Evaluation of Operational Model Cyclone Structure Forecasts during Extratropical Transition

JENNI L. EVANS AND JUSTIN M. ARNOTT

Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

FRANCESCA CHIAROMONTE

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 9 June 2005, in final form 7 February 2006)

ABSTRACT

Cyclone structure is known to be directly linked to the sensible weather effects produced by the weather system. The extratropical transition (ET) process leads to immense changes in cyclone structure and therefore to changes in the associated weather experienced. Although structure is clearly an important cyclone characteristic, validation of cyclone structure forecasts in operational numerical models has not been previously performed. In this study, short-term (12–36 h) forecasts of cyclone structure from tropical genesis to the completion of ET are validated using fields from the Navy Operational Global Atmospheric Prediction System and the NCEP Aviation model. The cyclone phase space (CPS) is used to quantify differences between forecast and analyzed storm structure, both on a point-by-point basis and through a cyclone-type-based comparison. This cyclone-type comparison exploits a previously defined breakdown of cyclone structure regimes in the CPS. The impacts of synthetic vortex insertion on the ensuing agreement between forecast and analyzed storm structure are explored. While the results show reasonable forecast skill for well-defined (i.e., nonhybrid) systems, cyclones in the process of ET are found to be poorly forecast, emphasizing the need for improved understanding and simulation of the structural changes experienced by ET cyclones.

1. Introduction

Tropical cyclones (TCs) have long been recognized as substantial hazards to life and property (e.g., Dunn 1961; Frank and Husain 1971). The concentrated region of intense winds and rainfall in TCs makes accurate track forecasts vital. Attempts to forecast TC track are documented back to at least the nineteenth-century studies of Padre Benito Viñes in Cuba (e.g., Viñes 1877, 1895). Advancements in TC forecasting over recent decades have been achieved in part through improved operational numerical modeling of these storms (McAdie and Lawrence 2000).

Model forecasts for TCs are traditionally assessed predominantly on track, with some attention paid to intensity validation on central pressure (McBride and Holland 1987; Fiorino et al. 1993; Surgi et al. 1998; Nagata et al. 2001; Goerss et al. 2004). Little if any attention is paid to the storm structure and validation of quantitative precipitation forecasts for TCs in their infancy.

More recently, the potentially extreme weather and wave activity resulting from TCs that undergo extratropical transition (ET) have become widely recognized (e.g., Jones et al. 2003). During ET, the TC loses its tropical characteristics—while the storm intensity (peak sustained winds) invariably decreases, the areal extent of gale force winds expands and the forward motion of the storm increases considerably. An exceptional example of gale force wind expansion occurred with the ET of Hurricane Irene in 1999 (Fig. 1). The
increasing speed of motion of the ET system results in pronounced asymmetries in both its wind and rain signatures (Willoughby 1995). The fast-moving, synoptic-scale ET system can drive “trapped-fetch” ocean wave growth, resulting in observed peak wave heights of up to 30 m (Bowyer 2000; Bowyer and MacAfee 2005; MacAfee and Bowyer 2005). The broad region of heavy rainfall that can accompany the ET system provides a substantial flooding threat (DiMego and Bosart 1982; Atallah and Bosart 2003; Colle 2003). Because 50% of landfalling Atlantic TCs undergo ET (Hart and Evans 2001), it is clear that these storms can have a substantial societal impact. Hurricanes Floyd (in 1999) and Juan (in 2003) are recent examples of such landfalling ET systems.

Recently, Hart (2003) has proposed a cyclone phase space (CPS) for characterizing the structure of synoptic cyclones. The CPS is spanned by three parameters: the (i) lower- and (ii) upper-tropospheric thermal winds (referred to as $-V_L^T$ and $-V_U^T$, respectively), and (iii) lower-tropospheric thermal asymmetry $B$. These three parameters are sufficient to distinguish when a TC commences and completes ET (Evans and Hart 2003).

Arnott et al. (2004) investigated the structural evolution of ET systems in the CPS using k-means clustering. Seven clusters were required to characterize the distinct structure regimes observed in the range of TC evolution from tropical cyclogenesis to ET. Arnott et al. (2004) obtained similar structure regimes using analyses from the National Centers for Environmental Prediction Aviation Model [AVN; now known as the Global Forecast System (GFS; Kanamitsu 1989)] and the Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991). This concordance suggests that the seven structure regimes documented by Arnott et al. (2004) are robust and define the major structural milestones of an ET system.

The dramatic structural changes experienced by cyclones undergoing ET imply that even with a correct track forecast, a poor forecast of these structural changes will translate into a poor forecast of the sensible effects of the storm. Thus, while accurate forecasts of track and intensity are a necessary first step, accurate forecasts of cyclone structure are a key to warning the public about the impending effects of an ET system.

One example of a poor structure forecast in conjunction with a reasonable track forecast is Tropical Storm Arthur from July 2002. Arthur was identified as a tropical depression off the Carolinas at 1800 UTC 14 July and upgraded to a tropical storm at 0600 UTC 15 July. Arthur continued to intensify to its peak intensity of 50 kt ($1 \text{ kt} = 0.5144 \text{ m s}^{-1}$) before being declared extratropical at 0000 UTC 17 July. A comparison of the NOGAPS forecast and verifying track locations (Fig. 2a) reveals track forecast errors comparable to the operational mean error at the same lead time (i.e., on the order of 200 km). However, plots of the forecast and verifying CPS values reveal that the storm was forecast to attain a more baroclinic structure than it actually attained in the 36-h forecast time period (Fig. 2b). In fact, while the forecast anticipated a growing baroclinic (i.e., cold core) cyclone, the analyses indicate a growing warm-cored cyclone. This complete difference in cyclone type (and likely, therefore, in the distribution of sensible weather effects) is troubling and suggests the need to verify forecasts of cyclone structure. A Geostationary Operational Environmental Satellite (GOES) image of Arthur (Fig. 3) provides further evidence that the storm still retained its tropical structure characteristics 27 h into the forecast. Arthur provides a concrete example of a reasonable track forecast but a poor ET forecast. [More cases of this type, such as Erin (in 2001), can be found online (available online at http://moe.met.fsu.edu/cyclonephase).] Thus, even with the generally impressive skill of modern track forecasts, forecasts of significant weather accompanying the storm passage could still be in error if the structure of the storm is not accurately predicted.

