Probabilistic Verification of Global and Mesoscale Ensemble Forecasts of Tropical Cyclogenesis

SHARANYA J. MAJUMDAR

Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida

RYAN D. TORN

Department of Atmospheric and Environmental Sciences, University at Albany, State University of New York, Albany, New York

(Manuscript received 2 March 2014, in final form 24 May 2014)

ABSTRACT

Probabilistic forecasts of tropical cyclogenesis have been evaluated for two samples: a near-homogeneous sample of ECMWF and Weather Research and Forecasting (WRF) Model–ensemble Kalman filter (EnKF) ensemble forecasts during the National Science Foundation's (NSF) Pre-Depression Investigation of Cloud-systems in the Tropics (PREDICT) field campaign (15 August–30 September 2010) and ECMWF ensemble forecasts during the 2010–12 Atlantic hurricane seasons. Quantitative criteria for tropical cyclone (TC) formation were first determined from model analyses based on threshold values of lower-tropospheric circulation, local thickness anomaly, and minimum sea level pressure. A binary verification was then performed for all ensemble forecasts with initial-time tropical disturbances. During the PREDICT period, the ECMWF and WRF–EnKF had similar verification statistics, with reliability diagrams of positive slope flatter than unity, and relative operating characteristic (ROC) curves that demonstrate skill. For the 2010–12 ECMWF ensemble forecasts, the equitable threat score was small and positive, with skill mostly lost after 5 days. The reliability diagrams for 1–5-day forecasts were monotonic increasing, though an overly large number of short-range ensemble forecasts predicted a low probability of a TC when a TC was verified. The ROC curves exhibited similar skill for forecasts out to 5 days. The reliability curves were sensitive to parameters such as time tolerance and threshold values, and insensitive to cases that originated from African easterly waves versus those that did not. Qualitative investigations revealed case-to-case variability in the probabilistic predictions. While the sample size was limited, the ensembles showed the potential for probabilistic prediction out to 5 days, though it appeared that the model struggled with developing a warm core in the short-range forecast.

1. Introduction

An outstanding challenge in numerical weather prediction is the accurate prediction of the formation of a tropical depression, commonly known as tropical cyclogenesis (or ''genesis''). In parallel, the scientific hypotheses behind genesis remain actively debated. Several theories have emerged that describe how a seemingly disorganized cluster of isolated thunderstorms amalgamates into a tropical cyclone (TC) with organized deep convection and a closed surface wind circulation pattern about a well-defined center (Tory and Frank 2010). A recently proposed hypothesis suggested that genesis occurs inside a recirculation region of cyclonic rotation and weak straining deformation within a parent tropical wave (Dunkerton et al. 2009). The convective- and mesoscale vorticity within this ''pouch'' region, situated in a lower-tropospheric layer where the wave speed equals the flow speed, is allowed to strengthen and organize into a coherent warm-core, subsynoptic-scale circulation. The testing of this hypothesis was the main premise of the National Science Foundation (NSF) Pre-Depression Investigation of Cloud-systems in the Tropics (PREDICT) field campaign in 2010 (Montgomery et al. 2012). The field campaign also provided the opportunity to conduct relevant predictability studies, via the development of new real-time ensemble predictions, as well as sensitivity and diagnostic products. Recent papers (Torn

Corresponding author address: Prof. Sharanya J. Majumdar, Department of Atmospheric Science, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149.
E-mail: smajumdar@rsmas.miami.edu

DOI: 10.1175/WAF-D-14-00028.1

and Cook 2013; Komaromi and Majumdar 2014) have used concepts developed during PREDICT to expand on earlier works (e.g., Sippel and Zhang 2008; Wang et al. 2010; Elsberry et al. 2010).

Given the uncertainty and limited information in deterministic predictions, particularly of events such as genesis, it is prudent to take a probabilistic approach. The National Oceanic and Atmospheric Administration's (NOAA) National Hurricane Center (NHC) has adopted this approach in forecasting the probability of genesis of any selected cluster of thunderstorms within the next 2 days; this forecast period was extended to 5 days in 2013 (Blake 2013). These forecasts have to date been prepared subjectively using mostly satellite observations and a combination of deterministic numerical weather prediction (NWP) models.

Another option is probabilistic forecasts based on ensemble prediction systems. Although most probabilistic verifications of TC forecasts have focused on track (e.g., Majumdar and Finocchio 2010; Dupont et al. 2011; Hamill et al. 2011; Yamaguchi et al. 2012), genesis is now being investigated. The European Centre for Medium-Range Weather Forecasts (ECMWF) has begun providing "strike probabilities" of weak TC-like vortices (Vitart et al. 2012). Using this ensemble, Elsberry et al. (2010, 2011, 2014) and Tsai et al. (2013) have conducted verifications and "forecastability" studies of genesis and track via metrics such as the "likelihood value." Snyder et al. (2010) evaluated the probability of genesis for seven tropical systems using the National Centers for Environmental Prediction (NCEP) Global Forecast System, and suggested that differences in the pregenesis environment may influence the accuracy. As part of NOAA's Hurricane Forecast Improvement Project (Gall et al. 2013), genesis probabilities using multiple real-time ensembles are being produced.

In this study, we conduct a probabilistic verification of genesis over 3 yr worth of ensemble predictions. To accomplish this, it is first necessary to define an objective metric that captures the onset of genesis. Several TC trackers have been used extensively in coarser-resolution models (e.g., Marchok 2002; Camargo and Zebiak 2002; Peng et al. 2012). More recently, Tory et al. (2013) developed a method for identifying circulation patterns that are favorable for TC development. Their method is intended to be independent of the models to which it is applied, by focusing on pouch-centric parameters that combine the Okubo–Weiss parameter, absolute vorticity, and environmental parameters. Our approach is similar to that of Tory et al. (2013) in that we identify regions of enhanced large-scale vorticity in the lower troposphere, though our method aims directly at identifying TCs via objective detection of a warm core and a local minimum of mean sea level pressure. We identify threshold values of these quantities that are consistent with genesis in numerical model analyses, intending them to be independent of the model, and offering the capability to provide and evaluate probabilistic predictions in pouch regions.

We verify the binary quantity of whether a TC has formed or not, using the ECMWF ensemble and a regional ensemble based on the Advanced Research core of the Weather Research and Forecasting (WRF) Model (ARW; Skamarock et al. 2008). Both ensembles are evaluated for a near-homogeneous sample during 2010, and the ECMWF ensemble is also evaluated over three Atlantic seasons (2010–12). The ensemble prediction systems are described in section 2, followed by an introduction of the new metric for a TC and especially genesis in section 3. The probabilistic verification is presented in section 4, followed by case examples in section 5 and conclusions in section 6.

2. Ensemble prediction systems

a. *ECMWF*

This study uses data from the operational ECMWF Ensemble Prediction System contained in The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) database (available online at http://tigge.ecmwf.int). During 2010–12, this system comprised a 50-member ensemble produced twice daily (0000 and 1200 UTC) at a horizontal resolution of T639 triangular truncation (32 km) with 62 vertical levels. There were no major changes to the model or ensemble perturbation method throughout this period. The ensemble was initialized through a combination of (i) initial-time dry global singular vectors (SVs) maximized over each hemisphere, (ii) an "ensemble of data assimilations" used to create initial spread in the ensemble members, and (iii) initial-time tropical SVs with a dry total energy norm and additional linearized physics relevant to tropical processes, optimized over TCs (not pregenesis disturbances) and over the Caribbean area (0°–25°N, 100°–60°W) when a TC is absent. In addition to the initial perturbations, the ensemble included two methods that account for model error, which were found to be important within the context of TC track forecasts (Lang et al. 2012). First, the stochastic kinetic energy backscatter scheme (SKEB) (Berner et al. 2009) was applied to represent model error associated with the upscale transfer of kinetic energy originating from unresolved subgrid-scale processes. Second, the stochastic perturbed parameterization tendencies scheme (SPPT; Palmer et al. 2009) was applied to represent model uncertainty associated

with physics parameterization schemes. Owing to the large data volume, this study uses 6-day ensemble forecast fields at 1° resolution saved daily (0000 UTC only) for the 2010–12 seasons, and twice daily (0000 and 1200 UTC) during the PREDICT field campaign period.

b. ARW

During the PREDICT field campaign, twice-daily (0000 and 1200 UTC) 96-member ensemble forecasts were produced using ARW. This ensemble prediction system comprised a 36-km horizontal-resolution domain that included much of the North Atlantic basin [Fig. 1 in Torn and Cook (2013)] and a 12-km vortex-following two-way interactive nest centered on one of the pregenesis systems being tracked by NHC in real time as a potential genesis candidate (''INVEST''). Ensemble initial conditions were taken from the cycling ensemble Kalman filter system. An analysis ensemble was generated every 6 h by combining conventional observations, including PREDICT and NOAA aircraft dropwindsondes, with a 6-h forecast ensemble from the previous analysis time using the Data Assimilation Research Testbed (DART; Anderson et al. 2009). Ensemble Kalman filter (EnKF) data assimilation was performed over both the large-basin domain and the nested domains that follow the TC. Unique lateral boundary conditions were obtained for each ensemble member using the fixed covariance perturbation technique of Torn et al. (2006), in which the ensemble mean was the corresponding NCEP Global Forecast System forecast and the perturbations were random draws from the NCEP error covariances in the WRF variational data assimilation (WRF-VAR) system (Barker et al. 2004). The interested reader is directed to Torn and Cook (2013) for a more complete description of this modeling setup.

3. Identifying genesis in models

a. Choice of metrics and threshold values

In designing a new objective, quantitative metric for TC detection and prediction, the following qualities are desired: (i) efficiency of automated use with a large number of ensemble members, (ii) applicability to a range of horizontal grid resolutions (i.e., from mesoscale to global model resolution), (iii) consistency among different models, (iv) ease of trackability in a model forecast several days prior to genesis, (v) ability to capture the timing of actual genesis, and (vi) minimization of erroneous cases that might qualify as a TC based on the metric(s) but in reality are not.

Given that a TC is a warm-core cyclonic disturbance with a closed circulation pattern around a local minimum of sea level pressure, we seek to use variables that are faithful to that definition, while satisfying the above rationale. The cyclonic part of the definition is determined by computing the circulation, or via Stokes's theorem, the area-averaged vorticity centered on the disturbance. This metric has the benefit that it relates to the meso-α-scale circulation that defines a predepression and it has the same meaning regardless of horizontal resolution. By contrast, vorticity is often contaminated by sub-meso-α-scale features at higher resolutions ($<$10 km); thus, one needs to determine an appropriate vorticity threshold each time the resolution is changed. Via a series of trials (not shown), the layer-averaged 700–850-hPa circulation (hereafter CIRC) computed over a disk of radius 200 km is found to be the least ambiguous in the selection of threshold values.

The second variable is a 200–850-hPa thickness anomaly (hereafter THICK), which is designed to measure the amplitude of the vortex warm core. This quantity, similar to that proposed by Cheung and Elsberry (2002), is defined as the difference between the average 200–850-hPa thickness within 100 km of the vortex center and the corresponding thickness within 1000 km of the vortex center. Computing an anomaly with respect to the environment precludes the possibility that this criterion could be met by a tropical system moving into the midlatitudes, and allows for the same definition regardless of basin.

The third and final variable is a local minimum of mean sea level pressure (MSLP). A value at any given location is deemed a local minimum if it is smaller than all values within 5° latitude or longitude of that location.

