and deliverables:

May 2006 – Provide the 2001-2005 error distributions (AL, EP, & WP) for the MC model (completed)
May 2006 – Perform evaluation of 2005 MC real-time runs (selected cases only, see section 5)
Dec 2006 – Perform verification on full 2006 season runs with preliminary best tracks (or final if available)
Jan 2007 - Deliver verification code to TPC and provide training
Feb 2007 – Semi annual report
Mar 2007 – Present progress at the 2007 IHC

May 2007 – Provide updated MC code to TPC for 2007 based upon 2005/2006 verification
Aug 2007 – Provide Year 2 final report/proposal renewal