In this study, we use the CPS of Hart (2003) to characterize and compare analyzed and forecast cyclone...
structures from operational numerical models. Quantification of the effects of the different synthetic vortices versus the vortex relocation technique used to initialize the tropical vortex in the analysis cycle of each model is one of the overall goals of this research. The CPS provides an objective measure of cyclone structure for model forecast validation and has been shown to capture the onset and completion of ET (Evans and Hart 2003). CPS-derived diagnoses of storm structure forecasts from the AVN and NOGAPS are compared with their verifying analyses out to 36 h. We employ the seven-cluster solution of Arnott et al. (2004) as a baseline for these comparisons, with the goal of developing an objective reference frame for the validation of numerical model forecasts of storm structure.

2. Data and k-means clustering of the analysis CPS set

a. Storm set and operational model fields

Nineteen TCs from the 1998–2002 Atlantic hurricane seasons form the base storm set (Table 1). All of these storms were declared extratropical at some stage of their life cycle in the National Hurricane Center (NHC) best track dataset (e.g., Neumann et al. 1993). Analyses from the AVN and NOGAPS are available every 12 h through the life of each storm, resulting in 387 individual storm analysis times for the full 19-storm set. All model fields used here were archived in near–real time at a reduced resolution of approximately 1° × 1°. This resolution has been found to be adequate for resolving the gross features associated with the warm core of TCs (Hart 2003).

Model fields used in this study are the analyzed and forecast mean sea level pressure (MSLP) and the geopotential height fields at 900, 600, and 300 hPa. The MSLP is used to track the motion of the vortex center. The height fields are used in the calculation of the three CPS parameters—the storm motion–relative 600–900-hPa thickness asymmetry $B$, the 600–900-hPa thermal wind $-V_T$, and the 300–600-hPa thermal wind $-U_T$ (Hart 2003). For each storm time, the structure of the forecast storm vortex is evaluated for 12-, 24-, and 36-h forecasts against each verifying analysis. Systematic changes in the forecast vortex, compared with its initializing analysis, are also assessed.

On 6 July 2000, TC initialization in the AVN model changed from synthetic vortex insertion to vortex relocation (Q. Liu 2003, personal communication). The relocation technique involves identifying the TC in the model initial guess fields and moving it to the observed cyclone location (Liu et al. 2000). Because there had been no named storms before this date in 2000 (Franklin et al. 2001), the AVN analysis and forecast data are partitioned into 1998–99 (synthetic vortex, AVN$_{SV}$) and 2000–02 (vortex relocation, AVN$_{VR}$) subsets to isolate the effects of the synthetic vortex. While the analysis fields for both NOGAPS and AVN are available for the full 5 yr of the study (1998–2002), a lack of NOGAPS forecast fields in 1998 reduced its forecast dataset to 15 cases, and only 284 of the 387 analysis times are available for AVN (Table 2). Thus, the model intercomparisons are between AVN$_{SV}$ (1998–99) and NOGAPS (1999), and between AVN$_{VR}$ and NOGAPS for the 2000–02 seasons (Table 2).
FIG. 3. Satellite observations of Tropical Storm Arthur (in 2002) as it commenced and underwent ET: (a) GOES-8 image at 1500 UTC 15 Jul 2002 (at this time, Arthur had sustained winds estimated at 45 kt, gusting to 55 kt); (b) Advanced Microwave Sounding Unit (AMSU) thermal anomaly at 1737 UTC 15 Jul 2002; (c) AMSU-derived gradient wind at 1737 UTC 15 Jul 2002; (d) AMSU thermal anomaly at 1726 UTC 16 Jul 2002 (just prior to being declared ET); and (e) AMSU-derived gradient wind at 1726 UTC 16 Jul 2002. The contour interval for the radial-height temperature anomaly cross sections in (b) and (d) is 0.5 K and the contour interval for the radial-height gradient wind cross sections in (c) and (e) is 5 m s$^{-1}$. 
TABLE 1. Nineteen tropical cyclones from the 1998–2002 Atlantic hurricane seasons that form the base storm set used in this study.

<table>
<thead>
<tr>
<th>Storm</th>
<th>Start time</th>
<th>End time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonnie</td>
<td>1200 UTC 26 Aug 1998</td>
<td>1200 UTC 03 Sep 1998</td>
</tr>
<tr>
<td>Danielle</td>
<td>1200 UTC 26 Aug 1998</td>
<td>1200 UTC 10 Sep 1998</td>
</tr>
<tr>
<td>Earl</td>
<td>1200 UTC 29 Aug 1998</td>
<td>1200 UTC 08 Sep 1998</td>
</tr>
<tr>
<td>Mitch</td>
<td>0000 UTC 22 Oct 1998</td>
<td>0000 UTC 09 Nov 1998</td>
</tr>
<tr>
<td>Floyd</td>
<td>0000 UTC 10 Sep 1999</td>
<td>1200 UTC 19 Sep 1999</td>
</tr>
<tr>
<td>Gert</td>
<td>0000 UTC 11 Sep 1999</td>
<td>1200 UTC 24 Sep 1999</td>
</tr>
<tr>
<td>Alberto</td>
<td>1200 UTC 06 Aug 2000</td>
<td>0000 UTC 24 Aug 2000</td>
</tr>
<tr>
<td>Gordon</td>
<td>0000 UTC 15 Sep 2000</td>
<td>0000 UTC 21 Sep 2000</td>
</tr>
<tr>
<td>Dean</td>
<td>0000 UTC 23 Aug 2001</td>
<td>0000 UTC 29 Aug 2001</td>
</tr>
<tr>
<td>Erin</td>
<td>0000 UTC 08 Sep 2001</td>
<td>1200 UTC 17 Sep 2001</td>
</tr>
<tr>
<td>Gabrielle</td>
<td>1200 UTC 10 Sep 2001</td>
<td>1200 UTC 24 Sep 2001</td>
</tr>
<tr>
<td>Karen</td>
<td>1200 UTC 11 Oct 2001</td>
<td>0000 UTC 15 Oct 2001</td>
</tr>
<tr>
<td>Michelle</td>
<td>1200 UTC 31 Oct 2001</td>
<td>0000 UTC 07 Nov 2001</td>
</tr>
<tr>
<td>Noel</td>
<td>0000 UTC 03 Nov 2001</td>
<td>0000 UTC 07 Nov 2001</td>
</tr>
<tr>
<td>Arthur</td>
<td>1200 UTC 10 Jul 2002</td>
<td>0000 UTC 18 Jul 2002</td>
</tr>
<tr>
<td>Gustav</td>
<td>1200 UTC 08 Sep 2002</td>
<td>0000 UTC 15 Sep 2002</td>
</tr>
<tr>
<td>Isidore</td>
<td>1200 UTC 17 Sep 2002</td>
<td>0000 UTC 03 Oct 2002</td>
</tr>
<tr>
<td>Josephine</td>
<td>1200 UTC 16 Sep 2002</td>
<td>0000 UTC 20 Sep 2002</td>
</tr>
</tbody>
</table>

