Validation Schemes for Tropical Cyclone Quantitative Precipitation Forecasts: Evaluation of Operational Models for U.S. Landfalling Cases

TIMOTHY MARCHOK

NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

ROBERT ROGERS

NOAA/AOML/Hurricane Research Division, Miami, Florida

ROBERT TULEYA

SAIC at NOAA/NWS/Environmental Modeling Center, Norfolk, Virginia

(Manuscript received 24 April 2006, in final form 21 December 2006)

ABSTRACT

A scheme for validating quantitative precipitation forecasts (QPFs) for landfalling tropical cyclones is developed and presented here. This scheme takes advantage of the unique characteristics of tropical cyclone rainfall by evaluating the skill of rainfall forecasts in three attributes: the ability to match observed rainfall patterns, the ability to match the mean value and volume of observed rainfall, and the ability to produce the extreme amounts often observed in tropical cyclones. For some of these characteristics, track-relative analyses are employed that help to reduce the impact of model track forecast error on QPF skill. These characteristics are evaluated for storm-total rainfall forecasts of all U.S. landfalling tropical cyclones from 1998 to 2004 by the NCEP operational models, that is, the Global Forecast System (GFS), the Geophysical Fluid Dynamics Laboratory (GFDL) hurricane model, and the North American Mesoscale (NAM) model, as well as the benchmark Rainfall Climatology and Persistence (R-CLIPER) model. Compared to R-CLIPER, all of the numerical models showed comparable or greater skill for all of the attributes. The GFS performed the best of all of the models for each of the categories. The GFDL had a bias of predicting too much heavy rain, especially in the core of the tropical cyclones, while the NAM predicted too little of the heavy rain. The R-CLIPER performed well near the track of the core, but it predicted much too little rain at large distances from the track. Whereas a primary determinant of tropical cyclone QPF errors is track forecast error, possible physical causes of track-relative differences lie with the physical parameterizations and initialization schemes for each of the models. This validation scheme can be used to identify model limitations and biases and guide future efforts toward model development and improvement.

1. Introduction

One of the most significant impacts of landfalling tropical cyclones (TCs) is the copious rainfall they often produce. Drowning from inland flooding in landfalling TCs was the leading cause of death from storms affecting the United States between 1970 and 2000 (Rappaport 2000). Such a significant impact highlights the importance of obtaining accurate rainfall forecasts for landfalling TCs. While significant improvements have been made in forecasts of TC track (Franklin et al. 2003; Aberson 2001) and, to a lesser extent, intensity (DeMaria and Gross 2003; DeMaria et al. 2005), much less attention has been focused on improving forecasts of rainfall (quantitative precipitation forecasting, or QPF) from TCs. Before TC rainfall forecasts can be improved, however, they must first be validated against observations to identify model limitations and biases and possible areas for improvement in the forecasts. Standard QPF validation techniques, such as bias and equitable threat scores, can assess some aspects of TC QPFs. However, an additional set of QPF validation techniques specific to TCs is needed in order to evalu-

Corresponding author address: Timothy Marchok, NOAA/Geophysical Fluid Dynamics Laboratory, 201 Forrestal Rd., Princeton, NJ 08540-6649.
E-mail: timothy.marchok@noaa.gov

DOI: 10.1175/WAF1024.1

© 2007 American Meteorological Society

ate the ability of the models to predict rainfall attributes unique to TCs, such as the extreme rain amounts so often responsible for the death and damage accompanying landfall.

Rainfall from a landfalling tropical cyclone is dependent on numerous storm-related and environmental factors. Tropical cyclone track is a significant determinant of the distribution of rainfall from the storm, with the heaviest rainfall occurring in a narrow swath close to the track of the storm (Lonfat et al. 2004). The translational speed of the storm plays an important role, both in creating azimuthal asymmetries in the rainfall field (Shapiro 1983) and in determining the duration of the rainfall. Another important determinant of TC rainfall is the presence of topography. The combination of strong winds, high moisture content, and sharp terrain gradients can create pronounced differences in rainfall on the windward and leeward sides of mountain slopes (e.g., Lin et al. 2001; Wu et al. 2002). The proximity of synoptic features, such as frontal boundaries and upper-level troughs (Bosart and Lackmann 1995; Atallah and Bosart 2003) can create significant bands of heavy rainfall at distances well removed from the TC. Vertical shear of the environmental wind can create asymmetries in the inner-core rainfall field that are related to the magnitude and direction of the shear vector (Bender 1997; Jones 2000; Frank and Ritchie 2001; Black et al. 2002; Corbosiero and Molinari 2002; Rogers et al. 2003; Lonfat et al. 2004). Finally, the intensity of the storm, the environmental humidity, and the properties of the underlying surface can impact the amount and distribution of rainfall received from a landfalling TC.

Tropical cyclone QPF techniques have been developed that account for various combinations of these factors. The simplest technique, which is known as Kraft's rule of thumb [attributed to R. H. Kraft by Pfost (2000)], consists of dividing a constant value by the translational speed of the storm to estimate the maximum rainfall that will be produced over a given location and time period traversed by the storm. While this technique accounts for the translational speed of the storm, it includes no information on the structure of the rainfall field. The Tropical Rainfall Potential (TRaP) method (Kidder et al. 2005) that was developed by the National Oceanic and Atmospheric Administration's (NOAA) Satellite Services Division (SSD) translates a satellite-estimated precipitation field to generate a 24-h rainfall accumulation. The Rainfall Climatology and Persistence (R-CLIPER) model is a climatology-based parametric model that has recently been developed (Marks et al. 2002; DeMaria and Tuleya 2001; Tuleya et al. 2007) to provide a benchmark against which forecasts of TC rainfall can be compared, similar to the way in which climatology and persistence-based CLIPER (Neumann 1972; Aberson 1998) and Statistical Hurricane Intensity Forecast (SHIFOR; Jarvinen and Neumann 1979; Knaff et al. 2003) predictions provide the benchmarks for track and intensity forecasts, respectively. The current operational version of the R-CLIPER, which is based on satellite-derived tropical cyclone rainfall observations (Marks et al. 2002), assumes a circularly symmetric distribution of rainfall and translates this distribution in time. It captures the dominant signals of translational speed and storm intensity, but it does not incorporate other processes that create asymmetries in the rain field. The most complex forecasting systems for producing TC QPFs are three-dimensional numerical models that produce spatially and temporally varying rainfall fields. The benefit of using numerical models is their ability to depict changes in the structure of tropical cyclones over time and how these changes are reflected in the rain field, both in a storm-relative sense and with accumulated rainfall swaths over a geographical area. Such models do suffer from deficiencies, related to resolution limitations and deficiencies in the representation of the initial state of the atmosphere and physical processes in the model. It is these deficiencies that can be identified by applying validation schemes specific for TC rainfall.

As an example of the varying abilities of numerical models to reproduce rainfall fields, storm-total rainfall fields for Hurricane Isabel (2003) produced by four models [i.e., Geophysical Fluid Dynamics Laboratory (GFDL), Global Forecast System (GFS), North American Mesoscale (NAM), and R-CLIPER] that have varying resolutions and complexities, are compared with observations in Fig. 1. All forecast and observed rainfall data in this figure have been interpolated onto a common 0.1° latitude–longitude grid. This case is highlighted here because the forecast track errors from the different model forecasts were relatively small. Therefore, it is likely that the differences in the distribution of rainfall are attributable to a variety of factors related to the handling of various physical processes by the models, and not just simply to differences in track forecasts. 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. Although the R-CLIPER model reproduces the general pattern of the rainfall, the amounts are smaller than observed and little of the structure in the rain field is predicted. The GFDL model predicts rain amounts and structures comparable to the observations, and the NAM and GFS models predict some structure to the rain field.

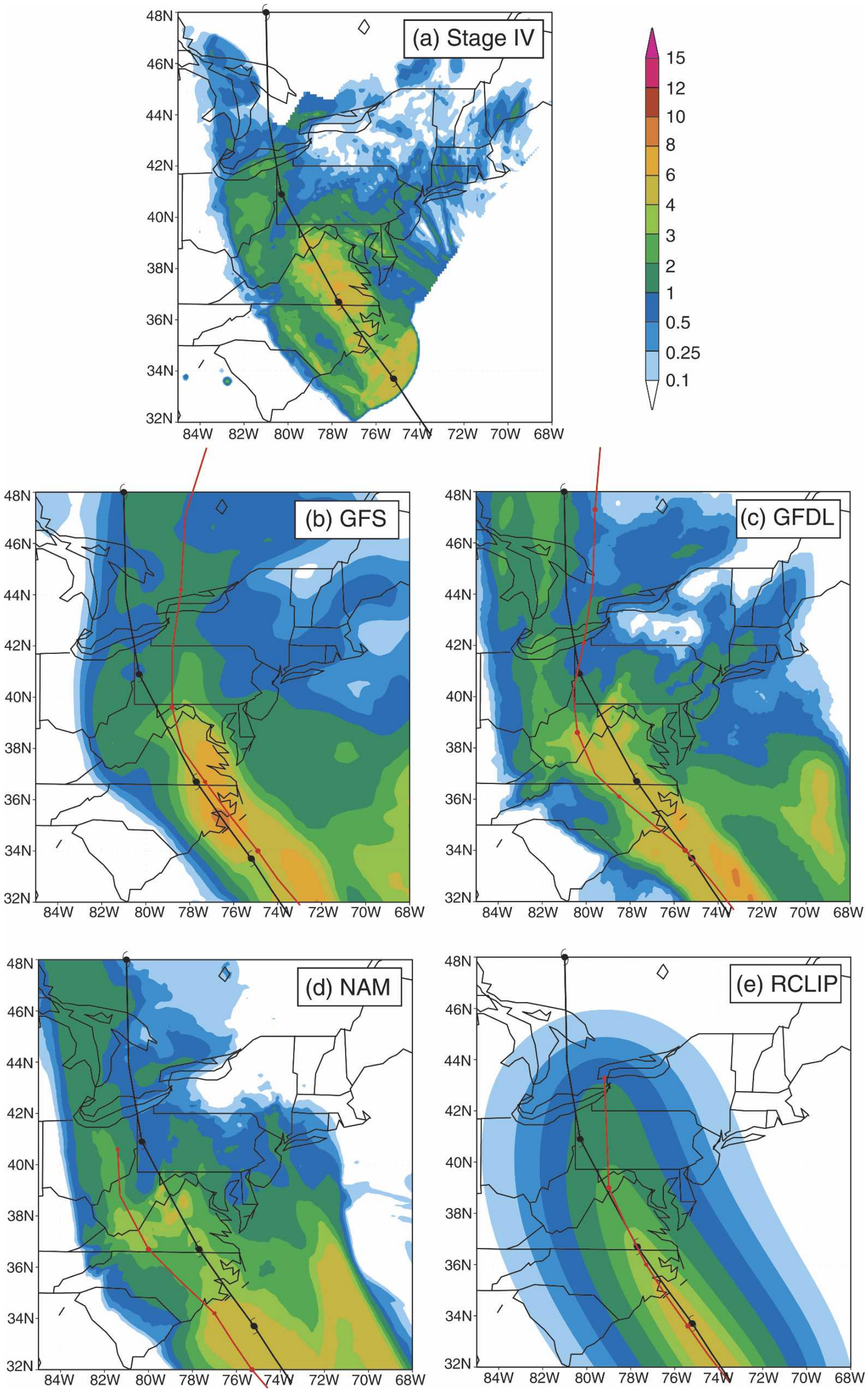


FIG. 1. Plot of 72-h accumulated rain (shaded, in.) from 1200 UTC 17 Sep to 1200 UTC 20 Sep 2003 for (a) stage-IV observations, (b) GFS, (c) GFDL, (d) NAM, and (e) R-CLIPER. The observed track is shown in black; each model's forecast track is shown in red.

Whereas the GFS predicts a larger area of maximum rain than was observed, the NAM predicts a smaller area of heavy rain. Farther inland, over Ohio and West Virginia, the GFDL and NAM models, and to a lesser extent the GFS model, predict a secondary axis of heavier rainfall to the left of the observed storm track that is consistent with the observations. However, the R-CLIPER produces only the main axis of heaviest rainfall that is aligned with the storm track.

