ABSTRACT

Several metrics are employed to evaluate predictive skill and attempt to quantify predictability using the ECMWF Ensemble Prediction System during the 2010 Atlantic hurricane season, with an emphasis on large-scale variables relevant to tropical cyclogenesis. These metrics include the following: 1) growth and saturation of error, 2) errors versus climatology, 3) predicted forecast error standard deviation, and 4) predictive power. Overall, variables that are more directly related to large-scale, slowly varying phenomena are found to be much more predictable than variables that are inherently related to small-scale convective processes, regardless of the metric. For example, 850–200-hPa wind shear and 200-hPa velocity potential are found to be predictable beyond one week, while 200-hPa divergence and 850-hPa relative vorticity are only predictable to about one day. Similarly, area-averaged quantities such as circulation are much more predictable than nonaveraged quantities such as vorticity. Significant day-to-day and month-to-month variability of predictability for a given metric also exists, likely due to the flow regime. For wind shear, more amplified flow regimes are associated with lower predictive power (and thereby lower predictability) than less amplified regimes. Relative humidity is found to be less predictable in the early and late season when there exists greater uncertainty of the timing and location of dry air. Last, the ensemble demonstrates the potential to predict error standard deviation of variables averaged in $10^8 \times 10^8$ boxes, in that forecasts with greater ensemble standard deviation are on average associated with greater mean error. However, the ensemble tends to be underdispersive.

1. Introduction

The prediction of the genesis of tropical cyclones (TCs) has become increasingly accurate in recent years, due to advancements in numerical modeling and data assimilation (Halperin et al. 2013). In 2013, the National Hurricane Center (NHC) introduced a new operational product issuing the probability of tropical cyclogenesis within 5 days, underlying the increased confidence in genesis prediction (genesis probabilities out to 5 days are provided every 6 h by the NHC online at http://www.nhc.noaa.gov/gtwo_atl.shtml). Additionally, several recent studies have assessed the ability of ensemble forecasts to accurately predict genesis in medium-range forecasts (e.g., Snyder et al. 2010, 2011). Tsai et al. (2013) even find that the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble demonstrates some skill in predicting genesis out to several weeks. Despite a recent uptick in studies that examine the predictability of whether or not genesis will occur, few if any examine the predictability of the pregenesis environment.

The predictive skill and the intrinsic predictability of tropical cyclogenesis (and intensity change) are largely influenced by a range of tropospheric conditions, both associated with the incipient vortex and the environmental flow with which it is interacting. To the best of our knowledge, no published study has explored the predictability of such variables directly relevant to cyclogenesis in the tropical Atlantic. One of the goals of the National Science Foundation’s Pre-Depression Investigation of Cloud-systems in the Tropics (PREDICT; Montgomery et al. 2012) field campaign and research initiative has been to identify the extent to which genesis is predictable and the factors that limit its predictability. In this paper, we use a variety of metrics as a proxy for quantifying predictability on a basinwide scale, using the ECMWF ensemble prediction system for the 2010
Atlantic hurricane season. The predictability associated directly with individual tropical waves and genesis events is the subject of a companion paper. While it is premature in these papers to offer rigorous conclusions about predictability with only a single hurricane season, we hope to offer a pathway toward understanding predictive skill and predictability through a first examination of a variety of ensemble-based metrics.

To attempt to quantify predictability, several metrics have been proposed in the literature, though all possess limitations and are dependent on the quality of the predictive model. One basic method was introduced in the famous study by Lorenz (1982), in which globally averaged errors of 500-hPa geopotential height forecasts from the ECMWF global model were computed to estimate lower and upper bounds of predictability. Several conclusions from this study and others by Lorenz (1963, 1965, 1969) are well-held tenets of predictability. Error growth was found to be nonlinear in a dynamical system, generally growing most rapidly during the first 24 h of model integration due to errors at small scales. Thereafter, error growth begins to slow as errors of increasingly larger scales begin to saturate. The total error finally begins to asymptote toward a constant value after errors at all scales have saturated. At this point, which may be several weeks, any individual forecast has become completely indistinguishable from a random state of the system/atmosphere, and predictability is said to have been lost. A related major conclusion that is of particular relevance to this paper is that the predictability of the flow is dependent on both the scales of motion and the flow regime (Lorenz 1965, 1969).

In practice, the growth and saturation of errors with respect to different phenomena with complex physics is not well established. For example, it may be feasible for initially large error structures to transfer to smaller scales with time, which would essentially be a reversal of the classical Lorenz conceptual models. The initial error structures may be a consequence of the data assimilation technique and the observational data assimilated, and the ensemble initialization and model perturbation techniques.

Even when errors have not yet saturated at all scales, a prediction is not useful to forecasters unless its errors are lower than climatology, which is typically employed as the default “no skill” forecast in the absence of any additional information. Several studies have attempted to quantify the net predictability of a system by taking the ratio between the root-mean-square (RMS) error of the model prediction and the climatology (Shukla 1981, 1985; Hayashi 1986; Murphy 1988; Griffies and Bryan 1997). Many of these studies assumed the error distribution to be Gaussian, although Anderson and Stern (1996) demonstrate that this is not always the case. Another caveat in estimating the predictability by these means is that results are dependent upon the choice of variable(s) (Palmer 1996; Palmer et al. 1998). While many classical studies quantify errors associated with geopotential height at 500 hPa, the use of height at a different pressure level or a different variable such as wind may yield a different interpretation.

