BAMS Article

On the Prospects for Improved Tropical Cyclone Track Forecasts

Feifan Zhou and Zoltan Toth

ABSTRACT: The success story of numerical weather prediction is often illustrated with the dramatic decrease of errors in tropical cyclone track forecasts over the past decades. In a recent essay, Landsea and Cangialosi, however, note a diminishing trend in the reduction of perceived positional error (PPE; difference between forecast and observed positions) in National Hurricane Center tropical cyclone (TC) forecasts as they contemplate whether "the approaching limit of predictability for tropical cyclone track prediction is near or has already been reached." In this study we consider a different interpretation of the PPE data. First, we note that PPE is different from true positional error (TPE; difference between forecast and true positions) as it is influenced by the error in the observed position of TCs. PPE is still customarily used as a proxy for TPE since the latter is not directly measurable. As an alternative, TPE is estimated here with an inverse method, using PPE measurements and a theoretically based assumption about the exponential growth of TPE as a function of lead time. Eighty-nine percent variance in the behavior of 36–120-h lead-time 2001–17 seasonally averaged PPE measurements is explained with an error model using just four parameters. Assuming that the level of investments, and the pace of improvements to the observing, modeling, and data assimilation systems continue unabated, the four-parameter error model indicates that the time limit of predictability at the 181 nautical mile error level (n mi; 1 n mi = 1.85 km), reached at day 5 in 2017, may be extended beyond 6 and 8 days in 10 and 30 years' time, respectively.

https://doi.org/10.1175/BAMS-D-19-0166.1
Corresponding author: Feifan Zhou, zhouff04@163.com
In final form 5 July 2020

AFFILIATIONS: Zhou—Laboratory of Cloud-Precipitation Physics and Severe Storms, Institute of Atmospheric Physics, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, Beijing, China; **Toth**—Global Systems Laboratory, NOAA/OAR/ESRL, Boulder, Colorado

Tropical cyclones (TCs) are large-scale (typically 100–2,000-km diameter), well-organized warm core circulation systems that form in certain tropical ocean basins (e.g., Vigh and Schubert 2009; Komaromi and Doyle 2017). Depending on their height and intensity, they are carried as natural "tracers" by the background low- to midtropospheric or deeper-layer environmental flow until their circulation is disrupted by either strong vertical wind shear or a limitation in available surface energy flux (as they move over either colder water or land surfaces). Under favorable conditions, TCs have been observed to last up to 30 days (Hurricane Research Division 2019). Landfalling TCs are arguably the most devastating naturally occurring meteorological phenomena.

Given their societal significance, success in the prediction of TCs has been noted as the crown jewel of the maturing practice of numerical weather prediction (NWP; Yamaguchi and Majumdar 2010). The error in the position of 3-day National Hurricane Center (NHC) TC forecasts for the Atlantic (northeast Pacific) basins, for example, were reduced from around 300 (225) nautical miles (n mi) in 1990 to 100 (67) n mi in 2018 (Lawrence 1990; Cangialosi 2019) (1 n mi = 1.85 km). The pace of track error reduction, however, was noted to lapse or even cease in recent years (Landsea and Cangialosi 2018, hereafter LC18). In their study, LC18 raised the question whether this decrease in the rate of error reduction should be interpreted as a sign that "the approaching limit of predictability for tropical cyclone track prediction is near or has already been reached."

In this paper, we define predictability as the time period an initial error in a forecast made with a perfect model (like nature itself in an analog approach; Lorenz 1969a) reaches a prespecified level. There is a general consensus that with any level of initial error variance the atmosphere has a finite time of predictability (Lorenz 1969b). Whether the predictability time limit as initial errors go to zero is bounded [i.e., would never exceed a certain time limit, as Lorenz (1969b) proposed], however, is still an open question.

The general question of atmospheric predictability has been probed with a wide array of approaches, using theoretical considerations (e.g., Leung et al. 2019), simple model simulations (e.g., Rotunno and Snyder 2008), and complex operational modeling systems (e.g., Zhang et al. 2019). The predictability of TCs as a specific phenomenon received less attention in the literature (see LC18). Of relevance here are studies by Plu (2011) and Aberson (1998). Using a lagged forecast difference [Lorenz 1982; applied globally, and reported here with the European Centre for Medium-Range Weather Forecasts (ECMWF) forecast system] and an analog approach (Fraedrich and Leslie 1989; applied over the Atlantic basin), respectively, they found that the doubling time of TC track error variance is around 42 h. While this estimate of linear error growth rate is a critical data point for TC track predictability, no attempts have been made to assess how the predictability time limit for TC track errors may expand as NWP, and in particular initial conditions influencing the evolution and transposition of TCs improve.

The present study revisits the question posed by LC18, whether the skill of current TC forecasts may be imminently curtailed by a firm bound on predictability. Examining the TC track error dataset used in LC18 (except the omission of tropical depressions), the following topics are explored: 1) How are TC track forecast errors expected to grow as a function of lead time? 2) Are operational TC track forecasts affected by model errors? 3) What is the trend in TC track analysis and forecast error reduction over the past decades? And 4) how may such errors behave in future decades?