b. K-means clustering of analysis-derived CPS locations

Arnott et al. (2004) generated CPS locations for each of the 387 storm analysis times for the 19 ET cases considered here. Because B is an order of magnitude smaller than \(-V_B^L\) and \(-V_B^U\), the CPS positions are normalized prior to clustering. The normalization both here and in Arnott et al. (2004) is

$$X(s; i)_k = \frac{Y(s; i)_k - \overline{Y}_k}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} [Y(s; i)_k - \overline{Y}_k]^2}},$$  \hspace{1cm} (1)

where \(k = 1, 2, 3\) and indexes coordinates in the CPS [so that \(Y(s; i)_k = B, -V_B^L, \text{ or } -V_B^U\)], \(n\) is the number of observations in the dataset (387), and the overbar indicates a mean value of the particular parameter (thus the quantity in the denominator is the parameter’s standard deviation). Also, \(s = \text{NOGAPS, AVN}_{SV}, \text{ and AVN}_{VR}\) indexes sources (i.e., model and initialization method), \(I = 1, \ldots, n(s)\) indexes observations (i.e., across all times for each storm in turn), and the 3D points \([X(s; i)_1, X(s; i)_2, \text{ and } X(s; i)_3]\) correspond to the normalized locations \([B(s; i), -V_B^L(s; i), \text{ and } -V_B^U(s; i)]\) in the CPS.

After normalization, the separation of two points in the CPS is quantified through their Euclidean distance. To partition the analysis-based, normalized CPS data objectively into distinct groups based on storm structure, Arnott et al. (2004) used the \(k\)-means clustering algorithm (Anderberg 1973; Hartigan and Wong 1979).

This algorithm finds a local minimum for a function summarizing within-cluster separation, and the degree to which a solution fits the data depends of course on the number of clusters being utilized. Combining various statistical assessment approaches, Arnott et al. (2004) established that a seven-cluster breakdown was preferred. This clustering methodology is also utilized here with the resulting cluster solutions for the NOGAPS and AVN analyses shown in Fig. 4.

A comparison of storm times with cluster membership for each source analysis reveals that storms progress in a consistent manner through the clusters. This path through the CPS characterizes the structural evolution of Atlantic storms undergoing ET (Arnott et al. 2004). The distribution of storm category assigned by NHC for each NOGAPS (AVN) cluster is plotted in Fig. 5a (Fig. 5b). The relative locations of the pie charts in these figures indicate the relative locations of cluster centroids. The analyzed storm structure in both models follows the typical path to transition documented by Arnott et al. (2004)—a weak TC begins in cluster 1 (possibly even prior to naming), strengthening (possibly to hurricane intensity) and moving into cluster 2. If the storm further intensifies and moves into cluster 3 in its tropical phase, it will weaken (return to cluster 2) prior to commencing transition. As transition commences, the system moves into cluster 4. The completion of transition most often results in progression into cluster 5, with storms that continue to strengthen baroclinically moving on to cluster 6. A minority of hybrid systems do not take this path, but instead finish in cluster 7. This correspondence of cluster membership to storm category is confirmed by the dominant NHC storm classifications for each cluster (Fig. 5).

While the NOGAPS and AVN analysis datasets span very similar domains in the CPS for midlatitude systems (negative values of \(-V_B^L\) and \(-V_B^U\)), these models differ in their representation of strong TCs—the median values of \(-V_B^L\) for the complete analysis sets are 43.0 (NOGAPS; 1999–2002), 30.1 (AVN; 1998–2002), 27.4 (AVN_{SV}; 1998–99), and 31.0 (AVN_{VR}; 2000–02), dem-

TABLE 2. Number of times for which both analysis and 12–36-h forecasts were available for the 19 Atlantic ET cases from 1998 to 2002. Partitioning of the study time was due to the change of AVN vortex initialization at the beginning of 2000 from synthetic vortex insertion (AVN_{SV}) to vortex relocation (AVN_{VR}).

<table>
<thead>
<tr>
<th>Model</th>
<th>Time period</th>
<th>No. of storm times</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVN</td>
<td>1998–99</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>2000–02</td>
<td>228</td>
</tr>
<tr>
<td>NOGAPS</td>
<td>1999</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>2000–02</td>
<td>236</td>
</tr>
</tbody>
</table>

Unauthenticated | Downloaded 12/28/22 02:52 PM UTC
onstrating the consistent analysis of stronger warm-cored TCs in the NOGAPS (Goerss and Jeffries 1994) compared with either version of the AVN initialization. Comparison of Fig. 4a (4b) with Fig. 4c (4d) confirms that this difference in the median analyzed warm-core strength is reflected by changes in cluster compactness as well as in centroid location. These differences are related to differences in the initialization of TCs in NOGAPS and AVN and between the two AVN initialization routines (Evans and Arnott 2004). Quantification of the effects of the different synthetic vortices versus the vortex relocation technique is one of the overall goals of this research.

3. Methodology for evaluating forecasts of storm structure

The objective of this study is to evaluate global model forecasts in reference to the parameters expressed by the positions in the CPS, and in particular to the structure regimes expressed by the clusters produced by the corresponding verifying analyses. We will consider both the CPS data overall, and the CPS data separated into forecast hour \( h \) and model initialization source \( s \).

For every analysis time for every storm, we compile a set of normalized CPS locations for 12-, 24-, and 36-h forecasts. These predicted CPS positions [normalized using Eq. (1), with the same mean and standard deviation values as for the analysis CPS described in section 2b] are denoted by PX\((s; i; h)\) = [PX\((s; i; h)_1\), PX\((s; i; h)_2\), PX\((s; i; h)_3\)]. Once again, the source index \( s \) indexes the model and initialization method (NOGAPS, AVN\(_{SV}\), or AVN\(_{VR}\)) and the point index, \( i \), indexes “storm times” (i.e., storms and times within each storm life cycle). The hour index \( h \) is used here to denote the forecast time (12, 24, or 36 h) being considered.