The evolution of these three variables in the ECMWF analysis through to the genesis of Tropical Storm Karl (2010) is presented in Fig. 1. Twelve hours prior to genesis (0000 UTC 14 September) the disorganized group of showers and thunderstorms depicted in the satellite imagery possessed a maximum of CIRC exceeding $6 \times 10^{-5}\,s^{-1}$, a thickness anomaly exceeding 15 m, and a broad area of low pressure over the Caribbean Sea with no distinct minimum (Figs. 1a–d). The CIRC and THICK maxima are situated on the northern edge of the broad low-level gyre in the streamline analysis, at the pouch center. Twelve hours later, which was retrospectively determined by NHC to be the genesis time, the thunderstorm cluster still appears disorganized in satellite imagery, though the CIRC and THICK maxima in the ECMWF analysis have strengthened gradually, and a closed low of $<$1010 hPa is now evident (Figs. 1e–h). By 0000 UTC 15 September, the CIRC associated with newly formed Tropical Storm Karl has exceeded $8 \times 10^{-5}\,s^{-1}$, the collocated warm core has strengthened rapidly to yield a thickness anomaly

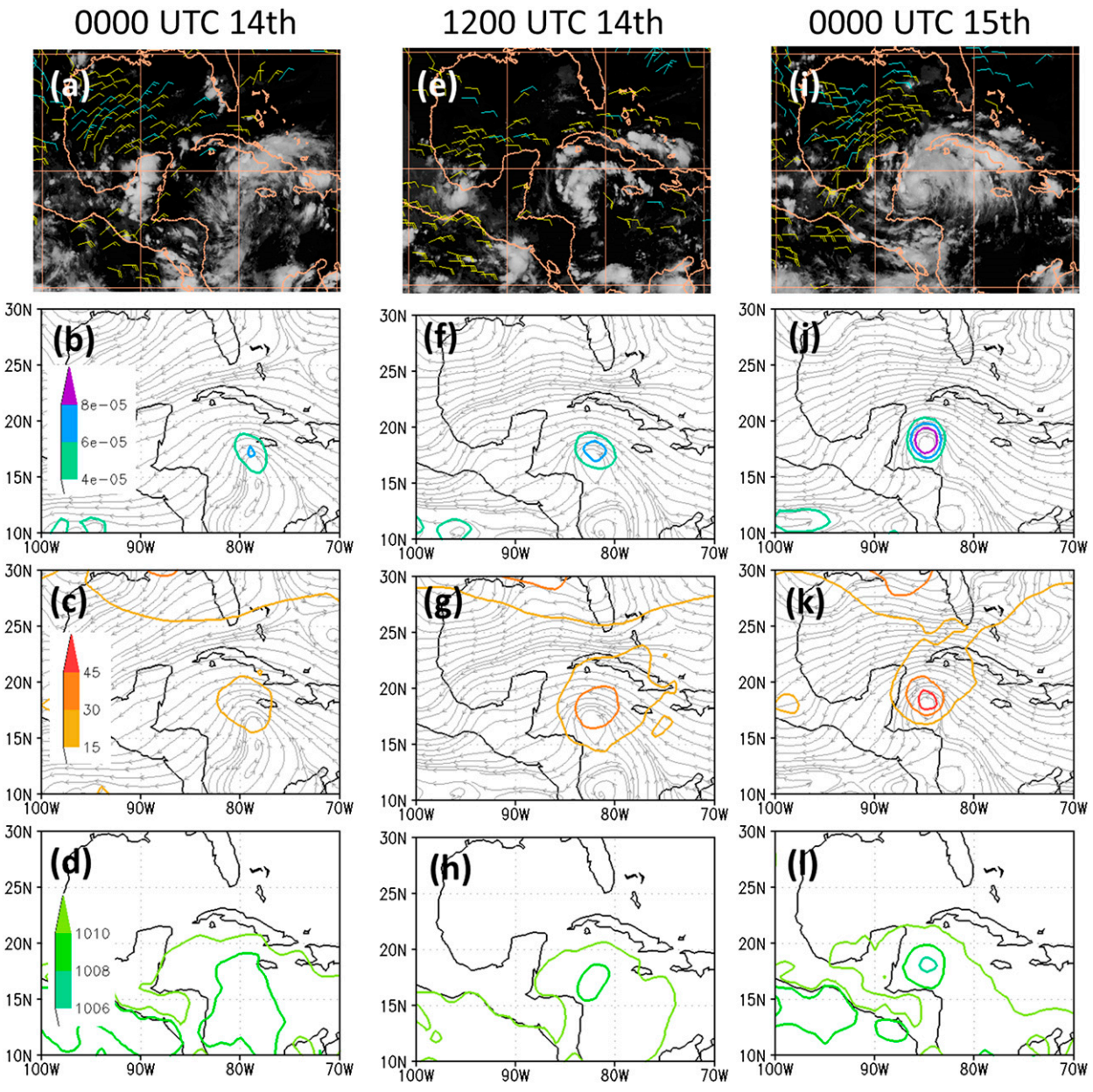

FIG. 1. (a),(e),(i) Infrared *Geostationary Operational Environmental Satellite-East* (*GOES-East*) imagery and lower-level cloud-track winds at 12-h intervals. Yellow barbs are for 700–850 hPa and cyan barbs for 851–950 hPa. (b),(f),(j) Corresponding ECMWF analysis contours of 700–850-hPa CIRC (s^{-1}) and 700–850-hPa streamlines. (c),(g),(k) Corresponding ECMWF analysis contours of THICK (m) and 700–850-hPa streamlines. (d),(h),(l) Corresponding ECMWF analysis contours on MSLP (hPa). [Satellite images courtesy Cooperative Institute for Meteorological Satellite Studies (CIMSS)/University of Wisconsin–Madison.]

exceeding 45 m, and the MSLP has deepened further (Figs. 1i–l). The evolution of these quantities can also be depicted as a time series in a circulation–thickness anomaly "phase space" diagram (green line in Fig. 2a).

To determine the appropriate values for TC genesis, the three variables are computed from analysis fields and compared against the NHC genesis timing. Given that not all depressions possess similar characteristics or even develop via similar mechanisms, we expect variability from case to case. However, a uniform set of threshold values is necessary in predicting genesis in models, since one does not know a priori what the values will be at the onset of genesis for any given disturbance. In addition to identifying the onset of genesis in the analysis fields, it is also necessary to minimize the number of cases in which "genesis' is identified too often in the model

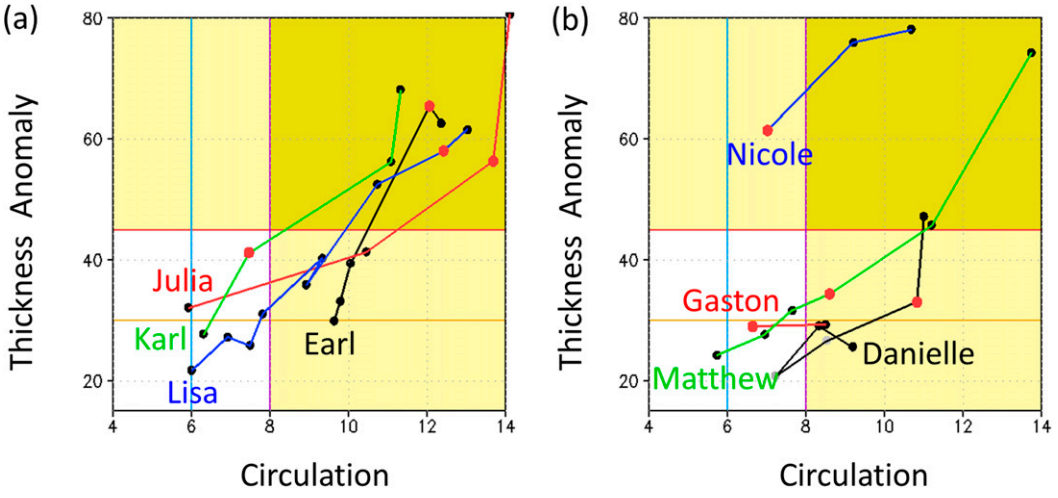

FIG. 2. "Phase space" of CIRC and THICK derived from ECMWF analyses, for selected PREDICT cases prior to genesis, and within 24 h of genesis. The dots are at 12-h intervals, with the red dots representing the time of genesis in the NHC Best Track dataset to the nearest 6 h. (a) Four cases illustrating a smooth development of circulation and a warm core. (b) Four cases showing irregular and diverse types of development. Dark yellow regions depict the space of values in which the criteria for genesis are exceeded. Light yellow regions correspond to where either the CIRC or THICK thresholds are exceeded, but not both. The colored lines of constant CIRC and THICK correspond to the contour values of CIRC = 6×10^{-5} and $8 \times 10^{-5}\,s^{-1}$ and THICK = 30 and 45 m, consistent with the colors of the spaghetti diagram in Fig. 1.

analysis (a false alarm) and cases in which the model analysis misses an actual genesis event.

The fields of CIRC and THICK are first computed from the twice-daily ECMWF analyses between 5°S–40°N and 110°W–0°, and over the WRF regional domain. Local maxima of these fields and minima of MSLP that exceed deliberately weak values are then identified in order to capture all disturbances. The respective values are CIRC $> 4 \times 10^{-5}\,s^{-1}$, THICK $>$ 20 m, and MSLP $<$ 1020 hPa. The CIRC maxima are the least ambiguous, and after finding these the common THICK maxima that are situated within 2° of the circulation maxima are tabulated. Following this, weak local minima of MSLP situated within 2° of the common CIRC and THICK maxima are identified. The 2° range is sufficiently generous to account for sheared disturbances, and/or disturbances containing a significant separation between the CIRC maximum and the weak MSLP minimum. This range is also strict enough to omit maxima over 200 km apart that are highly unlikely to belong to a feature that may shortly form as a depression. We emphasize that these weak common maxima are objectively identified and ordered into "model disturbances" solely from the analysis fields, with no human classification.

The next stage is to match these model disturbances with the NHC best tracks. The best-track data used here include all developing TCs, from the NHC genesis time to 12 h after this time (allowing for a "time tolerance"), and all INVESTs that were considered at some point by NHC to have a 30% chance or higher of genesis within 48 h. This sample is selected to provide a balance between the number of disturbances that develop into TCs and nondevelopers that have a plausible chance of development, while omitting trivial cases. Any model disturbance that meets or exceeds the specified threshold values in all three criteria is identified as a TC. For every analysis time up to and including the time 12 h after genesis, the following classifications are made:

- Hit *H*—a TC in both the model analysis and classified by NHC;
- Miss *M*—not a TC in the model analysis, but a TC classified by NHC;
- False alarm FA—a TC in the model analysis, but not classified by NHC; and
- Correct negative CN—neither a TC in the model analysis or NHC.

Additional measures of the appropriateness of a set of threshold values include the number of TCs for which the actual genesis is captured in the model analyses, and the "bias," which refers to whether genesis is captured too early or too late on average.

