9A.1 NEW METHODS FOR EVALUATING RAINFALL FORECASTS FROM OPERATIONAL MODELS FOR LANDFALLING TROPICAL CYCLONES

Timothy Marchok[1], Robert Rogers[2] and Robert Tuleya[3]

[1]NOAA / Geophysical Fluid Dynamics Laboratory, Princeton, NJ
[2]NOAA / AOML / Hurricane Research Division, Miami, FL
[3]SAIC at NOAA / NCEP / EMC, Old Dominion University, Norfolk, VA

1. INTRODUCTION

Over the past few decades, significant research has been conducted into improving tropical cyclone (TC) track and intensity forecasts while relatively little work has been done to improve tropical cyclone rainfall forecasts. This is at least partly due to a lack of rainfall forecast validation schemes designed specifically for landfalling tropical cyclones. We have addressed this issue by developing a set of validation techniques and objective skill indices specific for landfalling tropical cyclones that provide a baseline measure of quantitative precipitation forecasting (QPF) skill.

Rainfall in tropical cyclones is produced primarily through the evolution of convective features, with the heaviest rainfall occurring along the track of the storm near the intense core convection (e.g., Lonfat et al., 2004). This strong relationship between storm track and rainfall distribution can be exploited for the purpose of forecast validation, which is done in a post-mortem mode when the best track data are available. Indeed, several of the techniques developed in this study use track-relative methods that allow for model rainfall fields to be compared and validated against observations while reducing the impact of model track forecast errors on QPF skill statistics.

In this presentation, we describe a set of new TC QPF validation techniques. In addition, we describe the development of objective QPF skill indices based on these techniques that allow for objective comparison of TC QPF performance among dynamical models and against the benchmark R-CLIPER model. Finally, we apply these techniques and indices to the validation of QPF from operational National Weather Service models for landfalling U.S. TCs for the combined 1998-2004 Atlantic seasons as well as for the 2005 season.

2. VALIDATION TECHNIQUES

There are many features of a rainfall forecast that can be evaluated. The 72-h accumulated rainfall plots for Hurricane Jeanne (2004) shown in Figure 1 indicate wide variability among the observations and models in terms of volume, distribution and patterns of the rainfall.