Fig. 4. Distribution of analyzed storm structures in the CPS for the NOGAPS and AVN global model seven-cluster solutions: (a) NOGAPS B vs \(-V^L\), (b) NOGAPS \(-V^L\) vs \(-V^H\), (c) AVN B vs \(-V^H\), and (d) NOGAPS \(-V^H\) vs \(-V^L\). Here W and C denote warm- and cold-core signatures, respectively, with WW (CC) for deep warm (cold) core. Asymmetric (symmetric) structure is implied by residence above (below) the \( B = 10 \) line in (a) and (c). The raw CPS locations are plotted here. These are scaled prior to clustering to avoid the thermal wind components dominating the asymmetry, \( B \), in the clustering solution (Arnott et al. 2004).
To evaluate the structural evolution of the forecast vortex, we compare the forecast CPS location against the analysis CPS location at the validation time using a variety of measures. In particular, say for the ith storm time, we consider coordinate-wise Euclidian distances
\[ d_{ik} = \frac{|X_k(s; i) - PX_k(s; i; h)|}{X_k(s; i) - PX_k(s; i; h)} \quad k = 1, 2, 3, \] (2)
and the 3D Euclidean distance
\[ d_{i} = \sqrt{\sum_{k=1,2,3} [X_k(s; i) - PX_k(s; i; h)]^2}. \] (3)

Fig. 5. Distribution of storm category within each cluster as assigned by NHC for (a) NOGAPS (1999–2002) and (b) AVN (1998–2002). The locations of the pie charts in these schematics indicate the relative locations of the cluster centroids. Hence, the scaling in this figure differs from Fig. 4 to avoid overlapping pie charts.

a. Point-by-point comparisons

To evaluate the structural evolution of the forecast vortex, we compare the forecast CPS location against the analysis CPS location at the validation time using a variety of measures. In particular, say for the ith storm time, we consider coordinate-wise Euclidian distances

\[ d_{ik} = \frac{|X_k(s; i) - PX_k(s; i; h)|}{X_k(s; i) - PX_k(s; i; h)} \quad k = 1, 2, 3, \] (2)

and the 3D Euclidean distance

\[ d_{i} = \sqrt{\sum_{k=1,2,3} [X_k(s; i) - PX_k(s; i; h)]^2}. \] (3)
The $d_{ik(s,h)}$ measures deviations between analyses and predictions relative to each CPS parameter ($k = 1, 2, 3$) separately, while $d_{ik(h)}$ captures distances between the analysis and prediction in the three-parameter CPS simultaneously. Below, we will indicate with $\|X(s; i)\|$ the Euclidian norm of a CPS location; this is the distance between $X(s; i)$ and the origin of the CPS, as obtained from Eq. (3) substituting $PX(s; i; h)$ with the zero vector. Because there is some uncertainty in the details of the model analyses, it is not clear that these deviations between the forecast and analyses can simply be regarded as forecast errors. This assertion will be further explored below.

b. Cluster membership comparisons: Simple cluster assignment and randomization analysis

In addition to the point-by-point comparisons, the analysis-based clusters in the 3D CPS provide another reference frame for comparison between analysis points and forecasts. In the $k$-means algorithm, cluster centroids are iteratively repositioned and analysis points $X(s; i)$ are reassigned to clusters until within-cluster differences are minimized. The resulting solution comprises centroids $C(s; j) = [C(s; j)_1, C(s; j)_2, C(s; j)_3]$ (where $j = 1, \ldots, 7$) and cluster memberships $m(s; i)$ (where $m$ is an integer between 1 and 7) for each point. The centroids of the analysis clustering solution are also used to assign cluster memberships to the forecast CPS positions $PX(s; i; h)$. These memberships, indicated by $mp(s; i; h)$ ($1 \leq mp \leq 7$), are obtained by minimizing $d[PX(s; i; h), C(s; j)]$ over $j$ for each forecast point. While this assignment rule for the forecasts does not take into consideration any feature of the clusters (e.g., shape and size in 3D) except for their centroids, it mimics the logic of the $k$-means algorithm.

It must be remarked that cluster memberships $m(s; i)$ and $mp(s; i; h)$ may be misleading in that they do not provide a sense of how “well positioned” a point is in the cluster to which it has been assigned. In principle, $d[X(s; i), PX(s; i; h)]$ may be very small, and yet $m(s; i)$ and $mp(s; i; h)$ may be different because the two points reside at the “boundary” region between two clusters, so the analysis and forecast CPS points may be attributed to different clusters. Thus, we do not only use cluster membership comparisons to evaluate forecast structure accuracy, but include the point-by-point comparisons outlined above, along with a series of randomization analyses to be described below. A variant of the cluster-based $\chi^2$ testing (Wilks 1995) was inconsistent with the point-by-point comparisons, failing to distinguish between agreements and compensating disagreements between cluster memberships. Because the $\chi^2$ test is a commonly used statistical measure of agreement between partitions, the failure of this test is summarized in the appendix.

Randomization analyses: An alternative test statistic and three reference distributions

To evaluate the consistency between the forecast and analysis storm structure, as expressed through membership in one of the seven clusters, we have devised an alternative testing approach to the more traditional $\chi^2$ test. We proceed as follows: given a source $s$ and forecast time $h$, we count for how many of the $n$ available points the forecast and verifying analysis memberships coincide. Dividing this count by $n$ yields a success statistic $S$ (i.e., the percentage of successful cluster forecasts):

$$S = \frac{\text{count of coincident memberships}}{n}.$$

To evaluate the significance of $S$, instead of postulating a reference distribution, we calculate a variety of empirical p-values based on randomizations of the memberships. Keeping analysis memberships fixed, we simulate three different reference distributions by generating forecast memberships completely at random (RANDOM), at random, but reproducing the distribution of the forecast points across the clusters (RAN-FORC), and at random, but reproducing the distribution of the analysis points across the clusters (RAN-ANAL).

In the first scenario, artificial forecast memberships $mp(s; i; h)_{\text{RANDOM}}$ are generated by randomly and independently assigning an integer between 1 and 7 to each forecast $PX(s; i; h)$. In the second scenario, artificial forecast memberships are generated creating a random permutation of $i = 1, \ldots, n$, say $\pi(i), i = 1, \ldots, n$, and setting $mp(s; i; h)_{\text{RAN-FORC}} = mp[s; \pi(i); h]$. Unlike RANDOM, RAN-FORC preserves the overall distribution of forecast points across the clusters. However, it breaks the association between forecast and analysis memberships (note that RAN-FORC represents a typical random permutation analysis). In the third scenario, artificial forecast memberships are generated, permuting the corresponding analysis memberships; given a random permutation $\pi(i), i = 1, \ldots, n$, we set $mp(s; i; h)_{\text{RAN-ANAL}} = m[s; \pi(i); h]$. RAN-ANAL imposes the overall cluster distribution observed for the analysis points upon the forecast points, essentially requiring the randomized forecasts to have the same “climate” as the analyses. However, because forecast memberships are assigned at random, this overall “climate match”
4. Evaluation of operational model forecasts of storm structure in the CPS

To evaluate the global model forecasts of storm structure, we are characterizing storm structure in terms of the three CPS variables. Structure forecasts are assessed by two separate techniques: (i) point-by-point comparisons of location in the CPS and (ii) comparisons of cluster membership utilizing the three randomization analyses outlined in section 3b as well as the seven-cluster k-means solutions constructed from the verifying analyses. The first approach gives unambiguous information on the similarity of the two distributions, but the information is not distilled by storm type. In contrast, the second approach assesses similarity on the basis of storm type and provides a range of measures of statistical significance for such similarity.

