Improving the Validation and Prediction of Tropical Cyclone Rainfall

Joint Hurricane Testbed
First Year Annual Report
August 1, 2003 – July 31, 2004

Principal Investigators Timothy Marchok, NOAA/GFDL
Robert Rogers, NOAA/AOML/HRD
Robert Tuleya, SAIC at NCEP/EMC

TPC Contact Richard Pasch

Project summary and timeline

Through funding from the Joint Hurricane Testbed (JHT), this project proposed to improve the validation and prediction of tropical cyclone rainfall. Improved validation of rainfall will enable the forecaster to identify errors and biases in the models, which can aid the forecaster in interpreting numerical guidance of rainfall and adjusting their forecasts accordingly. An accurate diagnosis of rainfall forecast errors requires a validation scheme that accurately measures the performance of the forecast system. However, no standard technique has been developed to validate rainfall forecasts for tropical cyclones. Conventional measures of precipitation forecast skill, such as bias and threat scores, are difficult to interpret in the context of tropical cyclones due to the strong dependence of rain location and magnitude on the forecasted track of the storm and differences in the spatial and temporal sampling areas of rain gauge data compared to model output. Therefore, a key task in improving rainfall forecasts is to develop validation schemes for tropical cyclone rainfall that provide a baseline measure of forecast skill independent of track error and sampling issues.

To accomplish the goals stated above, several deliverables were proposed to be completed by the end of this 2-year project: 1) Development of new rainfall validation schemes that provide a baseline of comparison for different forecast systems; 2) Production of rainfall forecast error statistics for historic United States landfalling storms using traditional and new validation techniques for the operational GFDL, Eta and GFS models, and the benchmark Rainfall CLIPER (R-CLIPER) model; and 3) Design of a new forecasting tool based on the R-CLIPER model that incorporates information related to vertical shear and storm track.

With these goals in mind, the following tasks were proposed to be completed by the end of the first year:

- acquire National Precipitation Validation Unit (NPVU) and other historical rain datasets
- validate current & historic cases with operational and improved GFDL, Eta, GFS, and R-CLIPER models
- compare GFDL forecasts with NOAH LSM coupled model
- develop new verification techniques
- evaluate shear, track fields from GFDL runs on historical cases, quantify these relationships for incorporation into R-CLIPER

This document will report on the progress reached up to this point.

First year accomplishments

1) Summary of accomplishments

The following list summarizes the accomplishments for the previous year. Further explanations are provided in subsequent sections.

Data acquisition

- gridded multi-sensor (gauge and radar) rainfall observations, rain gauge observations, GFDL, GFS, and Eta operational models from all U.S. landfalling tropical cyclones from 1998-2003 acquired. R-CLIPER rainfall benchmark model run for each storm. GFDL operational model also available from 1995-1997.

Verification using conventional techniques

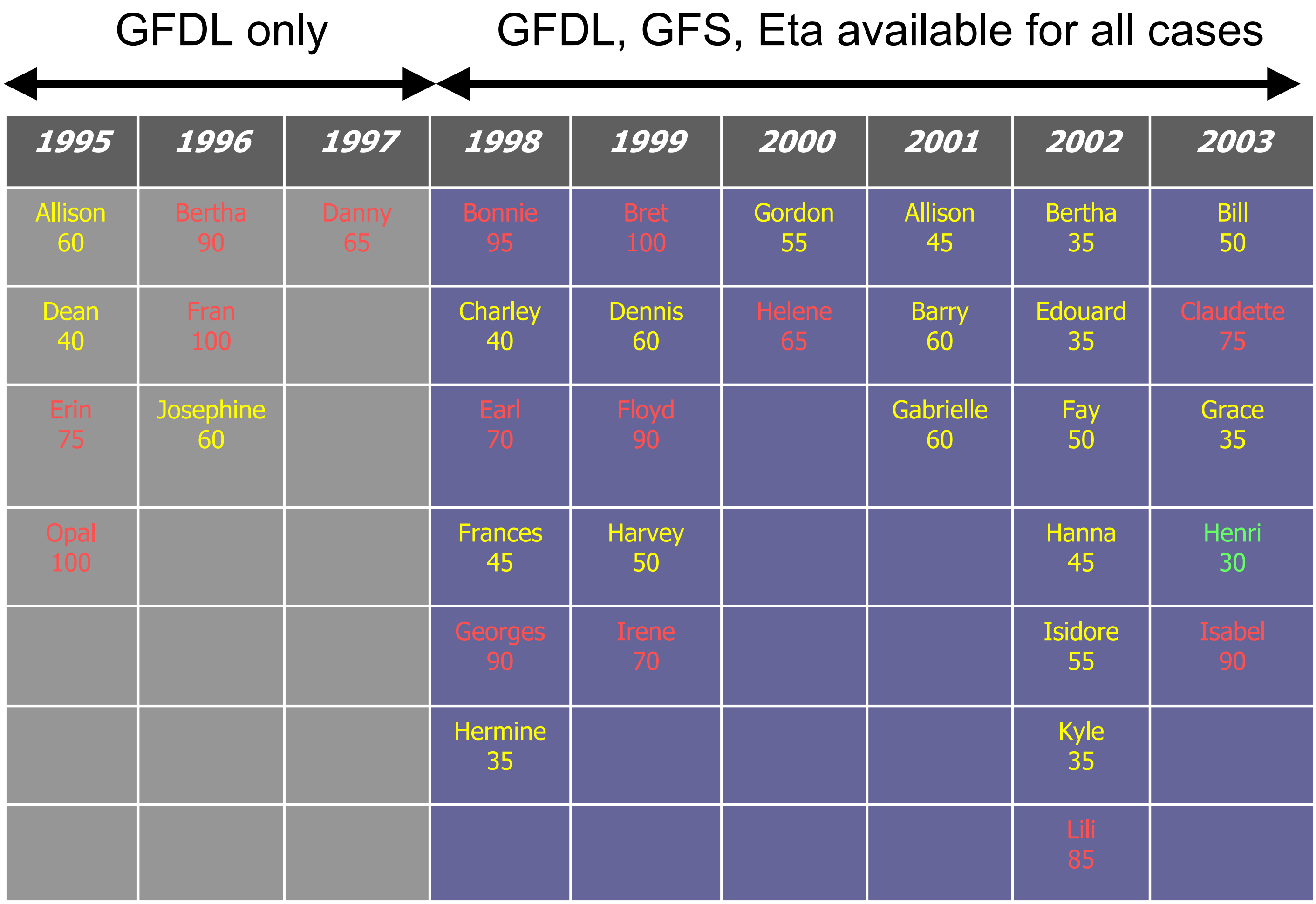
- Bias scores - all storms, 1998-2003 (GFDL, GFS, Eta, R-CLIPER, 2xR-CLIPER)
- Equitable Threat Scores (ETS) - all storms, 1998-2003 (GFDL, GFS, Eta, R-CLIPER, 2xR-CLIPER)
- Correlation coefficients – all storms, 1998-2003 (GFDL, GFS, Eta, R-CLIPER)
- Bias scores - all storms, 1998-2003 (GFDL, GFS, Eta, R-CLIPER, 2xR-CLIPER), stratified by intensity of storm at landfall (tropical storm vs. hurricane)
- ETS - all storms, 1998-2003 (GFDL, GFS, Eta, R-CLIPER, 2xR-CLIPER), stratified by intensity of storm at landfall (tropical storm vs. hurricane)