Positional error dataset

The center position of forecast cyclones is defined and uniquely determined using well-established formulas based on the minimum in forecast sea level pressure and other known parameters (Schenkel and Hart 2012). The center position of the corresponding TCs in nature, however, is not known exactly. Most TCs form over open oceans and are only remotely observed. Satellite, and if available, other observations leave the center position ambiguous, especially in case of weak systems. Traditionally, responsible forecast centers produce a subjective analysis (or "fix") for the center position based on all information available in real time, including, in rare occasions, even NWP guidance. After all observations are collected, the real-time position information is updated and a new, final subjective center position analysis is prepared that is called "best track" position (NHC 2019). It is well understood that errors remain in the final best track position; their exact level is unknown and has only been subjectively estimated (see Torn and Snyder 2012; Landsea and Franklin 2013).

Conventionally, TC track forecast performance is assessed as the distance between the exactly known forecast position (P_f) and the estimated best track position (P_b) of the corresponding TC observed in nature:

$$x_p = \left\| P_f - P_b \right\|, \tag{1}$$

where ||···|| stands for great-circle distance and x_p is the perceived track error. Note that due to the error present in the best track position (P_b), x_p differs from what we call true forecast error $\left(x_t^f\right)$, defined as the distance between the forecast (P_f) and the actual or true position (P_t) of TCs in nature:

$$x_t^f = \left\| P_f - P_t \right\|. \tag{2a}$$

Likewise, with P_b taking the place of P_f, true error can also be defined for the best track analysis position:

$$x_t^b = \left\| P_b - P_t \right\|. \tag{2b}$$

Since x_p is only an estimate (and as we will see in the next section, a biased estimate) of the true forecast error, following the nomenclature used by Leith (1978) and Pena and Toth (2014, PT14 henceforth), from now on we refer to the traditionally defined track error [Eq. (1)] as "perceived error."[1] Perceived error is calculated and averaged here only for cases where a storm is diagnosed both in the forecast and observations.

Since true errors cannot be measured, the present study, like others such as LC18, uses perceived error to assess the quality of TC track forecasts. In particular, perceived track error in 12-, 24-, 36-, 48-, 72-, 96-, and 120-h lead-time Atlantic basin "official" NHC forecasts is used over the years 2001–19. Data from the first 17 years (2001–17) are used for development, while data from 2018 and 2019 (yet unavailable at the time of development) are used for independent evaluation. As can be seen from Fig. 1a, the total number of cases for the developmental period decreases with longer lead times as storms identified at initial time may become too weak either (or both) in the forecast or respective observed data.[2] For visual clarity, data for 12- and 24-h lead times are omitted from Fig. 1. The number of cases per season varies between less than 20 at 120 h and close to 400 at 36 h, depending on how active each season was.

Consistent with results in Fig. 3 of LC18, perceived error exhibits an overall decrease over the years (Fig. 1c), while it

[1] For a broader discussion of various metrics assessing the quality of NWP analysis and forecast products, see, e.g., the introduction section in PT14.

[2] The main results of this study remain unchanged if the analysis is restricted to a smaller number of storms that persisted both in the forecast and observed data up until 5 days.