The above two measures can be applied to a single deterministic forecast with a point verification. However, given that the uncertainty of a forecast is an essential component of modern-day prediction and is also a direct estimate of the predictability of a given flow, it is imperative to devise and evaluate relevant measures. Predicted uncertainty in ensemble forecasts is commonly computed via the ensemble forecast variance and its skill is evaluated against the variance or standard deviation of forecast errors (e.g., Wang and Bishop 2003; Grimit and Mass 2007). In a recent study by McMurdie and Ancell (2014), predictability is defined as the ensemble spread of surface pressure at the final forecast time, where large spread means low predictability. However, the aforementioned study focuses on mid-latitude rather than tropical predictability.

Another metric that has been used more commonly in longer-range predictions is predictive power (PP; Schneider and Griffies 1999). The PP quantifies the forecast uncertainty by comparing the probability density function (PDF) based on an ensemble forecast with a climatological PDF, which represents a distribution of all historical outcomes. Predictability is said to be lost once the forecast PDF has become broader than the climatological PDF. Related metrics such as relative entropy (Kleeman 2002) and mutual information (Kleeman 2007) have also been devised. Several studies have utilized PP and similar methods to assess predictability (e.g., Stephenson and Doblas-Reyes 2000; Schneider and Held 2001; Waliser et al. 2003; DelSole 2004; Abramov et al. 2005; Giannakis and Majda 2012). However, in all the cited studies, computational constraints restricted the resolution at which PP and related metrics could be calculated. Also, the vast majority of these studies focused on more slowly varying systems than those considered in this study, such as the general circulation of the ocean or climate systems.

In this paper, we extend the error, error variance, and predictive power metrics into the domain of medium-range atmospheric predictions over the tropical Atlantic Ocean, using an operational ensemble of relatively high resolution (25 km) and relatively sophisticated physics compared with earlier studies. The primary purpose is to establish a basis for understanding predictability, via quantitative measures of a spectrum of kinematic variables.
relevant to tropical cyclogenesis. In section 2, the variables and metrics are formally introduced. In sections 3 and 4 the results using traditional metrics of RMS error and error growth and predictive power are presented, respectively. Conclusions are provided in section 5.

2. Methodology

a. ECMWF ensemble and climatology

Analyses and 10-day forecasts from the operational 50-member ECMWF Ensemble Prediction System at T639 L62 resolution, interpolated onto a horizontal grid with 0.25° spacing are used in this study. The period considered is from 1 June to 30 November 2010. The ensemble data, produced twice daily, were provided via the Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) database (Bougeault et al. 2010; http://tigge.ecmwf.int/). The ensemble is initialized through a combination of (i) initial-time dry global singular vectors (SVs) maximized over the Northern Hemisphere, (ii) an “ensemble of data assimilations” to create initial spread in the ensemble, and (iii) initial-time tropical SVs with a dry total energy norm and additional linearized physics relevant to tropical processes, optimized over the Caribbean area (0°–25°N, 100°–60°W). The SVs are not optimized on regions centered on tropical waves and disturbances prior to their existence as TCs. Model error perturbations are included via the stochastic kinetic-energy backscatter scheme (SKEB; Berner et al. 2009) and the stochastic perturbed parameterization tendencies scheme (SPPT; Palmer et al. 2009).

The choice of variables was restricted by availability within TIGGE, as well as available disk space. Fields of geopotential height (denoted by $Z$) at the 200-, 500-, and 850-hPa pressure levels, together with horizontal wind components at 200, 700, and 850 hPa, and 700-hPa relative humidity were archived. The following quantities are used in this paper:

1) Relative vorticity: 850 hPa;
2) Vertical shear of horizontal wind: 850–200 and 850–500 hPa;
3) Divergence: 200 hPa;
4) Relative humidity: 700 hPa;
5) Velocity potential: 200 hPa;
6) Circulation: average 850–700-hPa relative vorticity within 200 km of each grid point; and
7) Thickness anomaly, defined by

$$\Delta Z = Z_{r=100\, \text{km}} - Z_{r=1000\, \text{km}}. \quad (1)$$

The first four variables are conventionally used in various diagnostics relevant to tropical meteorology. Positive relative vorticity, low vertical wind shear, positive upper-level divergence/negative low-level divergence, and high ambient relative humidity have all been found to be associated with TC formation (Gray 1979). Specific threshold ranges have been mapped to genesis parameters that incorporate these variables either directly or indirectly (DeMaria et al. 2001), such as through instability calculations, and are included in operational products produced at the Colorado State Regional and Mesoscale Meteorology Branch (RAMMB) to predict genesis probabilities (Schumacher et al. 2009). The fifth, velocity potential, describes the irrotational flow, with the field being smoother and more slowly varying than divergence in both space and time. For these reasons, it is hypothesized that velocity potential will prove to be more predictable than divergence. Velocity potential $\phi$ is computed via

$$\delta = \nabla^2 \phi, \quad (2)$$

where $\delta$ is the horizontal divergence and the Laplacian is inverted using a successive overrelaxation technique (Bijlisma et al. 1986). The advantage of using velocity potential is that it preferentially emphasizes the planetary-scale aspects of the divergent circulation while smoothing theconvective-driven component (Ventrice et al. 2013).

The circulation and thickness anomaly were chosen since they were found to be relatively unambiguous in objectively identifying a tropical disturbance or cyclone in numerical model fields (Majumdar and Torn 2014, manuscript submitted to Wea. Forecasting). First, the circulation was found to be easier to unambiguously track temporally than relative vorticity, given its distinct local maxima in easterly waves and gyres compared with multiple filaments with adjacent local maxima of vorticity. Second, the 850–200-hPa thickness anomaly $\Delta Z$ yields a distinct and unambiguous signal of a local warm anomaly, which may correspond to a warm core building in a developing tropical cyclone. The exceedance of circulation and thickness anomaly by given threshold values in ECMWF analyses (investigated in Majumdar and Torn 2014, manuscript submitted to Wea. Forecasting) was found to be consistent with whether the tropical disturbance had attained tropical cyclone status (as declared by NHC), with relatively few “misses” (weak circulation and thickness anomaly, TC declared) or “false alarms” (strong circulation and thickness anomaly, no TC).