This example illustrates many aspects of TC-

Table 1. Storms included in this study by year. Boldface indicates the storm was of hurricane intensity at landfall, while lightface indicates tropical depressions and tropical storms. All cases used begin at 1200 UTC on the date indicated. Numbers in parentheses indicate the observed maximum wind speed (kt) at landfall.

1998	1999	2000	2001	2002	2003	2004
Bonnie, 26 Aug (95)	**Bret, 22 Aug (100)**	Gordon, 17 Sep (55)	Allison, 5 Jun (45)	Bertha, 4 Aug (35)	Bill, 30 Jun (50)	Bonnie, 12 Aug (45)
Charley, 21 Aug (40)	Dennis, 4 Sep (60)	**Helene, 21 Sep (65)**	Barry, 5 Aug (60)	Edouard, 4 Sep (35)	**Claudette, 15 Jul (75)**	**Charley, 13 Aug (125)**
Earl, 2 Sep (70)	**Floyd, 15 Sep (90)**		Gabrielle, 13 Sep (60)	Fay, 6 Sep (50)	Grace, 31 Aug (35)	**Frances, 4 Sep (95)**
Frances, 10 Sep (45)	Harvey, 21 Sep (50)			Hanna, 14 Sep (45)	Henri, 5 Sep (30)	**Gaston, 29 Aug (65)**
Georges, 27 Sep (90)	**Irene, 15 Oct (70)**			Isidore, 25 Sep (55)	**Isabel, 17 Sep (90)**	**Ivan, 15 Sep (110)**
Hermine, 19 Sep (35)				Kyle, 11 Oct (35)		**Jeanne, 25 Sep (105)**
				Lili, 2 Oct (85)		Matthew, 9 Oct (40)

produced rainfall that are desirable to incorporate into a validation scheme. The TC circulation dynamically constrains convective development to storm-relative locations that may persist for time periods from hours to days. The ability of a model to reproduce observed rainfall fields is dependent on its ability to capture these dynamical features (e.g., eyewall, rainband, and stratiform rain) and to accurately predict the track and intensity of the storm as well as interactions with topography and other environmental features. A validation scheme that can target these regions of a storm and account for the different abilities of the models to predict track and intensity is required to identify model limitations and biases in the forecasts. Furthermore, much useful information can be obtained by considering the performance of the model forecasts for the entire distribution of rainfall in a statistical manner, not just peak rainfall amounts or point comparisons with specific rain gauges. Focusing on the statistical properties of rainfall distributions is particularly important when comparing models of varying resolution to observations based on comparatively small sampling areas such as from radar data or rain gauges, since a spatially averaged field always has lower variability than point values (Tustison et al. 2001) and cannot reproduce the extreme amounts and high-frequency signal from the point values. Until recently, many QPF validation schemes for tropical cyclones have been run on fixed geographical domains. Notable exceptions include Ebert et al. (2005) and Ferraro et al. (2005), who utilized the technique of Ebert and McBride (2000) to evaluate TRaP forecasts for landfalling TCs. With this technique, validations were performed on bounded regions of significant rainfall that were identified and matched in both the forecast and observed fields. Another example is in the study of Tuleya et al. (2007), who targeted the validation domains to areas close to the storm track. Such schemes narrow the focus of the validation to rainfall that is more directly linked with the storm and, thus, make the validation storm specific.

In this paper, our primary goal is to develop and test a QPF validation scheme specifically designed to objectively evaluate model rainfall forecasts for landfalling tropical cyclones. We test this scheme by performing validations of all U.S. landfalling TCs from 1998 to 2004 using the operational GFS, NAM, and GFDL hurricane models. The skill of these models is measured relative to the benchmark R-CLIPER forecasting scheme and validated against multisensor gridded rainfall observations available online. The validation scheme accounts for the varying abilities of the models to reproduce elements of the storm (e.g., structure, track, and intensity), compares the entire rainfall distribution rather than just the peak rainfall, considers the total volume of rainfall for some of its metrics, and focuses on storm-related rainfall.

2. Data and methodology

A total of 35 U.S. landfalling storms between 1998 and 2004 (Table 1) are studied, with a range in intensities from a tropical depression (Henri of 2003) to a category 4 strength hurricane (Charley of 2004) at landfall. One forecast from each storm is included in the database. To coincide with the storm database used by Tuleya et al. (2007), the initial times were always from the last 1200 UTC time within 24 h of landfall. Forecast and observed data for each storm were included in the database until advisories from the National Hurricane Center (NHC) were no longer issued for the storm. The

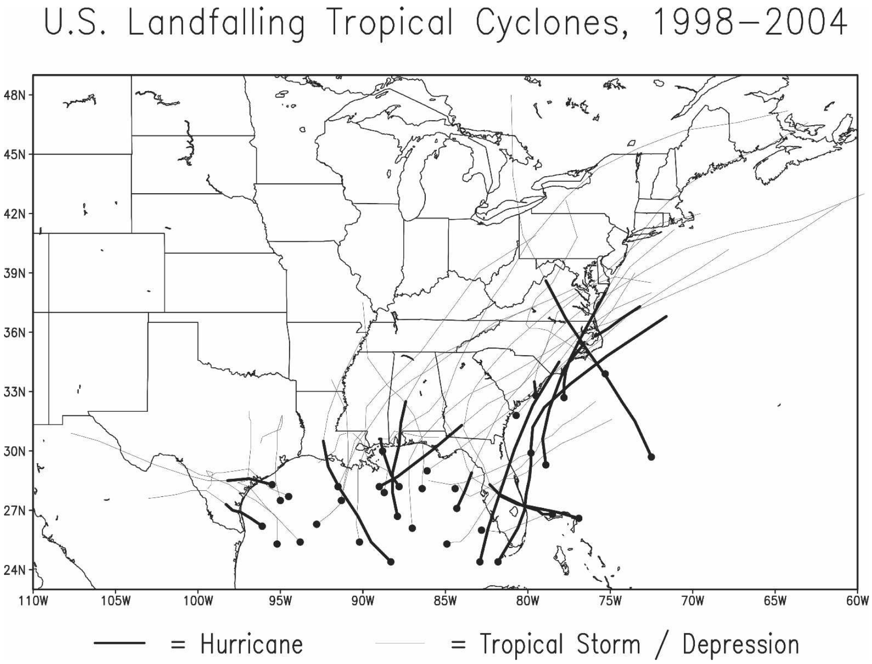

U.S. Landfalling Tropical Cyclones, 1998–2004
48N
45N
42N
39N
36N
33N
30N
27N
24N
110W
105W
100W
95W
90W
85W
80W
75W
70W
65W
60W
= Hurricane
= Tropical Storm / Depression

FIG. 2. Observed tracks for all tropical cyclones used in study. Dots indicate the starting positions for each forecast included in this study. Line thickness indicates storm intensity classification along the tracks.

tracks of the storms used in the study are shown in Fig. 2. While the storms took a variety of tracks over the Gulf and Atlantic coast states, the majority made landfall along the Gulf coast. Nevertheless, the storm tracks passed over a variety of topographies with different translational speeds and, thus, include a wide spectrum of conditions that produced many different rainfall distributions.

Rainfall observations were from the stage-IV hourly 4-km gridded rainfall data (stage II prior to 2002) provided by the Environmental Modeling Center (EMC) at the National Centers for Environmental Prediction (NCEP; Baldwin and Mitchell 1997). The 13 regional River Forecast Centers (RFCs) perform quality control on these data and then send them to EMC where they are combined into a unified analysis. These data consist of multisensor (i.e., rain gauges, radar) rainfall maps that cover the entire contiguous United States and are available on an hourly basis for all times back to 1998. In Tuleya et al. (2007), only rain gauges were used and hurricane-specific verifications were calculated at the gauge sites. In this paper, the more complete stage-IV multisensor rain observations were used and hurricane-specific verifications were calculated after interpolating the analyzed values to a 0.1° latitude–longitude grid. It should be noted that for proper radar estimation of rainfall from tropical systems, an adjustment to the reflectivity–rainfall rate (*Z–R*) factor is recommended (Fulton et al. 1998), and it is unclear how fully reliable the application of that adjustment is in practice.

The predictions used are from the real-time operational NCEP models used in forecasting hurricanes: the 2003 version of the GFDL hurricane model, the GFS, and the NAM model. The study was limited to these dynamical models due to the availability of archived synoptic forecast data from these models back to 1998. The GFDL model is a nested, hydrostatic regional model run with a minimum horizontal grid length of 1/6° (approximately 18 km), the GFS (previously known as the Aviation, or AVN, Model) is a global spectral model run at approximately 1/2° resolution (approximately 55 km), and the NAM (previously known as the Eta Model) is a limited-area model with a minimum grid length of 12 km. The climatology-based R-CLIPER model provides a benchmark to evaluate the skill of the dynamical models. The R-CLIPER is run at 4-km grid length and uses the NHC official forecast positions as a guide. For all of the analyses performed in this study, forecast data from the models were interpolated to the same 0.1° latitude–longitude grid as used for the observed stage-IV data. A common grid of such fine resolution was chosen in order to retain as much of the magnitude as possible of the extreme rainfall values frequently observed in TC rainfall data. The choice of resolution for the data and the verification grid box will have some impact on the validation statistics (Tustison et al. 2001; Gallus 2002). Tests were performed to determine the sensitivity of TC QPF validation techniques developed in this study to varying grid resolution, and results from these tests will be discussed below in section 3. It should be noted that the QPF validation techniques developed as part of this study are not restricted to use with only fine-resolution grids, and in fact they may be used with coarser-resolution data as well.

When calculating the rainfall forecast statistics, the areal domains of the forecasts and observations are restricted with a land–sea mask, and Canada and Mexico are excluded from the analyses. For most of the evaluations, only those areas within 600 km of the observed storm track are included. Some validations will be done in a track-relative manner (described below), and for these procedures, only data within 400 km of the forecast or observed track are included in the analyses. The purpose of these track-based domain restrictions is to limit the inclusion of rainfall that is not directly related to the tropical cyclone, such as rainfall from a frontal boundary or midlatitude cyclone that falls at radii well removed from the track of the TC.

All of the statistics in this study are for storm-total rainfall and include data up to a maximum lead time of 72 h, although the verification period for an individual case will be shorter if the storm dissipates or becomes extratropical prior to 72-h lead time. This use of storm-total rainfall, combined with the 600-km-track radius limit described above, means that all rainfall is included that falls anywhere within 600 km of the track, excluding the previously mentioned sea- and land-masked regions, from the beginning of the forecast through 72 h or the lead time at which the storm becomes nontropical.

Numerous observational and modeling studies have highlighted the importance of the extratropical transition (ET) process in modifying the structure of a tropical cyclone and its rainfall distribution (e.g., DiMego and Bosart 1982; Klein et al. 2000; Ritchie and Elsberry 2001; Atallah and Bosart 2003; Colle 2003). However, the databases used to develop the R-CLIPER model did not explicitly contain cases from nontropical systems (Tuleya et al. 2007), and since we are using R-CLIPER as our climatology benchmark model in this study for assessing the skill of the operational models for TC QPF, we will only perform validations up to the time at which the NHC discontinues its tracking of a storm as a tropical system. An interesting follow-up to the current study would be to extend the analysis to include storms that are undergoing extratropical transition.

3. Development of TC QPF metrics

In this section, we present a new set of techniques for validating TC rainfall forecasts as well as a set of metrics for objectively comparing operational numerical TC rainfall forecasts against one another and against the benchmark R-CLIPER forecasts. As discussed in the introduction, many aspects of TC rainfall forecasts should be compared to assess the skill of a particular forecast and identify possible model limitations and biases in the forecasts. For the Isabel forecasts in Fig. 1, some models showed a better ability to predict the overall pattern of rainfall, that is, the maximum along and to the right of the track and local maxima associated with the rainbands and topography, respectively, and local minima associated with topography (GFDL, NAM). Other models (e.g., R-CLIPER) were incapable of producing any such structures in the rainfall field. Some of the models also showed a better ability to produce the lighter rainfall amounts (GFS, GFDL, R-CLIPER), while others better produced the heavier rainfall amounts (GFS, GFDL). Rainfall amounts at the extreme end of the distribution [6–8 in. (152–203 mm) in this case] were better produced by some models (GFS and GFDL) and were not at all produced in others (NAM and R-CLIPER). Finally, the rainfall forecasts depend on the track of the storm predicted by the model (i.e., GFDL, GFS, and NAM) or provided by an external source (R-CLIPER). The GFDL and NAM predictions in Fig. 1 are well correlated with the observed fields despite track errors by both of those models after landfall. The GFS forecast track was very close to the observed track in the 12 h after landfall, and the forecasted rain field was well correlated with the observed rainfall.