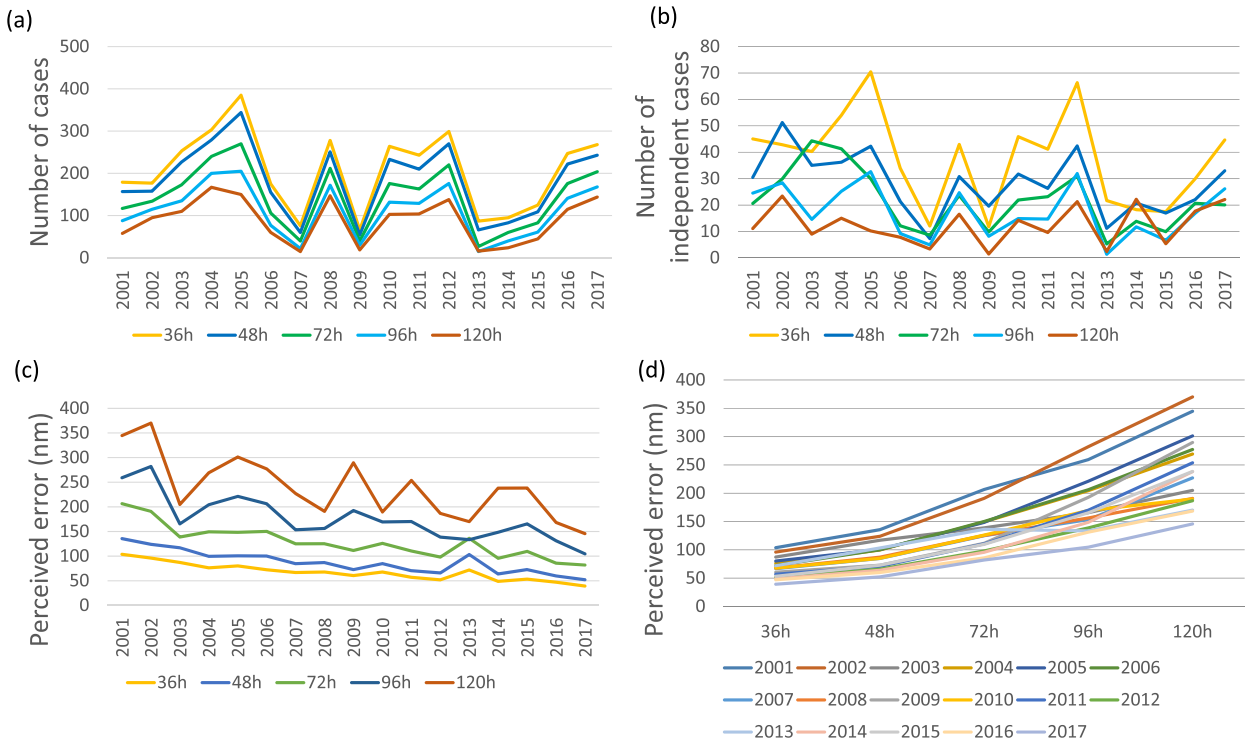

Fig. 1. (a),(b) The nominal (*M*) and independent [see Eq. (7)] number of verified TC forecasts at various lead times over the 2001–17 developmental period, respectively. (c),(d) Seasonal mean of perceived forecast error measurements displayed over the 2001–17 period and as a function of lead time, respectively.

shows an amplification as a function of lead time (Fig. 1d). The official NHC forecast (www.nhc.noaa.gov/modelsummary.shtml) well reflects the performance of state-of-the-art NWP systems as it is a composite of NWP track guidance from four to five numerical products available from leading international centers [such as the National Centers for Environmental Prediction (NCEP), ECMWF, and the Met Office] that perform best in a few seasons just prior to each year, with only minor subjective modifications. As with other phenomena and metrics, there is a tendency in Figs. 1c and 1d for reduced errors and a reduced slope of error growth in more recent years. This has been partly attributed to NWP model improvements and an associated reduction in model related forecast errors (Bauer et al. 2015; Zhang et al. 2019), a topic we will return to in the sections below.

Methodology

Ideally, one would use true error, the variance distance between an analysis or forecast field and reality on scales resolved by a numerical model, to evaluate the quality of NWP guidance. As reality is directly inaccessible, most studies instead use perceived error measurements as a proxy for true error. In a series of recent studies, PT14 and Feng et al. (2017, 2020) showed how true error (described analytically via a few unknown parameters) can be inferred from known perceived error measurements. Using an inverse approach called Statistical Analysis and Forecast Error estimation (SAFE), perceived error is modeled as a function of the unknown true error parameters, using a priori knowledge (in the form of assumptions) about the behavior of analysis and forecast error. A variational minimization of the difference between measured and modeled perceived error (i.e., fitting error) results in the best estimate of the unknown true error parameters, while validating the assumptions underlying the error behavior model. SAFE has been tested and evaluated in a variety of simulated and real life settings.

Our objective is to arrive at a bias-free estimate of the true error variance in both the initial and forecast conditions of the official forecast, as well as in the best track position analysis.

None of these quantities are measured; only perceived forecast error measurements are available. The method used here for the estimation of true TC track error is a simplified version of SAFE that we call SAFE-s. First, we insert the true position into Eq. (1):

$$x_p^2 = \left\| P_f - P_b \right\|^2 = \left\| \left(P_f - P_t \right) - \left(P_b - P_t \right) \right\|^2. \tag{3a}$$

Considering Eqs. (2a) and (2b), and exploiting the law of sum of variances, we then express perceived error variance $\left(x_p^2\right)$ as a sum of the variances of true best track analysis error and true forecast error minus their correlation term:

$$x_p^2 = \left(x_t^f - x_t^b \right)^2 = \left(x_t^f \right)^2 + \left(x_t^b \right)^2 - 2\rho x_t^f x_t^b. \tag{3b}$$

Here ρ is the correlation between the true errors in the forecast $\left(x_t^f\right)$ and the verifying best track analysis $\left(x_t^b\right)$. It is clear from Eq. (3b) that perceived error variance $\left(x_p^2\right)$ provides a biased estimate of the true error variance. Depending on the actual values of x_t^b and ρ, perceived error may under- or overestimate x_t^f.

For a bias-free estimate of the true error, we consider a set of equations like (3b) valid for each lead time i for which perceived error variance measurements are available. We note that the number of unknowns on the right side of the equations far outnumber the number of measured quantities on the left. To reduce the number of unknowns and facilitate the estimation of true error variance, we introduce a priori knowledge in the form of three basic assumptions, to be validated with the experimental data. As noted below, the assumptions listed here and in subsequent sections are motivated either theoretically, or by the scientific principle of parsimony, striving to find the simplest explanation for the behavior of perceived error measurements.