The domain used throughout this study comprises the Atlantic Ocean from 0°–25°N and 10°–100°W (Fig. 1a).
Attention in this study is emphasized on tropical weather, incorporating tropical waves, dry air masses originating from Africa, and gyres and inverted troughs. Less emphasis is given to conditions that may lead to cyclogenesis of disturbances of baroclinic origin or of a subtropical nature. As will be discussed later, the latitudinal extent is restricted further to prevent baroclinic processes from contributing substantially to the variables under investigation.

It should be noted that all calculations include a land mask to ensure that results are relevant to only the tropical Atlantic without contamination from South America or the east Pacific. The geographic area excluded by the land mask is depicted in Fig. 1a (anything west of the red line). However, it should be noted that results in sections 3 and 4 were not found to be particularly sensitive to whether or not the land mask was included.

Finally, the climatology is generated using 33 years of Interim ECMWF Re-Analysis (ERA-Interim) data from 1979 to 2011. ERA-Interim is a reanalysis dataset that uses the ECMWF data assimilation system and forecast model from 2006, but with T255 spectral resolution and 60 vertical levels (Dee et al. 2011). The climatological data, originally at the equivalent of 0.703° × 0.702° resolution on a Gaussian grid, are interpolated to a 0.25° fixed grid using spherical harmonics.

b. Metrics to quantify error growth and predictability
1) RMS ERROR AND ERROR VARIANCE

In this paper, the RMS error $e(t)$ of the control forecast $x^f(t)$ (an ensemble member integrated from the unperturbed analysis) is computed with respect to the corresponding ECMWF analysis $x^a(t)$ (as in Figs. 4 and 5).
The ensemble mean is also compared against the average RMS error of a corresponding “forecast” that simply uses climatology.

The ensemble offers additional information via the PDF. The most basic higher moment is the ensemble variance, which serves as a prediction of the forecast error variance $\sigma^2_{N-1}$:

$$
\sigma^2_{N-1}(t) = \frac{1}{N-1} \sum_{i=1}^{N} [x_i^f(t) - \bar{x}^f(t)]^2,
$$

where $N$ is the total number of ensemble members (50 members plus 1 control run here), and $\bar{x}^f(t)$ is the ensemble mean forecast of $x$ at time $t$. Rather than computing the forecast variance for individual tropical waves or cyclones (e.g., McMurdie and Ancell 2014), Eq. (3) is calculated over sixteen $10^\circ \times 10^\circ$ regions within $5^\circ$-25$^\circ$N and 100$^\circ$-10$^\circ$W.\footnote{The area from 100$^\circ$-80$^\circ$W south of 15$^\circ$N is excluded to ensure that the east Pacific does not factor into these calculations.}

Several different approaches have been taken to evaluate the reliability of ensemble predictions via their statistical consistency versus actual errors (reviewed in Grimit 2004). In other words, the ensemble needs to be able to discriminate between different error distributions, with forecasts of higher error corresponding to larger error spread in the ensemble predictions. A probabilistic approach that has been adopted over the past decade is to determine whether a relationship exists between the predicted error variance [Eq. (3), or the standard deviation] and the corresponding actual errors in the ensemble mean (e.g., Wang and Bishop 2003). This is accomplished via first plotting the distribution of forecast errors of the ensemble mean versus their corresponding predictions of the variance (or standard deviation) of the ensemble mean, for each forecast case. By averaging the data points in equally sized bins of increasing predicted forecast error variance, the variance (or standard deviation) of the actual errors is computed for each bin. Ideally, a linear, increasing relationship of slope 1 between the predicted and actual error variance (or standard deviation) is found. In this paper, we elect to plot the error of the ensemble mean, retaining the sign of the errors, versus the ensemble standard deviation, following the bottom panel of Fig. 5 of Grimit and Mass (2007). In this way, the distribution of forecast errors, including any biases and departures from Gaussianity are illustrated, while retaining the ability to formally evaluate the relationship between the predicted and actual error range.

2) PREDICTIVE POWER

The PP takes advantage of information contained in ensemble forecasts to quantify the difference between the ensemble distribution and a reference climatological distribution (Schneider and Griffies 1999). It is defined as

$$
PP = 1 - e^{-R_v},
$$

where the predictive information $R_v$ is calculated as

$$
R_v = -\frac{1}{2m} \log \left( \frac{\det \Sigma}{\det C_v} \right),
$$

and $m$ is the dimension of the state space and $C_v$ and $\Sigma$ are the sample covariance matrix of the ensemble perturbations and climatological covariance matrix, respectively. The sample covariance matrix of the ensemble perturbations is given by

$$
C_v = \frac{1}{M-1} E_v (E_v)^T,
$$

where $M = 51$ ensemble members, and $E_v = (e_1^v \, e_2^v \, \cdots \, e_M^v) = (x_1^v - \bar{x}, x_2^v - \bar{x}, \cdots , x_M^v - \bar{x})$ is a matrix of differences between state vector $x^m$ for each ensemble member $m$ and ensemble mean $\bar{x}$, all calculated at lead time $v$. Note that Schneider and Griffies (1999) compute $\Sigma M^{-1} E_v (E_v)^T$ rather than $E_v (E_v)^T$, but they are mathematically identical. Next, the climatological covariance matrix is computed as

$$
\Sigma = \frac{1}{N-1} E_v (E_v)^T,
$$

where $E_v = (e_1 \, e_2 \, \cdots \, e_N) = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots , x_N - \bar{x})$ is a matrix of differences between state vector $x^N$ for each day $n$ and the monthly mean $\bar{x}$. We use the monthly mean instead of the climatological mean so that the difference $x^N - \bar{x}$ does not contain interseasonal variability (here we are only interested in the short term through intraseasonal errors). As forecast errors grow, the numerical values contained in $C_v$ also grow. Schneider and Griffies (1999) suggest that once the determinant of this error covariance matrix exceeds the determinant of $\Sigma$, predictability has been lost.