The predictions in Fig. 1 illustrate three elements of TC rainfall forecasts that will be used as a basis for comparing the various models:

1) model ability to match the large-scale rainfall pattern,
2) model ability to match the mean rainfall and the distribution of rain volume, and
3) model ability to produce the extreme amounts often observed in TCs.

Methods for validating the forecasts that address each of these elements are described and presented in this section. In addition, since our goal is to use these new techniques to evaluate and compare the skill of operational TC rainfall forecasts, it is necessary to analyze results from the techniques using metrics that are as objective as possible. Since several of the techniques that will be described below are represented through

TABLE 2. Summary of individual TC QPF skill indices, whether they are dependent or independent of track error, and the primary QPF attribute described.

	Dependency on track error		QPF attribute described		
Index	Error dependent	Error independent	Pattern	Mean/volume	Max
Large-scale ETS	×		✓		
Pattern correlation	×		✓		
Mean rainfall error index		×		✓	
Large-scale CDF median value		×		✓	
Track-relative CDF median value		×		✓	
Large-scale CDF percentage at 95th percentile		×			✓
Track-relative CDF percentage at 95th percentile		×			✓

profiles and plots that can allow for some degree of subjective interpretation, comparisons of each new metric will be synthesized into a skill index to facilitate objective comparisons among the various models and against the benchmark R-CLIPER. The skill indices will rely upon algorithms that assign a value ranging from 0 for no skill to 1 for the most skill. Table 2 shows a list of the skill indices, which of the three elements they most directly address, and whether their values are predominantly track dependent or track independent. The formulation for each skill index algorithm is described in the appendix.

a. Pattern matching

For the pattern-matching techniques, two metrics that are commonly used in validations of QPFs are also used here to evaluate the ability of models to reproduce observed rainfall patterns produced by the landfalling TCs: equitable threat score (ETS; Schaefer 1990) and pattern correlation. The ETS is essentially the ratio of the number of forecast "hits" to the total number of forecast hits and misses, but it includes a "chance" factor to account for the number of hits that would be expected to occur purely due to random chance. This chance factor penalizes a model for erroneously overproducing rainfall amounts above a given threshold. Hits are defined as locations where the forecast rainfall amount matches or exceeds the observed rainfall amount for a given rainfall threshold. Pattern correlation is simply the correlation coefficient of the forecast rainfall and the observed rainfall at all grid points. Both the ETS and pattern correlation are dependent on the specific geographic location of the forecast and the observed amounts of rainfall and are thus sensitive to model track forecast errors.

The ETSs and the pattern correlations for the various models for the U.S. landfalling storms are provided in Fig. 3. For both analyses, all data points within 600 km of the best track are included. While the GFS model generally has the highest ETS across all rainfall thresholds, including the highest threshold of 9 in. (229 mm), the R-CLIPER model has the smallest ETS across all rainfall thresholds. The most significant ETS differences between the GFS and the other models occur at the low and high extremes of the rainfall distribution [i.e., <0.25 in. (6.4 mm) and >2 in. (51 mm)]. The cor-

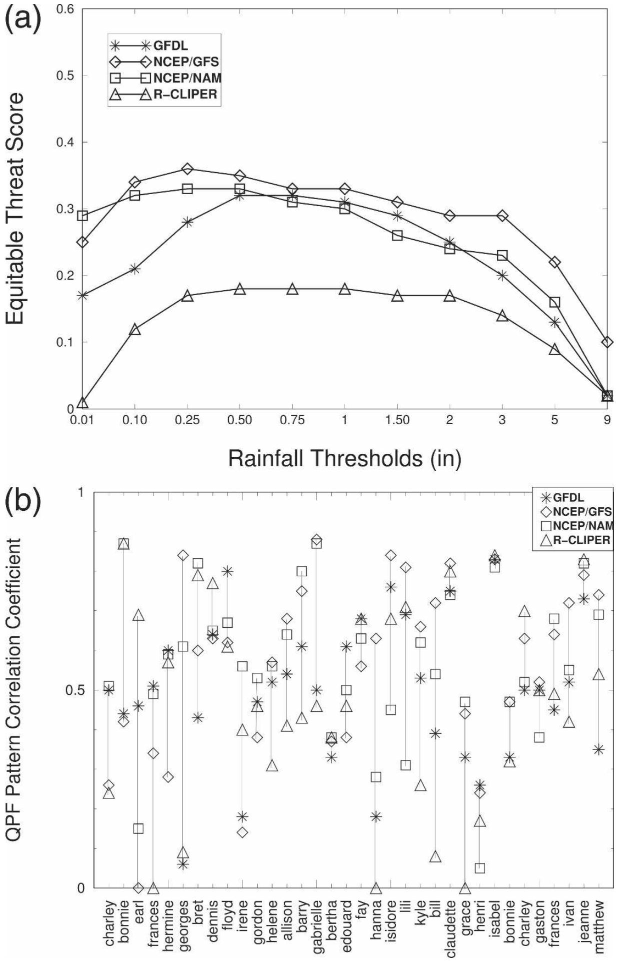


FIG. 3. (a) ETS and (b) QPF pattern correlations for storm-total rainfall for all models and all U.S. landfalling storms from 1998 to 2004.

relation statistics (Fig. 3b) show significant case-to-case variability in the performance of the models. In general, the GFS has the highest frequency of superior performance (highest r for 38% of the cases) for pattern correlations, while the GFDL and the R-CLIPER models have the lowest frequency of superior performance (highest r for 18% of the cases). Summed over all storms in this study, the average correlation coefficients (*r*) are as follows: GFS, 0.65; NAM, 0.56; GFDL, 0.50; and R-CLIPER, 0.40.

The pattern-matching verifications in this and the following sections are performed using data that have been interpolated to a grid with 0.1° latitude–longitude resolution. As described by Gallus (2002), the resolution of the verification grid box will have an impact on the equitable threat scores, with scores improving when the verification is performed on a coarser grid. We investigate this effect in the present study by interpolating all of the original forecast and observed data to a grid resolution of 0.5° latitude–longitude, or about the resolution of the coarsest model in the study (GFS), and then reevaluating the ETS and pattern correlations. The interpolations are performed using a budget interpolation, which is a nearest-neighbor averaging method that Accadia et al. (2003) have shown conserves total precipitation with more accuracy than bilinear interpolation.

The comparisons in Fig. 4a indicate a very small, but consistent, improvement in ETS for the GFDL, GFS, and NAM models when validating using the coarser-resolution data. For the heavy rain thresholds [5 in. (127 mm) and greater], the results for each of these models are nearly identical for the two different grid resolutions. The comparison of pattern correlation coefficients in Fig. 4b indicates similar incremental improvements for the GFDL, GFS, and NAM models when using the coarser-resolution data. The improvements accomplished by validating the coarser data from the GFDL, GFS, and NAM models are small and comparable in magnitude and direction among the three models; therefore, the change in verification grid resolution does not alter the conclusions derived from this pattern-matching segment of the verification. Finally, the improvements seen in ETS and pattern correlation for the R-CLIPER are negligible, likely due to the fact that the already very smooth fields of the R-CLIPER forecasts do not gain much of an advantage in the verifications by being interpolated to a coarser resolution.

b. Mean rainfall and rain flux distributions

The mean rainfall and the rain volume distributions provide useful indicators of the ability of the various models to produce all aspects of the rainfall distribution, that is, light, moderate, and heavy rain, and may better identify model limitations and biases in the forecasts than individual point comparisons with gauge measurements. Mean rainfall forecast storm totals in 20-km swaths centered on each model's forecasted storm track are compared in Fig. 5 with the mean observed rainfall centered on the best track. The observed rainfall profile for this sample of storms is similar to the radial distribution of mean tropical cyclone rain rates calculated from Tropical Rainfall Measuring Mission (TRMM) observations by Lonfat et al. (2004). Since similar TRMM observations from a global sample of TCs were the basis for the R-CLIPER technique, it is not surprising that the mean profile from the R-CLIPER forecasts is only slightly higher than observed (by <10%) in the innermost 90 km and slightly lower than observed (by ≤10%) outward from there. The

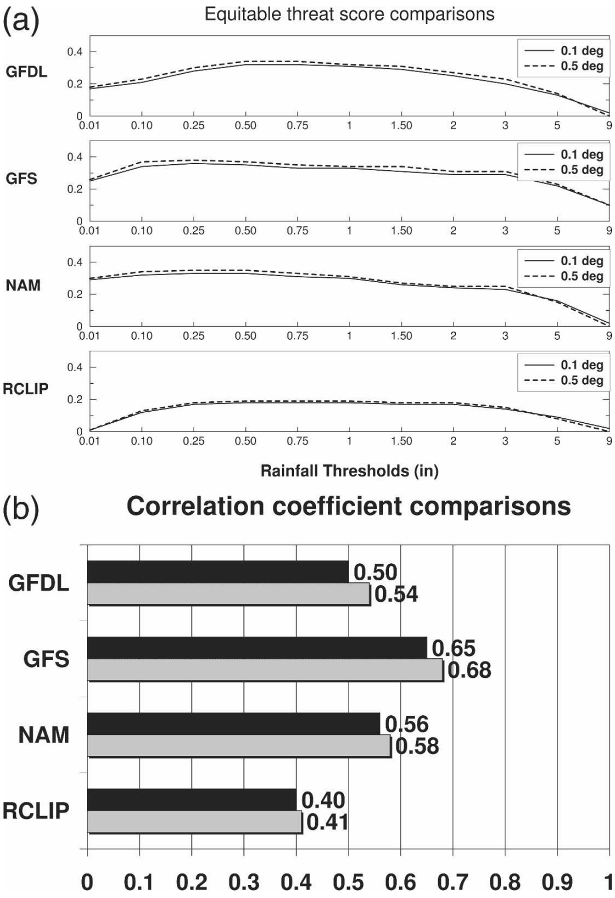


FIG. 4. (a) ETSs for storm-total rainfall for all storms in the study, comparing results using data interpolated to 0.1° grid resolution (solid line) with data interpolated to 0.5° grid resolution (dashed line). (b) Comparison of mean QPF pattern correlation coefficient for storm-total rainfall for all storms in the study, comparing results using data interpolated to 0.1° grid resolution (solid bar) with data interpolated to 0.5° grid resolution (dotted bar).

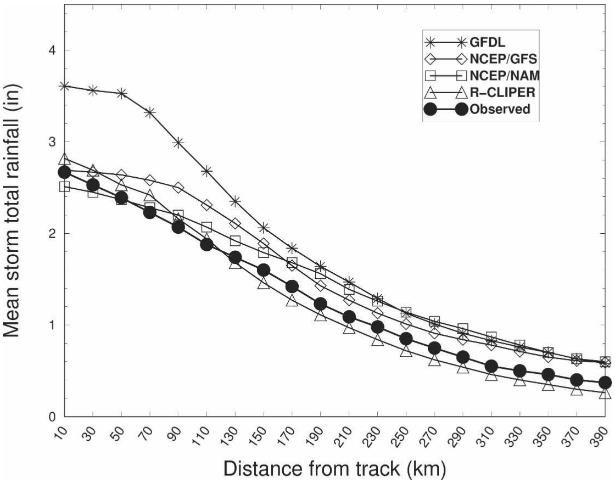


FIG. 5. Radial distribution of mean storm-total rainfall (in.), averaged over all storms in the study for all models and observations as a function of cross-track distance from the storm track.

TABLE 3. Rainfall volume statistics for all cases included in this study, computed over a domain that includes all points within 600 km of the best track.

	Stage IV	GFDL	GFS	NAM	R-CLIPER
Rainfall volume per case (km^3)	25.2	34.6	30.4	26.1	19.8
Mean rainfall bias (km^3)		9.4	5.2	0.9	−5.4
Rainfall bias (%)		37.4	20.5	3.6	−21.3

GFDL model has the largest bias in the radial distribution of rainfall, particularly in the regions closest to the path of the storm where the mean rainfall totals are about 40% higher than the observed mean. The GFDL rainfall profile remains higher than the observations out to 390 km. A similar comparison, although less pronounced, is seen with the GFS model, with mean rainfall about 10%–20% higher than the observations within 150 km of the center. Although the NAM model produces mean rainfall totals that are slightly lower than the observations within 50 km of the track, the predicted amounts are higher than the observations from 50 km outward.