For both ECMWF and WRF–EnKF ensemble analyses, the threshold value of CIRC is varied between 5×10^{-5} and $9 \times 10^{-5}\,s^{-1}$, THICK between 30 and 50 m, MSLP between 1008 and 1014 hPa, and the time tolerance between 0 and 12 h. The statistics are shown for

TABLE 1. Statistics of the detection of a TC as determined by threshold values in the ECMWF model analyses. The following columns are presented: TOL, time tolerance (h); *H*, number of hits/number of actual TC cases; CN, number of correct negatives/number of actual non-TC cases; *M*, number of misses/number of actual TC cases; FA, number of false alarms/number of model TC cases; GEN, number of model genesis cases/all actual genesis cases; and bias, average time (h) that actual genesis precedes model genesis (positive values suggest that the model genesis is on average later than the genesis time declared by NHC). For 2010, the statistics are also shown when the threshold values of CIRC and THICK are varied. The thresholds of CIRC = $8 \times 10^{-5}\,s^{-1}$ and THICK = 45 m, listed in boldface, were chosen.

Yr	CIRC	THICK	MSLP	TOL	*H*	CN	*M*	FA	GEN	Bias (h)
2010	7	45	1012	12	53/63	210/221	10/63	9/62	18/22	+4.7
2010	8	40	1012	12	51/63	207/221	12/63	10/61	17/22	+3.9
2010	**8**	**45**	**1012**	**12**	**51/63**	**212/221**	**12/63**	**7/58**	**17/22**	**+5.3**
2010	8	50	1012	12	46/63	213/221	17/63	5/51	17/22	+13.1
2010	9	45	1012	12	45/63	217/221	18/63	6/51	15/22	+6.8
2011	**8**	**45**	**1012**	**12**	**40/56**	**165/177**	**16/56**	**15/55**	**15/19**	**+9.6**
2012	**8**	**45**	**1012**	**12**	**31/48**	**136/143**	**17/48**	**8/39**	**12/16**	**+7.5**

CIRC = 7×10^{-5}, 8×10^{-5}, and $9 \times 10^{-5}\,s^{-1}$ and THICK = 40, 45, and 50 m for the 2010 season in Table 1. With stricter criteria (higher CIRC and THICK, lower MSLP, no time tolerance), the fraction of false alarms is reduced but at the expense of increasing the fraction of misses. With more relaxed criteria, the fraction of misses is reduced but a higher fraction of false alarms occurs. To strike a balance with similar small fractions of misses and false alarms, the following criteria are judged to be optimal for ECMWF over the 2010–12 seasons: CIRC $8 \times 10^{-5}\,s^{-1}$, THICK 45 m, MSLP 1012 hPa, and a 12-h time tolerance (Table 1). These optimal threshold values are generally consistent through the three seasons. There is a general "late" bias in the model analyses by a few hours, though the bias is closer to zero when the calibration is performed against NHC's working best track during the season.[1] Although the WRF system bears little resemblance to the ECMWF system, similar optimal threshold values are found, thereby lending credibility to the metrics employed here.

Examples of cases during the 2010 PREDICT field campaign in which a steady progression toward genesis is realized in the ECMWF analysis are shown in Fig. 2a. While some cases exhibit a gradual development (Lisa), others are characterized by a rapid increase in circulation (Julia). In contrast, other cases such as pre-Gaston (Fig. 2b) are clear "misses" that are not captured by these metrics. In both the ECMWF and WRF–EnKF analyses, these misses often correspond to similar weak, short-lived TCs. Developers such as Danielle (Fig. 2b) show an irregular pathway to genesis, with an abrupt building of the warm core in the ECMWF analysis 12 h after genesis has been declared. An unusual developer is Nicole (Fig. 2b), which possesses a strong warm core but a weak circulation in the analysis at genesis. Matthew (Fig. 2b) shows a steady pathway toward genesis, but the actual genesis occurs when the thickness anomaly in the ECMWF analysis is still weak (about 34 m). Figure 2 underlines the variety of values of CIRC and THICK at the time of genesis for each system, often with significant differences in their values 12 h earlier or later.

Several other variables have been considered in this study, including the maximum 10-m wind speed, change in thickness anomaly over 12 h (signifying a rapidly organizing warm core), radial gradient of MSLP near the center, 925- and 1000-hPa geopotential heights, and the Okubo–Weiss parameter. However, the misses and false alarms are not reduced with their inclusion. In the remainder of this paper, only CIRC, THICK, and MSLP are considered.

b. *Identifying genesis in ensemble predictions*

The next stage is to identify and track disturbances in the ensemble forecasts. Only those cases with an identifiable disturbance at the initial time are considered. An "identifiable disturbance" here refers to an NHC INVEST with at least a 30% chance of genesis within 48 h at some point during its existence. Cases in which genesis occurs at some later forecast time, but in which no initial disturbance is present, are omitted. This is due in part to the ambiguity in objectively identifying common weak disturbances in different ensemble members, and to retaining the emphasis on the probability that a presently identified disturbance will develop. A disadvantage of this approach is that it limits the sample size beyond 2 days. Additionally, nondeveloping disturbances that had been assigned by NHC a <30% probability of development throughout their existence are omitted. An inspection of the values of CIRC and THICK for these weak disturbances over our full period of investigation

[1] The genesis time in the retrospective best track often precedes that in the working best track by ~6 h.

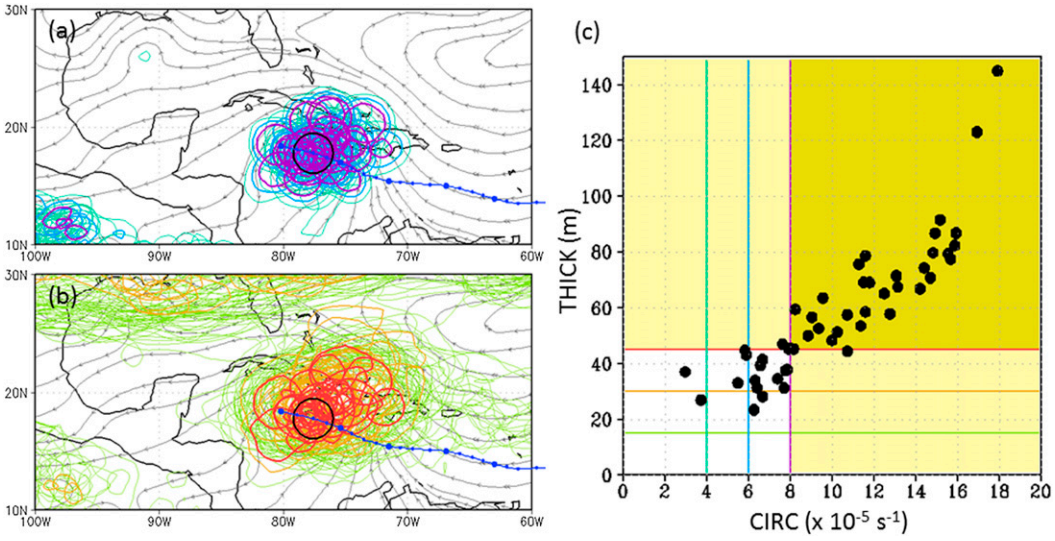

FIG. 3. The 108-h ECMWF ensemble forecast of the disturbance that developed into Tropical Storm Karl, initialized at 0000 UTC 10 Sep 2010. (a) Spaghetti diagram of CIRC at 4×10^{-5}, 6×10^{-5}, and 8×10^{-5} s^{-1} for control plus 50 ensemble members, with 700–850-hPa ensemble mean streamlines and the ECMWF deterministic pouch track forecast shown for reference. (b) Spaghetti diagram of THICK at 15, 30, and 45 m for control plus 50 ensemble members, with the same streamlines and pouch track as in (a). (c) Phase space for CIRC, THICK, and MSLP. In this ensemble forecast, MSLP $<$ 1012 hPa in all the ensemble members, and all are therefore shaded black. The colored lines of constant CIRC and THICK correspond to the contour values of CIRC = 4×10^{-5}, 6×10^{-5}, and 8×10^{-5} s^{-1} and THICK = 15, 30, and 45 m, consistent with the colors of the spaghetti diagrams in (a) and (b).

reveals a negligible number of false alarms, in other words, ensemble members erroneously predicting genesis (not shown).

For each ensemble member, the initial location of all disturbances is stored from the pouch best track or the NHC INVEST best track (if no pouch is classified). Local maxima in CIRC and THICK, as well as local minima of MSLP within 2° of each other, are then identified at successive 12-h forecast intervals, within boxes of 3° latitude and 5° longitude of the previous position. After some experimentation, this box size was found to be sufficiently small to exclude any neighboring disturbances (which are usually at least 1000 km away), and sufficiently large to track excessively fast-moving disturbances. If the track moves outside the 5°–30°N, 100°–10°W domain for a given ensemble member, that member is excluded from the probabilistic forecast and the probability is adjusted.

As an example, Figs. 3a and 3b show the 108-h ECMWF ensemble prediction of CIRC and THICK initialized at 0000 UTC 10 September 2010 for the pouch (PGI44L) that became Tropical Storm Karl at the verification time. Variance in the positions, strengths, and sizes of the local maxima are readily evident. The three aforementioned criteria for a TC are concurrently satisfied for 34 of the 50 ensemble members, suggesting a 68% chance that pouch PGI44L develops into a TC by the verification time (Fig. 3c).

An alternative means of visualizing the distributions of the variables is via probability density functions (PDFs), as illustrated in Figs. 7b and 7c of Montgomery et al. (2012). However, for brevity and to identify how the variables are correlated over the ensemble, the phase space of Fig. 3c is used here. In this case, CIRC and THICK are positively correlated. In other words, the ensemble members that possess a stronger (weaker) circulation also possess a stronger (weaker) warm core. This linear relationship is found to occur for most ensemble forecasts, although there are exceptions, particularly when the predepression does not exhibit typical tropical characteristics. A few cases exist in which CIRC and THICK appear sufficient, but the surface low is not yet well developed enough.

4. Verification

a. *Verification methods*

Binary forecasts (yes–no) of a TC in each ensemble member are verified. A deterministic forecast would represent an entry of 1 in one of the four contingency table categories of hits, misses, false alarms, or correct negatives. For a probabilistic forecast (e.g., Fig. 3), it is possible to derive percentages for each of these categories: $H = 68\%$, $M = 32\%$, and FA $-$ CN $= 0\%$. For each forecast time, these contingency tables are populated by all forecasts in the sample. Given that a TC

developing prior to the model verification time is always awarded a "hit" when the model is too late with its genesis prediction, a "time tolerance" is included in order to account for cases in which the model predicts genesis slightly too early, thereby reducing the likelihood of the statistics being skewed toward a slow bias in the model. Through a series of trials summarized in section 4c, we decided upon a time tolerance of 1 day.

Several common statistics are computed: those employed here are the bias = $(H + \mathrm{FA})/(H + M)$ and the equitable threat score ETS = $(H - H_{\mathrm{random}})/(H + M + \mathrm{FA} - H_{\mathrm{random}})$, where $H_{\mathrm{random}} = (H + M)(H + \mathrm{FA})/\mathrm{Total}$. While other statistics have been computed [such as the probability of detection POD = $H/(H + M)$], they do not yield additional insights.