Scatterplots of the analysis clusters for both NOGAPS and AVN (Figs. 4 and 5) display the reference storm structure distributions against which we will now compare the forecast results. Designating these analysis-based distributions as references is not intended to imbue them with the mantle of “truth.” The interpretation of the differences between analysis and forecast structures will consider both analysis and model biases as potential sources of these differences.

Results from each of the forecast assessment techniques are presented in this section and the discussion is expanded further in section 5.

a. Point-by-point comparisons

To evaluate the structural evolution of the forecast vortex, we compare the forecast CPS location at the validation time using 1D separation and 3D Euclidean distance [Eqs. (2)-(5) above]. A graphical visualization of the coordinate-wise distances in the NOGAPS model is provided in scatterplots of the normalized CPS values from the forecast storm structure \( P_X(s; i; h) \) versus the analyzed storm structure \( X(s; i) \). A scatterplot of the 3D displacement between the forecast and analysis CPS locations \( d_{(s,i)} \) versus the 3D analysis CPS magnitude \( \| X(s; i) \| \) is also shown (Fig. 6d). These diagnostics are designed to investigate whether and how errors depend on the storm structure expressed in terms of the three parameters of the CPS, as well as the distance from the origin in such space.

A cursory examination of these scatterplots shows that the forecast structures appear to be in good agreement with the NOGAPS analyses on a variable-by-variable basis (Figs. 6a,c), however, the comparison of \( d_{(s,i)} \) versus \( \| X(s; i) \| \) (Fig. 6d) demonstrates that errors of up to 100% are possible. Further inspection reveals that these differences derive predominantly from deviations in the forecast storm symmetry \( d_{(s,i)} \) and lower-tropospheric thermal wind \( d_{2(s,i)} \) (Hart et al. 2006) record increasing structure variability with increasing \( \nabla L \), \( \nabla T \), and (consequently) 3D magnitude in the CPS \( \| X(s; i) \| \) (Fig. 6d). This result is consistent with theirs, since increasing 3D forecast errors are observed with increasing distance from the CPS origin.

Comparison of the trend lines in Figs. 6a,b demonstrates that the forecast storm structure becomes increasingly asymmetric and has a weaker lower-tropospheric thermal signature with forecast time compared with the verifying analyses. Thus, the model tends to forecast storms with a weaker tropical structure than is analyzed. Comparing slopes of trend lines based on their standard errors reveals that there is a statistically significant shift in the lower-tropospheric forecast storm structure \( (\nabla L) \) between the 12-h and 24-h forecast times, but that the change from 24 h to 36 h is not significant.

It is possible that the trend in the NOGAPS forecasts of the storm structure toward a more asymmetric and weaker cold core (cf. the NOGAPS analyses) is due to the analysis method; once a tropical storm is identified, the NOGAPS model initialization is enhanced by the insertion of a synthetic vortex. This was also true of the AVN model in the first 2 yr of our period of interest. Thus, we now examine point-by-point errors in both the AVN and AVNVR.

Comparison of the two recent AVN initialization methodologies—synthetic vortex initialization (1998–99) and vortex relocation (2000–02)—has the advantage that the vast majority of the model structure is consistent between the two datasets (certainly more than between the NOGAPS and AVN results). Scatterplots of the seven-cluster solutions for each model...
version are given in Fig. 7. Although both datasets have the same percentage of warm-cored storms (69%), the AVN<sub>VR</sub> results have more points in the \((-V_L^1 > 0, B > 10\) quadrant, the region identified by Arnott et al. (2004) as representing transitioning TCs. This implies that initialization with the synthetic vortex acts to reduce the signature of the ET phase. Evans and Hart (2003) noted that storms classified as tropical by NHC invariably fell in the range of \(-10 < B < 10\), where \(B\) is the storm symmetry parameter of the CPS. The above suggestion of enhanced symmetry in the synthetic vortex results is further supported by comparing the percentage of times in which \(B\) falls within this range: 65% of AVN<sub>SV</sub> storm times fall in this range compared with only 33% of AVN<sub>VR</sub> cases (Fig. 7).

Comparisons of the normalized CPS values from the forecast storm structure \(PX(s; i; h)_k\) versus the analyzed \(X(s; i)_k\) CPS measures of storm structure from the complete NOGAPS dataset: a) normalized \(B\); b) normalized \(-V_L^1\); c) normalized \(-V_U^1\); and d) the normalized 3D displacement between the forecast and analysis CPS locations vs \(\|X(s; i)\|\). In (a)–(d), data relating to 12-h forecasts are dark circles with a short dashed trend line, data relating to 24-h forecasts are light shaded squares with a dot-dashed trend line, and data relating to 36-h forecasts are open diamonds with a long dashed trend line. A solid 1:1 trend line is plotted on (a)–(c) for reference.
difference in the initial synthetic vortex strength between the datasets. This difference is further explored in section 4c. Comparing the slopes of trend lines for the AVN$_{VR}$ forecasts (Fig. 9b) based on their standard errors reveals that there is a statistically significant shift in the lower-tropospheric forecast storm structure between the 24- and 36-h forecast times, but that the change from 12 to 24 h is not significant. This hints at a difference in the vortex structure evolution between the vortex relocation (AVN$_{VR}$) and synthetic vortex (NOGAPS and AVN$_{SV}$) initializations. Further exploration of this result requires the inclusion of longer forecast lead times; this is the focus of an ongoing study.

Finally, individual CPS parameter forecasts from AVN$_{VR}$ show no particular trends across forecast times (Figs. 9a,c), strongly suggesting that synthetic vortex insertion is affecting the results.

### b. Cluster membership comparisons: Randomization analysis

None of the “randomized” success statistics described in section 3b ever surpass the success percentage $S$ of the actual forecast cluster assignment (Table 3). Thus, our empirical $p$-values are all $<0.001$ (1000 randomizations were run in each case) and represent very strong statistical evidence—the chances of the observed $S$ occurring under any of the “null” scenarios we considered are extremely small.