Development of new verification techniques

- Rain flux PDFs – all storms, 1998-2003 (GFDL, GFS, Eta, 2xR-CLIPER)
- Track-relative Rain flux PDF in swaths surrounding the storm – all storms, 1998-2003 (GFDL, GFS, Eta).

2) Acquisition of observational and forecast datasets

The first task consisted of gathering datasets to be used in the validation. The main rainfall observational dataset used in this work is hourly gridded rainfall data provided by the National Precipitation Validation Unit. This data is available online from NCAR, and it consists of multi-sensor (i.e., rain gauges, radar) rainfall maps that include areas impacted by landfalling tropical cyclones back to 1998. Before 1998, there is rain gauge data. Models evaluated in this work are the NCEP operational models: GFS, Eta, and GFDL. Table 1 shows the cases for which forecasts from these models are available. A total of 28 cases, spanning a range of intensities at landfall, are available for all three models from 1998-2003 (an additional 8 cases are available for the GFDL model back to 1995). An additional model, R-CLIPER, is also available. The R-CLIPER model (Marks et al. 2002, DeMaria and Tuleya 2001) is a simple scheme that has been developed to provide a benchmark against which forecasts of rainfall can be compared, similar to the way in which CLIPER and SHIFOR predictions provide the benchmarks for track and intensity forecasts, respectively. R-CLIPER is also run for each case used in the evaluations.



GFDL only

GFDL, GFS, Eta available for all cases

1995	1996	1997	1998	1999	2000	2001	2002	2003
Allison 60	Bertha 90	Danny 65	Bonnie 95	Bret 100	Gordon 55	Allison 45	Bertha 35	Bill 50
Dean 40	Fran 100		Charley 40	Dennis 60	Helene 65	Barry 60	Edouard 35	Claudette 75
Erin 75	Josephine 60		Earl 70	Floyd 90		Gabrielle 60	Fay 50	Grace 35
Opal 100			Frances 45	Harvey 50			Hanna 45	Henri 30
			Georges 90	Irene 70			Isidore 55	Isabel 90
			Hermine 35				Kyle 35	
							Lili 85	

Table 1. List of U.S. landfalling cases and model availability for the rainfall evaluation. Colors represent different intensities at landfall: Green – tropical depression; Yellow – tropical storm; Red – hurricane.

3) Calculation of error statistics using conventional techniques

One of the primary tasks in this work is to evaluate the performance of the various forecast models using techniques commonly used in the operational community. This can provide a baseline against which any new validation techniques can be compared. Two such techniques are the bias score and the equitable threat score. The bias score is obtained by the formula (Ebert et al. 2003):

$$BIAS = \frac{F + H}{M + H} \qquad (1)$$

where F = "false alarms", or predictions of rain where no rain occurred
H = "hits", or correct predictions of rain occurrence
M = "misses", or rain occurrences that were not predicted

while the equitable threat score (ETS) is obtained by the formula:

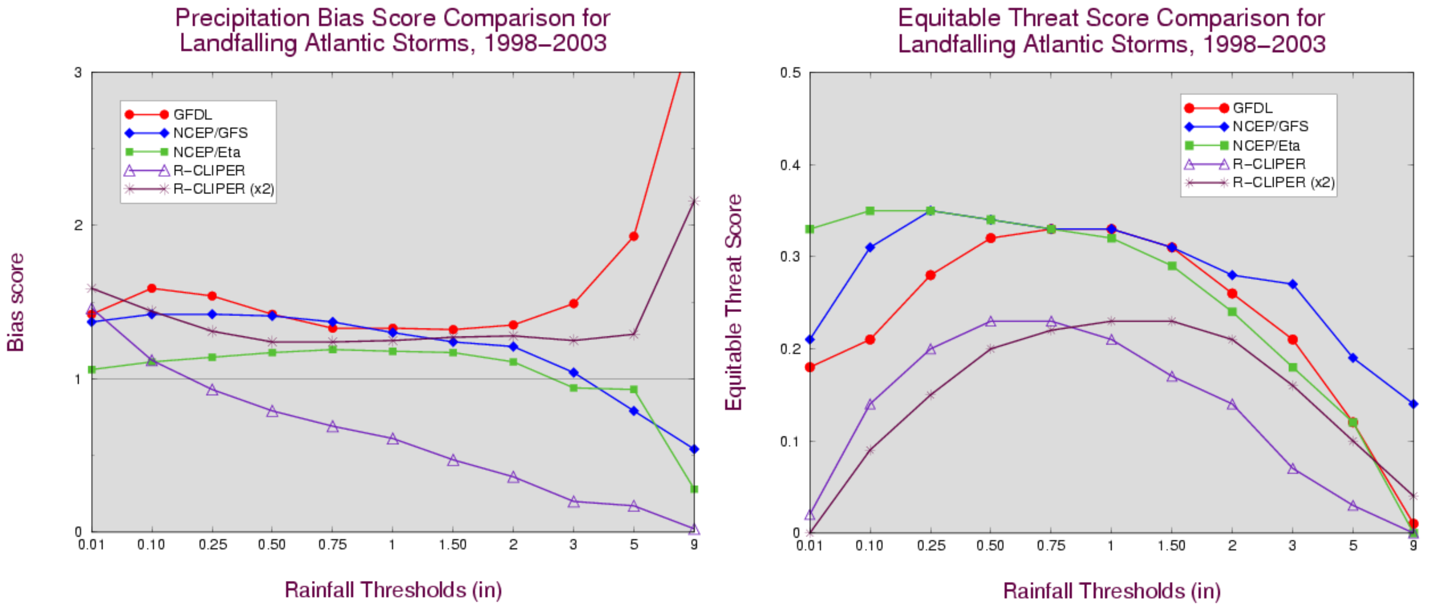
$$ETS = \frac{H - H_{random}}{H + F + M - H_{random}} \qquad (2)$$

where $H_{random} = \frac{(H+M)(H+F)}{N}$ represents the number of random hits expected due to chance for a given number of forecasts N

The bias score essentially compares the number of grid points (or area) within a forecast receiving rainfall exceeding a given threshold with the number of points (or area) in an observational dataset receiving rainfall exceeding that same threshold, independent of location errors. A value of 1 means the same number of points (or area) in the forecast exceed the given threshold amount as in the observations. Values greater than 1 indicate a forecast bias toward greater areal coverage for that rainfall amount than was observed, while values less than 1 indicate a forecast bias toward less areal coverage. The ETS counts the number of forecast "hits," i.e., the number of locations where the forecasted rain field matches the observed rain field within a given threshold. For this reason the ETS is dependent on location errors. A value of 1 indicates an exact overlap with the forecasted and observed area receiving rainfall of a given amount, while a value of 0 indicates that there is no overlap in space with forecasted and observed rainfall amounts of a given amount.

Figure 1 shows a plot of precipitation bias score and equitable threat score of 72-h rainfall for all 28 cases for each of the models (GFDL, GFS, Eta, R-CLIPER, and R-CLIPERx2). (The R-CLIPERx2 run is used as a result of conclusions drawn from Marks and DeMaria (2003) that it is necessary to double the R-CLIPER rain fields to produce reliable rain fields across the distribution.) As can be seen from the bias score (Fig. 1a), there is a fairly close agreement between most of the models (except for R-CLIPER) for rainfall amounts between 0.75 and 3 inches. For lighter rainfall amounts, there is a notable high bias for the GFDL and GFS models. For heavier rainfall amounts, there is a pronounced high bias for the GFDL and R-CLIPERx2 models, and a pronounced low bias for the Eta and GFS. The R-CLIPER shows a low bias across almost all of the distribution. For the ETS (Fig. 1b), all models except the Eta show the highest accuracy in the 0.5 to 2-inch rainfall band. For lighter and heavier amounts, they all perform worse. For the ligher amounts, the Eta does better than the GFDL and GFS, while for the heavier amounts, the GFS does best. Interestingly, both versions of the R-CLIPER perform worst across nearly all of the distribution.

Not surprisingly, the models exhibited significant case-to-case variability. This variability is seen in Fig. 2, which shows correlation coefficients for forecasted and observed 72-h rainfall for all 28 cases. There is significant case-to-case variability both in the average correlation coefficient among all models and in the spread in correlation coefficients for each case. For example, all models fared poorly for Helene, Bertha, and Harvey, while they all performed well for Floyd, Fay, and Isabel. For Georges, the GFDL performed poorly, but the GFS performed well. For Irene, the GFS and the GFDL both did poorly, while the Eta performed comparatively well. For Isabel, however, all of the models did very well. These differences are likely largely due to the track errors for each of the models. For those cases where one model (or all models) forecasted track well after landfall, that model forecasted rainfall well. Conversely, if the model forecasted track poorly, then it forecasted rain poorly.



(a) (b)

Figure 1. (a) Bias Score comparisons of 72-h rainfall for all models for 28 cases shown in Table 1; (b) Equitable Threat Score comparisons of 72-h rainfall for all models for 28 cases shown in Table 1.