The reliability diagram is also employed to determine whether an X% probability of genesis verifies on X% of all occasions. The range of predicted forecast probabilities is divided into bins, and the percentage of cases in each bin that verify as a TC is then plotted against the midrange forecast probability in each bin. A diagonal line that passes through the origin and is of unity slope implies perfect reliability. The Brier skill score is not considered here, since it relies on a climatology, which is nontrivial for discrete events such as genesis. Finally, the relative operating characteristic (ROC) uses a set of increasing probability thresholds to make a yes–no decision to quantify the ability of the forecast to discriminate between genesis and nongenesis. The ROC is plotted as the POD versus the probability of false detection POFD = FA/(CN + FA).

b. *2010 PREDICT period: ECMWF and WRF–EnKF*

The PREDICT field campaign (15 August–30 September 2010) occurred during an active period, in which 11 of the 12 candidate systems developed into TCs. These included 11 named TCs from Danielle to Nicole, plus one nondeveloper (PGI39L) that exceeded 30% probability of genesis within 48 h in NHC's forecasts. According to the NHC Tropical Cyclone Reports (available online at http://www.nhc.noaa.gov/data/#tcr), 8 of the 11 named TCs originated from tropical waves, while one (Karl) formed from the interaction between a tropical wave and an elongated trough of low pressure, and the remaining two (Hermine and Nicole) formed from low pressure areas not associated with waves. PGI39L was associated with a tropical wave. For this period, ECMWF and WRF–EnKF ensemble forecasts that initialized every 12 h are included. The sample is mostly homogeneous (sample sizes given in bars in Fig. 4a), except for a few cases where a pouch existed at 0 h but was outside the WRF–EnKF domain over Africa.

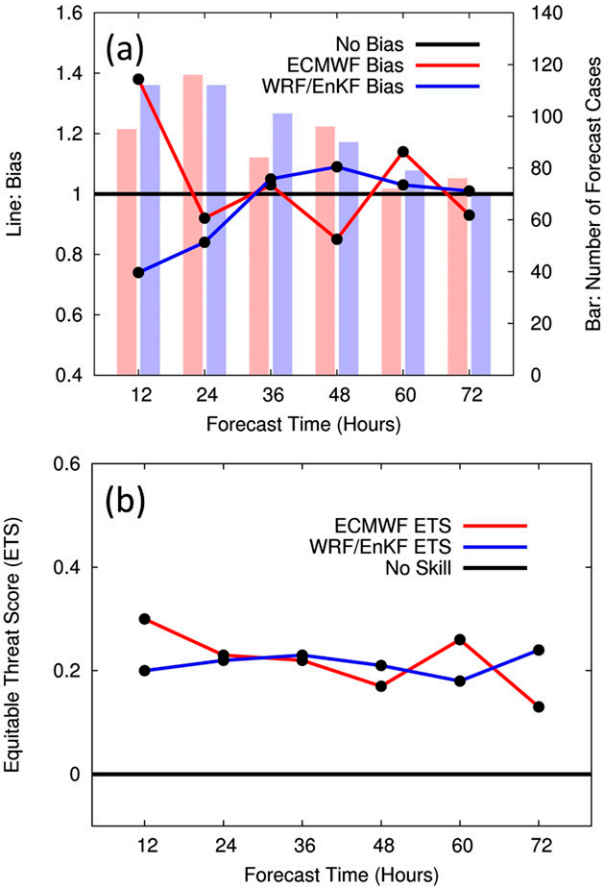

FIG. 4. (a) Bars show the numbers of ECMWF (red) and WRF–EnKF (blue) forecast cases during the PREDICT period (from 0000 UTC 15 Aug to 1200 UTC 30 Sep 2010). Lines show the average biases of the respective ensemble forecasts. (b) ETS. All lines are for ECMWF (red) and WRF–EnKF (blue) ensemble forecasts of 12–72 h.

The statistics for the ECMWF and WRF–EnKF ensemble forecasts are largely similar for 24–72-h forecasts (Fig. 4). Neither ensemble exhibits a clear bias (Fig. 4a). The equitable threat score is modest for both ensembles, with values close to 0.2 for most forecast times (Fig. 4b). Somewhat surprisingly, the statistics in Fig. 4 do not deteriorate noticeably as the forecast lead time increases.

For the probabilistic verification via the reliability diagram, six bins of forecast probability ranges are ordered as follows: 0%, 1%–10%, 11%–20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100% (Fig. 5a). As expected, the vast majority of cases lie in the first two bins of <10% probability of genesis. A difference exists in the distribution of probabilistic forecasts between the two ensembles, with ECMWF being more dispersive with a larger number of cases in the middle probability range (1%–80%), while WRF–EnKF has a relatively

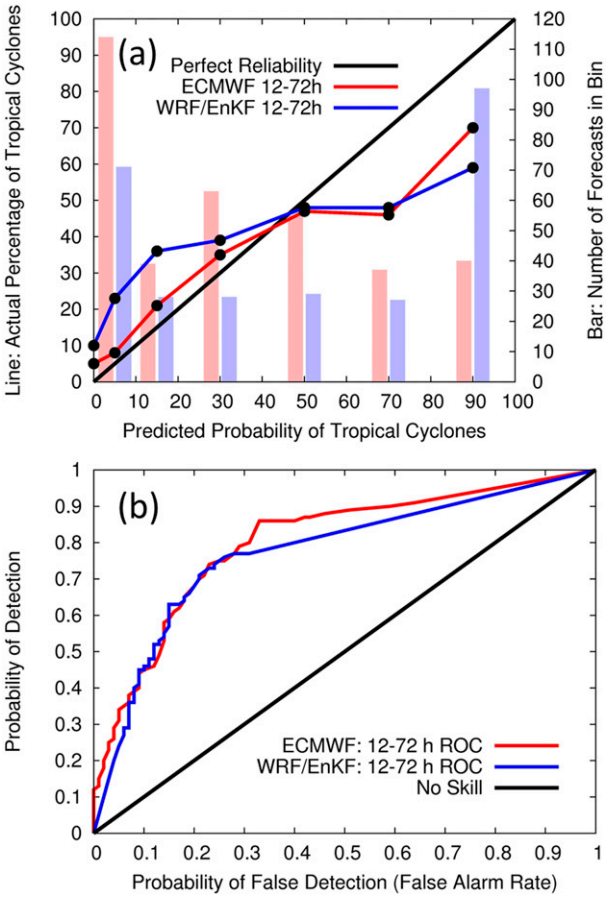

FIG. 5. (a) Bars show the numbers of ECMWF (red) and WRF–EnKF (blue) 12–72-h probabilistic forecasts in probability bins of 1%–10%, 11%–20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100% during the PREDICT period (from 0000 UTC 15 Aug to 1200 UTC 30 Sep 2010). The numbers of cases in the 0% bin (bars not plotted) are 191 for ECMWF and 284 for WRF–EnKF. Lines show the reliability curves for ECMWF (red) and WRF–EnKF (blue) ensembles, with a black dot plotted at the midrange of each probability bin. (b) Corresponding ROC curves.

high number of cases in the bins of zero probability and highest probability. The ECMWF ensemble produces a near-unity slope over all 12–72-h forecasts for probabilities up to 60%, while the WRF–EnKF ensemble on average underpredicts the probability in the ranges up to 40%. For the large number of 0% forecast probabilities, the fraction of cases of genesis is nontrivial (5% for ECMWF, 10% for WRF–EnKF). For both ensembles, when the probability of genesis is predicted to be high (suggesting high confidence that genesis will occur), the actual frequency of occurrences is too low. Finally, both ensembles exhibit similar skill in their 12–72-h forecasts based on the ROC, given that the curve lies far above the line of the unity slope (Fig. 5b). In other words, for given genesis and nongenesis events, the ensemble forecasts are able to discriminate between the two outcomes.

TABLE 2. Number of tropical disturbances in the ECMWF sample that exist between 0000 UTC 1 Jun and 0000 UTC 30 Nov during each year. The sample comprises NHC INVESTs given a probability of genesis within 48 h exceeding 30%. The number of disturbances that developed into tropical cyclones is given in parentheses.

	Total cases	Pure waves	Hybrid waves	Not waves
2010	27 (21)	15 (9)	4 (4)	8 (8)
2011	28 (17)	14 (11)	2 (1)	12 (5)
2012	17 (16)	13 (13)	0 (0)	4 (3)

c. 2010–12 Atlantic hurricane seasons: ECMWF

Similar verifications are now performed for the 2010, 2011, and 2012 Atlantic hurricane seasons. All 12–144-h ECMWF ensemble forecasts initialized at 0000 UTC each day between 1 June and 30 November are included for each season.

The number of cases listed in Table 2 comprises all INVESTs within our domain that were declared by NHC to possess >30% chance of genesis within 48 h during its existence. The disturbances are divided into those that originated from African easterly waves (AEWs), those that were a "hybrid" of a wave and another disturbance (e.g., a trough of low pressure), and those that were not waves. In 2010 (2011 and 2012), the sample of 27 (28 and 17) disturbances includes 15 (14 and 13) events that originated from AEWs, 9 (11 and 13) of which developed into TCs. In 2011, some disturbances were associated with a quasi-stationary trough situated offshore of the eastern United States, a few of which underwent genesis (Bret, Gert, and Sean). In 2012, there were several developers north of 30°N, which are excluded from our sample, and all but one of our 17 disturbances developed into a TC. In each year, most of the nondevelopers were in the early and late parts of the season, and most of the nonwave cases occurred during the late season. The sample size used in the verification for each forecast lead time is given in Fig. 6a.

For forecasts up to 1 day, the ETS in 2010 is inferior to that in 2011 and 2012 (Fig. 6a). For 36-h forecasts, the ETS is around 0.2 for all 3 yr. For longer forecast times, the ETS decreases on average, approaching zero (no skill) after 5 days. Though not shown here, the bias is found to be minimal up to 5 days in 2010, but on the low side in 2011 and 2012. In the reliability diagram, the number of 12–72-h forecasts with high probabilities is particularly small in 2011 and 2012, and therefore the final bin contains all forecasts with a predicted probability exceeding 60% (Fig. 6b). The reliability diagram