As would be expected, the success statistics from completely random assignments $S_{\text{RANDOM}}$ are sizably lower on average than those from the other two randomizations. In fact, $S_{\text{RANDOM}}$ carries no information about the climate of the system (the likelihood of realizing particular structure types in the CPS).

The success statistics from the “forecast match” randomization $S_{\text{RAN-FORC}}$ and the “analysis match” randomization $S_{\text{RAN-ANAL}}$ are very similar on average, ranging from 19.9% to 23.2% (Table 3). This suggests that the distribution of structure types in the analysis and forecast sets represents very similar “structure climates.” Interestingly, the average $S_{\text{RAN-ANAL}}$ does not always exceed the average $S_{\text{RAN-FORC}}$; that is, the likelihood of two points being grouped in the same cluster in both the analysis and forecast is not always increased (on average) by imposing the analysis cluster distribution. While this result is somewhat surprising, the averages are uniformly close, and the slight differences shown in Table 3 may have little significance. Indeed, across models, forecast times, and the RAN-ANAL and RAN-FORC randomization scenarios, averages never differ by more than 3%–4%.

Results from $S_{\text{RAN-ANAL}}$ are presented graphically in Fig. 10. The skill of the 12–36-h forecast of cluster membership (based on CPS structure) compared with the randomized cluster membership is apparent for all three model versions. Also apparent is an expected gradual decrease in cluster forecast skill with increasing forecast time.

At 12 and 24 h, the NOGAPS and AVN$_{SV}$ forecasts are the most skillful with the AVN$_{VR}$ slightly behind. This trend reverses itself by 36 h, when the AVN$_{VR}$ forecast skill is the highest of all three model versions. One interpretation of this skill reversal is that the synthetic vortex maintains a more realistic vortex structure at short lead times, but that at long times (36 h or more), the structure imposed in the initial conditions may be inappropriate because the storm may evolve away from this imposed “tropical” structure. This could be of particular importance in our dataset because every cyclone underwent ET. To test this hypothesis, we...
measured forecast and analysis agreement partitioning the storm dataset into pre-ET onset ($B < 10$) and post-ET onset ($B > 10$). We again use the success statistic (i.e., percentage of successful cluster forecasts) $S$ described in section 3b. The results are recorded in Table 4. A considerable decrease (approximately 10%–15%) in cluster forecast success is noted after ET onset in the AVN$_{SV}$. While the 36-h AVN$_{VR}$ forecasts decrease in skill (7%) after ET onset, the same is not true at 24 h, when the AVN$_{VR}$ forecast skill actually increases slightly (3%). Unfortunately, because of the relatively small sample sizes in this partition, these results are not statistically significant. Thus, while this analysis supports our hypothesis that the synthetic vortex delayed the simulation of ET-related storm structural changes (thus leading to larger post-ET onset forecast errors), it is not conclusive. Perhaps the best evidence for the need for change here was in the actions of the operational center in replacing this initialization procedure.

c. Examination of synthetic vortex magnitudes

To explore the source of the consistent forecast error difference between the NOGAPS and the two AVN initialization approaches, we compared the structure and strength of the initial vortex for individual storms. East–west cross sections of analyzed potential vorticity are used to contrast the strength of the NOGAPS and AVN$_{SV}$ initial vortices for Hurricanes Bonnie (in 1998) and Gert (in 1999; Figs. 11a–d). Hurricane Isidore (in 2002) provides a comparison between the NOGAPS and AVN$_{VR}$ initial vortices (Figs. 11e,f). While only a small sample of storms is shown here to illustrate the relative strengths of the initial vortices for each initialization procedure, these examples are consistent with a larger set of storms examined. Clearly, the AVN$_{SV}$ initial vortex is considerably stronger for both Bonnie (Figs. 11a,b) and Gert (Figs. 11c,d). The stronger initial vortex in AVN$_{SV}$ is likely more able to sustain itself (i.e., remain more symmetric) further into the forecast.

Fig. 8. Same as in Fig. 6, but from the AVN$_{SV}$ dataset.
than the NOGAPS synthetic vortex, helping to explain the discrepancy between Fig. 6a and Fig. 8a (see section 4a). After the change in the AVN vortex initialization procedure, the AVN\textsubscript{VR} initial vortex was weaker and spatially larger than the NOGAPS initial vortex (Figs. 11e,f). These trends in the initial vortex strength are consistent with the relativities in forecast structure error between model sources discussed above. In particu-

| Table 3. Forecast success statistics (percentages) for the actual forecast cluster membership $S$, and for the three randomization analyses used to construct empirical $p$-values (see section 3b for a complete explanation of the derivation of these statistics). The numbers for $S\text{RANDOM}$, $S\text{RAN-FORC}$, and $S\text{RAN-ANAL}$ are averages from 1000 randomizations performed for each model and forecast time with the first–third quartile ranges provided in parentheses. |
|-----------------|-----|----------------|----------------|----------------|
| Forecast time   | Model | $S$           | $S\text{RANDOM}$ (avg) | $S\text{RAN-FORC}$ (avg) | $S\text{RAN-ANAL}$ (avg) |
| 12 h            | NOGAPS | 81.1          | 14.3 (12.6–15.7) | 23.1 (21.7–24.4) | 22.3 (20.5–24.0) |
|                 | AVN\textsubscript{SV} | 80.3          | 14.4 (12.2–16.3) | 21.2 (19.0–23.1) | 22.2 (20.4–24.5) |
|                 | AVN\textsubscript{VR} | 79.8          | 14.3 (12.7–16.0) | 19.9 (18.3–21.6) | 19.9 (18.3–21.6) |
| 24 h            | NOGAPS | 71.9          | 14.3 (12.6–15.8) | 23.2 (21.3–24.9) | 22.0 (20.6–23.7) |
|                 | AVN\textsubscript{SV} | 71.2          | 14.2 (12.3–15.8) | 21.3 (19.2–23.3) | 22.0 (19.9–24.0) |
|                 | AVN\textsubscript{VR} | 70.8          | 14.3 (12.7–16.0) | 19.9 (18.4–21.7) | 20.4 (18.4–22.2) |
| 36 h            | NOGAPS | 69.2          | 14.3 (12.8–15.6) | 22.6 (20.8–24.0) | 22.3 (20.4–24.0) |
|                 | AVN\textsubscript{SV} | 65.8          | 14.4 (12.3–16.4) | 20.7 (18.5–22.6) | 22.2 (19.9–24.0) |
|                 | AVN\textsubscript{VR} | 70.2          | 14.2 (12.5–16.0) | 20.5 (18.8–22.1) | 20.4 (18.8–22.1) |

Fig. 9. Same as in Fig. 6, but from the AVN\textsubscript{VR} dataset.
lar, because the set of storms being examined here all underwent ET, it is not surprising that the forecast structure errors increase by 36 h. However, the AVNVR has considerably more skill in its 36-h structure forecast than either of the other initialization procedures. We argue that the weaker (but still realistic) initial vortex present in AVNVR more readily evolves extratropical characteristics than the vortices generated in either of the other initializations. Furthermore, the stronger synthetic vortices are more skillful at shorter lead times because the vortex is more likely to still be tropical and to sustain this tropical nature in the forecast.

d. Cluster membership comparisons: Forecast versus analysis biases

Secure that overall similarities between forecast and analysis memberships in the CPS cluster framework are meaningful and robust, we now move to a more detailed examination of the structure forecast results by cluster and forecast time. Once again, consistency between forecast and analysis memberships is our criterion for forecast skill.