To compute PP with so many data points (due to a combination of high resolution and large domain size), it is necessary to truncate the computations to find the determinants of $C_v$ and $\Sigma$. Since $E_v (E_v)^T$ is a $[28800 \times 28800]$ matrix, computation of the determinant is expensive and prone to large rounding errors. Instead, we calculate $(E_v)^T E_v$ to produce a $[51 \times 51]$ solution (50 perturbed members plus 1
control), and compute the determinant by taking the product of the eigenvalues \( \lambda \) given by \( \det[-A + (E_y)^T E_y] = 0 \). This can be done because the nonzero eigenvalues of \( E_y (E_y)^T \) and \( (E_y)^T E_y \) are equal (e.g., Golub and Van Loan 1996), and there are only 50 modes of variability in \( C_y \). Therefore, \( m = 50 \) in Eq. (5).

3. Error growth

a. Monthly evolution of variables

To give perspective to the results, we first document the evolution of the relevant variables through the 2010 Atlantic hurricane season, based on ECMWF analysis fields. The evolution of the environment during 2010 is overall consistent with the climatological progression, although conditions were slightly more favorable during peak season than the climatology, likely contributing to anomalously high TC activity that year.

In June, much of the central Atlantic is dominated by relatively high values of 850–200-hPa vertical wind shear (exceeding 15 m s\(^{-1}\)) (Fig. 1b). The shear decreases rapidly through July and remains favorable in the 7.5–12.5 m s\(^{-1}\) range for the majority of August and September (Figs. 1c–e). By October, shear again increases rapidly, especially north of 15°N (Fig. 1f), and high shear values dominate much of the tropical Atlantic in November (Fig. 1g). The 850–500-hPa vertical wind shear exhibits a similar spatial distribution and evolution (not shown).

Comparison to the 1979–2011 climatology suggests that vertical wind shear was anomalously low in the central Caribbean during the 2010 hurricane season. Despite generally more favorable conditions, only one genesis event, Matthew, occurred in this region. However, this area is generally not associated with many genesis events, as persistent low-level mass divergence in this region induces a subsidence regime that weakens convection (Shieh and Colucci 2010).

An examination of 700-hPa relative humidity reveals that the northeast quadrant of the tropical Atlantic is generally driest throughout the season (Fig. 2). This minimum in moisture is particularly pronounced during June and July (Figs. 2a,b) when the Saharan air layer (SAL) is most prominent (Dunion 2011). The basin, and in particular the main development region (MDR) and the Caribbean, possesses the greatest average 700-hPa RH during August–October (Figs. 2c–e), with noticeable drying in the Gulf of Mexico and western Atlantic during October–November as midlatitude intrusions of dry air become increasingly frequent (Figs. 2e,f).

Last, the evolution of 850–700-hPa circulation is examined. Negative values of circulation, corresponding to anticyclonic vorticity, dominate the northern half of the domain during June–August (Figs. 3a–c). Positive values of circulation progressively build with time during the first half of hurricane season, particularly within 10°–20°N, 15°–50°W (Figs. 3c,d), before diminishing rapidly in October–November (Figs. 3e,f). The African
Easterly wave train and tropical cyclones contribute substantially to the mean circulation field. Monthly-mean circulation values at or above $1.5 \times 10^{-5} \text{ s}^{-1}$ in August and September 2010 exceed climatological values as a result of easterly waves that tended to be larger and stronger than average (P. Klotzbach 2014, personal communication). A maximum in cyclonic circulation in the western Caribbean is also evident throughout the 2010 season. The genesis of Tropical Cyclones Alex, Karl, Matthew, Nicole, Paula, and Richard all occurred in this region.

**b. Evolution of errors**

The next stage is to document errors in ensemble-mean forecasts between 0 and 10 days. Between days 2 and 4, errors in circulation tend to grow most quickly in the eastern MDR along the preferred path of easterly waves (Figs. 4a,b). Circulation errors throughout the rest of the basin begin to “catch up” by day 8, with errors along the path of recurving tropical cyclones becoming pronounced (Fig. 4c). Meanwhile, errors in both 700-hPa relative humidity (Figs. 4d–f) and deep-layer shear (not shown) are generally maximized north of 15°N. Differences in the phase, strength, and evolution of midlatitude dry air intrusions, upper-level troughs, and SAL outbreaks are likely culprits. Such events introduce errors in the forecast that rarely occur equatorward of 15°N, where the shear is more persistently low and moisture is more persistently high.