Assumption 1. First, we assume that the dynamics of the large-scale environmental circulation responsible for the transposition of TCs is well simulated in modern NWP models. Note that the failure of numerical models to faithfully represent finer-scale motions would not violate this assumption, as long as associated forecast errors do not project onto TC positional errors, or are already saturated (thus not growing any further). If assumption 1 is valid, it follows that the true error in the official forecasts is not affected by model induced errors but originates solely from initial (or analysis) error. Following the parsimony principle, we would introduce numerical model error related parameters only if an error behavior model without them fails to describe the experimental perceived error data.

Assumption 2. Next, we assume that the error in the initial condition (analysis) of NWP forecasts considered in the official forecast amplifies exponentially:

$$\left(x_t^i \right)^2 = x_0^2 e^{\alpha(i-1)dt}, \tag{4}$$

where x_0 and x_t^i are the true analysis error and true forecast error at lead time i, and α is the exponential rate of forecast error growth in a time unit of a day. The exponential growth of linearly evolving perturbations and errors (i.e., the difference between two diverging trajectory segments) is theoretically well established (Lorenz 1963) and can be interpreted as a response to a constant force associated with instabilities present in dynamical systems. Ever since Lorenz's early paper, the exponential formula has been used in a long series of studies to describe forecast error growth (e.g., Leith 1978; Lorenz 1982; Dalcher and Kalnay 1987; Schubert and Suarez 1989; Peña and Toth 2014; Feng et al. 2017, 2020), sometimes along with other terms supposed to represent model related error.

The growth of error in nonlinear systems, however, is generally found to be first tempered, then inhibited by nonlinear processes related to the finite size of such systems (e.g., Lorenz 1982). Using the root-mean-square (rms) difference between gridded forecasts and verifying analysis fields, slower than exponential growth is observed as early as at 24-h lead time (Gilmour et al. 2001). This is because small, subsynoptic-scale forecast features become decorrelated from their observed counterpart early in the forecast process. Hence, error on small scales does not grow any longer, so that part of the error spectrum saturates early on in the forecast, leading to subexponential growth.

Surprisingly, no sign of nonlinear saturation emerges in the TC track forecast error data even up to day 5 (see Fig. 1d). This is due to the different nature of the track error metric [Eq. (1)]. With traditional gridpoint-based metric, once a forecast TC no longer overlaps with its observed counterpart, rms error may no longer grow, as beyond this point in time, rms error is insensitive to how far the forecast and observed storms drift from each other. In sharp contrast, the TC track error metric is sensitive only to the distance between the forecast and observed storms. On a planet with a single ocean basin at the latitudes of TC activity, for example, the initially exponentially growing errors in the position of long-lived forecast TCs would be nonlinearly bounded only by the finite size of the planet, the maximum positional error being half of the circumference of a latitude belt. TC track error thus appear to reflect and emphasize the largest-scale components of the global circulation.

Assumption 3. Finally, we assume that the error in the forecast and verifying analysis positions do not correlate beyond 24-h lead time [i.e., $\rho = 0$ in Eq. (3)]. PT14 found that with increasing lead time, ρ decreases but in 24 h it drops only to a value around 0.3. Note, however, that unlike in SAFE, here we use a verifying analysis (best track position) that is made *independently* from the NWP data assimilation system. Hence the true error in the verifying analysis (i.e., best track position; x_t^b) and in the NWP forecast initial condition (x_0) could be considered independent. To be conservative, we assume independence only after 24-h lead time, considering that NWP analysis and forecast fields are available at the time TC specialists define the best track position. If this simplifying assumption 3 is valid, ρ at and beyond 36-h lead time can be assumed zero, so the third term in the right hand side of Eq. (3b) vanishes. This reduces the number, and facilitates the estimation of the unknowns in the simplified SAFE-s methodology used here.

Estimation algorithm. With the above assumptions, perceived error at and beyond 36-h lead time can be simulated with just three unknown quantities that describe the true error in the official initial (analysis, x_0) and forecast guidance (α, growth rate of forecast error), as well as in the best track analysis $\left(x_t^b\right)$. Inserting error growth Eq. (4) into, and dropping the correlation term from the rightmost part of Eq. (3b) yields

$$\left(x_p^i\right)^2 = x_0^2 e^{\alpha(i-1)dt} + \left(x_t^b\right)^2. \tag{5}$$

To solve for the three unknowns, we consider a set of equations like Eq. (5) valid for five lead times: 36, 48, 72, 96, and 120 h, and in the first application of SAFE-s, minimize the distance between the seasonal average of perceived error *measurements* [lhs of Eq. (5)] and the perceived error *simulated* (or modeled) by the unknowns [rhs of Eq. (5)]:

$$J = \min\left(\max_{i>2}\left\{\left[x_0^2 e^{\alpha(i-1)dt} + \left(x_t^b\right)^2 - \left(x_p^i\right)^2\right]^2 w_i^{-1}\right\}\right). \tag{6}$$

The interpretation of x_0 in Eq. (6) is discussed in the appendix. As explained by PT14, the use of the maximum (max, or L_∞) norm ensures that a comparably good fit is achieved at all

lead times, while the coefficients $w_i = \mathrm{SEM}_i \big/ \sum_i \mathrm{SEM}_i$ guarantee that each lead time is given appropriate weight in the minimization [Eq. (6)].