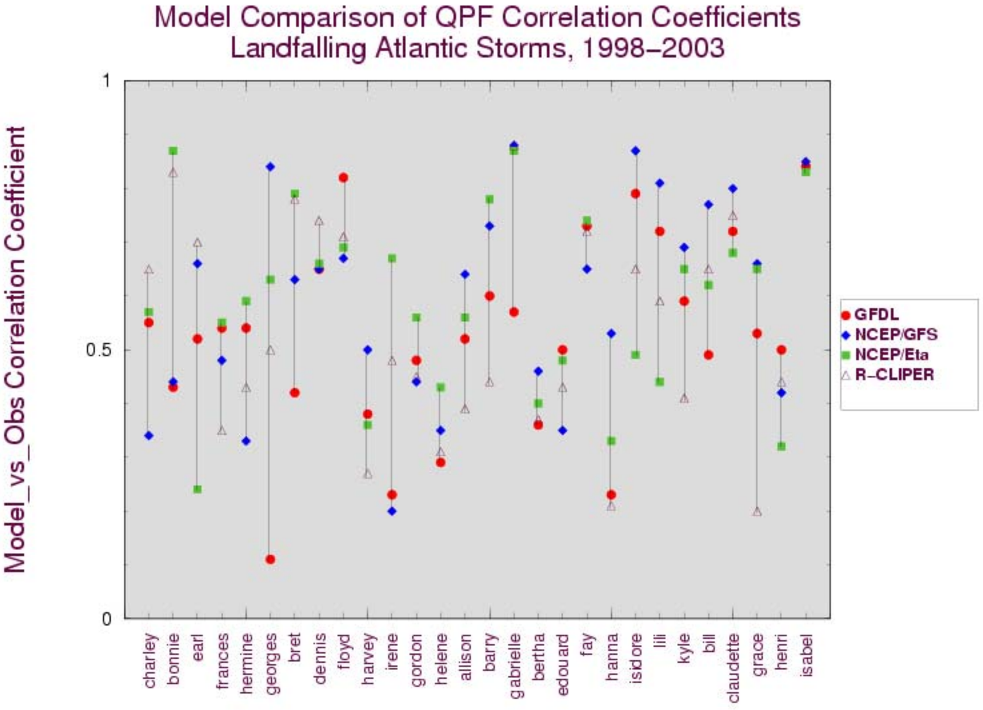
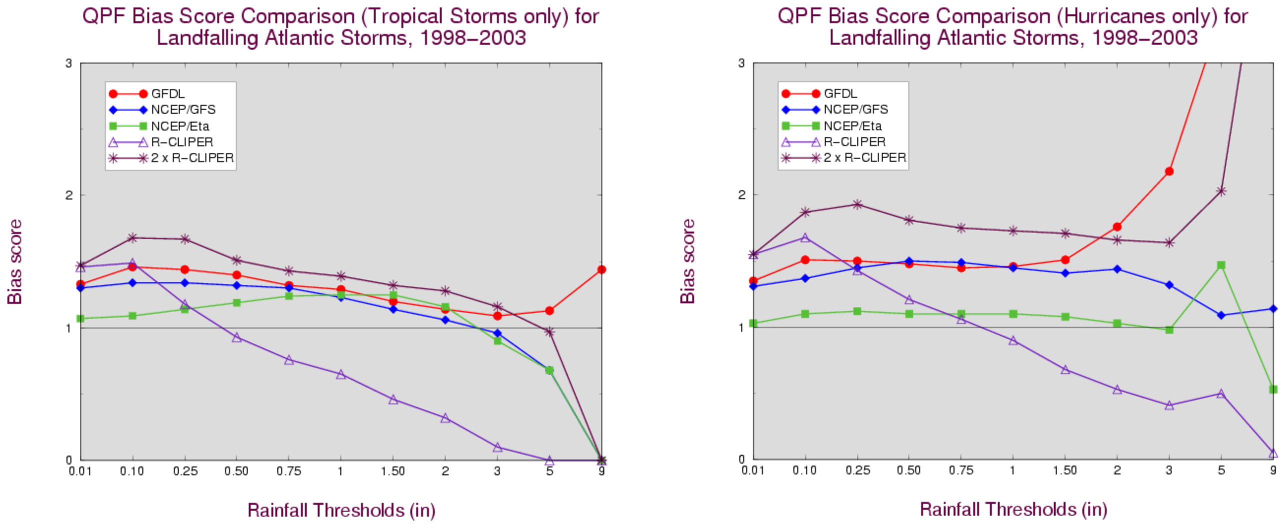


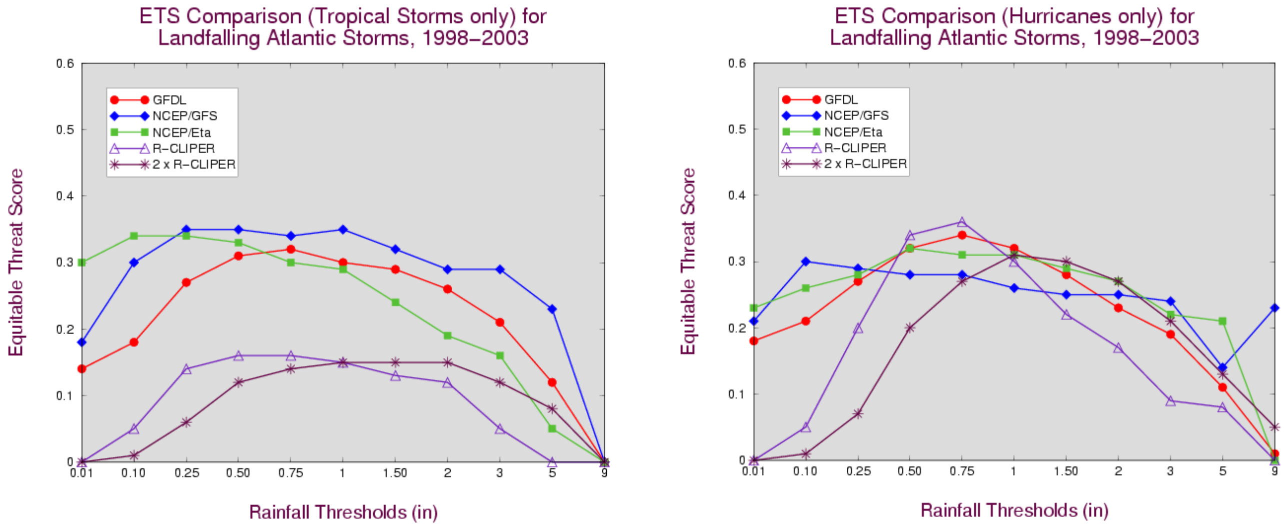
Figure 2. Correlation coefficients of forecasted vs. observed 72-h rain for all 28 cases shown in Table 1.