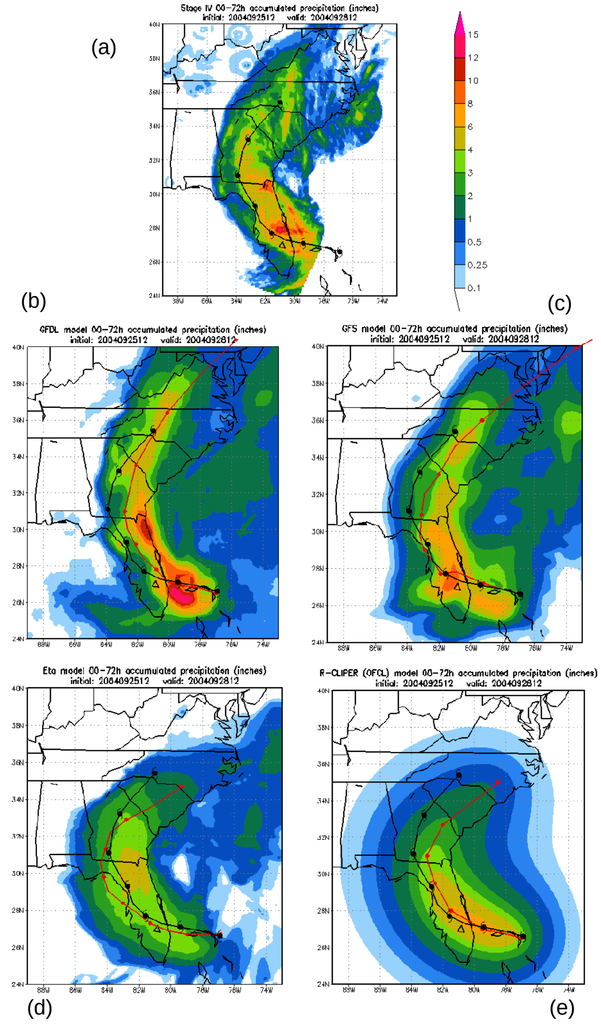

(a)
Stage IV 00-72h accumulated precipitation (inches)
initial: 2004092512 valid: 2004092812
15
12
10
8
6
4
3
2
1
0.5
0.25
0.1
(b)
GFDL model 00-72h accumulated precipitation (inches)
initial: 2004092512 valid: 2004092812
(c)
GFS model 00-72h accumulated precipitation (inches)
initial: 2004092512 valid: 2004092812
Eta model 00-72h accumulated precipitation (inches)
initial: 2004092512 valid: 2004092812
R-CLIPER (OFCL) model 00-72h accumulated precipitation (inches)
initial: 2004092512 valid: 2004092812
(d)
(e)

Figure 1. Plots of 72-h accumulated rain (shaded, in) from 12 UTC 25 to 12 UTC 28 September, 2004 for (a) Stage IV observations; (b) GFDL; (c) GFS; (d) NAM; (e) R-CLIPER. The observed track is shown in black; each model's forecast track is shown in red. R-CLIPER uses the NHC Official track.

* *Corresponding author address:* Timothy Marchok, NOAA / GFDL, Princeton, NJ. Email: timothy.marchok@noaa.gov

For example, while the GFDL and GFS models both produce the fairly large areas of moderate (4"+) rainfall, their axes of heaviest precipitation in Georgia are too far to the east, consistent with both models' eastward track forecast bias at that point in the forecast. The NAM and the R-CLIPER models both under produce the volume of rainfall that fell over land and fail to produce the extreme amounts that were observed in central and northeastern Florida.

In order to evaluate the unique characteristics of TC rainfall as shown in this example, we have developed validation techniques that address the following four critical aspects of TC rainfall forecasts: (1) A model's ability to match the large-scale rainfall pattern; (2) A model's ability to match the mean rainfall and the distribution of rain volume; (3) A model's ability to produce the extreme rainfall amounts often observed in TCs; and (4) The impact of a model's track forecast error on its QPF skill.

All validations are performed on storm-total rainfall accumulations out to a maximum lead time of 72 hours. For the 1998-2004 Atlantic seasons, a total of 35 landfalling storms were used, with one forecast included from each storm. The case selected for each storm is from the last 12 UTC cycle prior to landfall. A mask was used to exclude all non-CONUS data from all validations.

2.1 Pattern Matching

Two metrics commonly used in validations of QPF are used to evaluate the ability of models to reproduce rainfall patterns produced by the landfalling TCs: equitable threat score (ETS) and pattern correlation. For these analyses, the validation grid was restricted to include only those points within 600 km of the best track.

The ETS results in Figure 2 show that the GFS outperformed the other models at almost all rainfall thresholds for the 1998-2004 storm sample. The GFDL and NAM were similar for all amounts greater than 0.25 inches, and all three dynamical models outperformed R-CLIPER by a significant amount.

For the pattern correlations (not shown), there was wide case-to-case variability, but the GFS had the highest frequency of superior performance (highest correlation for 38% of the storms), while the GFDL and R-CLIPER had the lowest frequency of superior performance (highest correlation for 18% of the storms). The computed mean correlation coefficients over all cases are as follows: GFS (0.65), NAM(0.56), GFDL(0.50) and R-CLIPER (0.40).