The uncertainty in the mean of a sample of error measurements for each storm in a season $\left[\overline{x}_i = (1/M)\sum_{j=1}^{M} x_{i,j}\right]$ can be quantified as the difference between the sample-based mean and the expected value of the measured quantity, that is called *standard error of the mean* or SEM_i. The introduction of weight coefficients w_i is necessary as SEM_i depends on both the number of sample points (M) and the standard deviation in the measurements (sd_i), both of which vary greatly over lead time (i):

$$\mathrm{SEM}_i = \frac{\mathrm{sd}_i}{\sqrt{M}} f_i, \tag{7}$$

where $\mathrm{sd}_i = \sqrt{\sum_{j=1}^{M} \left(x_{i,j} - \overline{x}_i\right)^2 \big/ (M-1)}$, and f_i is an adjustment factor to reflect the autocorrelation (r_i) in the sample: $f_i = \sqrt{(1+r_i)/(1-r_i)}$. The corresponding number of independent sample points $\left(M/f_i^2\right)$ is shown in Fig. 1b.

As a quantitative estimate of the uncertainty in the seasonal mean of perceived error measurements, SEM_i can also be used to assess the quality of perceived error simulations. Assuming that all three assumptions used in the formulation of Eq. (5) are valid and that fitting error (i.e., the difference between simulated and measured perceived error) follows a Gaussian distribution, the absolute value of fitting error, for example, must be below SEM_i^{68} for 68% of the measurements [which is the probability that a Gaussian variable falls in a plus/minus one standard deviation interval around its expected value, $\left| x_0 e^{(1/2)\alpha(i-1)dt} - x_p^i \right| \leq \mathrm{SEM}_i^{68}$]. It follows that as long as the percentage of fitting errors violating the above inequality is below/above an error level SEM_i^s, that is the estimation uncertainty associated with a preselected statistical significance level s [which is SEM_i multiplied with the coordinate value of the Gaussian density function evaluated at $1 - (s/2)$], the overall error simulation is consistent/inconsistent with the experimental measurement data. For further details, see PT14.

As an inverse method, SAFE-s uses perceived forecast error measurements to infer true initial (x_0) and forecast error $\left(x_t^i\right)$. As discussed earlier, among many possible choices, we use TC track error as a metric of perceived forecast error. Since TCs are large-scale features, the error in their forecast position depends not only on the quality of the initial TC center position, but also of other features in NWP analyses (feeding into the official forecast; Plu 2011). So x_0 reflects the general quality of NWP analysis fields in areas affecting the future position of TCs, in a measure equivalent to initial TC positional error distance in the unit of nautical miles.

Error behavior in individual seasons

In this section SAFE-s is used to estimate the behavior of true error separately for each season over the 2001–17 period. To solve the minimization problem of Eq. (6) (and its variant in the upcoming sections), we use the spectral projected gradient 2 (SPG2) method of Birgin et al. (2001). Results for two selected (the first and last in the developmental record) seasons are shown as examples in Fig. 2. In both cases, the seasonal mean of measured and simulated perceived errors (Figs. 2a,b) are closer than expected by chance at the 97% or lower levels of significance (Figs. 2c,d). This also holds true for all the other 15 seasons. In every season, true positional forecast error appears to grow exponentially. It follows that the behavior of perceived error measurements can be explained without invoking model related errors. In fact, 97% of the variance in the seasonal mean perceived error measurements (17 seasons × 5 lead times = 85 points) is explained by the 51 estimated independent parameters (17 seasons × 3 = 51 unknowns; see first column in Table 1). We conclude that the experimental measurement data are consistent with assumptions 1–3 (no model error, exponential growth of true error, and no correlation between true best track analysis and forecast errors beyond 24-h lead time).