Because of the threat of inland flooding from tropical cyclone rainfall, it is important to evaluate how well the models can forecast the volume of water that will fall over a given region. Rainfall volume statistics computed over a domain restricted to all points within 600 km of the best track are compared in Table 3. All three dynamical models (GFDL, GFS, NAM) have a positive bias compared to the observed rainfall, while the R-CLIPER model has a pronounced negative bias. The GFDL has the largest positive bias, while the NAM has a much smaller bias in that the mean volume per case is quite close to the observed rainfall volume.

To facilitate comparisons between the observations and the models of how the volume of water is distributed across the analysis grid, a variable called rain flux is used. This flux is simply the product of the rainfall value at a grid point and the representative areal coverage of that point (units of in. km^2). This calculation is made for two primary reasons. First, it can account for the dependency of rainfall volume on the resolution of the model grid. For example, an inch (25.4 mm) of rainfall produced at a grid point in the GFS model, which represents a roughly 55 km × 55 km (1/2° × 1/2°) area, is a much greater total water amount than an inch of rainfall produced at a grid point in the GFDL model, which represents a roughly 18 km × 18 km (1/6° × 1/6°) area. To facilitate consistent comparisons between output files, however, the rainfall values from all models and observations in this study are interpolated to a latitude–longitude grid with a fixed resolution of 0.1°. While this masks some of the impacts of the varying resolution described above, the impact of varying resolution is still accounted for indirectly. The second reason for using rain flux as a variable is that the rain flux is plotted as a function of rainfall amount. This is in contrast to many standard precipitation verification techniques, which simply account for the number of occurrences of exceeding various rainfall amount thresholds, but do not factor in the volume of water when evaluating QPFs (e.g., the bias score). The rain flux values are kept in mixed units of in. km^2 in order to facilitate categorizing them based on the intensity of rainfall within each grid box.

Probability distribution functions (PDFs) of rain flux for each of the models are compared in Fig. 6a for all of the storms, using all points within 600 km of the observed storm track. This figure shows the comparison of how rain flux is distributed by rain amount for each of the models. Because rainfall intensity is nearly logarithmically distributed, the rain flux values are categorized into 27 thresholds that are defined by using the following relationship for the decibel rain rate (dBR):

$$\mathrm{dBR} = 10 \log(R), \qquad (1)$$

where R is the value of the rain flux threshold and dBR is in the range of $\{-30, -10, -9, -8, -7, \ldots, 13, 14, 15\}$. These rain flux thresholds provide for a broader range of rainfall intensities than in the equitable threat score analysis in Fig. 3, especially for the heavy to extreme amounts.

Compared to the observations, a larger proportion of

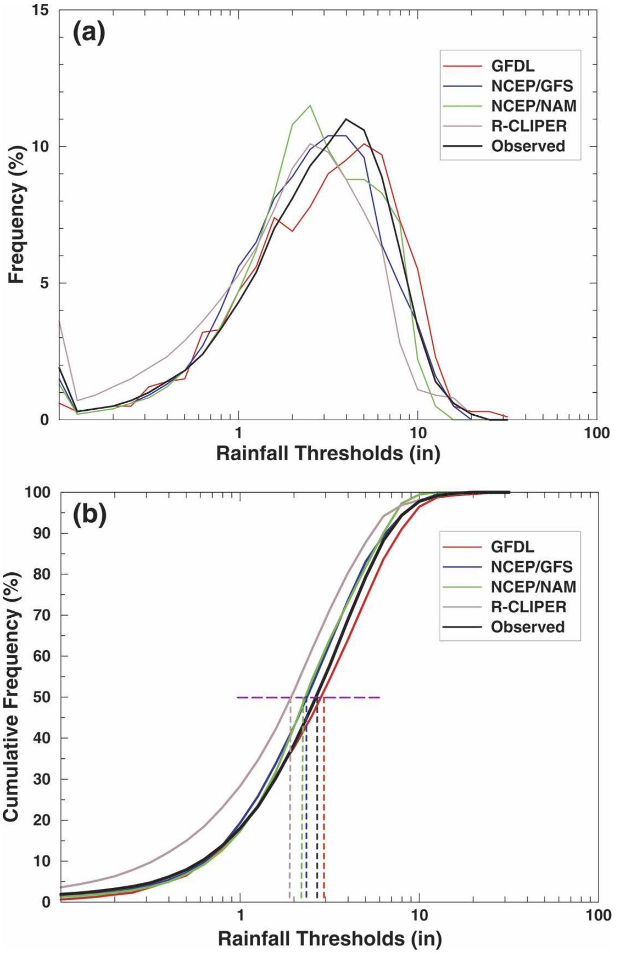


FIG. 6. (a) PDF of rain flux within 600 km of the observed storm track for all storms in this study for all models and observations. (b) As in (a), but for the CDF. The median (50%) level is indicated by the horizontal dashed line. Vertical dashed lines indicate, for each model and the observations, the rainfall threshold associated with the median rain flux value.

the total rain flux for the GFDL model in Fig. 6a is accomplished at the high rain amounts [i.e., values larger than 6 in. (152 mm)], while a smaller portion of the rain flux occurs in the light-to-moderate rain amounts [i.e., 1–3 in. (25–76 mm)]. The inverse is true for the NAM and R-CLIPER models; that is, 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. For the GFS model there is a slight overrepresentation of rain flux for the light-to-moderate rain amounts, and the rain flux in the extreme rain amounts [>10 in. (254 mm)] compares very well with the observations.

Cumulative distribution functions (CDFs) of rain flux, derived directly from the PDFs shown in Fig. 6a, are shown in Fig. 6b. The median value of the rain flux represents the point on the CDF at which 50% of the rain flux occurs in rain amounts greater than the indicated threshold rain amount. For the observations, the 50th percentile occurs at 2.8 in. (71 mm). For the GFDL model, the 50th percentile occurs at a slightly higher value [i.e., 3 in. (76 mm)], which indicates that a slightly smaller proportion of rain flux is occurring in the light-to-moderate rain amounts than was observed. For the NAM and GFS, the 50th percentile is at a smaller value [2.2 in. (55.9 mm)] than the observations, indicating that a larger proportion of rain flux is occurring in these light-to-moderate values compared to the observations. This bias toward lighter rain amounts is most evident in the R-CLIPER, where the 50th percentile rain flux is at 1.9 in. (48.3 mm).

By comparing track-relative distributions of rain flux in bands surrounding the forecast and observed tracks, we can reduce the impact of track forecast errors on QPF validation statistics. An example from Hurricane Isidore (2002) illustrates the setup of these bands (Fig. 7a). Distributions of model forecast rain flux are calculated within 100-km-wide bands surrounding each model's forecast track and are compared against distributions of observed rain flux calculated within bands surrounding the best track (Fig. 7b). The innermost 100-km band focuses on rain within the core of the storm, while the outer bands correspond to rain in the outer rainbands and stratiform areas. The distributions shown for each of the bands in Fig. 7b are approximately lognormal, with the modal values of the observed rain flux (peak in the distribution) occurring at 5.5 in. (140 mm) for the inner-core band, 4 in. (102 mm) for the 200–300-km band, and 1–2 in. (25.4–50.8 mm) for the 400–500-km band.

The PDF of rain flux for the GFDL, NAM, and observed rainfall fields are shown in Fig. 8a for the 0–100-km band around the storm track, where rainfall from the eyewall (or eyewall remnants) would tend to predominate. The GFDL has a clear tendency to produce too much rain flux in the high-to-extreme rain amounts, while the NAM produces too much rain flux in the light-to-moderate rain amounts, which suggests that the GFDL tends to overpredict eyewall rain while the NAM tends to underpredict eyewall rain, which is consistent with the results shown in Fig. 5. The GFS and R-CLIPER comparisons in this swath (Fig. 8b) show that the GFS slightly overproduces rain flux for the moderate-to-heavy rain range [<10 in. (254 mm)], but it underproduces rain flux for the extreme rain amounts [>10–15 in. (254–381 mm)]. The R-CLIPER has the closest resemblance to the observed flux distributions in the inner core (Fig. 8b), showing the ability to pro-

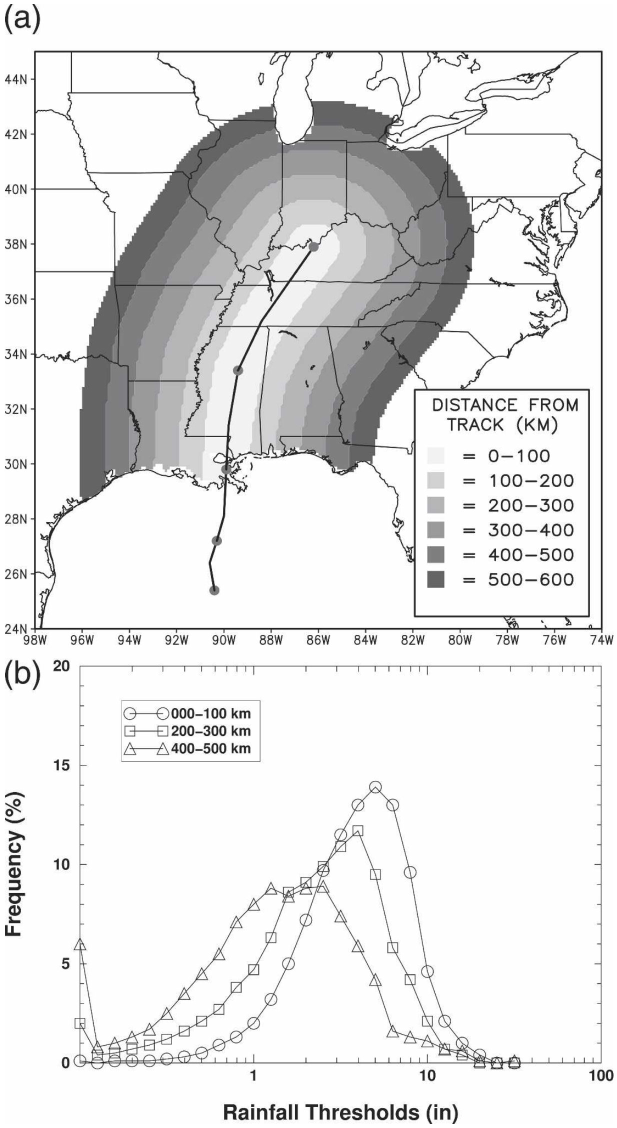


FIG. 7. (a) Schematic of 100-km-wide bands surrounding observed track for Hurricane Isidore (2002) within which track-relative rain flux PDFs are calculated. Storm-track positions are marked every 12 h, beginning 1200 UTC 25 Sep 2002. (b) Rain flux PDF from stage-IV data averaged over all cases in this study for the 0–100-, 200–300-, and 400–500-km bands.

duce rain flux that matches the observed for light, moderate, heavy, and extreme rain amounts.

The corresponding rain flux PDFs are shown in Figs. 8c and 8d for the 300–400-km swath, where a mixture of rainband and stratiform rain would likely predominate. While the rain flux PDF from the GFDL model (Fig. 8c) agrees well with the observations in this swath, the NAM model has a tendency to produce peak rain flux values in higher rain amounts than the observed distribution. The GFS (Fig. 8d) shows a slight tendency to overpredict rain flux for the light-to-moderate amounts and underpredict rain flux for the heavy rain amounts. Whereas the R-CLIPER 0–100-km rain flux PDF agrees very well with the observations, the 300–400-km rain flux PDF is significantly skewed toward lighter rain rates. That is, the R-CLIPER significantly overproduces rain flux for the light rain amounts in the 300–400-km swath and significantly underproduces rain flux for the moderate-to-heavy rain amounts. A likely explanation is that R-CLIPER is based on azimuthally averaged rainfall amounts that at outer radii are often composed of a few relatively large amounts and other areas of little or no rain. At these radii, the mean rainfall rate produced by R-CLIPER is probably a poor estimate of the rainfall at any given point.