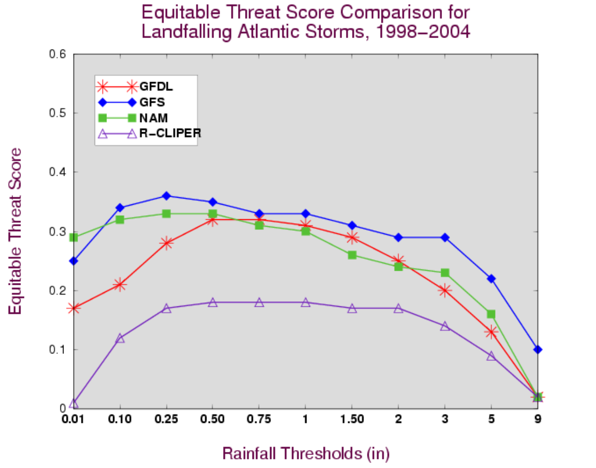


Figure 2. Plot of equitable threat score (ETS) for storm total rainfall for all models and all U.S. landfalling storms from 1998-2004.

2.2 Mean rainfall and rain volume distributions

Due to the threat of inland flooding, it is important to evaluate the ability of the models to forecast the volume of water that will fall over a given region. Figure 3 shows the mean storm-total forecasted rainfall in 20-km swaths centered on each model's forecasted storm track, as well as the observed rainfall centered on the best track.

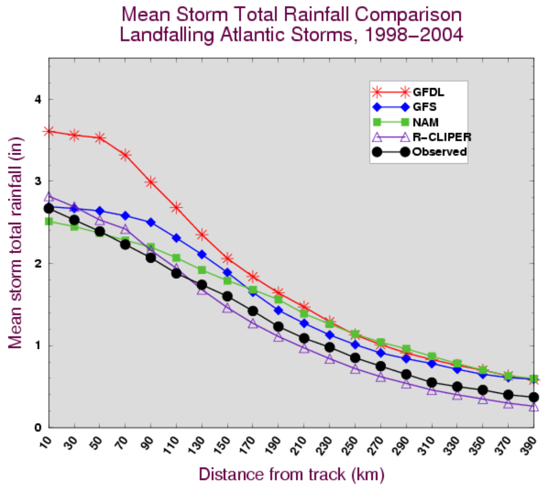


Figure 3. Radial distribution of mean storm rainfall (in) for all 1998-2004 storms for all models and observations, plotted as a function of across-track distance from the storm track.

Throughout most of the radial range, R-CLIPER closely approximates the observed mean rainfall totals, followed by the NAM and GFS. The GFDL has a pronounced over forecast bias, with mean rainfall totals about 40% higher than the observed mean within 150 km of the storm. An evaluation of total accumulated rain volume within 600 km of the best track confirms this bias in the GFDL, with a mean percent volume bias of 37.4% greater than the observations. The GFS also has a significant over-forecast bias (20.5%), while the NAM has a much lower bias (3.6%). The R-CLIPER, due to poor rain production at large radii, has a negative mean volume bias (-21.3%).

To further refine the analysis of rain volume distribution across a grid, a variable called rain flux is introduced. This is simply the product of the rainfall value at a grid point and the representative areal coverage of that point. In contrast to most standard QPF verification techniques which simply account for numerical occurrences of exceeding various thresholds, an analysis of rain flux provides a volume variable that can still be categorized by rainfall amount. For this reason, the rain flux values are kept in mixed units of in-km^2. Figure 4 shows PDFs of the rain flux for each of the models, using all points within 600 km of the best track. For the GFDL model, compared to the observations, a larger proportion of the total rain flux is accomplished at the high rain amounts (i.e., >6 inches). The inverse is true for the NAM and R-CLIPER models, i.e., a larger proportion of the rain flux is accomplished at the light-to-moderate rain amounts and a smaller proportion of the flux occurs for the heavy rain amounts.