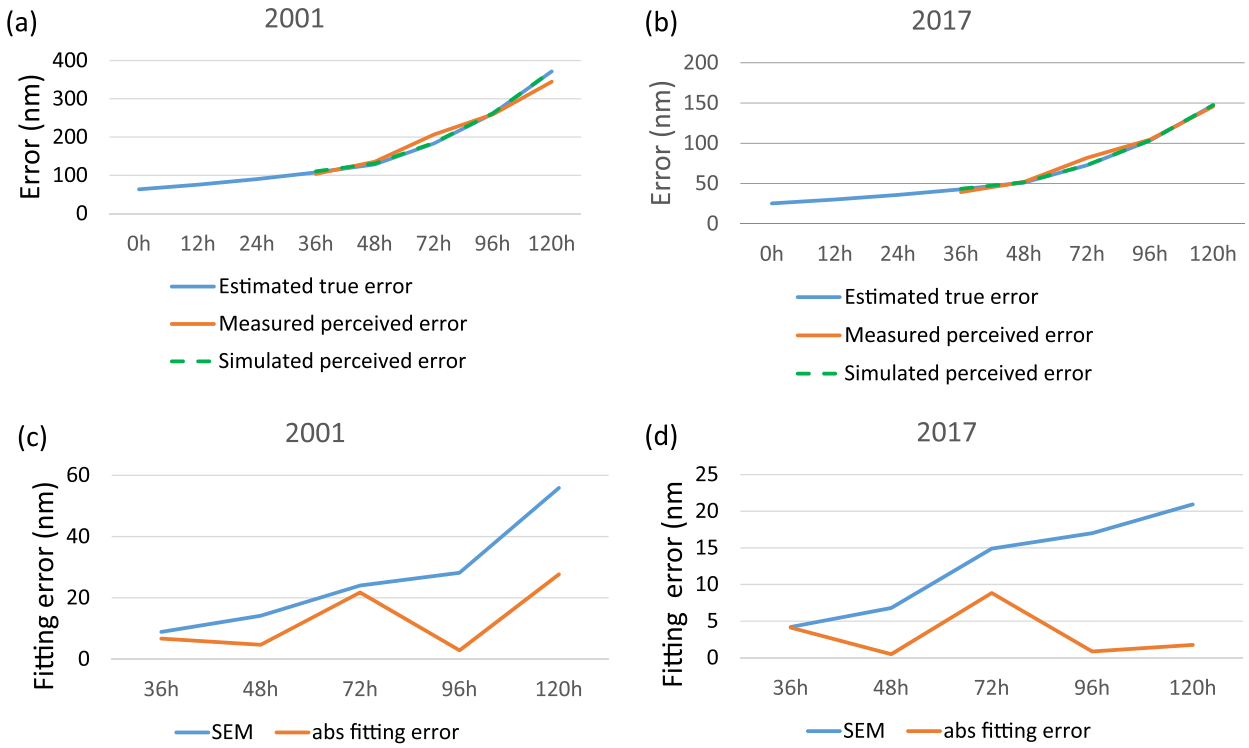

Fig. 2. (a),(b) Measured (red line) and simulated (green dashed) perceived and estimated true error (blue; n mi) for the 2001 and 2017 seasons, respectively, as a function of lead time. (c),(d) As in (a) and (b), but for SEM[97] (blue line) and the absolute fitting error (red). Note the difference in vertical scale between the 2001 and 2017 panels.

The estimated unknown parameters for the 2001–17 seasons, displayed in Fig. 3, can be summarized as follows. 1) True initial (analysis) error decreases over the years with a somewhat diminishing rate, from above 60 n mi to just below 30 n mi; 2) the error in the best track analysis is only about 40% of that in the NWP analysis (and even a smaller percentage of that in the forecasts) and shows a similar decrease over the years from above 20 to below 10 n mi;[3] and 3) the 24-h forecast error growth rate appears to be rather stable over the years, showing only small year-to-year fluctuations in the 1.35–1.61 range. Based on these

[3] These estimates reflect seasonal mean values; the error for weaker and stronger than average storms, as discussed by Torn and Snyder (2012) and Landsea and Franklin (2013), may vary significantly.

Table 1. Parameters, fitting error, uncertainty, and effectiveness indicators for various estimation configurations. For further details, see text.

	Number of parameters				Uncertainty range
	51	35	20	4	
Analysis error in 2017 (n mi)	25.23	24.98	27.71	27.84	5.1%
24-h growth rate	1.452	1.440	1.418	1.454	1.2%
1-yr reduction rate of analysis error	0.951	0.952	0.954	0.961	0.5%
Best track analysis error fraction	0.416	0.299	0.299	0.385	15%
Best track error in 2017 (n mi)	6.944	7.468	8.284	10.72	—
Number of outliers expected by chance (at 97% significant level)	2	2	2	2	—
Actual number of outliers for original (and derived) parameter estimates	0 (3)	0 (3)	0 (1)	2	—
Variance explained in 85 measurement data points by original (and derived) parameter estimates	97.1% (88.4%)	95.0% (88.5%)	93.3% (88.7%)	88.9%	—

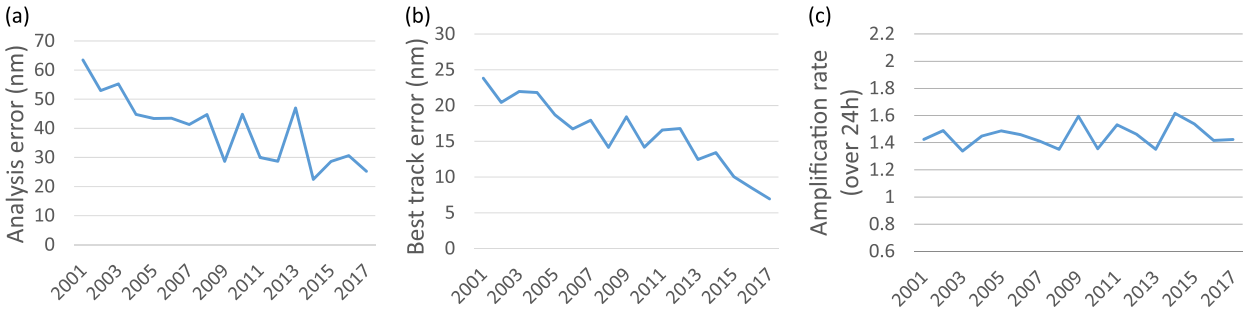

Fig. 3. The three true error parameters as estimated independently for each season by SAFE-s: (a) analysis (initial) error, (b) best track error, and (c) 24-h amplification rate.

observations, and in an attempt to describe the behavior of perceived and true errors with a vastly reduced set of unknown parameters, we introduce further assumptions (to be validated with experimental measurements in the next section).