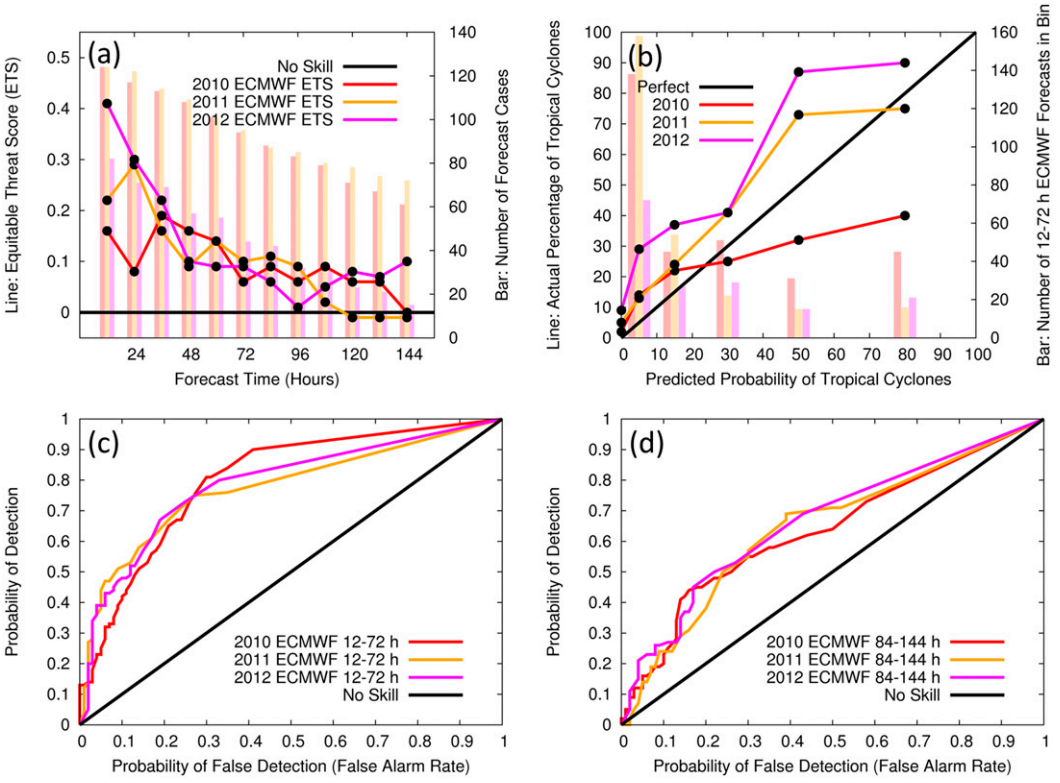

FIG. 6. Evaluations of daily ECMWF ensemble forecasts over the 2010–12 Atlantic hurricane seasons. (a) Bars show the numbers of cases in 2010 (red), 2011 (orange), and 2012 (pink) ECMWF 12–144-h forecasts. Lines show the ETS for each forecast time. (b) Bars show the numbers of probabilistic forecasts in probability bins of 1%–10%, 11%–20%, 21%–40%, 41%–60%, and 61%–100% for each of the 3 yr. The numbers of cases in the 0% bin (bars not plotted) are 347 in 2010, 400 in 2011, and 214 in 2012. Lines show reliability curves for 12–72-h forecasts for each year, with a black dot plotted at the midrange of each probability bin. Corresponding ROC curves are shown for (c) 12–72- and (d) 84–144-h forecasts.

for 12–72-h forecasts is monotonic increasing, though with differences from year to year. In 2010, the slope is flattest, suggesting a limited ability to discriminate between forecasts of low and high probability of genesis over the larger sample. In particular, as is similar for the subsample during the PREDICT period, ECMWF tends on average to overpredict the probability of genesis when the probability exceeded 40%. This can be largely attributed to a few cases in which the disturbance appeared likely to develop for a few days, but did not (e.g., PGI24L and PGI39L). In contrast, in 2011 and 2012, the slope of the reliability curve is steeper and closer to unity. However, there is underprediction in most bins. In particular, in 2012, genesis actually occurred in 30%–40% of those instances when a nonzero probability of less than 20% had been predicted for 12–72-h forecasts.

For probabilistic forecasts beyond 3 days, the size of the bin containing probabilities exceeding 40% is too small to compute a meaningful observed relative frequency for individual years; thus, only ROC curves are shown. For 12–72-h forecasts, the ROC illustrates a similar level of skill in each of the 3 yr (Fig. 6c). For 84–144-h forecasts, the skill diminishes but the ROC curves for each year remain noticeably above the line of the unit slope (Fig. 6d). In other words, the hit rate exceeds the false alarm rate over the full range of probability thresholds for medium-range probabilistic forecasts.

When the reliability diagram is computed over all 3 yr combined, the slope is on average near unity for probabilities up to 60% for 12–72-h forecasts, though with the aforementioned bias of underprediction when the probability is low (Fig. 7a). In the 60%–100% bin, there is a slight tendency to overpredict the probability of genesis, though this sample is dominated by the 2010 nondevelopers. Combining the 3 yr, an attempt is made at a reliability diagram for 84–144-h forecasts. A monotonic increasing reliability curve is found. The gradient of the slope is roughly 0.5, suggesting a limited ability to discriminate between the range of probabilities of genesis.

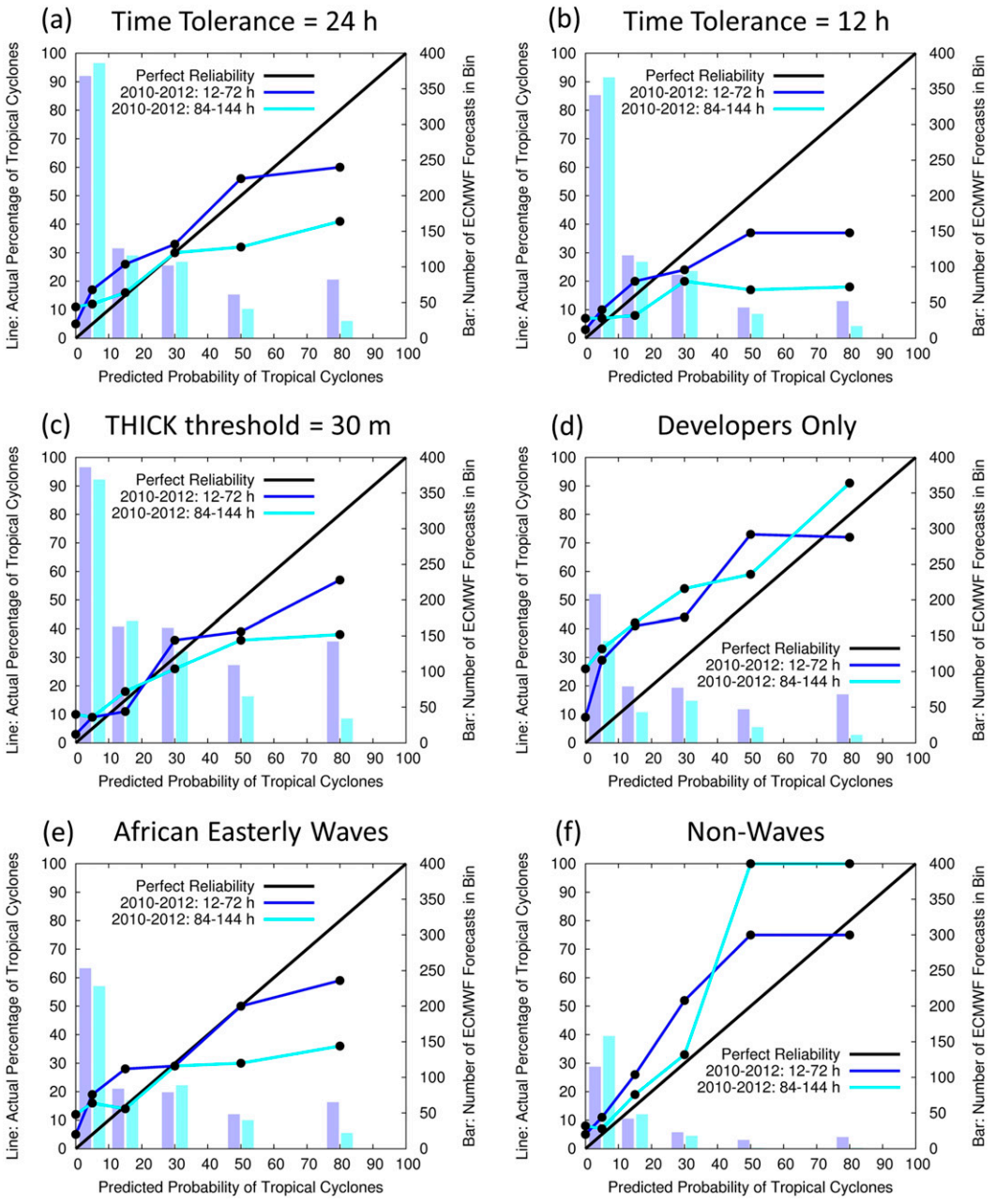

FIG. 7. Evaluations of daily ECMWF ensemble forecasts, cumulative over the 2010–12 Atlantic hurricane seasons. (a) Bars show the numbers of 12–72- (dark blue) and 84–144-h (light blue) probabilistic forecasts in probability bins of 1%–10%, 11%–20%, 21%–40%, 41%–60%, and 61%–100%. Lines show reliability curves for the standard sample. (b) As in (a), but for a time tolerance of 12 h instead of 24 h. (c) As in (a), but for a THICK threshold value of 30 m instead of 45 m. (d) As in (a), but for only those disturbances that developed into TCs. (e) As in (a), but for disturbances associated with AEWs. (f) As in (a), but for disturbances clearly not associated with AEWs. The numbers of 12–72-h cases in the 0% bin (bars not plotted) are (a) 961, (b) 946, (c) 739, (d) 579, (e) 630, and (f) 331. The corresponding numbers of 84–144-h cases are (a) 415, (b) 397, (c) 322, (d) 170, (e) 271, and (f) 144.

The statistics in the reliability diagrams are expected to be sensitive to the parameters selected in our study. First, it is reasonable to suggest that our time tolerance is too generous, given that too many cases with a low predicted probability underwent genesis. If the time tolerance is reduced from 24 to 12 h, this underforecasting bias is diminished (cf. Figs. 7b and 7a). However, the number of cases with a high predicted probability of

genesis that actually undergo genesis becomes unacceptably low. While it is subjective, we suggest that a time tolerance of 24 h is the most acceptable. Next, one may also question whether the results are sensitive to the threshold values of CIRC, THICK, or MSLP. For example, the calibrated threshold value of THICK = 45 m may be too strict. If the threshold is reduced to 30 m, meaning a weaker warm core, more cases of higher probability are added to the sample. The corresponding reliability diagram shows no bias in the low-probability predictions for 12–72 h, though the observed frequency is too low for the 40%–60% bin (cf. Figs. 7c and 7a). In contrast, the reliability diagram for 84–144-h forecasts is largely insensitive to the modified threshold value. The slight overall superiority of these results with the more relaxed threshold may imply that our threshold was too strict, though a 30-m threshold also introduces an unacceptable number of "false alarms" in the analysis (e.g., 22/76 instead of the 7/58 listed in Table 1 for 2010).

An argument can be made that the results are sensitive to the types of disturbances. If only those disturbances that actually developed are included in the sample, the slope of the reliability diagram for both 12–72- and 84–144-h forecasts is approximately one, but the underprediction bias is overly strong (Fig. 7d). This result also emphasizes the sensitivity of the reliability diagrams to the few nondevelopers in the sample. There is a level of subjectivity about how to select nondevelopers, which would bring the reliability curve downward whenever any ensemble members predict genesis. Additionally, since we restrict our sample to disturbances that are already an INVEST, the sample of 84–144-h forecasts can be skewed toward long-lived disturbances that do not develop, along with a small number of disturbances that may move out of the domain before they develop. In the future, it will be desirable to add the verification for systems that do not exist at 0 h.

It has been suggested by Halperin et al. (2013) that numerical models may produce better predictions of genesis from an AEW than disturbances of different origins. One may accordingly expect the probabilistic skill to be higher for AEWs. Dividing the sample into waves versus nonwaves (the latter including the "hybrids" in Table 2), the reliability diagram in Fig. 7e is slightly inferior to that in Fig. 7a, suggesting that the removal of nonwaves from the sample does not improve the reliability. The sample of nonwaves is too small to draw conclusions, except that the probability of genesis is underpredicted for the small number of cases (Fig. 7f).