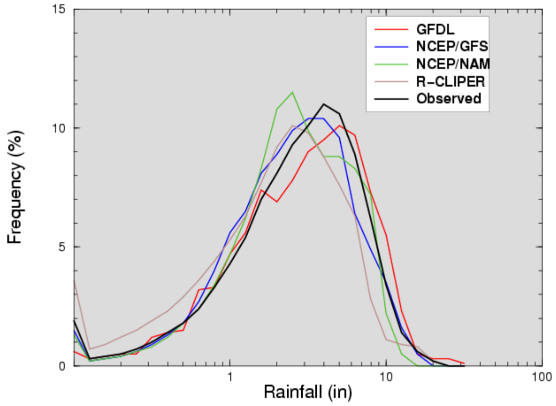


Figure 4. Probability distribution function (PDF) of rain flux for all 1998-2004 storms, using data within 600 km of best track for all models and observations.

From the PDFs of rain flux shown in Fig. 4, we can derive a cumulative distribution function (CDF) of rain flux for each of the models and the observations. From this CDF, we can compare the median value of rain flux among the various models, i.e., the point on the CDF at which 50% of the rain flux occurs in rain amounts greater than the indicated threshold rain amount. Figure 5 indicates that the 50th percentile occurs at 2.8 inches for the observations. The median for the GFDL is at a slightly higher value (3 inches), while the median for the NAM and GFS is at a lower value (2.2 inches) than the observations. The bias toward lighter amounts is most evident in R-CLIPER, where the 50th percentile is at 1.9 inches.

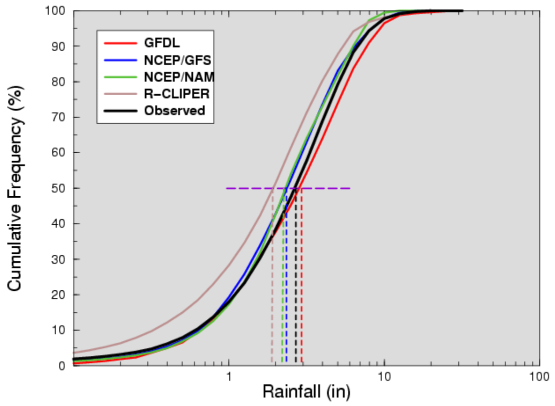
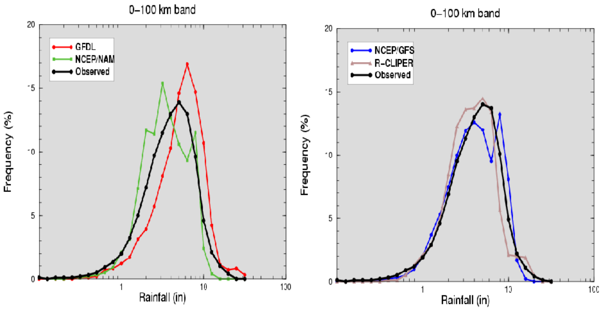


Figure 5. As in Fig. 4, but for Cumulative distribution function (CDF). Median (50%) level is indicated by the horizontal dashed line. Vertical dashed lines indicate, for each model and the observations, the rainfall threshold associated with the median rain flux value.

2.2.1 Track-relative analyses

In order to reduce the impact of a model's track forecast error on its QPF validation statistics, we can compare track-relative distributions of rain flux in bands surrounding the forecast and observed tracks. Distributions of rain flux are calculated within 100-km wide bands surrounding each model's forecast track and are compared against distributions of observed rain flux that are calculated within bands surrounding the best track. Figure 6a indicates that for the core region band (0-100 km), the GFDL has a pronounced bias towards producing too much rain flux in the high to extreme amounts, while the NAM produces too much rain flux in the light to moderate amounts.



(a) (b)

Figure 6. PDFs of rain flux for all models and observations for all 1998-2004 storms. (a) PDFs of rain flux within 0-100 km track-relative swath for GFDL, NAM and Stage IV; (b) As in (a), but for GFS, R-CLIPER and Stage IV.

The GFS core region PDF profile matches most of the observed profile well, but it underproduces rain flux in the extreme amounts (>10-15 in). The R-CLIPER core region profile offers the closest match to the observed, matching light, moderate, heavy and extreme amounts.