Rainfall forecasts were also verified after stratifying landfalling tropical cyclones based on their intensity at landfall. Figure 3 shows the bias scores separately for tropical storms and for hurricanes. All of the dynamical models show a high bias for rain amounts up to 5 inches for tropical storms. Above that amount, all of the models have a low bias except for the GFDL. The Eta model performs the best up to about 3 inches, above which point the GFDL performs best. In contrast, for hurricanes there is a more significant high bias, especially for the R-CLIPER models. Furthermore, for high rain amounts there is a pronounced high bias for both the GFDL and the 2xR-CLIPER models. The GFS also maintains a slight high bias, and the Eta no longer has the low bias it had for the tropical storms sample. The Eta performs the best of all models up to 5 inches for hurricanes, above which the GFS performs the best. Figure 4 shows a similar



(a) (b)

Figure 3. Bias Score comparisons for storms stratified by landfalling intensity. (a) tropical storms (b) hurricanes.



(a) (b)

Figure 4. Equitable Threat Score comparisons for storms stratified by landfalling intensity. (a) tropical storms (b) hurricanes.

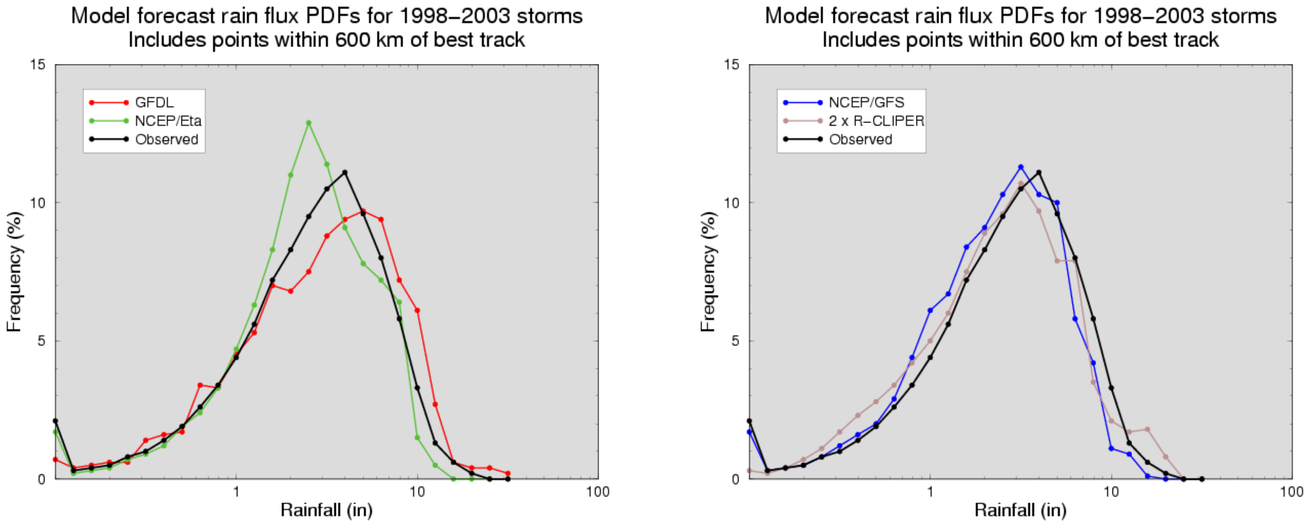
comparison using the equitable threat score. The GFS performs the best for tropical storms using this metric, and the two R-CLIPERs perform the worst. For hurricanes, the GFS and the Eta perform the best for light and heavy rain amounts, while the GFDL and 2xR-CLIPER are the best for the 0.5 to 1 inch range. It is interesting to note that there is a significant increase in the equitable threat scores for the R-CLIPER models in the hurricane sample compared to those for the tropical storms sample.

4) Development of new verification techniques

While the verification techniques shown above yield valuable information regarding the errors of the models, there are limitations in what these techniques can reveal. For example, some standard verification techniques do not account for the significant error that can arise simply from having an incorrectly-forecasted storm track (e.g., the ETS and correlation coefficient). Furthermore, a great deal of useful information can be obtained from considering the performance of the forecasts for the entire distribution of rainfall, not just peak rainfall amounts or point comparisons with specific rain gauges. This latter point is particularly important when comparing models of varying resolution to observations based on comparatively small sampling areas such as radar data or rain gauges, since a spatially averaged field always has lower variability than point values (Tustison et al. 2001). As a result of these limitations, work has begun in developing new validation techniques that better account for such factors as track error, sampling size discrepancies, and comparing the entire distribution of rainfall rather than peak rainfall amounts and point comparisons.

One technique that has been developed involves comparing the probability distribution functions (PDFs) of rain flux for each model with the observations. Calculating the rain flux consists of multiplying the rain amount by the resolution of the grid being considered to yield a total volume of rain falling on the grid. Using this technique can account for the differences in variability that arise due to averaging scale discrepancies (Tustison et al. 2001), though differences that arise in models due to the ability to resolve different features remain. Furthermore, this technique is more amenable to other types of track-relative verification schemes (discussed below).

Figure 5 shows rain flux PDFs for all 28 cases for 72-h rain falling within 600 km of the storm track. The observed rain flux shows a log normal distribution, with peak proportions of rain flux (about 11%) centered at 4-inches and returning back to 0% at about 15 inches. The Eta and GFDL models (Fig. 5a) accurately reproduce the flux distribution for the light rain amounts (R < 1 inch). However, the Eta model shows a maximum in the distribution of about 13% at 2 inches and lower frequency values for the higher rain values. This suggests that too much of the rain flux is concentrated in the lighter rain amounts than what is observed. The GFDL shows an opposite relationship: the peak in the flux distribution is at about 6 inches, and above that value the distribution is higher than the observations, indicating that too much of the rain flux is concentrated in the heavier rain amounts. These relationships are roughly consistent with the bias score relationships seen in Fig. 1a. A comparison of the observations with the GFS and 2xR-CLIPER models (Fig. 5b) shows that both of these models produce distributions close to what was observed, with peak values of the rain flux occurring at approximately the same rain amount as the observed distribution.