While there are geographically preferred regions for slightly faster error growth, the errors are growing everywhere with increasing lead time. By computing the mean of the forecast errors from June to November at every grid point in our Atlantic domain, we compare error growth characteristics against the hallmark study of error growth in the atmosphere (Lorenz 1982). Our results in Fig. 5 appear qualitatively similar to those of Fig. 1 of Lorenz (1982) who examined globally averaged 500-hPa geopotential height (Fig. 5a), in contrast to our regional averages of circulation, relative humidity, and vertical wind shear (Figs. 5b–d). In each case, the most rapid error growth occurs during the first 24 h of model integration due to the rapid growth of errors on the small scales, then declines as these small-scale errors saturate and/or energy is transferred to larger scales. In both the case of Lorenz (1982) and our study, the error growth rate becomes quasi-linear from around day 2. The error doubling time from this point is about 4 days in Lorenz (1982), whereas using our metrics the errors may saturate beyond 10 days without having doubled. In other words, errors are growing considerably more slowly in the tropics as expected. Beyond 8 days, as the errors at

---

**Fig. 3.** Monthly-mean 850–700-hPa circulation (s$^{-1}$) for 2010 for the months of (a) June, (b) July, (c) August, (d) September, (e) October, and (f) November.
all spatial scales approach a maximum, the error curve slowly begins to saturate and asymptote toward a constant error. Note that there is significant variability in the error growth rate from one forecast to the next, indicated by the large 95% confidence interval accounting for the error growth curve for all forecasts during June–November 2010. If forecast errors were calculated beyond 10 days, eventually a maximum error state would be reached. Beyond this timeframe, a particular deterministic forecast becomes no more skillful than a forecast made with arbitrary initial conditions, and all predictability has been lost (Lorenz 1965).

A linear best-fit trend line is extrapolated from the day 4 to 6 error growth in each curve. By 240 h, the error growth rate has slowed below the linear rate for all our metrics as large-scale errors saturate. It has dropped the most substantially for circulation (Fig. 5b), while it remains close to linear for shear (Fig. 5d). One may interpret this result as suggesting that basinwide shear on average possesses a more extended range of predictability than circulation.

The error growth rates also vary from month to month (Fig. 5). For circulation, the greatest errors occur during August and September, where higher uncertainty associated with an abundance of mature TCs contributes to larger errors. It is hypothesized that circulation errors for 2010 are overall greater than average due to heightened wave activity. For shear and RH, the greatest errors occur during October and November, where uncertainty that is likely associated with fronts and midlatitude systems affecting the tropics results in greater forecast uncertainty. Also note that since the much larger 95% confidence interval is for the full spread of all days within the 6-month period and not just the monthly averages, the variability from one forecast to the next is even greater than what the variability of the monthly means suggests. Therefore, it appears that total error and error growth rates are highly regime dependent.

The philosophy introduced by Lorenz (1982) abides by one particular definition of predictability, in which one may interpret a forecast as being useful if errors have not yet saturated. However, even if errors are still growing, the forecast may have lost skill compared against climatology, thereby eliminating its practical utility. To quantify this, we compare ECMWF control and ensemble-mean forecasts for August–October 2010 against “forecasts” during that same time frame using the 1979–2011 climatology as the forecast (Fig. 6). It is found that the point at which the ECMWF control run forecast errors exceed errors of the climatological forecast varies greatly depending upon choice of variable. By
this criterion, forecast skill as a proxy for predictability is lost in the medium range, or 120 h for circulation (Fig. 6a). For some variables, predictability is lost much more rapidly, with forecast error of the control run exceeding that of the climatological forecast within 48 h for 850-hPa vorticity (Fig. 6b) or only 24 h for 200-hPa divergence (Fig. 6c). Meanwhile, other variables are associated with much greater estimated predictability, with forecast error not exceeding climatological error until >240 h for 850–200-hPa thickness anomaly (not shown), 240 h for 200-hPa velocity potential (Fig. 6d), 192 h for 850–200-hPa wind shear (Fig. 6e), 180 h for 700-hPa relative humidity (Fig. 6f), and 108 h for 850–500-hPa wind shear (not shown). The much lower errors for circulation versus vorticity forecasts in comparison to climatology are likely a result of the fact that it is much easier to forecast an area-average quantity than it is to correctly predict a value at every possible grid point. The lower predictability associated with 850–500-hPa shear than 850–200-hPa shear is likely due to the greater uncertainty in the overall steering pattern at 500 hPa (along with, to some degree, the fact that forecasting climatology for midlevel shear in 2010 actually made a good forecast: only a 2.6 m s\(^{-1}\) error on average). As expected, forecasts of velocity potential are associated with much lower error than divergence forecasts. The order of magnitude difference between the predictability “limit” (by this definition) of divergence and velocity potential is a direct consequence of the variability of their respective spatial and temporal scales.

The ensemble mean forecast performs notably better on average than the control forecast. For example, using the ensemble mean forecast rather than the control forecast greatly extends the predictive skill for circulation from 108 to 240 h (Fig. 6a), or from 180 to >240 h for 700-hPa relative humidity (Fig. 6f). The lead time at which a skillful forecast can be made versus climatology only increases by a small amount for 850–200-hPa wind shear, although the control forecast is already very skillful to begin with (Fig. 6e). These results suggest that the ensemble mean forecast could be useful in predicting
environments favorable for tropical cyclones to develop with greater than one week lead time.