2.3 Extreme rain amounts

The ability of models to produce the extreme amounts that are often observed in TCs is another factor that is critical for inland flooding applications. Because these extreme amounts are highly localized, it is unrealistic to expect current models to accurately predict the location of such events. Rather, we focus our analysis on the overall probability distribution of these extreme amounts. Using the CDF analyses described in the previous section, we develop two techniques.

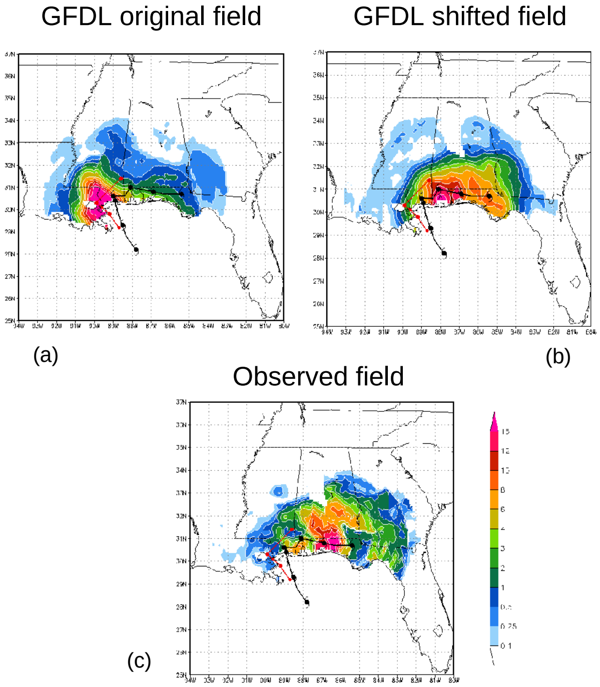
From the rain flux CDF shown in Fig. 5, we focus on the extreme end of this profile and determine how far the model-produced rain flux CDF curve deviates from the observed rainfall's 95th percentile. In Fig. 5, the 95th percentile in the observed rain flux distribution corresponds to a rainfall threshold of 8.3 inches, i.e., 5% of the observed rain flux is accomplished in amount thresholds greater than 8.3 inches. For the GFDL model, the 8.3 inch threshold falls at 92%, meaning that 8% of the GFDL rain flux occurs in values greater than 8.3 inches. Thus, more of the rain flux in the GFDL occurs in rain amounts greater than 8.3 inches, compared to the observations. By contrast, the 8.3 inch threshold for the NAM and R-CLIPER both fall at the 97-98% mark, meaning that a smaller proportion of their rain flux occurs at rain amounts above 8.3 inches. The 8.3 inch threshold for the GFS falls at 95%, exactly matching the observed value.

The second technique for extreme amounts employs the same technique just described, but instead of including all data points within 600 km of the best track, it utilizes the same breakdown into 100-km wide bands described above for examining the mean rainfall. This track-relative analysis is done for all 100-km wide bands out to 600 km from the storm track.

2.4 Sensitivity to track error

In previous sections, we have used track-relative analyses to compare rainfall data along two different storm tracks. Here we use a more direct approach and shift the model forecast rainfall fields in order to explicitly remove the impact of track forecast error on QPF performance.

Track-error removal is accomplished by shifting each 6-hour forecasted rainfall field by a distance equal to the difference in position of the forecasted vs. the observed storm location. The field of rainfall that is selected to be shifted includes only those grid points that are within 400 km of the location that is the midpoint between two successive 6-hourly forecast positions. These shifted 6-hourly rainfall fields are then summed over the lifetime of the storm, producing storm-total shifted rainfall analyses. Figure 7 shows an example of a shifted GFDL storm-total rainfall field for Hurricane Georges (1998). For this specific case, shifting the rainfall fields results in an increase in the correlation coefficient from 0.14 to 0.73, indicating a significant contribution of track error in this case.