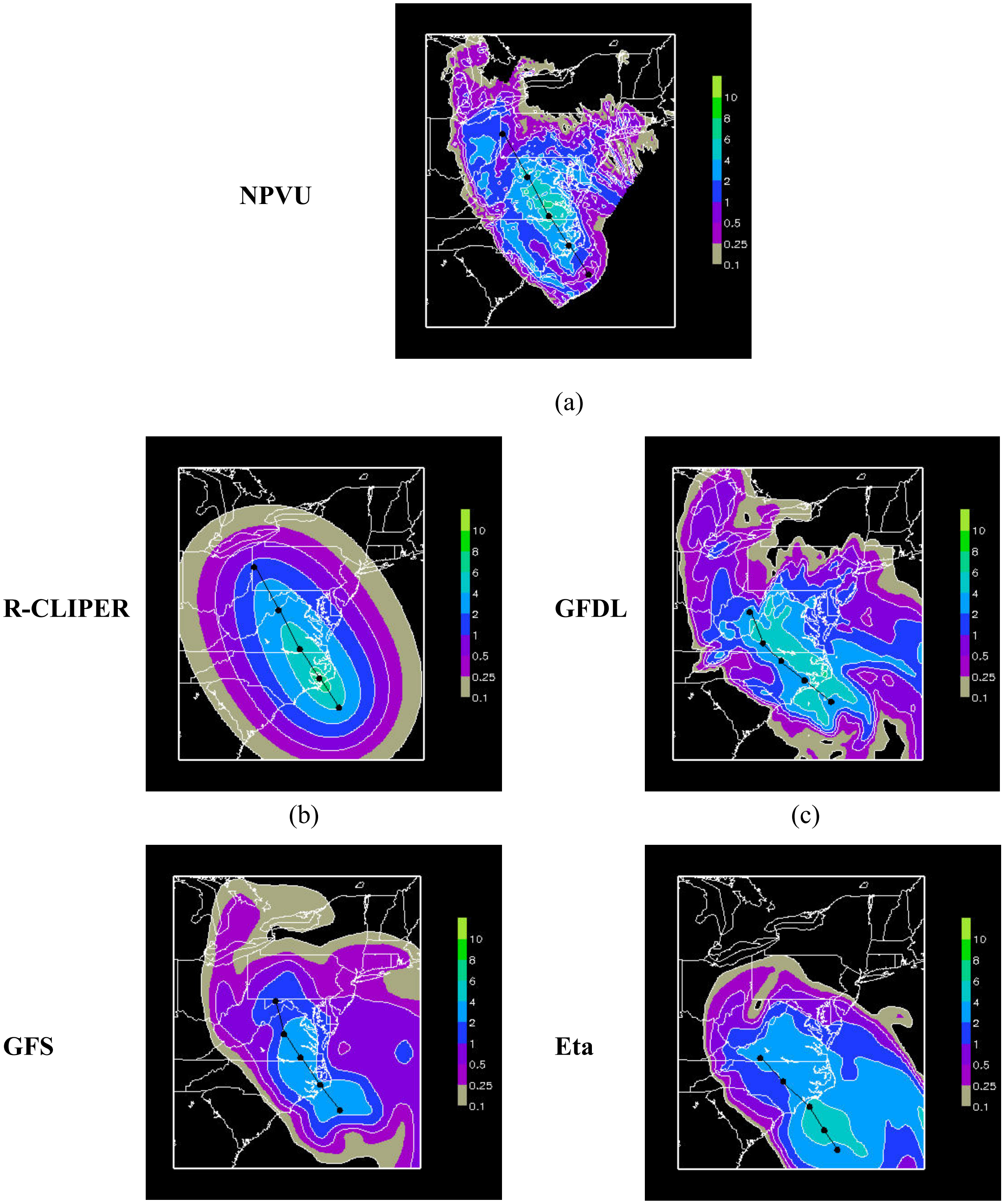
(a) (b)

Figure 5. Probability Distribution Functions of 72-h rain flux for all 28 cases for (a) observed, GFDL, and Eta fields; (b) observed, GFS, and 2xR-CLIPER fields.

With the PDFs thus calculated, cumulative rain flux distributions of the models can be easily performed and compared against the observations. As an example, Fig. 6 shows a plot of 24-h accumulated rain for the observations, R-CLIPER, GFDL, Eta, and GFS forecasts from 12 UTC 18 to 12 UTC 19 September of Hurricane Isabel. The observed rain maximum stretches along and just to the right of the storm track, and there is significant structure in the rain field, corresponding to rainbands and topographic effects (e.g., the maximum in Delaware and the minimum in southwestern Pennsylvania). The R-CLIPER produces comparable amounts, but little structure in the rain field. The GFDL produces rain amounts and structures comparable to the observations, while the Eta and GFS show some structure to the rain field, but the GFS produces a smaller area of maximum rain. The Eta, by contrast, produces a broad area of heavy rain. The cumulative frequency distributions (CDFs) for each dataset can be computed from the PDFs, and these CDFs can be compared using the probability matching method (PMM; Calheiros and Zawadski, 1987; Rosenfeld et al 1993). The PMM finds the set of pairs of observed and forecast CDFs at which the cumulative probabilities of the two are equal, assuming that the area covered by that cumulative probability rain amount is equivalent for both.