This example of R-CLIPER highlights one of the advantages of this new validation scheme: by examining the data from a variety of different perspectives, model deficiencies and biases in the forecasts can be isolated. The analysis of mean rainfall rate (cf. Fig. 5) indicated that the R-CLIPER does an excellent job of approximating the mean rainfall rate out to large radii, and in fact this matching of the observed mean rainfall rate was a specific design consideration for R-CLIPER (Tuleya et al. 2007). However, in reality at large radii the rainfall from tropical cyclones is largely determined by rainbands that produce strong asymmetries (rainfall maxima) over only relatively small regions of the rain field. The result in this case is that the R-CLIPER produces unrealistically large areas of small rainfall amounts due to its assumption of an azimuthally symmetric distribution. Thus, the profile of the R-CLIPER rain flux in the 300–400-km band (Fig. 8d) is characterized by an overabundance of rain flux in small rainfall thresholds and a lack of rain flux in the moderate-to-heavy thresholds. This same line of reasoning helps to explain why the R-CLIPER performs so poorly in ETS compared to the dynamical models (cf. Fig. 3), despite having a smooth field of rainfall that would normally be considered favorable for the equitable threat score diagnostic. In this example, while some methods in this validation scheme help to isolate highlights of the R-CLIPER forecasts (i.e., mean rainfall rate), other methods (ETS) help point to a problem, and still other methods (track-relative profiles of rain flux in outer radial bands) help to isolate the nature of the problem.

c. Extreme rain amounts

It is also important to evaluate how well each model produces the extreme rain events. Two evaluation techniques are developed for this attribute. The first technique compares the rain flux CDF (cf. Fig. 6b) for the observed rainfall within 600 km of the best track against that of each model and determines how far the model-

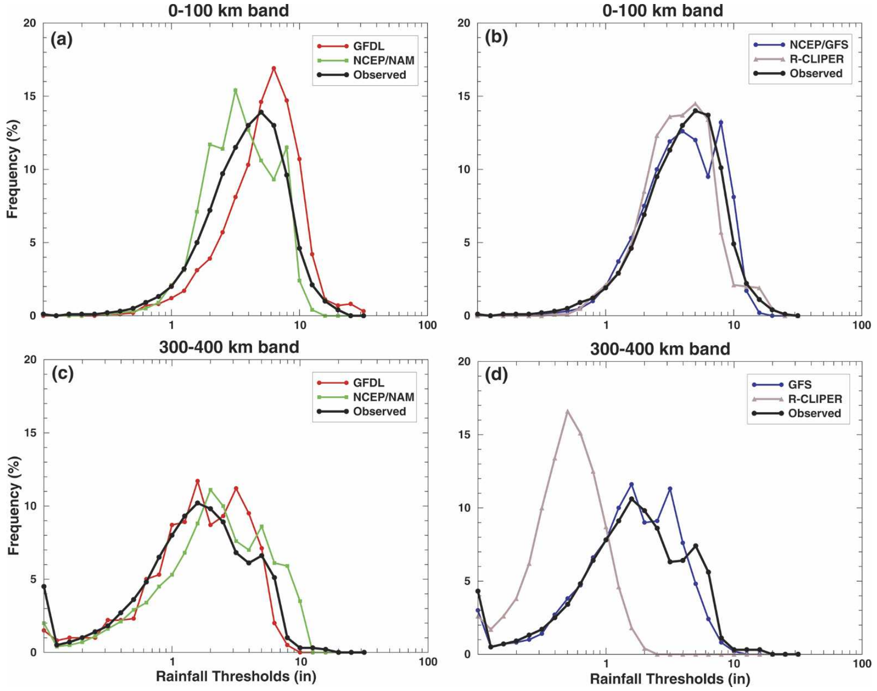
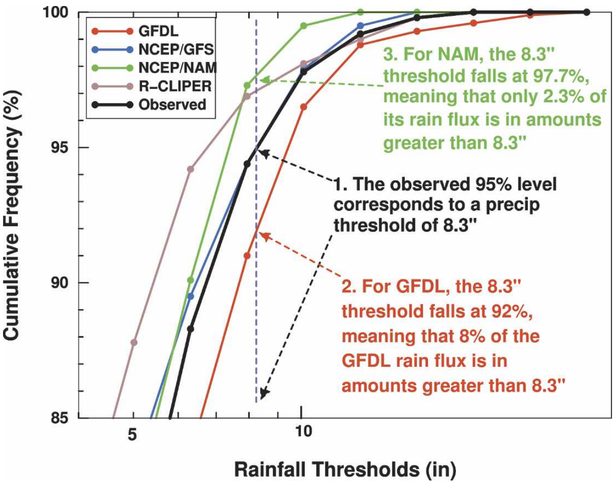


FIG. 8. PDFs of rain flux for all models and observations for all storms in this study. (a) PDFs of rain flux within 0–100-km track-relative swath for GFDL, NAM, and stage IV. (b) As in (a) but for GFS, R-CLIPER, and stage IV. (c) PDFs of rain flux within 300–400-km track-relative swath for GFDL, NAM, and stage IV. (d) As in (c) but for GFS, R-CLIPER, and stage IV.

produced CDF curve deviates from the observed rainfall's 95th percentile. For the storms in this study (Fig. 9), the 95th percentile in the observed rain flux distribution corresponds to a rainfall threshold of 8.3 in. (211 mm). For the GFDL model, the 8.3-in. threshold falls at 92%, which means that 8% of the rain flux occurs in values greater than 8.3 in. (compared with 5% from the observations). Thus, more of the rain flux predicted by the GFDL model is in rain amounts greater than 8.3 in., which is consistent with the comparisons shown above (cf. Figs. 6 and 8). By contrast, the 8.3-in. threshold for the NAM and R-CLIPER models both fall at the 97%–98% mark, which means that a too small fraction of the rain flux occurs at rain amounts above 8.3 in. The 8.3-in. threshold for the GFS falls at 95%, which exactly matches the observed value.

A similar comparison of the rain flux distributions for the extreme rain amounts is made for 100-km-wide bands surrounding the observed and forecast tracks (cf. Fig. 7a). The 95th percentile for the observed rain flux distribution in the 0–100-km band corresponds to a rainfall threshold of 9.3 in. (Fig. 10a). For these extreme amounts, the R-CLIPER and GFS rain flux CDF curves are close to the observed 9.3-in. (236 mm) threshold (95% and 96%, respectively). The 9.3-in. threshold for the GFDL model falls near the 90% mark, which again indicates that proportionately too much of the GFDL rain flux in the core region is occurring at these extreme rain amounts. By contrast, the



3. For NAM, the 8.3" threshold falls at 97.7%, meaning that only 2.3% of its rain flux is in amounts greater than 8.3"

1. The observed 95% level corresponds to a precip threshold of 8.3"

2. For GFDL, the 8.3" threshold falls at 92%, meaning that 8% of the GFDL rain flux is in amounts greater than 8.3"

FIG. 9. Top 15% of CDFs of rain flux within 600 km of best track for all models and observations for all storms in this study. Positioning of the vertical dashed line that intersects the observed profile indicates the rainfall threshold matching the 95th percentile level for observed rain flux.

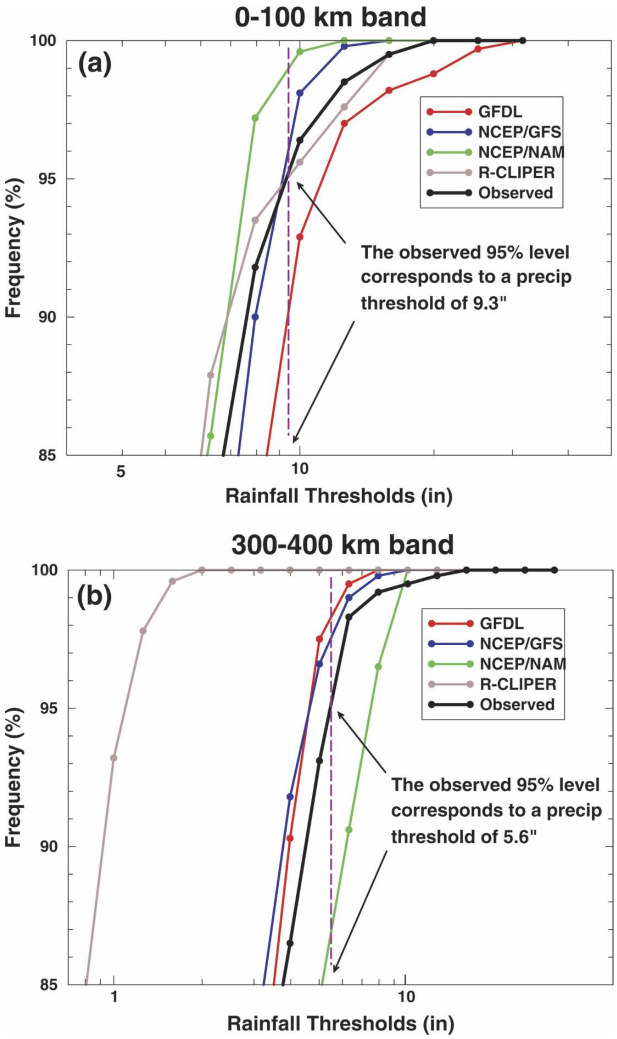


FIG. 10. Similar to Fig. 9 except that data are restricted to the track-relative (a) 0–100- and (b) 300–400-km bands surrounding respective model forecast and observed storm tracks.

NAM model rain flux CDF curve crosses the 9.3-in. threshold at 98%, which suggests a deficiency in predicting extreme rain amounts in the 0–100-km band.

For the 300–400-km band (Fig. 10b), the 95th percentile for the observed rain flux distribution corresponds to a rainfall threshold of 5.6 in. (142 mm). Compared to the observed rain flux CDF, the GFDL and GFS models both have proportionately too little rain flux in the extreme amounts in this 300–400-km band, with the CDF curves for both models crossing the 5.6-in. rainfall threshold at the 97%–98% mark. The CDF curve for R-CLIPER indicates this model has no skill in predicting extreme rain amounts in these regions well removed from the storm track. The CDF curve for the NAM model crosses the 5.6-in. threshold at 87%, which indicates that the NAM model produces a disproportionately large amount of rain flux at these high rain thresholds compared to observations. This characteristic in the outer regions contrasts with the NAM model's tendency to underpredict extreme rain amounts in the core region.

As described above for the pattern-matching techniques, tests were performed to determine the sensitivity of both of these volume-related rain flux distribution and extreme amount techniques to grid resolution. Similar to the results from the grid resolution sensitivity tests for the pattern-matching techniques, plots of the PDF of rain flux for the various models (not shown) indicated only insignificant differences between analyses done using data interpolated to 0.1° and 0.5°. The lack of differences, both for the pattern-matching techniques and for the rain-volume related techniques, is likely due to the fact that tropical cyclones dynamically constrain rainfall to storm-relative locations that can persist for several hours or longer, as mentioned in the introduction. This constraint typically produces patterns of rainfall that are significantly smoother and more uniform than those from, for example, continental summertime convection, and it is one feature that distinguishes tropical cyclone rainfall from these other rain-producing events. Further smoothing of the rain field also occurs as we evaluate rainfall accumulated over a period of 72 h.

4. Evaluation of TC QPF skill indices for recent U.S. landfalling storms

Techniques presented in the previous section describe various methods for validating TC QPFs in three main elements: the ability to match observed rainfall patterns, the ability to match the mean value and volume of observed rainfall, and the ability to produce the extreme amounts often observed in tropical cyclones. In addition, QPF skill indices based on these techniques are outlined in Table 4 and specific details of the formulations for the skill indices are provided in the appendix. These skill indices may be useful in validations of operational forecasts, as they allow for objective comparisons of QPF skill among the various numerical models and against the benchmark R-CLIPER. In this section, we provide values for the QPF skill indices for the storms in the 1998–2004 sample as well as discussion of some of the highlights. Values for each of the QPF skill indices are shown in Table 4. The QPF skill indices in the appendix are formulated so that the numbers closest to 1 (0) indicate the most (least) skill for that index. The values for the indices within each of the three attributes shown in Table 4 are combined into one summary comparison for each attribute and pre-

TABLE 4. Value of each of the TC QPF skill indices for each model. A value of 0 indicates no skill; a value of 1 indicates the most skill. The most skillful score for each metric is set in boldface.