Error behavior in past decades

Given the somewhat regular behavior of the three unknown parameters over the 17-yr experimental period noted above, here we formulate additional assumptions about relationships between, and the behavior of the unknown parameters over the years. The introduction of these assumptions into SAFE-s simplifies the estimation and description of the unknown quantities by reducing the overall number of unknowns from the 51 used in the previous section.

Assumption 4. As pointed out earlier, the subjective best track and objective NWP analysis methods are mostly independent. The two methods, however, utilize mostly the same set of observations (gradually enhanced and enlarged over the years), while both benefitting from improvements in data processing algorithms. Therefore, for simplicity, we assume that the true error variance in the best track analysis is a given fraction of the error variance in the NWP analysis (initial condition):

$$\left(x_t^b\right)^2 = \gamma x_0^2, \tag{8}$$

where γ is the best track error variance expressed as a fraction of the NWP analysis error variance. Variability unexplained by the above equation will be assumed to be associated with sampling noise. With this assumption, the number of unknown quantities drops by 16, as the new ratio parameter replaces the 17 seasonal best track analysis error variance parameters.

Assumption 5. It is well understood that the quality of NWP analyses is a function of the effectiveness of the observing, data assimilation, and modeling systems (e.g., Toth and Buizza 2018). Admittedly, even with a sustained effort, improvements in any of the three systems at any NWP center may be introduced at an uneven pace over the years [see Fig. 2.12 in Toth and Buizza (2018) as an example from ECMWF]. Recall, however, that the official TC forecast evaluated in the current paper is a subjectively weighted composite of forecasts from four to five different NWP centers. As changes introduced at various centers are usually uncoordinated, fluctuations are expected to be smoothed out. Consequently, we assume that the level and efficiency of NWP research and development affecting the official forecast is constant (i.e., unchanged) over the years. Theoretically, this implies that initial error in the official forecast improves each year by the same fraction, and deviations in the data from this rule are due to sampling noise. The constant rate of error reduction amounts to an exponential decrease of initial error over the years (with an asymptotic value of zero at time infinity):

$$x_{0_j}^2 = x_{0_{2001}}^2 e^{\beta(j-1)}, \tag{9}$$

where β is the exponential rate of decrease in analysis error variance in a time unit of a year, $x_{0_{2001}}^2$ is the analysis error variance in the first season in the record, and $x_{0_i}^2$ is the analysis error variance in the jth season.

With the addition of assumption 5 [captured in Eq. (9)], the number of unknowns is further reduced by 15 as the new exponential analysis error decrease parameter replaces 16 of the 17 original seasonal analysis error variance parameters. Given assumption 4 above, assumption 5 implies that the best track analysis error variance also undergoes an exponential decrease over the years.

Assumption 6. The growth rate of true error in a given, possibly wide amplitude range is a reflection of the level of instabilities present in a dynamical system. Barring any noticeable change from a slowly evolving climate, the expected value of error growth rate for any forecast system is constant. Though seasonal variations in the circulation such as the phase of El Niño–Southern Oscillation (ENSO; Wang et al. 2017) or the Madden–Julian oscillation (MJO; Madden and Julian 1971) may influence instabilities affecting the divergence of TC trajectories, for simplicity, here we assume that the TC positional error growth rate is constant over the 17 seasons studied. In any case, since MJO or ENSO variations are predictable only on seasonal time scales at most, for the decadal projection of TC error behavior, we have no better choice than setting the error growth rate at its climatologically expected value.

With assumption 6, the 17 seasonal growth-rate values are replaced with just one parameter, resulting in an additional reduction of the total number of unknowns by 16, down to just 4: the analysis error in 2001 $\left(x_{0_{2001}}\right)$, the exponential rate of the decrease in analysis error variance (β), the exponential rate of growth in true forecast error variance (α), and the best track analysis error variance as a fraction of the NWP analysis error variance (γ).

Parameter estimates. For simplicity, in the previous section true error evolution was described by 3 independent parameters, estimated in separate minimization procedures for each of the 17 seasons [51 total number of independent parameters; see Eq. (6)]. In contrast, here the unknowns for the same 17-season developmental sample are estimated in a *single minimization*. With the successive introduction of assumptions 4–6, the 85 seasonal mean perceived error measurements are simulated with a reduced number of unknowns (35, 20, and 4), as a function of not only lead time i = 3, ..., 7, but also years j = 2001, ..., 2017:

$$J = \min\left(\max_{i=3,7;j=1,17}\left\{\left[x_0^2 e^{\beta(j-1)} e^{\alpha(i-1)dt} + \gamma x_0^2 e^{\beta(j-1)} - x_{pi,j}^2\right]^2 w_{i,j}^{-1}\right\}\right). \tag{10}$$

The main results in the form of the 4 summary parameters and related information are reported in Table 1 using 35, 20, and 4 unknown parameters.[4] For reference, results from the year-by-year estimates reported in the previous section (51 unknown parameters) are also listed in Table 1. For cases when more than 4 unknowns are estimated, assumptions 4–6 are used in a posteriori manner to derive the missing summary parameters based on the estimated parameters (corresponding results are listed in parentheses). Specifically, the growth of the forecast error as a function of lead time, and the decrease of analysis error over the years is estimated by fitting the exponential formulas in Eqs. (4) and (9), respectively, while the overall ratio of the best track and NWP analysis is estimated as the mean of the ratios estimated independently for each year.