Figure 7 shows PMMs for each of the models compared with the observations. From this figure it is clear that the rain flux occurs at lower rain amounts (i.e., a low bias) across the entire distribution for the GFS compared to the observations. The rain flux occurs at higher rain amounts for the Eta (i.e., a high bias) up to about 60% threshold, then it approaches the observations for the upper end of the distribution (no bias). For the GFDL and R-CLIPER, the rain flux occurs at lower rain rates for initial 30% of distribution (i.e., a low bias), then occurs at higher rain rates for the top 50% of distribution (high bias)

Another verification technique that has been developed involves calculating rainfall statistics (e.g., rain flux PDFs) within predefined swaths around the storm track (either observed or forecasted). Figure 8 shows a schematic of the calculation areas. This technique of calculating track-relative accumulated rainfall exploits the fact that for tropical cyclones, the



(a)

(b)

(c)

(d)

(e)

Figure 6. Plot of 24-hr accumulated rainfall (in) from 12 UTC 18 to 12 UTC 19 September 2003 for Hurricane Isabel for (a) NPVU data; (b) R-CLIPER; (c) GFDL; (d) GFS; and (e) Eta models. Dynamical forecast models (c, d, and e) were initialized at 12 UTC 17 September. Dark solid line denotes best track or forecast position, with position of storm every 6 h denoted.

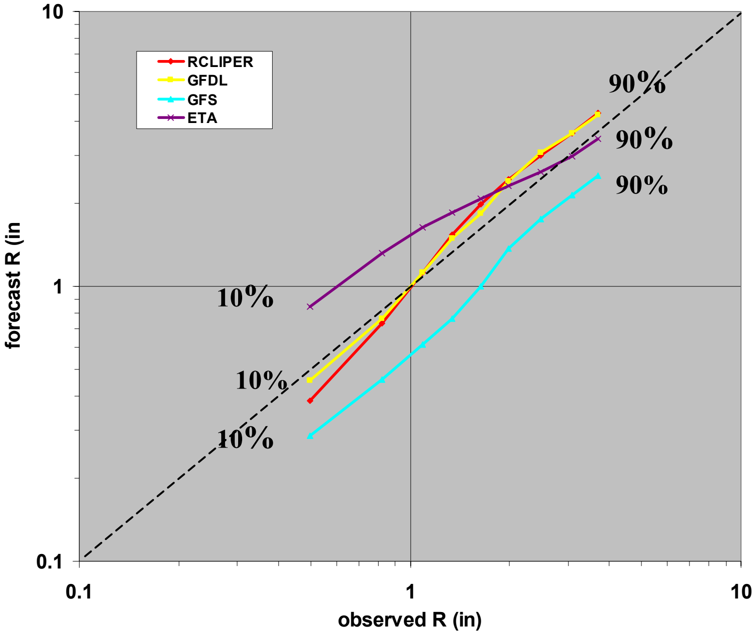
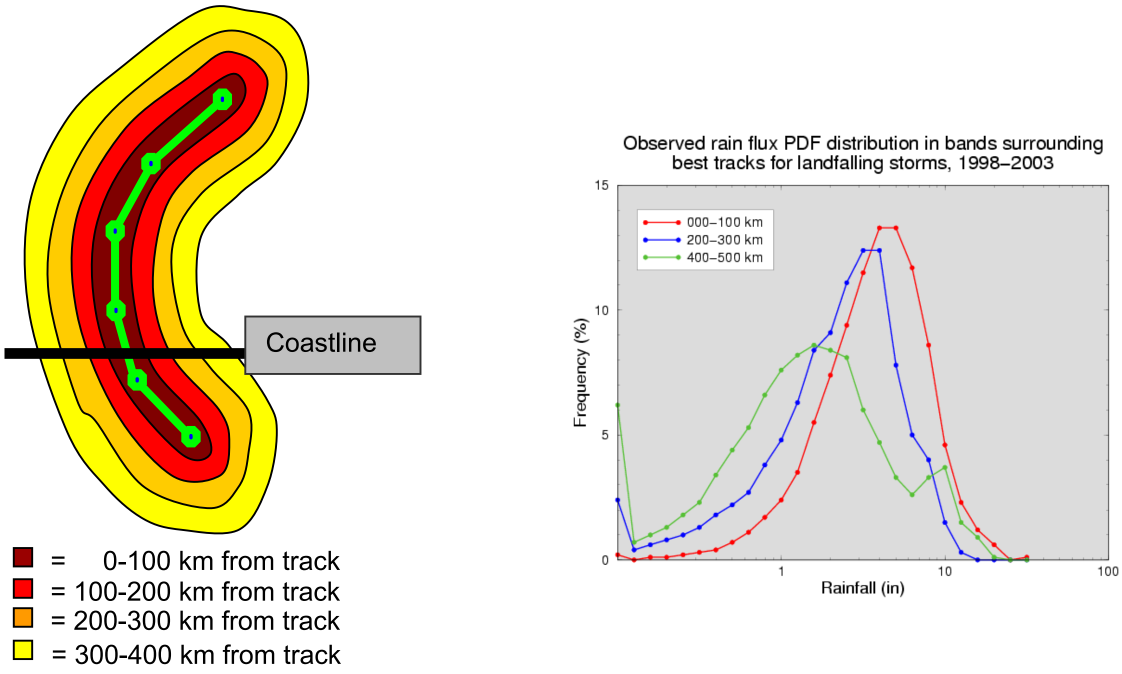


Figure 7. Probability-matched 24-h rain fluxes from observations and forecasts from Fig. 6. Each point presents the probability-matched value at 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% from left to right, respectively.