Two questions that stand out from Figs. 6 and 7 are whether the reliability diagrams are significantly different between the 3 yr and whether the difference in reliability is significant when the forecast time is extended from the 12–72- to the 84–144-h window. The variation in skill that exists within the forecast ensemble is also of interest. To address these questions, a conventional bootstrapping method is employed in which the ensemble of 50 ECMWF members for each forecast case is randomly resampled 1000 times. The forecast probabilities and reliability diagrams are then computed for each of the 1000 resampled 50-member ensembles (gray lines in Fig. 8). For each bin of predicted probability in the reliability diagram, the mean value together with the 5% and 95% percentiles of the percentage of verifying tropical cyclones in each bin are then deduced. The distinct lack of overlap between the confidence intervals for the bins centered on probabilities of 30% or

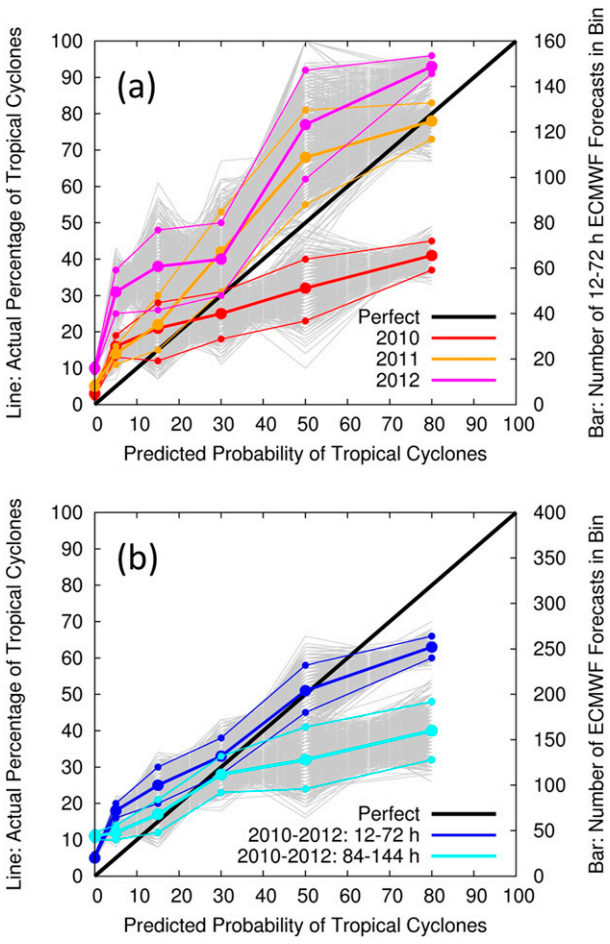

FIG. 8. Reliability diagrams including bootstrap confidence intervals. Gray lines represent reliability curves for 1000 random samples of the 50 ECMWF ensemble members for all forecast cases. The thick colored line is the corresponding mean reliability curve, and the thin colored lines represent the 5th and 95th percentile values. (a) As in Fig. 6b, but for 12–72-h forecasts in each year. (b) As in Fig. 7a, but for 12–72- and 84–144-h forecasts cumulative over the 3 yr.

higher suggests that the results for 2010 are significantly different from those of the two successive years (Fig. 8a). It is also evident from Fig. 8a that the underpredictive 2012 forecasts in the lower probability bins (up to 20%) are significantly inferior to the preceding 2 yr. The corresponding confidence intervals in these bins are also wider in 2012. Otherwise, the variation in skill due to the resampling is generally consistent between the 3 yr. Combining the 3 yr, the minimal overlap between the confidence intervals in the lower probability bins suggests that the 12–72-h probabilistic forecasts are significantly more underpredictive than the 84–144-h forecasts (Fig. 8b). On the other hand, for probabilities exceeding 40%, the 12–72-h forecasts are significantly superior and possess less variation of skill compared with the 84–144-h forecasts.

The samples in our reliability diagrams are grouped into 12–72- and 84–144-h forecasts, both for brevity and to increase the sample size in the bins of larger probabilities (while keeping in mind that the elements in the sample are not independent). The question remains whether the reliability diagrams are dominated by forecasts of any particular time range (e.g., 12–24 h), and whether there is variability for different forecast times. By considering subsamples of 12–24-, 36–48-, and 60–72-h forecasts, their respective reliability diagrams are mostly similar (Fig. 9a). The fact that even for 12–24-h forecasts, about one-third of all forecasts in the 10%–20% probability bin verify as TCs warrants further investigation. In the medium range, the reliability diagrams appear to possess skill up to the 40% probability range for 84–96- and 108–120-h forecasts, but are more or less flat (no skill) for 132–144-h forecasts (Fig. 9b).

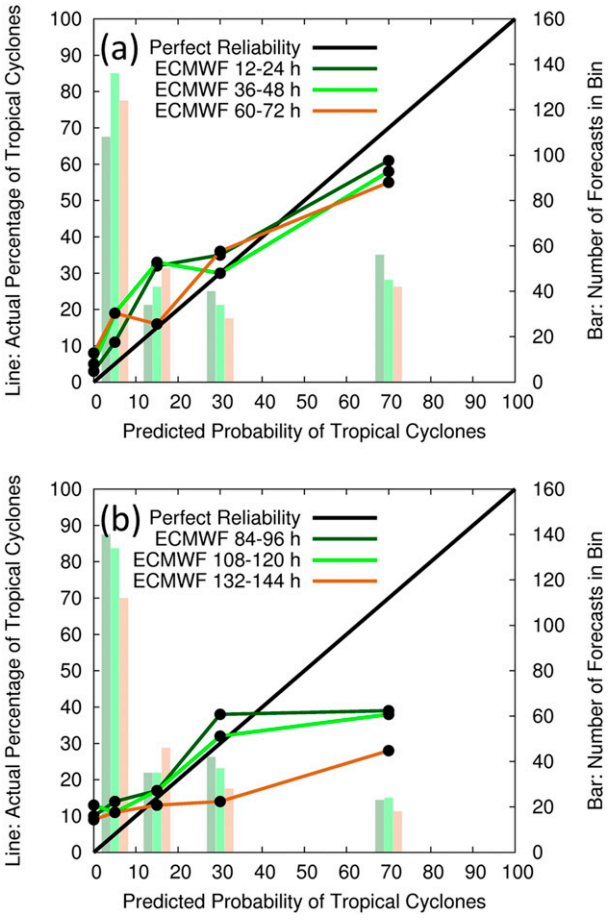

FIG. 9. Evaluations of daily ECMWF ensemble forecasts, cumulative over the 2010–12 Atlantic hurricane seasons, stratified by forecast time. (a) Bars show the numbers of 12–24- (dark green), 36–48- (green), and 60–72-h (bright green) probabilistic forecasts in probability bins of 1%–10%, 11%–20%, 21%–40%, and 41%–100%. Lines show the reliability curves. (b) As in (a), but for 84–96- (dark green), 108–120- (green), and 132–144-h (bright green) forecasts.

5. Investigation of individual cases

The results in the previous sections bring into question the characteristics of the probabilistic predictions on a case-by-case basis. In this section, we examine ensemble forecasts initialized at different times, valid at the 0000 UTC time immediately after genesis had occurred in NHC's best track (within our time tolerance of evaluation).

It is reasonable to expect that an ensemble forecast of a TC valid at its time of genesis would comprise several members with a weak value of circulation (and perhaps a corresponding weak thickness anomaly), while other members would possess values of CIRC, THICK, and MSLP that signify a mature tropical cyclone. The spread of values is expected to be proportional to lead time. One might assume that, on average, the values will be distributed normally about the verifying value, as is the underlying assumption in many ensemble applications such as data assimilation. While this may be a reasonable assumption for variables such as 500-hPa geopotential height in the midlatitudes, the inherent nature of genesis being an extreme event even after an INVEST has been declared suggests that the verification may fall near the tail of the ensemble distribution. As we shall see in the remaining figures, this is mostly the case. Regardless of the assumed distribution, we expect that a prediction of an X% chance of a TC valid at a given time should verify X% of the time, as has been evaluated in the reliability diagrams.

a. "Classical" case: Pre-Isaac (2012)

The 5-day ECMWF ensemble forecast of the disturbance that developed into Tropical Storm Isaac, valid at 0000 UTC 22 August 2012 (18 h after genesis), shows a spread of solutions (Fig. 10a). Several members possess minimal CIRC and THICK, and a corresponding

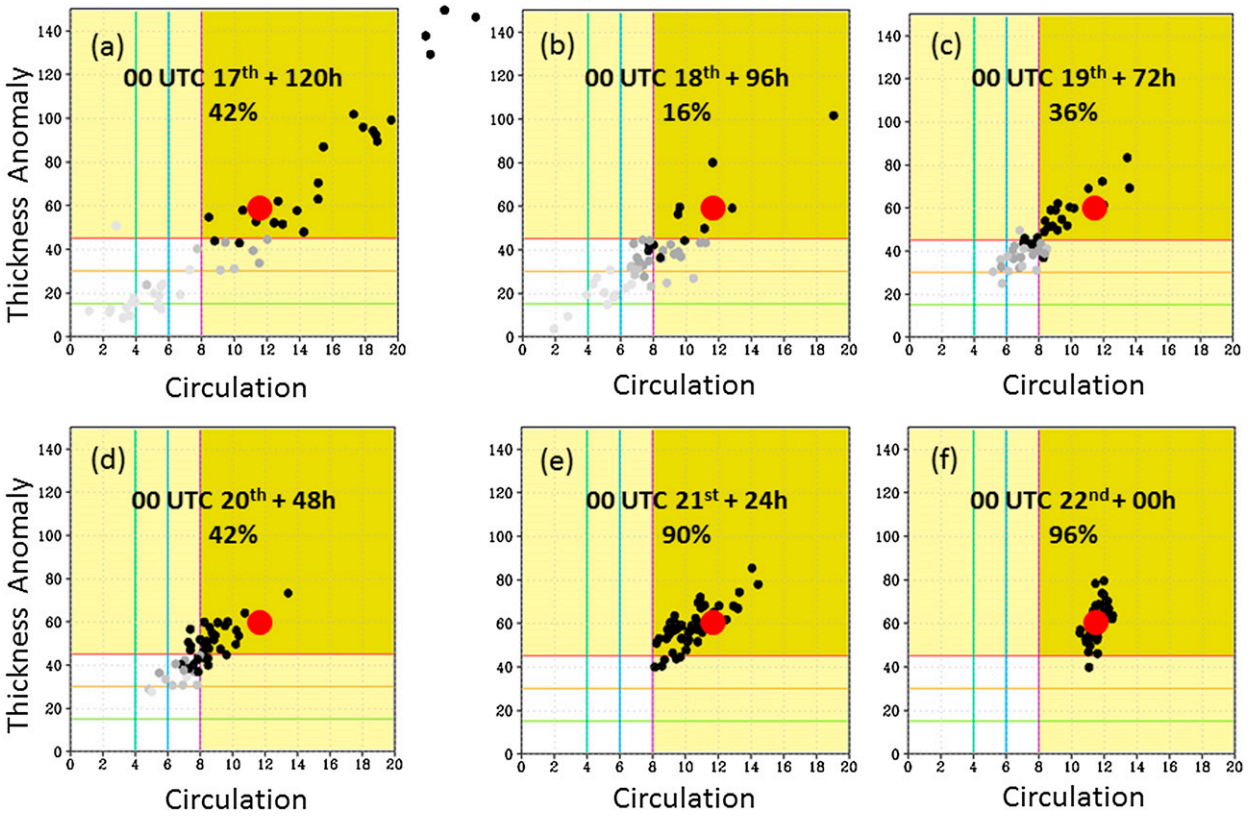

FIG. 10. Phase space for CIRC, THICK, and MSLP for ECMWF ensemble predictions of pre-Isaac (2012): black dots, MSLP < 1012 hPa; medium gray dots, 1012 < MSLP < 1014 hPa; and lightest gray dots, MSLP > 1014 hPa. The red dot shows a verifying value in the ECMWF analysis, valid at 0000 UTC 22 Aug 2012. Forecasts of 0–5 days valid at 0000 UTC 22 Aug 2012 are shown, with the corresponding probabilities of genesis (percentage of black dots within dark yellow area).