Forecast and analysis cluster memberships for each forecast time (12, 24, and 36 h) and source (NOGAPS, AVNSV, and AVNVR) are summarized in Fig. 12. Each of these panels comprises a 7 × 7 grid of possible outcomes, with the size of the circle at each grid node being proportional to the percentage of forecasts residing in that cluster (given the analysis cluster). Trend lines have been added to help visualize patterns in memberships for various forecast times and sources. In each case, a perfect forecast would produce a trend line coinciding with y = x (i.e., 100% of the circles residing along the pale solid line).

Cluster membership (and hence, storm structure) is most reliably forecast for strong baroclinic cyclones (cluster 6). Strong tropical cyclones (clusters 2 and 3), while well handled by AVNSV and, to a lesser extent, AVNVR, show significant forecast difficulty in the NOGAPS. The dominant trend in the NOGAPS is for cluster 3 (cluster 2) cyclones to evolve into the weaker

<table>
<thead>
<tr>
<th>Forecast time</th>
<th>24 h</th>
<th>36 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-ET onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOGAPS</td>
<td>72.6 (164)</td>
<td>72.2 (162)</td>
</tr>
<tr>
<td>AVNSV</td>
<td>77.8 (137)</td>
<td>72.5 (136)</td>
</tr>
<tr>
<td>AVNVR</td>
<td>69.3 (90)</td>
<td>72.8 (91)</td>
</tr>
<tr>
<td>Post-ET onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOGAPS</td>
<td>70.0 (90)</td>
<td>64.0 (89)</td>
</tr>
<tr>
<td>AVNSV</td>
<td>61.4 (76)</td>
<td>62.5 (73)</td>
</tr>
<tr>
<td>AVNVR</td>
<td>72.3 (57)</td>
<td>65.8 (56)</td>
</tr>
</tbody>
</table>
Fig. 11. East–west cross sections of Ertel’s potential vorticity [in potential vorticity units (PVU); 1 PVU = \(10^{-6} \text{ m}^2\text{ kg}^{-1} \text{s}^{-1}\)] through (a), (b) Hurricane Bonnie (in 1998); (c), (d) Hurricane Gert (in 1999); and (e), (f) Hurricane Isidore (in 2002) using the analysis fields from (a), (c) AVN_{SV}; (e) AVN_{VR}; and (b), (d), (f) NOGAPS.
cluster 2 (cluster 1; Fig. 12). This result is consistent with that from the point-by-point comparisons shown in Fig. 6b, which indicates that forecast vortices in the NOGAPS tend to evolve into a weaker lower-tropospheric warm core with time.

For all forecast times, the error in cluster membership is largest for weaker systems. This is especially evident for cluster 4, which has forecasts for all other clusters except 6 (the strong baroclinic structure type). In fact, cluster 4 is the only cluster in which 24- and 36-h forecasts can be more likely to fall in a different cluster than what is analyzed (AVN_{SV}, middle row). Arnott et al. (2004) demonstrated that storms in cluster 4 were predominantly undergoing ET. Before entering cluster 4, these storms had typically evolved to a hurricane (classified in either clusters 2 or 3), then weakened in intensity (typically dropping one cluster back to cluster 1 or 2), and then entered cluster 4 as ET commenced. As shown by Evans and Hart (2003), storms typically take about 24–36 h to complete transition, with a corresponding residence time in cluster 4. In cluster space, if a storm does not complete transition it will return to clusters 1–3; the completion of transition corresponds to membership in clusters 5–7, with cluster 5 being the most common next step of the ET path (Arnott et al. 2004). A similar spread of forecast structure types is evident in cluster 5, the cluster representing weaker baroclinic lows and most often signifying the end of transition. Thus, the ambivalence of the structure forecasts for storms in clusters 4 and 5 is another concrete demonstration of the lack of operational model forecast skill for TCs undergoing ET (e.g., Jones et al. 2003).

The spread of cluster membership errors evident at 12 h usually represents “shades of gray”—that is, incorrect forecasts of the strength of a tropical or an extratropical low. By 36 h, these errors in cluster membership translate into incorrect forecasts of system type, once again highlighting the model difficulty in dealing with the end of the TC life cycle as either tropical decay or ET (recall Tropical Storm Arthur; Figs. 2 and 3).

Fig. 12. Forecast vs analyzed cluster membership for 12-, 24-, and 36-h storm structures for the (top) NOGAPS, (middle) AVN_{SV}, and (bottom) AVN_{VR} initializations. The size of the circle at each grid point is proportional to the percentage of forecasts falling in that cluster. A trend line (dashed) and 1:1 line (solid) are provided for reference.
Such structure errors can occur even when the storm track forecast validates well.

5. Summary and conclusions

The goals of this study were to develop and demonstrate a new methodology for the objective evaluation of operational global model forecasts of storm structure and to assess potential biases resulting from model initialization type or forecast situation. Because the initialization methodology for the AVN model changed substantially early in 2000, results from this model were partitioned into two groups: AVN$_{SV}$ (1998–99; synthetic vortex) and AVN$_{VR}$ (2000–02; vortex relocation). The NOGAPS vortex initialization procedure used synthetic vortex insertion throughout the period of interest.

Storm structure was characterized in terms of the three CPS variables, and 12-, 24-, and 36-h forecasts of storm structure were evaluated and compared with the verifying analysis from the same model. Structure forecasts were assessed on a point-by-point basis and in reference to the seven-cluster $k$-means solutions constructed from the verifying analyses.