GFDL original field

GFDL shifted field

(a)

(b)

Observed field

(c)

Figure 7. Example of storm-relative grid-shifted rainfall fields for GFDL forecast of Hurricane Georges (initial time of 12 UTC 27 Sep 1998) compared with observed fields. (a) Original GFDL 0-72h forecast of rainfall; (b) Shifted GFDL 0-72h forecasted rainfall; (c) Observed 0-72h rainfall. Amounts are in inches. GFDL forecast track is shown in red; The best track is shown in black.

Equitable threat scores are computed for all cases before and after the grid-shifting is done. The ETS comparisons shown in Figure 8 indicate that the unshifted scores for the GFS are the highest across all rainfall thresholds, while the R-CLIPER performs the worst. Once the fields are shifted, however, the GFDL and NAM models show comparable skill to the shifted GFS model over almost all rainfall thresholds. Similar improvements are noted in the pattern correlation coefficients (not shown). The GFS shows the least improvement from the grid-shifting and is the least sensitive to track error. This is likely at least partially due to the fact that the GFS had the lowest 48-h forecast track error for this sample of storms.

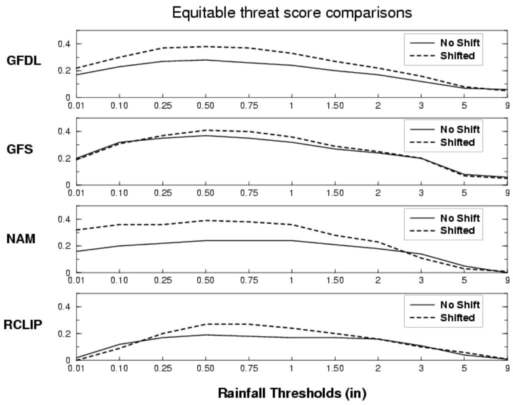


Figure 8. Comparison of ETS for all models before (solid line) and after (dashed line) performing grid shift for all 1998-2004 storms.

3. SKILL INDICES

The techniques outlined in Section 2 allow for evaluation of several different aspects of TC rainfall forecasts. In order to make evaluative statements on the relative performance of various models being compared, however, further synthesis of the results is required. This is especially critical for application to the validation of operational model forecasts, where an end-of-hurricane season analysis requires definitive statistics on model performance. Therefore, we develop skill indices that assess the performance of TC rainfall forecasts for these critical forecast attributes: (1) Pattern matching; (2) Mean rain and rain flux (volume); (3) Extreme rain; and (4) Impact of track error. Table 1 presents a summary of the various components that contribute to the skill indices. These indices rely upon algorithms that assign a value of 0 for no skill and 1 for most skill. The indices can be computed for all dynamical models as well as for the R-CLIPER model, thereby offering a means of evaluating skill relative to a benchmark climatological model. Details on the formulation for each algorithm will be presented in a forthcoming paper. Results (not shown) from the initial study of the 1998-2004 storms indicate superior performance by the GFS in 7 of the 9 indices shown.

		QPF attribute described			
Index		Pattern	Mean / Volume	Maximum	Impact of track error
Large-scale ETS		✓			
Pattern Correlation		✓			
Mean rainfall error index			✓		
Large-scale CDF median			✓		
Track-relative CDF median			✓		
Large-scale CDF (95th %)				✓	
Track-relative CDF (95th %)				✓	
Grid-shifted pattern correlation					✓
Grid-shifted ETS					✓

Table 1. Table summarizing individual TC QPF skill indices and the primary QPF attribute described.

Acknowledgments: This research was funded by a grant from the Joint Hurricane Testbed.

4. REFERENCES

Lonfat, M., F. Marks, Jr., S. Chen, 2004: Precipitation distribution in tropical cyclones using the Tropical Rainfall Measuring Mission (TRMM) microwave imager: A global perspective. *Mon Wea Rev*., **132**, 1645-1660