Index	QPF attribute described	GFDL	GFS	NAM	R-CLIPER
Large-scale ETS	Pattern	0.42	**0.54**	0.44	0.27
Pattern correlation	Pattern	0.50	**0.65**	0.56	0.40
Mean rainfall error index	Mean/volume	0.80	0.93	0.94	**0.95**
Large-scale CDF median value	Mean/volume	**0.83**	0.71	0.65	0.23
Track-relative CDF median value	Mean/volume	0.58	**0.82**	0.65	0.17
Large-scale CDF percentage in 95th percentile	Max value	0.90	**1.00**	0.85	0.91
Track-relative CDF percentage in 95th percentile	Max value	0.80	**0.93**	0.71	0.66

sented in Fig. 11. As can be seen from both Fig. 11 and Table 4, the GFS performs the best of the models for all three categories of QPF attributes listed in column 2 of Table 4. All of the numerical models (GFS, GFDL, and NAM) show skill relative to R-CLIPER for all of the TC QPF attributes, with the exception of the NAM for the extreme rain skill index.

a. Pattern matching

For the pattern matching (Fig. 11), the GFS has the highest skill, although all of the numerical models (GFS, GFDL, NAM) have skill relative to R-CLIPER. The GFDL has the lowest skill among the dynamical models for pattern matching. The NAM model scores better than the GFDL in this metric due to the higher mean correlation coefficient for the NAM model described above. In addition, the NAM has a higher ETS than the GFDL for lighter rain rates (cf. Fig. 3a). It is not clear why the NAM model performs better than the GFDL, especially for the lighter rain rates. Since the pattern-matching metric is highly dependent on track error, the comparative skill of the GFS for this metric is likely largely attributable to the fact that the GFS had higher track forecast skill for the landfalling TCs analyzed in this study, particularly at the 48-h lead time (Fig. 12).

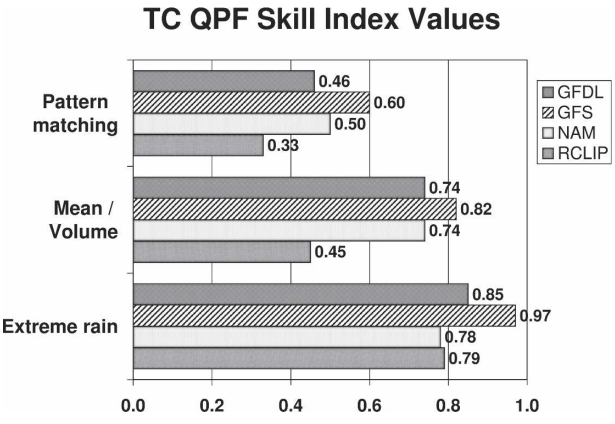


FIG. 11. Combined TC QPF skill indices from different models for all cases in this study showing degree of skill in the three TC QPF attributes of pattern matching, mean value and distribution of rainfall volume, and extreme rainfall. Scores range from 0 (no skill) to 1 (most skill).

One possible reason for why the R-CLIPER performed so poorly in this metric is that the R-CLIPER model is based on azimuthally averaged rainfall amounts. In reality, at outer radii these averages are composed of a few relatively large amounts and other areas of little or no rain. Thus, the mean rainfall rate at these distances is probably a poor estimate of the rainfall.

b. Mean rain and distribution of rain flux

For the mean rainfall and distributions of rain flux (Fig. 11), the skill indices for all of the dynamical models were very similar, but the GFS had a slight edge. All of the models have skill relative to the R-CLIPER. While the combined index for volume and distribution shown in Fig. 11 indicates a near equivalence among the different dynamical models, detailed comparisons of the entire distribution and track-relative locations (cf. Figs. 5, 6, and 8) show marked differences. The GFDL (NAM) produced too much (too little) rain in the inner core (Fig. 5), and the distribution of the rain flux was skewed toward the heavier (lighter) rain rates for the GFDL (NAM) model (Fig. 8a). By contrast, the GFS is better at predicting the amount and the distribution of inner-core rain. There are several possible explanations for these differences. While the GFDL and the GFS models both employ the same simplified Arakawa–Schubert (Pan and Wu 1995) convective parameterization scheme (CPS), the models differ both in spatial resolution and in how they handle microphysical processes. The finer resolution of the GFDL model would be associated with stronger vertical motions in the core region than the coarser-resolution GFS. Furthermore, the GFDL model has only a simplified method for handling microphysical processes in which it rains out all supersaturation (minus evaporation as

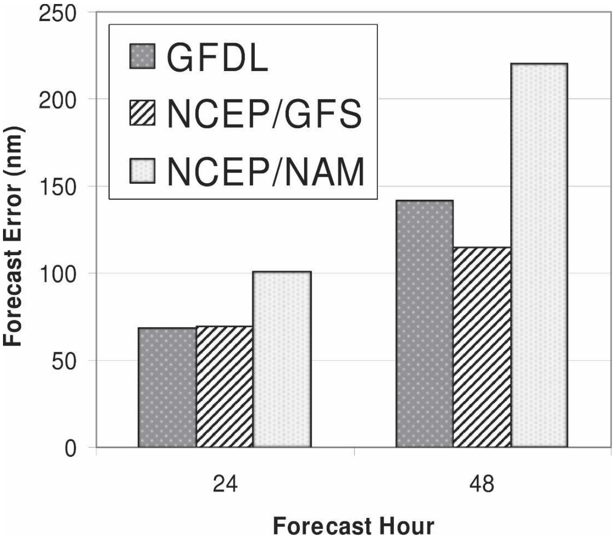


FIG. 12. Mean track forecast errors (n mi) at 24- and 48-h lead times for the GFDL, GFS, and NAM models for the specific set of landfalling cases included in this study.

the rain falls) in a layer when the relative humidity of that layer exceeds 100% (M. Bender 2006, personal communication), while the GFS model employs an explicit prognostic equation to track cloud water (Moorthi et al. 2001). These differences in the resolution and microphysical parameterization likely account for many of the discrepancies between the two models in their distributions of QPFs in the near-core region, although differences in boundary layer parameterizations may also play a role. Much of this reasoning is speculative, and more research is needed to investigate these hypotheses.

By contrast, the NAM model uses a different convective scheme, that of Betts–Miller–Janjić (Janjić 1994); a different microphysical parameterization; and it also does not include a bogus vortex in its initial conditions. All of these factors may explain why it produces too little rain in the inner core. The dependence of convectively produced rainfall on the Betts–Miller–Janjić convective scheme has been investigated in previous studies using the NCEP Eta Model (e.g., Mesinger 1998; Gallus 1999), with results also highlighting the inability of the Eta Model to produce the heaviest amounts of rainfall from convective events. Additionally, with no bogus vortex in the initial conditions, the inner core is less well defined (i.e., weaker) prior to landfall. A weaker inner-core circulation would be associated with weaker vertical motion, which would result in lighter rainfall. These hypotheses also require further research to investigate their validity.

The R-CLIPER produced too little rain at distances far removed from the track of the system. The reason for this problem is likely similar to the problem with pattern matching: calculating rain rates from an axisymmetric distribution produces a poor estimate of total rainfall at large distances from the center. Because a symmetric distribution of rain is assumed in R-CLIPER, no far-field influences in the environment are predicted that may produce asymmetries in the rain field, such as frontal boundaries.

c. *Extreme rain amounts*

For the extreme rain amounts (Fig. 11), the GFS has the highest skill. The GFDL produces too much of the heaviest rain (Fig. 6a), but both the GFDL and GFS have skill relative to R-CLIPER. The NAM has no skill relative to R-CLIPER, due mainly to its inability to produce the extreme rain amounts observed in the core. Since the highest rain rates are likely to occur in the inner core (cf. Fig. 7b), the explanation for these discrepancies is likely similar to that described above for the rain flux distribution. That is, the GFDL produces too much rain in the inner core (Fig. 10a) due to a finer grid resolution and a primitive handling of the microphysical processes. Even though the GFS has coarser resolution, the cloud water is handled explicitly, which may explain the higher skill for inner-core rainfall. The NAM model, which does not have a bogus vortex in its initial conditions, produces too little rain in the inner core.

5. Assessing the impact of TC track forecast error on QPF skill

In this section, we investigate the impact of model TC track forecast errors on QPF skill. Lonfat et al. (2004) showed that the distribution of rainfall from a TC is closely related to the track of the storm. In the discussion of the QPF validation scheme presented previously in this paper, the impact of track forecast error was indirectly accounted for in some of the metrics. For example, as part of the techniques for examining the distribution of the rain flux and the extreme rain amounts, we use a track-relative analysis so that we can compare the distribution of forecast rain flux along the model forecast track with the distribution of observed rain flux along the best track. This helps to reduce the impact of track forecast error, but it does not eliminate it, since significant differences in terrain and proximity to environmental atmospheric features may exist along the two different tracks. The pattern-matching techniques involving ETS and pattern correlations that were described previously are particularly sensitive to track forecast errors, especially for higher-resolution data. Here, we will focus specifically on the impact of track forecast error on these pattern-matching QPF metrics.

Differences in track forecast errors among the models can be significant, especially at lead times beyond 24 h. For the sample of landfalling cases in this study, the GFDL and GFS models have comparable track forecast skill at 24 h, and the error from both of these models is about 30% less than that of the NAM (cf. Fig. 12). At 48 h, the GFS was the best model for predicting the tracks of the landfalling storms in this study, with mean track forecast errors 19% less than those of the GFDL and 48% less than the NAM.

By removing the track error, one can quantify the impact of track error on track-dependent validation algorithms such as the ETS and pattern correlations. Such modified forecasts can also be used to more directly compare the contributions to the rainfall forecast error from other sources, such as resolution and parameterization deficiencies. Track-error removal is accomplished by shifting each 6-h rainfall forecast pattern by a distance equal to the difference in position of the forecasted versus the observed storm location. The field of rainfall that is shifted includes only those grid points that are within 400 km of the midpoint between two successive 6-hourly forecast positions. We use a 400-km threshold here as opposed to a 600-km threshold since we want to limit the areal coverage of rainfall that is being shifted in order to more effectively isolate rainfall associated with the storm at each 6-hourly location. These shifted rainfall predictions are then summed over the lifetime of the storm at each 6-hourly interval to produce storm-total shifted rainfall predictions. An example of a shifted GFDL storm-total rainfall field for Hurricane Georges of 1998 is shown in Fig. 13. The original GFDL forecasted rainfall field (Fig. 13a) and the resultant storm-total GFDL forecasted rainfall field after the 6-hourly rainfall fields are shifted (Fig. 13b) may be compared with the observed (stage IV) field (Fig. 13c) during the same 72-h period. For this case, shifting the rainfall field results in an increase in the correlation of storm-total rainfall from 0.14 to 0.73, which indicates a significant contribution due to the track error.

The ETS and pattern correlations are compared in Fig. 14 for all of the storms from 1998 to 2004, before and after the rainfall fields are shifted. Caution should be exercised when comparing the ETS and correlation results from Fig. 14 with those from Figs. 3 and 4, as they will be different. The analyses for Figs. 3 and 4 use all available storm-total rainfall data within 600 km of the best track, while the analyses for Fig. 14 use only data within 400 km of the midpoint between two successive 6-hourly storm locations. Since this latter method focuses only on near-storm data during each 6-h forecast period, the ETS and pattern correlations will be higher than those for the analyses presented in Figs. 3 and 4. It is still worthwhile, however, to compare ETS and correlation scores among the different models for the shifted rainfall fields.

As shown earlier (Fig. 3) and again in Fig. 14a, the ETS for the GFS model was the highest across all rainfall thresholds, while the R-CLIPER had the lowest ETS. When the rain fields are shifted to account for track error, the NAM model ETS is significantly improved, while the GFS model ETS is only slightly improved. After the shift, the GFDL and NAM models have comparable skill to the shifted GFS model over almost all rainfall thresholds. Whereas the R-CLIPER ETS is also improved, it is still lower than that of the other three models. These results suggest that the lower ETS for the NAM and GFDL compared to the GFS are mostly due to the deficiencies in their track forecasts. Once the track forecasts errors are accounted for, the remaining deficiencies are attributable to other aspects of the forecasting system, such as improper vortex initialization or deficient physical parameterizations. However, it is worth noting that even after the shifting is done, the equitable threat scores for the NAM still cannot match those of the GFS for amounts greater than 2 in. (50.8 mm). Similar results can be seen for the changes in the pattern correlations (Fig. 14b); that is, correlations improve the least for the GFS once the rain fields are shifted to account for track errors. The NAM performs almost as well as the GFS once the fields are shifted, while the GFDL and R-CLIPER also experience significant improvements in correlations.