The effect of synthetic vortex insertion in the model initial conditions was not conclusively resolved here. Indeed, the NOGAPS and AVN$_{SV}$ both have a version of this initialization approach, with the NOGAPS generally showing slightly superior skill over all forecast times examined (Tables 3, 4, and Fig. 10). Both the NOGAPS and AVN$_{SV}$ demonstrated slightly better skill than the AVN$_{VR}$ early in the forecast time period (12 and 24 h), with the AVN$_{VR}$ becoming increasingly reliable at 36 h. Partitioning of the three datasets into pre-ET onset ($B < 10$) and post-ET onset ($B > 10$) revealed that the forecast cluster membership was more reliable for the tropical phase of AVN$_{SV}$. However, once ET had commenced, AVN$_{VR}$ was superior (Table 4, Fig. 12). Because of the small sample sizes in each subgroup, these results were not shown to be statistically significant, yet they support the hypothesis that synthetic vortex insertion can improve the tropical phase structure forecasts while degrading the forecasts once ET begins. This assertion gains further support from the analysis of the forecast agreement with the verifying analyses at different lead times. The AVN$_{SV}$ was superior at shorter forecast lead times (12 and 24 h), but by 36 h the AVN$_{VR}$ showed better agreement with the verifying analyses. Once again, this result is consistent with the ET onset later in the forecast period being contaminated by the synthetic vortex.

As demonstrated by the variation in the “validating” structures present in the pairs of model analyses (either AVN$_{SV}$ and NOGAPS or AVN$_{VR}$ and NOGAPS) available for each storm, while the model analysis fields have been used as a reference here, they cannot be regarded as truth. However, no observational dataset presently exists from which to derive CPS analyses. Lacking this observational resource, we have chosen to compare the model forecasts against their own analyses. Because no model can be regarded as truth and all models have their own inherent biases, the choice of any of these models—or of a model distinct from these—as a single reference would have introduced complications due to intermodel differences. The results presented here illustrate structure forecast biases that are consistent with findings by other investigators without introducing intermodel complexities.

Results from the cluster membership comparisons are revealing in that they allow us to identify major structural groups (tropical, transitioning, and baroclinic) and associated forecast biases without getting tied up in the detail of the point-by-point comparisons. Weak tropical and baroclinic lows are poorly forecast compared with the more intense systems of the same type. In particular, storms that are undergoing ET (cluster 4) have the worst forecast skill, with forecasts in every cluster except intense hurricanes being evident. This is in agreement with Jones et al. (2003, 2004), who demonstrate a broad decrease in model forecast skill when a tropical storm is undergoing ET. The result here is more specific, showing that the storm itself and not only its environment is poorly forecast.

As demonstrated with the case of Tropical Storm Arthur (in 2002), this poor structure forecast need not be accompanied by a poor track forecast. In such cases, the agreement of the track forecast with other models may not alert the forecaster that the forecast storm structure is in error. This has the potential to result in poor significant weather forecasts, because the patterns and intensity of surface winds and intense rain are coupled to the storm structure.

Acknowledgments. The authors benefited from many lively discussions with Mark Guishard, Bob Hart, Jessica Higgs, and Adam Moyer. Mike Fiorino’s assistance in acquiring the operational model fields is gratefully acknowledged. Charles Pavloski’s aid with model data management was of great help. Thanks also go to Mark DeMaria for the AMSU analyses of Arthur. Many analyses and figures were created using the Grid Analysis and Display System (GrADS). The comments of Sarah Jones, Ron McTaggart-Cowan, and one anonymous reviewer improved the final paper. This work was supported by the National Science Foundation under Grant ATM-0351926. The second author was partially...
funded by an American Meteorological Society graduate fellowship sponsored by the Office of Naval Research and by a National Defense Science and Engineering Graduate (NDSEG) Fellowship. This financial assistance is greatly appreciated.

APPENDIX

A Caution on the $\chi^2$ Test

The $\chi^2$ test (e.g., as described in Wilks 1995) is a useful tool to assess the consistency between the analysis cluster membership distribution and the corresponding distribution for a model forecast (the lead times we considered were 12, 24, and 36 h). In our application, the analysis provides “expected counts,” so that the $\chi^2$ statistic is defined as follows:

$$\chi^2 = \sum_{\text{clusters}} \frac{(# \text{ forecast} - # \text{ analyzed})^2}{# \text{ analyzed}}, \quad (A1)$$

where # analyzed indicates the number of members in each cluster in the analysis distribution and # forecast indicates the number of members in each cluster in the forecast distribution. The summation is over all clusters (i.e., 1–7). One would expect the value of $\chi^2$, which measures discrepancy between the two distributions, to be inversely proportional to membership forecast accuracy. As shown below in a simple example, however, care must be taken when defining exactly what “accuracy” means.

Imagine a dataset of 100 points divided into two clusters (A and B) in the analysis distribution, with each cluster containing half of the observations (i.e., 50 points). In this example, model membership forecasts are incorrect for every forecast. That is, every observation that belonged to cluster A in the analysis is forecast to reside in cluster B and vice versa. The $\chi^2$ value in this case is

$$\chi^2 = \frac{(50 - 50)^2}{50} + \frac{(50 - 50)^2}{50} = 0 + 0 = 0, \quad (A2)$$

indicating a perfect consistency between the two distributions. However, this consistency is not reflective of accurate membership forecasts; in fact, the forecast here is completely erroneous. In this extreme example, while the $\chi^2$ test “does its job” in detecting a perfect match between forecast and analysis distributions, it carries no information about the accuracy of membership forecasts. In practical applications, the situation will likely be less extreme, and the $\chi^2$ test will carry some information about membership forecast accuracy. However, aggregate compensation effects in the cluster counts can still lead to similar distributions, “masking” the presence of erroneous forecasts. This was the case in our data, where $\chi^2$ results were in disagreement with the mean 3D CPS forecast errors. This was the motivation behind the similarity statistic $S$ that we devised.

In summary, care should be used when employing a $\chi^2$ test; while such a test is designed to assess the consistency between an observed (e.g., forecast) and an expected (e.g., analysis) distribution, it may fail as a means to evaluate membership forecasts. In our study, we pursued two parallel data analysis strategies. On the one hand, we considered point-by-point forecast errors in the 3D CPS (for individual parameters and jointly in 3D). On the other hand, to evaluate cluster membership forecasts, we considered an alternative statistic (the “success” statistic $S$) and developed empirical $p$-values for it with computational techniques (randomizations of forecast memberships under various null scenarios). It is advisable to perform at least a cursory point-by-point check of the $\chi^2$ outcome if member matching is a requirement of the analysis being performed.

REFERENCES

Fiorino, M., J. S. Goerss, J. J. Jensen, and E. J. Harrison Jr., 1993: An evaluation of the realtime tropical cyclone forecast skill...