(a) (b)

Figure 8. (a) Schematic showing swaths for computation of track-relative statistics (e.g., PDFs, bias score); (b) PDF distributions of observed 72-h rain flux in 0-100 km, 200-300 km, and 400-500 km swaths for 28 cases shown in Table 1.

precipitation intensity is highly correlated with distance from the storm track. This provides a real advantage over schemes that validate rainfall for regular extratropical midlatitude systems, since a mid-latitude storm track is not nearly as well correlated with rainfall and so does not provide a solid reference point around which to do any type of track-relative rainfall analysis. To illustrate this correlation for tropical cyclones, Fig. 8b shows observed 72-h rain flux PDFs for several bands of increasing distance from the storm track for all 28 cases. For the innermost 100 km, the rain flux is maximized at around 8 inches, while for the 200-300 km swath, the rain flux is maximized at about 2 inches. For the outermost swath the primary peak is around 1 inch (though there is a smaller secondary peak at 10 inches)

Each of the dynamical forecast models was compared against the observations for the innermost 100 km, providing a comparison of the performance of each model in producing rainfall within the inner core of a storm after it makes landfall (Fig. 9). From this comparison it can be seen that the Eta model has a tendency to produce too much rain for the lower rain amounts within the cyclone core compared to the observations. By contrast, the GFDL and the GFS models produced too much rain for the higher end of the distribution.

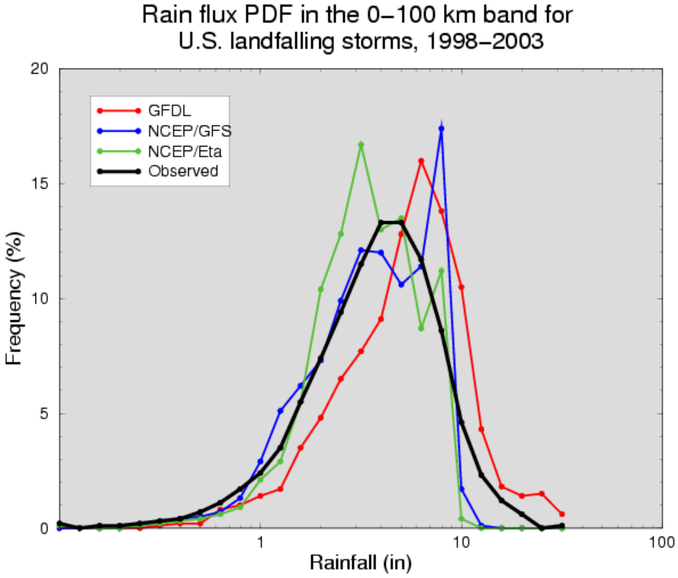


Figure 9. PDF distributions of 72-h rain flux for all cases in 0-100 km swath for observations, GFDL, GFS, and Eta models.

Pending items

This project is progressing well. Comparisons have been made between GFDL model running the NOAH Land-surface model with the GFDL model running a standard slab model, but those comparisons were not presented in this report due to space limitations. The only thing that has not yet been accomplished during year 1 are the evaluations of the relationship between vertical shear and convective asymmetries in the GFDL model for incorporation into a new model analogous to R-CLIPER. That is beginning now, and should be completed in a timely manner. Once this quantification occurs, it should be an easy task to incorporate it into the new model. Another possibility to using GFDL shear fields is to use GFS shear fields, since the new operational algorithm will use SHIPS shear fields which are also derived from the GFS.

Second-year activities

There are no significant changes in the time line, deliverables, or budget for the second year from what was stated in the original proposal. Work will continue on developing and refining the verification techniques presented here. A method of calculating areal rainfall averages following the storm will be developed. These averages will require output at a high time-resolution (e.g., hourly) to accurately capture the storm-relative rain fields. This requirement will be further explored. For the PDFs, a more easily quantifiable method for evaluating the performance of the models will be developed. For example, this may include stating the bias in the models as a difference in the mode of the distributions (i.e., the value of rain at which the peak flux occurs), after testing for statistical significance. In addition, storms will be stratified by other parameters, such as translational speed, proximity to topography, and vertical wind shear, and verified against observations. Such comparisons will allow for an assessment of the performance of the models in these varying conditions.

By the end of the second year, a full set of verification statistics, based on the conventional and the newly-developed verification techniques, will be provided to TPC. In addition, the relationship between vertical shear, storm motion, and accumulated rainfall will be incorporated into a rainfall forecast model that is similar in concept to R-CLIPER. This model will be designed to run operationally at TPC.

References

Calheiros, R. V., and I. Zawadski, 1987: Reflectivity-rain rate relationships for radar hydrology in Brazil. *J. Climate and Appl. Meteor.*, **26**, 118-132.

DeMaria, M. D., and R. Tuleya, 2001:Evaluation of quantitative precipitation forecasts from the GFDL hurricane model. Reprints *Symposium on Precipitation Extremes: Predictions, Impacts, and Responses*, AMS, Albuquerque, NM, 340-343.

Ebert, E.E., U. Damrath, W. Wergen, and M.E. Baldwin, 2003: Supplement to The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Met. Soc.*, **84**, 10-11.

Marks, F.D., G. Kappler, and M. DeMaria, 2002: Development of a tropical cyclone rainfall climatology and persistence (R-CLIPER) model. Preprints, *25th Conference on Hurricanes and Tropical Meteorology,* AMS, San Diego, CA, 327-328.

Rosenfeld, D., D. B. Wolff, and D. Atlas, 1993: General probability relations between radar reflectivity and rain rate. *J. Appl. Meteor.*, **32**, 50-72.

Tustison, B. D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106**, 11,775-11,784.

Presentations resulting from this work

Marchok, T., R. Rogers, and R. Tuleya, 2004: Improving the validation and prediction of tropical cyclone rainfall. 58th Interdepartmental Hurricane Conference, Charleston, SC, March 1-5.

Marchok, T., R. Rogers, and R. Tuleya, 2004: A comparison of GFDL, GFS, and Eta rainfall forecasts for U.S. landfalling storms, 1998-2003. 26th Conference on Hurricanes and Tropical Meteorology, Miami Beach, FL, May 3-7.

Rogers, R., T. Marchok, and R. Tuleya, 2004: The development of a new validation technique for tropical cyclone rainfall. 26th Conference on Hurricanes and Tropical Meteorology, Miami Beach, FL, May 3-7.