While comparing errors in the ensemble mean forecast to climatological errors addresses the mean error and error growth rate, it does not address the fact that there is significant variability from the climatological mean state. The ensemble forecast, with each member constructed to represent an equal-likelihood event, is also associated with an increasing variance with increasing lead time. Therefore, it is perhaps more appropriate to be comparing the probability distribution, or more specifically, the PDF of the ensemble forecast to the PDF of all possible states comprising the climatology. We address these issues by evaluating forecast error variance and predictive power.

c. Forecast error variance

Compiling all forecasts of 850–200-hPa wind shear during August–October 2010 in 10° × 10° boxes and plotting the absolute error of the ensemble mean versus

Fig. 6. Comparison of error growth rates for (a) 850–700-hPa circulation, (b) 850-hPa vorticity, (c) 200-hPa divergence, (d) 200-hPa velocity potential, (e) 850–200-hPa wind shear, and (f) 700-hPa relative humidity. Forecasts are made using the ECMWF control forecast (blue), the ECMWF ensemble mean (red), and a climatological forecast (green).
the square root of the predicted forecast error variance (the standard deviation), a modest but unconvincing positive relationship is observed (Fig. 7a). Low values of predicted error standard deviation should yield a low error of the ensemble mean. High values of predicted error standard deviation can yield low and high errors of the ensemble mean, though the average error of the ensemble mean over a large sample should be larger than when the predicted error standard deviation is low. Not surprisingly, the distribution changes as a function of forecast time, with generally lower (higher) predicted error standard deviations and errors of the ensemble mean at shorter (longer) forecast times. This relationship is mostly monotonic increasing and possesses a slope near one for 0–72-h forecasts (including the final bin containing the largest forecast standard deviations) and flattens to a slope less than one as the forecast time increases. Averaging within 10° × 10° boxes tends to filter out the small scales in the variance predictions as well as mean error, such that the results for vorticity and circulation are nearly identical. Despite overall positive results, the ensemble is found to be underdispersive for all variables.

4. Predictive power

The PP is developed such that it is invariant under linear coordinate transformations and applies to multivariate predictions irrespective of whether or not error distributions are Gaussian (Schneider and Griffies 1999). However, for the form used in this study, along with confidence intervals and significance levels, it is
assumed that error PDFs are approximately Gaussian. Climatological error distributions, $x_i - x$ from each day during June–November 1979–2011 (Figs. 9a,b) are compared with 120-h forecast error distributions, $x_{120} - x_{120}$ from each ensemble member forecast during June–November 2010 (Figs. 9c,d) in a 5° × 5° box in the MDR from 55°–60°W and 15°–20°N. Overall, error distributions for both 850–700-hPa circulation (Figs. 9a,c) and 700-hPa relative humidity (Figs. 9b,d) are approximately Gaussian. Note that the best-fit Gaussian curves are forced to be centered on zero, while the peaks in the actual distributions may be “biased” slightly above or below zero.

From Fig. 9, it is apparent that the overall variance or width of the climatological distribution is greater than that of the 120-h forecast error distribution. However, if error PDFs at multiple lead times were plotted, they would gradually widen with time from $t = 0$ to $t = 240$ h, eventually reaching or exceeding the width of the climatological PDF. There are also several cases in which the circulation error significantly exceeds the climatological mean, corresponding to a peak in the distribution along the tails of the PDFs in Figs. 9a,c. The rightmost peak along the positive tail in Fig. 9a corresponds to times when tropical cyclones pass through the domain, as the circulation becomes much greater than climatology. The positive (negative) peak at the end of the PDF tail in Fig. 9c corresponds to when a TC forms (fails to form) inside the domain when the ensemble does not predict it (predicts it), or when a TC inside the specified domain is stronger (weaker) than predicted. Since the width of the plotted Gaussian fit curve is based upon the variance of the distribution, and the variance is somewhat skewed by high and low outliers, much of the PDF in Fig. 9c lies “inside” the bell curve.

To estimate the predictability of 850–700-hPa circulation versus climatology using the ECMWF ensemble, PP is computed from $t = 0$ to $t = 240$ h in 12-h increments twice per day during August–October 2010 (Fig. 10a). Just as Schneider and Griffies (1999) run a Monte Carlo simulation to produce a confidence interval for the predictive power given a small number of actual forecasts, each of these independent forecasts from 2010 is treated as a sample of the true PP. The resulting mean PP crosses the 5% significance level, or the point at which the forecast error distribution and the climatological error distribution are no longer statistically distinguishable datasets at 180 h (Fig. 10b). This implies
a predictability limit for circulation at the previous forecast time of 168 h. This is shorter than the predictability for circulation found when using the saturation of error growth to determine predictability, but longer than the limit of predictability found when comparing the RMS forecast error to the RMS climatological error. However, there is significant variability within the 95% confidence interval for PP, and the true predictability for circulation could range between 84 and 240 h. This broad range can potentially be reduced in future work if additional years are included.

The PP calculation can be understood further by examining the eigenvalue spectrum (Fig. 10c). The PP becomes negative once the product of the first 50 eigenvalues (since we only have 51 ensemble members) of the forecast error exceeds the product of the eigenvalues of climatological error. This corresponds to a broadening of the forecast PDF beyond the width of the climatological PDF. All eigenvalues grow as errors grow with time from day 0 (blue) through day 10 (red), while the climatological eigenvalue distribution remains fixed.

The PP for relative vorticity is also included for comparison (Fig. 10d). Consistent with our results from section 3, PP suggests a lack of predictability for relative vorticity in <24 h, suggesting again that area-averaged quantities are substantially more predictable than non-averaged quantities.

Predictive information is calculated in order to quantify the broadening of the forecast PDF with increasing forecast time and compare it to the relative width of the climatological PDF (DelSole 2004). As such, some insight should be provided by examining the structure of the forecast error variance over the tropical Atlantic. Recall that while the mean PP for circulation became negative at 180 h, there were individual forecasts where the PP became negative at <48 h while other forecasts had positive PP through 240 h (Fig. 10a). The 156-h predicted forecast error variance for the highest PP valid at 0000 UTC 5 September 2010 forecast (Fig. 11a, denoted “A”) is compared with the 48-h predicted forecast error variance for the lowest PP forecast valid at 0000 UTC 14 September 2010 forecast.