6. Summary and concluding remarks

The main purpose of this paper was to design a scheme for validating the QPF for landfalling tropical cyclones that best accounts for their unique attributes and provides a framework for future validation efforts. Because the distribution of TC rainfall is so strongly dependent on storm track, a QPF validation scheme for tropical cyclones has requirements that are different from those for nontropical, continental summertime rainfall. Three characteristics of the models' performance were identified as critical to the evaluation of TC rainfall. These characteristics include the ability of the models to match QPF patterns, the ability to match the mean value and volume of observed rainfall and reproduce the distribution of rain, and the ability to produce the extreme amounts of rain often observed in TCs. A validation scheme was developed that employs

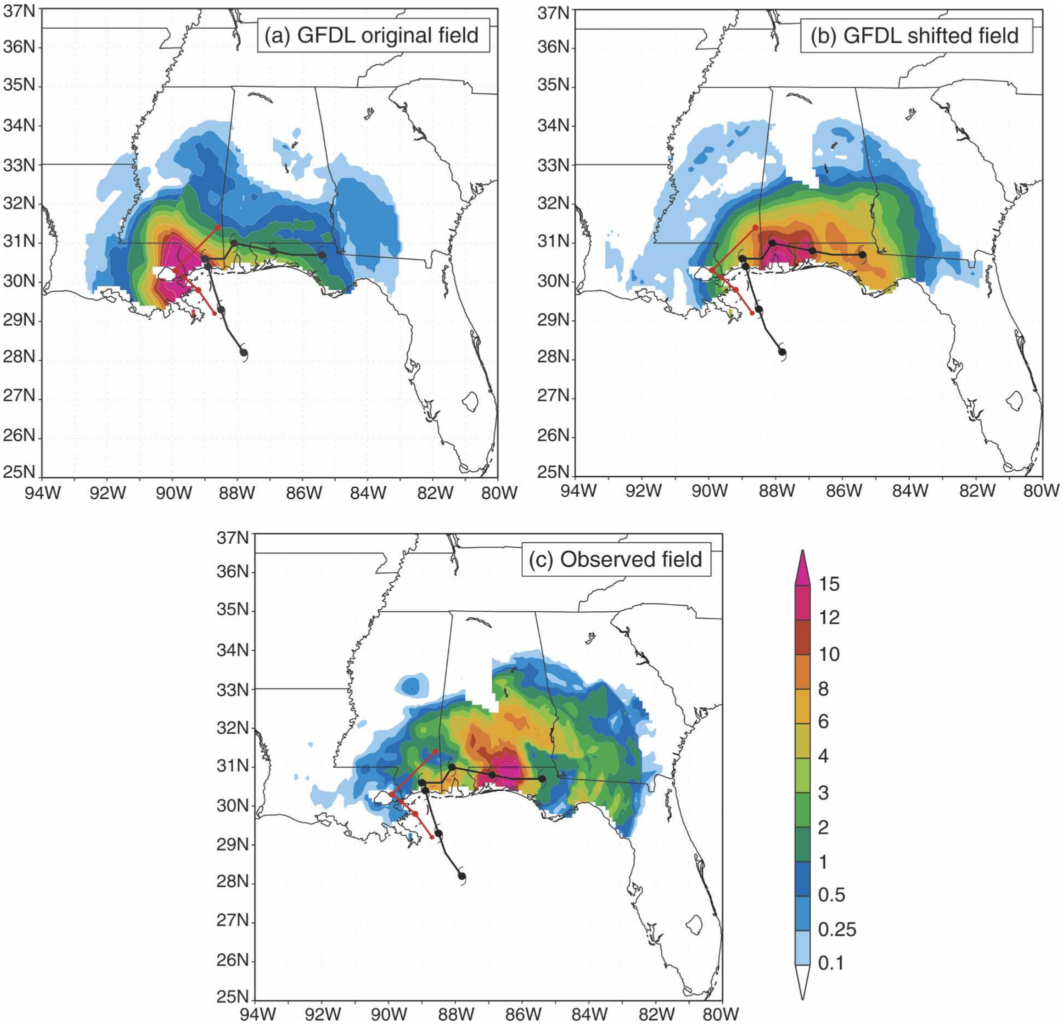


FIG. 13. Example of storm-relative grid-shifted rainfall fields (in.) for GFDL forecast of Hurricane Georges (initial time of 1200 UTC 27 Sep 1998): (a) original GFDL 0–72-h rainfall forecast, (b) shifted GFDL 0–72-h rainfall forecast, and (c) observed (stage IV) 0–72-h rainfall. GFDL forecast track is shown in red and the best track is shown in black.

traditional, commonly used QPF validation techniques (equitable threat score and pattern correlations) in combination with new techniques in order to evaluate those characteristics of QPFs that are most important to tropical cyclones.

To test the new validation scheme, evaluations of storm-total rainfall forecasts were performed for U.S. landfalling tropical cyclones from 1998 to 2004 for the following operational NWS models: NCEP/GFS, NCEP/GFDL, NCEP/NAM, and the benchmark R-CLIPER model. A summary of the results from these validations is shown in Table 5. Compared to R-CLIPER, all of the dynamical models have comparable or greater skill for all of the attributes. For the pattern-matching comparison, the GFS model performed the best across a broad range of rainfall thresholds, while the R-CLIPER performed far worse than all of the dynamical models. The NAM and the GFDL were comparable across all thresholds except for the very lightest amounts, where the NAM outperformed the GFDL. Comparisons of predictions of mean rainfall and distributions of rain volume showed that the GFDL produced too much of its rain flux in the higher range of rain rates. This bias was especially evident in the inner core. By contrast, the NAM model produced too little of its rain flux at the higher rain rates, especially in the inner core. The GFS model predicted the best distribution of rain flux, while the R-CLIPER produced significantly less rain flux in the higher rain rates at large distances from the storm track. The rainfall amount

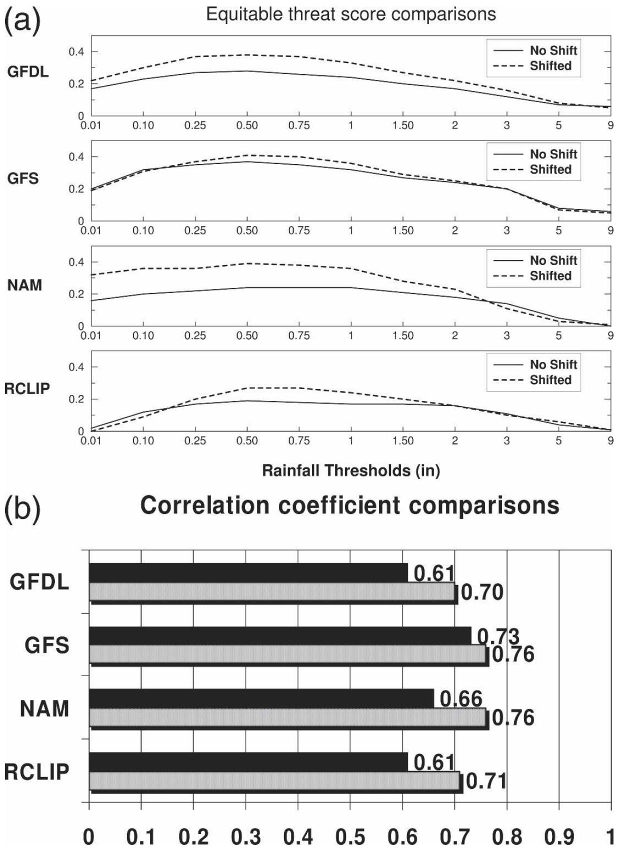


FIG. 14. (a) Comparison of the ETS for all models before (solid line) and after (dashed line) the grid of rainfall is shifted to account for track error. (b) Comparison of the mean QPF pattern correlation coefficient before (solid bar) and after (dotted bar) performing a grid shift.

corresponding to the top 5% of the observed rain flux distribution, which is indicative of extreme rain amounts, was larger for the GFDL model than the observations, which again indicates that the GFDL model predicted too much of the extreme rain. By contrast, both the R-CLIPER and the NAM model predicted too little of the extreme rain, while the GFS corresponded closely with the observed rain flux distribution.

The likely reasons for the biases in the QPF forecasts documented here may include deficiencies in the physical parameterizations, such as the microphysical parameterization, the planetary boundary layer parameterization, and the convective parameterization. The biases may also be related to deficiencies in the specification of the initial conditions, such as vortex initialization. A multitude of experiments, such as varying the physical parameterizations and the vortex initial conditions, can be performed to explore the sources of these biases. This is left for future work.

In addition to the development of the TC QPF validation scheme, results were presented that evaluated the impact of track forecast error on a model's QPF skill. We used a technique to shift 6-hourly accumulations of predicted rainfall by a distance equal to the difference in position of the forecasted versus the observed storm location, and then compared error statistics from the pattern-matching techniques before and after the shift. The ETS and pattern correlations for the GFDL and NAM experienced a substantial increase in scores after the rainfall fields were shifted, and the R-CLIPER also showed a modest improvement. After the shift, the GFDL and NAM models had skill comparable to the GFS model, suggesting that the lower scores from the GFDL and NAM models using the original, unshifted data are mostly due to the deficiencies in their track forecasts.

One of the most intriguing results from this work is that the GFS model, despite being the coarsest-resolution model, performed the best in all three of the metrics. This is due to several factors. First, for this sample of storms, the GFS had higher skill in predicting the track of landfalling tropical cyclones. Second, the GFS uses a more sophisticated microphysical parameterization scheme than the GFDL. Additionally, the higher resolution of the GFDL and the NAM models may be better for predicting more detailed rainfall structures, but these structures will be misplaced if the storm track is wrong. The impact of this effect is likely compounded here by our use of a 72-h forecast cutoff time, since it becomes increasingly difficult to predict the location and timing of such small-scale features at

TABLE 5. Summary of TC QPF skill index comparisons from Fig. 13.

TC QPF attribute	Best performer(s)	Worst performer(s)	Comments
Pattern matching	GFS	R-CLIPER	All dynamical models show considerable skill relative to R-CLIPER
Mean/volume	GFS	R-CLIPER	GFDL produces too much inner-core rain, NAM produces too little inner-core rain, R-CLIPER produces too little rain far from track of center
Extreme rain	GFS	NAM, R-CLIPER	GFDL overproduces the heaviest rain rates; GFS nearly exactly matches observations

extended forecast times. Thus, while higher resolution is necessary to predict the maximum rain rates and to improve intensity forecasts, it does not necessarily lead to improved rainfall forecasts. The performance from the GFS for tropical cyclone QPFs suggests that over the forecast time scale of a TC landfall event (2–3 days), current operational models have enough resolution to predict the distribution of large-scale, storm-total rainfall, and that perhaps it is of less importance to perfectly resolve the eyewall and rainbands in order to obtain some measure of QPF skill. Indeed, a recent study comparing the microphysics fields (and by extension the rainfall fields) of 1.67-km fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al. 1994) simulations of Hurricanes Bonnie (1998) and Floyd (1999) with a microphysics observational database indicates that, while there were notable differences between the simulated and observed microphysics fields, both the track and intensity of the simulated storms were reasonably well reproduced (Rogers et al. 2007). This suggests that, for those cases at least, accurately reproducing the microphysics fields was not crucial to obtaining accurate track and intensity forecasts. More testing of these issues is needed, however, including sensitivity tests involving different physical parameterizations and different horizontal and vertical resolutions, for both operational models at coarser resolution and research models at higher resolution (e.g., Zhang et al. 2000; Braun 2002; Rogers et al. 2003). Using a validation scheme such as the one presented here will be a key component enabling the identification of model limitations and biases in the forecasts from these models and guide further efforts toward model development and improvement.