weak MSLP (light gray). On the other hand, four solutions suggest values consistent with a strong hurricane. For the 4-day ensemble forecast initialized 1 day later, this probability decreases to 16%, with many ensemble members narrowly failing to meet the three criteria (e.g., the cluster of gray dots; Fig. 10b). For ensembles initialized on subsequent days and valid at the same time, the forecast probability of a TC increases, while the spread of the solutions is diminished (Figs. 10c–e). Between the 2- and 1-day lead times, the probability increases abruptly from 42% to 90%, as many solutions shift from values slightly lower than the thresholds to values slightly exceeding the thresholds, converging toward the verifying value in the ECMWF analysis (red dot). In this case, the abrupt change in probability is not due to a drastic change in the ensemble forecasts initialized 1 day apart, but as a consequence of the ''on–off'' nature of the thresholds. At the verifying time, the ECMWF ensemble members are centered on the value in the verifying analysis (Fig. 10f), though with a narrow spread of CIRC and a large spread of THICK, which differs from the shallower slope of correlation between the two quantities in the ensemble forecasts. Though not shown, the ensemble forecasts initialized at all times prior to genesis exhibited a steady development of a low-level circulation and warming of the inner core with time. This case represents one of a handful of TCs in our sample in which the genesis metrics in the ensemble members progress mostly steadily toward the value of the verifying analysis as the initial time approaches the verification time.

b. *Variability in 2012 cases*

A concern in 2012 was the underprediction of the probability of genesis, as illustrated in the reliability diagrams in Fig. 6b. An illustration of several cases, all of which developed from AEWs, highlights the challenges. First, for the disturbances that developed into Florence and Helene, the analysis values suggest a weak warm core shortly after genesis, with the thickness anomaly not exceeding 45 m (Figs. 11a,b). The majority of 1-day (and even 2 day) ensemble forecast values of THICK are smaller than this weak verifying value. Accordingly, even though genesis was declared by NHC within a day,

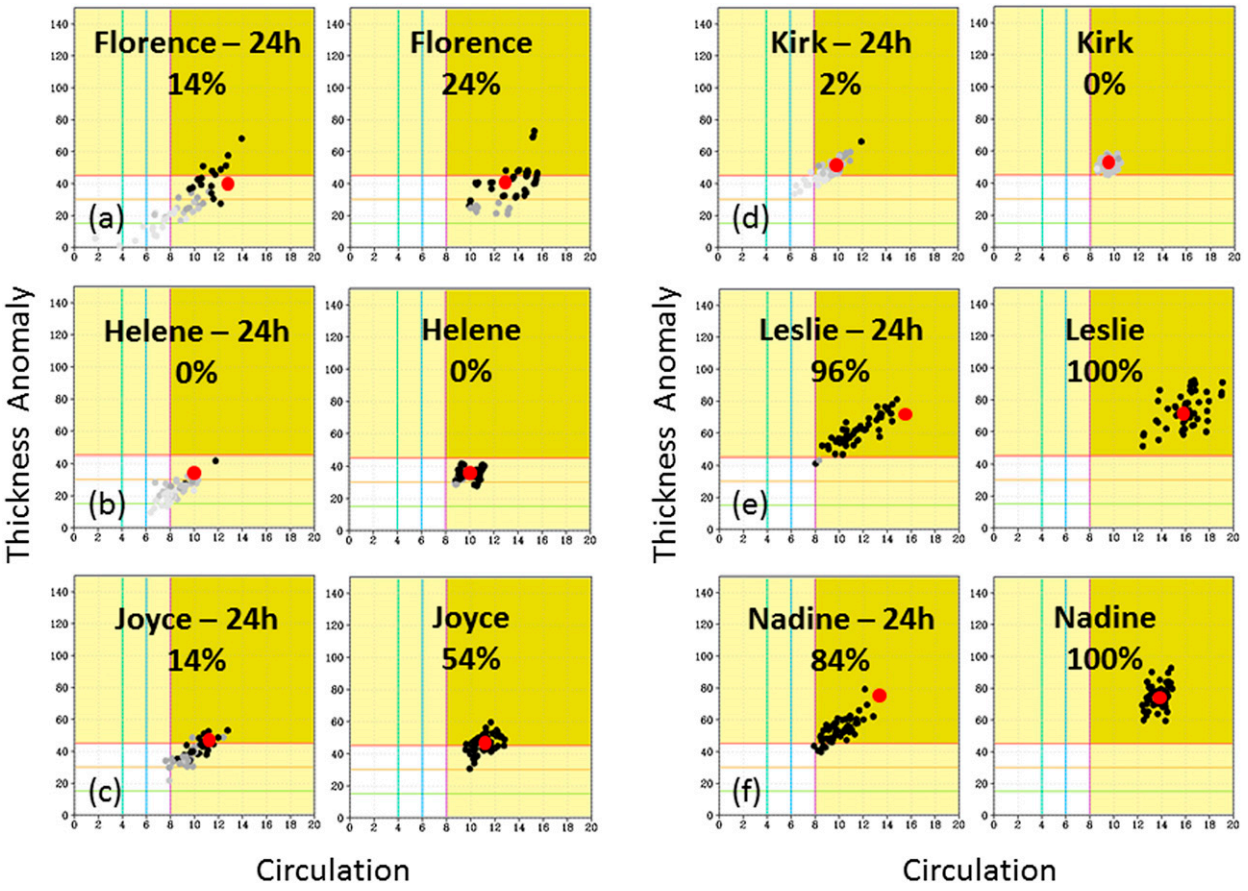

FIG. 11. Phase space for CIRC, THICK, and MSLP for ECMWF ensemble predictions of six disturbances that developed into tropical cyclones in 2012. Red dot shows the verifying value in the ECMWF analysis, valid at 0000 UTC at or immediately after genesis. Ensemble analyses and 24-h forecasts valid at this verification time are shown, with the corresponding probabilities of genesis (percentage of black dots within the dark yellow area). The colored lines of constant CIRC and THICK correspond to the contour values of CIRC = 4×10^{-5}, 6×10^{-5}, and $8 \times 10^{-5}\,s^{-1}$ and THICK = 15, 30, and 45 m, consistent with the colors of the spaghetti diagrams in Figs. 1 and 2.

very low probabilities of genesis were forecast. The case for Joyce is similar to those for Florence and Helene, with a slightly stronger warm core but similarly low 1-day (and 2 day) forecast probabilities (Fig. 11c). For Kirk, the probabilities are tiny due to the weak value of MSLP in virtually all the forecast and analysis ensemble members (Fig. 11d). In contrast, the analysis values in Leslie and Nadine easily exceed the threshold values (Figs. 11e,f). The 1-day ensemble forecast members, while again distributed toward weaker values, are sufficiently well developed to provide a high probability of the existence of a TC, resulting in many hits. Finally, an examination of the radius of the last closed isobar in the NHC Best Track dataset for each case reveals that the anomalously low probabilities are often associated with smaller cyclones at the time of genesis [Helene, Karl, and Michael, all <160 nautical miles (n mi; 1 n mi = 1.852 km)], while the more reliable cases are larger (Isaac, Leslie, and Nadine, exceeding 200 n mi). An accurate representation of the smaller developing cyclones is likely compromised substantially by the coarse grid spacing of the ensemble dataset.

c. *Nondeveloper: Pouch PGI24L (2010)*

A memorable nondeveloping AEW is pouch PGI24L, for which the NHC had predicted a chance of genesis as high as 70%. Many members of the ECMWF ensemble concur (Fig. 12a). A majority of the ensemble exceeds the threshold values of CIRC (Fig. 12b), though several members are less aggressive with the thickness anomaly (Fig. 12c). Contrary to the 2012 cases in Figs. 10 and 11, the analyzed values of CIRC and THICK are weaker than many ensemble members.

6. Conclusions

Probabilistic forecasts of tropical cyclogenesis were evaluated for

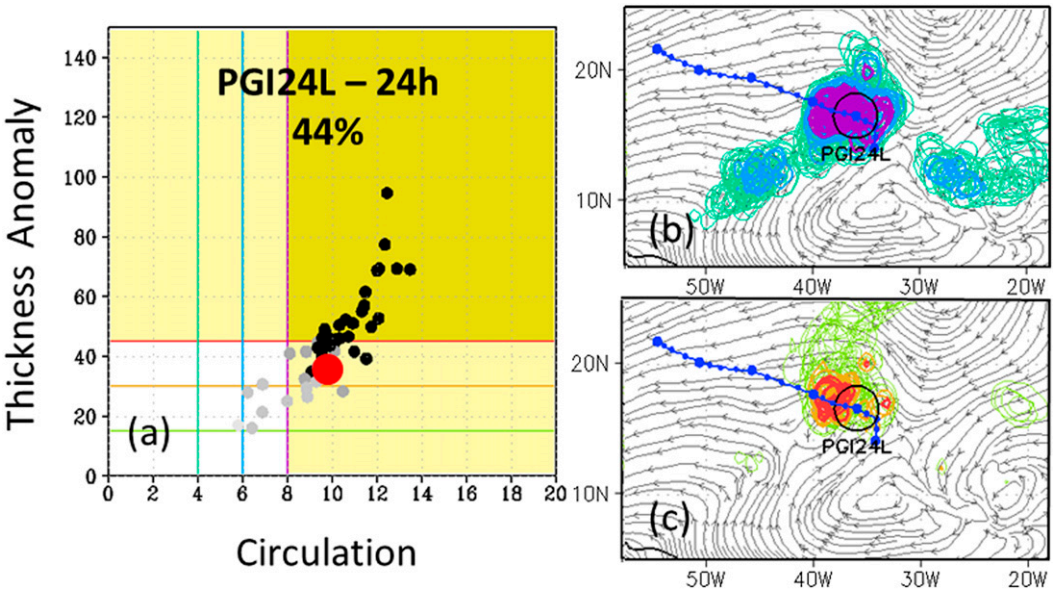

FIG. 12. (a) Phase space for CIRC, THICK, and MSLP for 24-h ECMWF ensemble prediction of nondeveloping pouch PGI24L, initialized at 0000 UTC 6 Aug 2010. Red dot shows the verifying value in the ECMWF analysis, valid at 0000 UTC 7 Aug 2010. Shown are corresponding ensemble spaghetti diagrams of (b) CIRC and (c) THICK.

- a near-homogeneous, twice-daily sample of ECMWF and WRF–EnKF ensemble forecasts during the PREDICT period of 15 August–30 September 2010 and
- daily ECMWF ensemble forecasts during 1 June–30 November for 2010–12.

Prior to the evaluation, quantitative criteria for TC genesis were deduced in numerical model analysis fields, whose genesis times based on these criteria were most consistent with the actual genesis times declared by NHC. The classification of a TC based on threshold values of three variables was found to be sufficient: 700–850-hPa circulation (CIRC), a local 200–850-hPa thickness anomaly (THICK), and an SLP minimum (MSLP). Although the values of these variables at genesis vary from case to case, fixed threshold values were necessary since it is impossible to know the values at the time of a future genesis event. A binary verification was performed for all ensemble forecasts, for a disturbance for which NHC had issued a 48-h genesis probability exceeding 30% at any point during its existence.