![Probability density functions for circulation and relative humidity](https://example.com)
It is apparent that, despite the fact that A is from a much longer-range forecast than B, there is overall much lower forecast error variance in A than in B. On 5 September, only a small error variance exists associated with the location and strength of a tropical wave over the Leeward Islands and Puerto Rico, which happens to be the remnants of Tropical Storm Gaston, along with another wave that has yet to emerge off of Africa. On the other hand, on 14 September, there is a high amount of error variance associated with three TCs across the basin: Igor, Julia, and Karl. This comparison illustrates a common example of when forecasts associated with greater (lesser) error variance are associated with a greater (lesser) determinant of the forecast error covariance matrix and thereby lower (greater) PP. Note that a high ensemble mean circulation does not guarantee a high forecast error variance, as Karl in the Bay of Campeche was associated with much lower forecast error variance than Julia (near 42°W), despite the fact that Karl had a stronger ensemble-mean forecast circulation than Julia (Fig. 11b). Forecast error variance is instead a more direct measure of the spread, and thereby the uncertainty, in the ensemble forecasts.

The above example demonstrates the utility of PP in individual test cases. However, in order to abide by the Gaussian assumption and for the 5% confidence interval to have meaning, it is necessary to examine the season as a whole and compare against the 33-yr climatology. The 33-yr August–October climatological error variance for circulation shows a local maximum right off the coast of Africa near the Cape Verde islands, as well as along the dominant recurving TC track within the northern MDR (Fig. 11c). The mean August–October 2010 0-h forecast error variance (i.e., analysis error variance) for circulation indicates overall low uncertainty across the basin, but a maximum near the Cape Verde islands (Fig. 11d). The mean forecast error variance increases rapidly through 120 h (Fig. 11e), continuing to increase but at
a slightly lower growth rate through 240 h (Fig. 11f). The overall structures of the climatological and forecast error variance plots are similar, but with greater error variance closer to 25°N in the forecast distribution and greater error variance closer to 10°N in the climatological error variance (along and to the north of the mean September circulation exceeding $5 \times 10^{-6}\text{s}^{-1}$ from Fig. 3d). This is likely a result of there being only a few strong low-latitude tropical waves, and numerous stronger waves and tropical cyclones farther north in 2010. The basin mean forecast error variance does not exceed the climatological error variance at 0 h, is of comparable magnitude at 120 h, and definitely does so by 240 h. Therefore, it is not surprising that the determinant of the forecast error covariance matrix exceeds the determinant of the climatological error covariance matrix in the PP calculation at 180 h.

Next, PP for 850–200-hPa wind shear is examined. Overall, the estimated predictability limit for shear is very similar to that of circulation, approximately 168 h (Fig. 12a). However, the upper and lower bounds of the PP estimate for shear are significantly narrower than for circulation. This is likely due to the fact that, while uncertainties associated with tropical waves and cyclones introduce extreme local maxima in error variance for circulation (and pronounced minima in PP) there is no analog for shear. However, there is still some flow-dependent variability associated with PP for shear. For example, during September 2010, most ensemble forecasts associated with greatest PP are during 1–10 September 2010, and those ensemble forecasts associated with the lowest PP are during 21–30 September 2010. Correspondingly, the average 500-hPa geopotential height pattern during the first 10 days of September (Fig. 12b) was much less amplified than the average pattern during the final 10 days (Fig. 12c). Similar results were found for various 10-day periods during August and October 2010. These results demonstrate that PP is highly dependent on the flow regime, with a general trend that PP for shear and thereby predictability of shear decrease during more amplified flow regimes. The greater forecast error variance for shear is found in the northern half of the domain (not shown), where greater day-to-day variability will lead to greater uncertainty in

Fig. 11. Ensemble mean 850–700-hPa circulation (magenta contours, $2.5 \times 10^{-5}\text{s}^{-1}$ increments) and circulation forecast error variance (shaded) for (a) a low-uncertainty 156-h forecast valid at 0000 UTC 5 Sep 2010 and (b) a high-uncertainty 72-h forecast valid at 0000 UTC 14 Sep 2010. Error variance for circulation from (c) ERA-Interim 1979–2011 is compared against the mean (d) 0-h, (e) 120-h, and (f) 240-h forecast error variance from the ECMWF ensemble for August–October 2010.
the forecast. This is also the region associated with the greatest month-to-month variability for shear (Fig. 1).

Predictive power suggests a predictability limit of <144 h for 700-hPa relative humidity (Fig. 13a), lower than that for either circulation or shear. It is likely that smaller-scale variability associated with regions of convection lead to faster growth of forecast error variance for RH than for either circulation or shear, for which the error variance is dominated by mesoscale or synoptic scale variability. However, in practice, it can be very difficult to separate the component of error variance for RH associated with convection from that which is associated with the synoptic-scale movement of moist and dry air masses. The overall structure of the August–October 2010 forecast error variance (Fig. 13b) closely resembles the August–October 1979–2011 climatological error variance across the basin (Fig. 13c). The greatest error variance for RH is generally confined to the northern half of the domain, where there are alternating moist and dry periods. Farther south, particularly along the ITCZ into the Caribbean, error variance for RH is very low. This is due to the fact that these regions are almost always very moist (Fig. 2), leading to both a relatively uniform climatology as well as low uncertainty in the forecast.

Predictive power for 200-hPa divergence drops sharply and drastically within the first 24 h of the ensemble

Fig. 12. (a) Predictive power for 850–200-hPa wind shear for consecutive 10-day time intervals from September 2010. 500-hPa geopotential height anomalies (m) for (b) 1–10 Sep and (c) 21–30 Sep 2010.