Acknowledgments. The authors thank Frank Marks, Mark DeMaria, and Morris Bender for their support of this project and helpful discussions; Beth Ebert and Ying Lin for their insightful help on QPF validation techniques; and Russ Elsberry, Tom Knutson, Beth Ebert, and two anonymous reviewers for their critical reviews of a previous version of this manuscript. This work was supported by a grant from the U.S. Weather Research Program (USWRP) Joint Hurricane Testbed (JHT).

APPENDIX

Quantification of TC QPF Skill Indices

The validation metrics for TC QPFs outlined in section 3 each address one or more of the following attributes of TC rainfall: 1) pattern matching, 2) mean rain and rain flux (volume), and 3) extreme rain. Here, we provide details on the formulation of the algorithms that are used to compute skill indices for each of these QPF attributes. For each skill index, the associated algorithm will assign a value ranging from 0 for no skill to 1 for the most skill.

a. Pattern matching

The index corresponding to the ability of the models to reproduce patterns of rainfall is derived from the ETS and the pattern correlations. To obtain an index for pattern matching, a mean ETS over all thresholds is calculated by weighting the ETS values according to the relative distribution of the observed rain flux, or rain volume, in these threshold bins. For each model, a pattern-matching metric with values ranging from 0 to 1 is calculated by combining the mean ETS averaged over all thresholds with the mean correlation coefficient and taking an average of those two values.

b. Mean rain and distribution of rain flux

The mean rain and rain flux distribution index consists of equally weighted contributions from the radial distribution of the mean rainfall (cf. Fig. 5), large-scale CDF median value (cf. Fig. 6b), and the track-relative CDF median value derived from the track-relative PDF swaths (cf. Fig. 8). The contributions are calculated as follows: for the radial distribution of mean rainfall, a mean rainfall error index (MREI) is calculated by scaling the differences between the mean rainfall in the forecasts and the observations by the maximum observed mean value found in any band and then summing over all bands out to 400 km. This is denoted by

$$\mathrm{MREI} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|\overline{R}_{\mathrm{fi}} - \overline{R}_{\mathrm{oi}}|}{R\,\mathrm{max}}\right), \qquad \text{(A1)}$$

where n is the number of radial bins, R_{fi} is the mean forecast rain for the ith radial bin, R_{oi} is the mean observed rain for the ith radial bin, and R_{max} is the maximum observed mean rain found in any band. In this formulation, the index is high (low) when the integrated difference between the mean rain from the forecasts and the observations is small (large).

For the large-scale CDF median value index (LS_CDF_MVI), the following formulation is used:

$$\mathrm{LS\_CDF\_MVI} = 1 - |R_{\mathrm{f}50\%} - R_{\mathrm{o}50\%}|, \qquad \text{(A2)}$$

where $R_{\mathrm{f}50\%}$ is the rainfall threshold corresponding to the 50th percentile on the rain flux CDF for each model and $R_{\mathrm{o}50\%}$ is the threshold corresponding to the observed 50th percentile. In this formulation, the index is high (low) when the difference in rain flux medians is small (large). If the difference in rain flux medians exceeds 1 in. (25.4 mm), the index is set to a no-skill score of zero.

A track-relative median index is calculated using the same method as in (A2), but here the method is applied separately for each of the four bands from 0–100 km out to 300–400 km from the storm track. The indices for each of the four bands are averaged together to calculate the track-relative median index.

c. Extreme rain amounts

The index for comparing the ability of the models to match observed extreme rainfall amounts is calculated by equally weighting the contributions from the 95th percentile of the rain flux CDF for the large-scale fields (cf. Fig. 9) and the 95th percentile of the rain flux CDF from the track-relative 100-km bands (cf. Fig. 10). The formulation for the large-scale CDF maximum index (LS_CDF_MI) is given by

$$\text{LS\_CDF\_MI} = 1 - (\text{CDF}_{\text{m95th}} - 95)^2, \qquad \text{(A3)}$$

where $\text{CDF}_{\text{m95th}}$ is the percentage of each model's rain flux CDF profile less than the rainfall threshold value associated with the 95% value on the observed CDF. The difference between the CDF from the model and from the observations is squared to give more weight to deviations from the observations in order to provide stronger differentiation between those models that closely approximate the observed extreme amounts and those that do not.

The contribution from the track-relative bands is calculated in an identical fashion to that for the large-scale distributions, except that the value for each of four 100-km-wide bands surrounding either the best track or a model's forecast track is calculated. The values for the four bands are averaged together to create an average value for each model for this track-relative index, and this index is averaged with the large-scale index to produce a comparison among models for the extreme rain events.

REFERENCES

Aberson, S. D., 1998: Five-day tropical cyclone track forecasts in the North Atlantic basin. *Wea. Forecasting,* **13,** 1005–1015.

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

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting,* **18,** 918–932.

Atallah, E. H., and L. F. Bosart, 2003: The extratropical transition and precipitation distribution of Hurricane Floyd (1999). *Mon. Wea. Rev.,* **131,** 1063–1081.

Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multisensor U.S. precipitation analysis for operations and GCIP research. Preprints, *13th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 54–55.

Bender, M. A., 1997: The effect of relative flow on the asymmetric structure in the interior of hurricanes. *J. Atmos. Sci.,* **54,** 703–724.

Black, M. L., J. F. Gamache, F. D. Marks, C. E. Samsury, and H. E. Willoughby, 2002: Eastern Pacific Hurricanes Jimana of 1991 and Olivia of 1994: The effect of vertical shear on structure and intensity. *Mon. Wea. Rev.,* **130,** 2291–2312.

Bosart, L. F., and G. M. Lackmann, 1995: Postlandfall tropical cyclone reintensification in a weakly baroclinic environment: A case study of Hurricane David (September 1979). *Mon. Wea. Rev.,* **123,** 3268–3291.

Braun, S. A., 2002: A cloud-resolving simulation of Hurricane Bob (1991): Storm structure and eyewall buoyancy. *Mon. Wea. Rev.,* **130,** 1573–1592.

Colle, B. A., 2003: Numerical simulations of the extratropical transition of Floyd (1999): Structural evolution and responsible mechanisms for the heavy rainfall over the northeast United States. *Mon. Wea. Rev.,* **131,** 2905–2926.

Corbosiero, K. L., and J. Molinari, 2002: The effects of vertical wind shear on the distribution of convection in tropical cyclones. *Mon. Wea. Rev.,* **130,** 2110–2123.

DeMaria, M., and R. E. Tuleya, 2001: Evaluation of quantitative precipitation forecasts from the GFDL hurricane model. Preprints, *Symp. on Precipitation Extremes: Prediction, Impacts, and Responses,* Albequerque, NM, Amer. Meteor. Soc., 340–343.

——, and J. M. Gross, 2003: Evolution of tropical cyclone forecast models. *Hurricane! Coping with Disaster,* R. Simpson, Ed., Amer. Geophys. Union, 103–126.

——, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting,* **20,** 531–543.

DiMego, G. J., and L. F. Bosart, 1982: The transformation of Tropical Storm Agnes into an extratropical cyclone. Part I: The observed fields and vertical motion computations. *Mon. Wea. Rev.,* **110,** 385–411.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

——, S. Kusselson, and M. Turk, 2005: Validation of Tropical Rainfall Potential (TRaP) forecasts for Australian tropical cyclones. *Aust. Meteor. Mag.,* **54,** 121–136.

Ferraro, R., and Coauthors, 2005: The Tropical Rainfall Potential (TRaP) Technique. Part II: Validation. *Wea. Forecasting,* **20,** 465–475.

Frank, W. M., and E. A. Ritchie, 2001: Effects of vertical wind shear on the intensity and structure of numerically simulated hurricanes. *Mon. Wea. Rev.,* **129,** 2249–2269.

Franklin, J. L., C. J. McAdie, and M. B. Lawrence, 2003: Trends in track forecasting for tropical cyclones threatening the United States, 1970–2001. *Bull. Amer. Meteor. Soc.,* **84,** 1197–1203.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting,* **13,** 377–395.

Gallus, W. A., Jr., 1999: Eta simulations of three extreme precipitation events: Sensitivity to resolution and convective parameterization. *Wea. Forecasting,* **14,** 405–426.

——, 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting,* **17,** 1296–1302.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 138 pp.

Janjić, Z., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.,* **122,** 927–945.

Jarvinen, B. R., and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for the North Atlantic basin. NOAA Tech. Memo. NWS NHC-10, 22 pp. [Available from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161.]

Jones, S. C., 2000: The evolution of vortices in vertical shear: III: Baroclinic vortices. *Quart. J. Roy. Meteor. Soc.,* **126,** 3161–3185.

Kidder, S. Q., S. J. Kusselson, J. A. Knaff, R. R. Ferraro, R. J. Kuligowski, and M. Turk, 2005: The Tropical Rainfall Potential (TRaP) technique. Part I: Description and examples. *Wea. Forecasting,* **20,** 456–464.

Klein, P. M., P. A. Harr, and R. L. Elsberry, 2000: Extratropical transition of western North Pacific tropical cyclones: An overview and conceptual model of the transformation stage. *Wea. Forecasting,* **15,** 373–396.

Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting,* **18,** 80–92.

Lin, Y.-L., S. Chiao, T.-A. Wang, M. L. Kaplan, and R. P. Weglarz, 2001: Some common ingredients for heavy orographic rainfall. *Wea. Forecasting,* **16,** 633–660.

Lonfat, M., F. D. Marks Jr., and S. 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.

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

Mesinger, F., 1998: Quantitative precipitation forecasts of the "early" Eta model: An update. Preprints, *16th Conf. on Weather Analysis and Forecasting,* Phoenix, AZ, Amer. Meteor. Soc., 184–186.

Moorthi, S., H.-L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN Global Analysis/Forecast System. NCEP Tech. Procedures Bull. 484, 14 pp. [Available from NCEP/EMC, W/NP23, World Weather Building, Washington, DC 20233; also available online at http://www.nws.noaa.gov/om/tpb/484.htm.]

Neumann, C. J., 1972: An alternate to the Hurran (hurricane analog) tropical cyclone forecasting system. NOAA Tech. Memo. NWS SR-62, 23 pp. [Available from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161.]

Pan, H.-L., and W.-S. Wu, 1995: Implementing a mass flux convection parameterization package for the NMC Medium Range Forecast Model. NMC Office Note 409, 40 pp. [Available from NCEP/EMC, W/NP23, World Weather Building, Washington, DC 20233.]

Pfost, R. L., 2000: Operational tropical cyclone quantitative precipitation forecasting. *Natl. Wea. Dig.,* **24** (1–2), 61–66.

Rappaport, E. N., 2000: Loss of life in the United States associated with recent Atlantic tropical cyclones. *Bull. Amer. Meteor. Soc.,* **81,** 2065–2074.

Ritchie, E. A., and R. L. Elsberry, 2001: Simulations of the transformation stage of the extratropical transition of tropical cyclones. *Mon. Wea. Rev.,* **129,** 1462–1480.

Rogers, R. F., S. S. Chen, J. E. Tenerelli, and H. E. Willoughby, 2003: A numerical study of the impact of vertical shear on the distribution of rainfall in Hurricane Bonnie (1998). *Mon. Wea. Rev.,* **131,** 1577–1599.

——, M. Black, S. S. Chen, and R. A. Black, 2007: An evaluation of microphysics fields from mesoscale model simulations of tropical cyclones. Part I: Comparisons with observations. *J. Atmos. Sci.,* **64,** 1811–1834.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Shapiro, L. J., 1983: The asymmetric boundary layer flow under a translating hurricane. *J. Atmos. Sci.,* **40,** 1984–1998.

Tuleya, R. E., M. DeMaria, and R. Kuligowski, 2007: Evaluation of GFDL and simple statistical model rainfall forecasts for U.S. landfalling tropical storms. *Wea. Forecasting,* **22,** 56–70.

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

Wu, C.-C., T.-H. Yen, Y.-H. Kuo, and W. Wang, 2002: Rainfall simulation associated with Typhoon Herb (1996) near Taiwan. Part I: The topographic effect. *Wea. Forecasting,* **17,** 1001–1015.

Zhang, D.-L., Y. Liu, and M. K. Yau, 2000: A multiscale numerical study of Hurricane Andrew (1992). Part III: Dynamically induced vertical motion. *Mon. Wea. Rev.,* **128,** 3772–3788.