For the 6-week sample during the PREDICT period, the ECMWF and WRF–EnKF ensembles produced similar results. The predictions for this sample were largely unbiased. The reliability diagrams were similar for both ensembles, with a slope that was generally increasing for greater probabilities but less than unity. The ROC curves were similar and demonstrated skill.

For the ECMWF ensemble during 2010–12, the ETS was comparable in each year with a value of around 0.2, decreasing to near zero (no skill) after 5 days. The reliability diagram for 12–72-h forecasts was monotonic increasing in each year, though the slope was significantly flatter in 2010. Particularly in 2012, there were too many cases in which a low probability of a TC was predicted, yet genesis occurred in reality. The ROC curves in each year exhibited skill for 12–72-h forecasts, and diminished yet positive skill for 84–144-h forecasts, suggesting that the hit rate exceeded the false alarm rate over all probability thresholds for medium-range forecasts. Combining the 3 yr into one sample, the reliability diagrams for 12–72-h forecasts were found to be superior to those for 84–144-h forecasts in the range of higher probabilities, though marginal skill in reliability was still achieved for forecasts out to 120 h.

The ECMWF reliability curves were sensitive to various parameters. One parameter was the time tolerance, the choice of which is admittedly subjective. Sensitivity to the threshold values was also found, and it can be argued that the ideal threshold in the analysis field may not directly equate to the ideal threshold in the model forecast. There was also sensitivity to the number of subjectively chosen developers and nondevelopers. However, there was not a distinct difference in the reliability diagrams between those cases that were associated with AEWs, and those that were not. Assessing the sensitivity to the forecast times, the reliability diagrams were mostly similar for forecasts of 1–3 days, and while they declined afterward, they still possessed skill out to 5 days.

An examination of the individual cases revealed strengths and deficiencies in the probabilistic predictions. For several cases, the threshold values were not met for all three variables in many short-range forecasts, yielding a forecast of low probability even shortly after genesis. This suggested an inability of the ensemble members to build the circulation and warm core over a short period, and an associated lack of dispersiveness in the ensemble. In some cases, the analysis values were weak at and after genesis. In contrast, other cases showed a gradually increasing probability of genesis, approaching 100% as the initial time approached the genesis time. In the majority of cases, the distribution of CIRC and THICK in the ensemble was skewed toward values lower than the verifying analysis. This may be a consequence of the model struggling to develop a warm core and a sufficient circulation, and/or the notion that genesis is an extreme event, and therefore it should always be near the tail of the distribution even in a perfect-model scenario.

It is important not to draw overly broad conclusions from these results, given the sensitivity to the cases in the small sample and the threshold values. Overall, both the ECMWF and WRF–EnKF ensembles showed potential for probabilistic prediction out to 3–5 days, though there is room for improvement, particularly in building a warm core in the short-range forecast, which may be a consequence of the model resolution and physics. The results demonstrated the inherent difficulty in defining and predicting genesis in models, particularly the higher moments that require dispersiveness in the ensemble. While we believe that we have made progress in this direction, further work is necessary to determine whether improved metrics that are insensitive to the resolution and physics of the model can be devised. After such improvements have taken place and been evaluated rigorously over a large sample (including cases in which no disturbance is detected at the initial time), newly calibrated ensemble predictions are expected to contribute toward increasingly robust operational predictions of genesis probabilities.

Acknowledgments. The authors gratefully acknowledge funding from National Science Foundation Grants ATM-0848753 and ATM-0848409. We thank Mark Boothe for providing pouch track data, and the PREDICT team, several staff members at ECMWF, and Will Komaromi for helpful discussions. The first author thanks ECMWF for permission to use the ensemble data during PREDICT, Bill Lapenta and NCEP for facilitating access to these data, and TIGGE for archived data. The second author would like to acknowledge high-performance computing support provided by the National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory, which is sponsored by the National Science Foundation. Comments from three anonymous reviewers improved the paper.

REFERENCES

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.,* **90,** 1283–1296, doi:10.1175/2009BAMS2618.1.

Barker, D. M., W. Huang, Y.-R. Guo, A. Bourgeois, and Q. Xiao, 2004: A three-dimensional variational data assimilation system for MM5: Implementation and initial results. *Mon. Wea. Rev.,* **132,** 897–914, doi:10.1175/1520-0493(2004)132<0897:ATVDAS>2.0.CO;2.

Berner, J., G. Shutts, M. Leutbecher, and T. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.,* **66,** 603–626, doi:10.1175/2008JAS2677.1.

Blake, E. S., 2013: Experimental 5-day genesis forecasts. *2013 National Hurricane Conf.,* New Orleans, LA, NOAA/NHC. [Available online at http://www.nhc.noaa.gov/outreach/presentations/nhc2013_nhcTWO5day.pdf.]

Camargo, S. J., and S. E. Zebiak, 2002: Improving the detection and tracking of tropical storms in atmospheric general circulation models. *Wea. Forecasting,* **17,** 1152–1162, doi:10.1175/1520-0434(2002)017<1152:ITDATO>2.0.CO;2.

Cheung, K. K. W., and R. L. Elsberry, 2002: Tropical cyclone formations over the western North Pacific in the Navy Operational Global Atmospheric Prediction System forecasts. *Wea. Forecasting,* **17,** 800–820, doi:10.1175/1520-0434(2002)017<0800:TCFOTW>2.0.CO;2.

Dunkerton, T. J., M. T. Montgomery, and Z. Wang, 2009: Tropical cyclogenesis in a tropical wave critical layer: Easterly waves. *Atmos. Chem. Phys.,* **9,** 5587–5646, doi:10.5194/acp-9-5587-2009.

Dupont, T., M. Plu, P. Caroff, and G. Faure, 2011: Verification of ensemble-based uncertainty circles around tropical cyclone track forecasts. *Wea. Forecasting,* **26,** 664–676, doi:10.1175/WAF-D-11-00007.1.

Elsberry, R. L., M. S. Jordan, and F. Vitart, 2010: Predictability of tropical cyclone events on intraseasonal timescales with the ECMWF monthly forecast model. *Asia-Pac. J. Atmos. Sci.,* **46,** 135–153, doi:10.1007/s13143-010-0013-4.

——, ——, and ——, 2011: Evaluation of the ECMWF 32-day ensemble predictions during 2009 season of western North Pacific tropical cyclone events on intraseasonal timescales. *Asia-Pac. J. Atmos. Sci.,* **47,** 305–318, doi:10.1007/s13143-011-0017-8.

——, H.-C. Tsai, and M. S. Jordan, 2014: Extended-range forecasts of Atlantic tropical cyclone events during 2013 using the ECMWF 32-day ensemble predictions. *Wea. Forecasting,* **29,** 271–288, doi:10.1175/WAF-D-13-00104.1.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.,* **94,** 329–343, doi:10.1175/BAMS-D-12-00071.1.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting,* **28,** 1423–1445, doi:10.1175/WAF-D-13-00008.1.

Hamill, T. M., J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011: Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Mon. Wea. Rev.,* **139,** 668–688, doi:10.1175/2010MWR3456.1.

Komaromi, W. A., and S. J. Majumdar, 2014: Ensemble-based error and predictability metrics associated with tropical cyclogenesis. Part I: Basinwide perspective. *Mon. Wea. Rev.,* **142,** 2879–2898, doi:10.1175/MWR-D-13-00370.1.

Lang, S. T. K., M. Leutbecher, and S. C. Jones, 2012: Impact of perturbation methods in the ECMWF Ensemble Prediction System on tropical cyclone forecasts. *Quart. J. Roy. Meteor. Soc.,* **138,** 2030–2046, doi:10.1002/qj.1942.

Majumdar, S. J., and P. M. Finocchio, 2010: On the ability of global ensemble prediction systems to predict tropical cyclone track probabilities. *Wea. Forecasting,* **25,** 659–680, doi:10.1175/2009WAF2222327.1.

Marchok, T., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology,* San Diego, CA, Amer. Meteor. Soc., P1.13. [Available online at https://ams.confex.com/ams/pdfpapers/37628.pdf.]

Montgomery, M. T., and Coauthors, 2012: The Pre-Depression Investigation of Cloud Systems in the Tropics (PREDICT) Experiment: Scientific basis, new analysis tools, and some first results. *Bull. Amer. Meteor. Soc.,* **93,** 173–187, doi:10.1175/BAMS-D-11-00024.1.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp.

Peng, M. S., B. Fu, T. Li, and D. E. Stevens, 2012: Developing versus nondeveloping disturbances for tropical cyclone formation. Part I: North Atlantic. *Mon. Wea. Rev.,* **140,** 1047–1066, doi:10.1175/2011MWR3617.1.

Sippel, J. A., and F. Zhang, 2008: A probabilistic analysis of the dynamics and predictability of tropical cyclogenesis. *J. Atmos. Sci.,* **65,** 3440–3459, doi:10.1175/2008JAS2597.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp. [Available online at http://www.mmm.ucar.edu/wrf/users/docs/arw_v3_bw.pdf.]

Snyder, A., Z. Pu, and Y. Zhu, 2010: Tracking and verification of the east Atlantic tropical cyclone genesis in NCEP global ensemble: Case studies during NASA African Monsoon Multidisciplinary Analyses. *Wea. Forecasting,* **25,** 1397–1411, doi:10.1175/2010WAF2222332.1.

Torn, R. D., and D. Cook, 2013: Role of vortex and environment errors in genesis forecasts of Hurricanes Danielle and Karl (2010). *Mon. Wea. Rev.,* **141,** 232–251, doi:10.1175/MWR-D-12-00086.1.

——, G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.,* **134,** 2490–2502, doi:10.1175/MWR3187.1.

Tory, K. J., and W. M. Frank, 2010: Tropical cyclone formation. *Global Perspectives on Tropical Cyclones: From Science to Mitigation,* J. C. L. Chan and J. D. Kepert, Eds., World Scientific, 55–91.

——, S. S. Chand, R. A. Dare, and J. L. McBride, 2013: The development and assessment of a model-, grid-, and basin-independent tropical cyclone detection scheme. *J. Climate,* **26,** 5493–5507, doi:10.1175/JCLI-D-12-00510.1.

Tsai, H.-C., R. L. Elsberry, M. S. Jordan, and F. Vitart, 2013: Objective verifications and false alarm analyses of western North Pacific tropical cyclone event forecasts by the ECMWF 32-day ensemble. *Asia-Pac. J. Atmos. Sci.,* **49,** 409–420, doi:10.1007/s13143-013-0038-6.

Vitart, F., F. Prates, A. Bonet, and C. Sahin, 2012: New tropical cyclone products on the web. *ECMWF Newsletter,* No. 130, ECMWF, Reading, United Kingdom, 17–23.

Wang, Z., M. T. Montgomery, and T. J. Dunkerton, 2010: Genesis of pre–Hurricane Felix (2007). Part II: Warm core formation, precipitation evolution, and predictability. *J. Atmos. Sci.,* **67,** 1730–1744, doi:10.1175/2010JAS3435.1.

Yamaguchi, M., T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quart. J. Roy. Meteor. Soc.,* **138,** 2019–2029, doi:10.1002/qj.1937.