Fig. 13. (a) Predictive power for 700-hPa relative humidity during August–October 2010, with the 95% confidence interval and the 5% significance level. (b) The 180-h forecast error variance during those 3 months as well as (c) the 1979–2011 climatological error variance for relative humidity are also indicated.
forecast (Fig. 14a). Beyond 24 h, PP for divergence is relatively steady. This is in agreement with the fact that forecast error variance for divergence grows very large by \( t = 24 \) h (Fig. 14b), particularly along the ITCZ and within the southern MDR, but barely amplifies at all and even decreases in some regions by 240 h (Fig. 14c). Since convective-scale predictability is closely tied to the small scales (Zhang et al. 2007) and errors at the small scales grow and saturate the most rapidly (Lorenz 1982), it is not surprising that divergence becomes unpredictable in less than 24 h due to the natural association between divergence and areas of convection.

In sharp contrast to the results for divergence yet consistent with the aforementioned results examining error growth (Fig. 6), the 200-hPa velocity potential is found to have much greater predictive power than divergence (Fig. 15a). Velocity potential is found to be predictable out to 180 h, on average. Regions of maximum forecast error variance for velocity potential (Figs. 15b,c) correspond with regions that had the greatest forecast error variance for divergence (Figs. 14b,c), which also happen to be regions that are the most persistently convectively active. However, unlike 200-hPa divergence, forecast error variance for velocity potential is extremely small at 24 h in comparison with what it grows to be by 240 h.

Last, 850–200-hPa thickness anomaly is associated with the greatest predictive power of all the chosen variables, to beyond 240 h (Fig. 16a). Correspondingly, there is a remarkable increase in forecast error variance.
from $t = 24$ (Fig. 16b) to $t = 240$ h (Fig. 16c). Similar to both RH and shear, the greatest forecast error variance for thickness anomaly is confined to the northern half of the domain, where greater uncertainty associated with synoptic features decreases predictability. Unlike for circulation, in which tropical cyclones are a significant contributor to the total error variance, the majority of the error variance for thickness anomalies does not appear to be associated with the warm cores of tropical cyclones.

5. Conclusions

Several metrics to evaluate predictive skill and attempt to quantify predictability have been explored using the ECMWF Ensemble Prediction System during the 2010 Atlantic hurricane season, in the context of large-scale variables in the tropical Atlantic basin that are relevant to tropical cyclogenesis. These metrics include 1) growth and saturation of RMS error, 2) RMS errors versus climatology, 3) predicted forecast error standard deviation (or variance), and 4) predictive power.

When error growth was computed in the same vein as the classical study of Lorenz (1982), basinwide errors were found to grow linearly beyond day 2, and began to saturate slowly after day 8. On average, the errors in the ensemble mean forecast were smaller than those from the corresponding control forecast, thereby extending the range of predictability. For variables including 850–700-hPa circulation, 200-hPa velocity potential, 700-hPa relative humidity, and 850–200-hPa vertical wind shear, the ensemble mean still exhibited skill (cf. climatology) beyond 9 days. In contrast, the 200-hPa divergence and 850-hPa relative vorticity were virtually indistinguishable from climatology beyond 2 days. Similar results were obtained using predictive power, in that the most predictable variables were 850–200-hPa thickness anomaly, 850–200-hPa wind shear, and 200-hPa velocity potential, with predictability estimated to be greater than one week. Relative humidity at 700- and 850–700-hPa circulation were also found to have predictive power out to about a week. Last, 200-hPa divergence and 850-hPa vorticity were found to be the least predictable, only to about 12 h using predictive power. An evaluation of the ability of the ensemble to predict the standard deviation of the forecast error in $10^5 \times 10^5$ boxes across the domain revealed a mostly linear increasing relationship between the predicted and actual standard deviations for all investigated variables, for forecast times up to and beyond a week.

There exists a considerable degree of variability from day-to-day and between successive initial conditions. This is the result of the evolving flow, with some flow regimes being more predictable than others as first put forth by Lorenz (1965). The degree of variability also depends on the atmospheric variables under consideration. While the predictive power of metrics such as wind shear exhibited relative low variability, the predictive power for circulation was found to decrease considerably in the presence of multiple tropical waves and tropical cyclones. Whether the relationship between predictability of variables relevant to tropical cyclogenesis and the subseasonal potential predictability may be connected to the Madden–Julian oscillation (Madden and Julian 1971) or other equatorial waves requires investigation. Future studies that estimate large-scale predictability in the tropical Atlantic using additional
hurricane seasons are encouraged to determine the robustness of the presented results.

Although the conclusions may differ as a result of the metric used and the flow, the predictability is generally dependent on the scales of motion (following Lorenz 1969). Those metrics typically associated with larger spatial scales (and slower temporal scales of variability) were generally found to be more predictable than those associated with smaller spatial scales (and faster temporal scales). For these reasons, properties of the atmospheric flow that are mostly driven by the large scales (such as tropical cyclone track) are more predictable than those quantities that are mostly driven by the small spatial scales (such as tropical cyclone intensity). For tropical cyclogenesis itself, the local scales of motion most directly relevant to the genesis process remain to be quantified and understood, together with the associated predictability in a frame of reference centered on the tropical disturbance under investigation. A forthcoming companion paper will explore these issues within a storm-relative framework.

Acknowledgments. The authors gratefully acknowledge funding from the National Science Foundation Grant ATM-0848753. The authors thank Rich Rotunno and Chris Davis of NCAR, Ryan Torn of SUNY at Albany, Brian Mapes and David Nolan of the University of Miami, Philip Klotzbach of Colorado State University, and two anonymous reviewers for their comments and suggestions, which have certainly improved this study. The first author would also like to thank and acknowledge support from the ASP student visitor program at NCAR. Last, we thank TIGGE for the availability of the data used in this study.

REFERENCES