[4] All results remain the same up to the fourth digit when the L_∞ norm is replaced with the L_2 norm.

Importantly, for all new parameter configurations (35, 20, or 4 parameters), the number of actual outliers at the 97% significance level is below that expected by chance. This is an

indication that the experimental perceived error measurements are consistent with assumptions 4–6: the ratio of the error in the best track and NWP analysis, and the growth rate of forecast error can be considered stationary, while the analysis error exhibits an exponential decrease over the years. The validity of the assumptions is also indicated by the relatively small amount of variance lost by their introduction. The 97% variance explained by the 51-variable error model drops only 2%–4% as each of assumptions 4–6 is introduced while the number of independent parameters drops from 51 to 35, 20, and 4. Using all assumptions and only 4 parameters describes 89% of the total variance in the experimental data.

Based on a subjective review of fitting error and SEM outlier statistics for experiments with different norms and number of unknowns (listed in Table 1) we conclude that the minimization with the maximum metric and using only 4 parameters may provide the best estimate of the unknown summary variables. Therefore, in the remainder of the study error estimates from this configuration will be used. The uncertainty in the estimated four parameters is quantified as half of the range of estimates listed in Table 1, expressed as a percentage of (and centered around) the best estimate. As seen from Table 1, the uncertainty is 1.2%, 0.5%, 5.1%, and 15% in the growth rate of forecast error, reduction rate of analysis error, the 2017 analysis error, and the best track fraction estimates, respectively. Note that the simulation of the 2018 and 2019 perceived error measurements (independent of the developmental period) by the four-parameter error model is statistically acceptable (i.e., fitting error smaller than associated SEM).

Projection into the future

In the previous sections, a compact, four-parameter model was introduced to describe TC track errors as a function of lead time and year. The six assumptions underlying the error behavior model have been validated by experimental data from the past. In this section, we explore how the same model can be used to quantify the expected quality of TC track forecasts in future years. Any extrapolation into the future, however, involves an additional, unverifiable assumption, that the relationships established on past data remain valid in the future. What does this entail for the six assumptions used above?

Assumptions for future projection. First, we note that assumptions 3 and 4 relate to the error in the best track analysis position. Since the best track analysis has no bearing on the quality (i.e., true error) of TC forecasts, and we are primarily interested in the true and not perceived performance of future forecast systems, these assumptions are irrelevant here[5] (except in the display of Fig. 4). Assumptions 2 and 6 pertain to the type of, and possible trend in error growth. As track error was found unaffected by model related errors, error growth is independent of forecast systems, hence reflects only the qualities of nature. As discussed earlier, the exponential growth of forecast error is theoretically founded that we expect to hold even under a potentially changing climate [assumption 2 for projection (P-2)]. For simplicity, we also assume that the error growth rate observed over the past decades will hold into the future (assumption P-6). In reality, the rate of growth, however, may exhibit (presumably) small changes in response to possible changes in the level of atmospheric instabilities.

[5] Note also that the relatively large (15%) uncertainty in the estimate of best track analysis error (Table 1) thus have no bearing on estimates of future error/predictability.

Unlike assumptions 2 and 6, the two remaining assumptions, 1 and 5, are contingent on engineered, and not on natural systems. Again, for simplicity, we assume that the quality of NWP systems will continue to improve, and in particular, analysis errors will continue to decrease into the future at the same rate observed over the past decades (assumption 5). We argue that maintaining the models' quality of not affecting track forecast error (assumption 1) is a corollary to continued future analysis (and forecast) improvements (assumption P-5/1). Past

reductions in initial error have been partially accomplished by periodic increases in the spatial resolution of NWP models. At each upgrade, the physical parameterization schemes representing the statistical effect of natural processes unresolved by the coarser-resolution model (e.g., convection) on resolved scale motions are carefully replaced by the explicit handling of such processes (e.g., the inclusion of nonhydrostatic processes). The validity of assumption P-5/1 thus appears to be contingent on whether the effective resolution of global NWP models can be further expanded in the coming decades. Given the blossoming research into fine-resolution limited-area modeling, and ongoing work on massively parallel and GPU computing, incorporating ever finer-scale processes in the dynamics and physics of numerical models, though not guaranteed, is a real possibility. Steady improvements in NWP performance (assumption P-5/1) is of course contingent on continued material and intellectual investments into the development of NWP observing, data assimilation, and modeling systems at rates of the past.

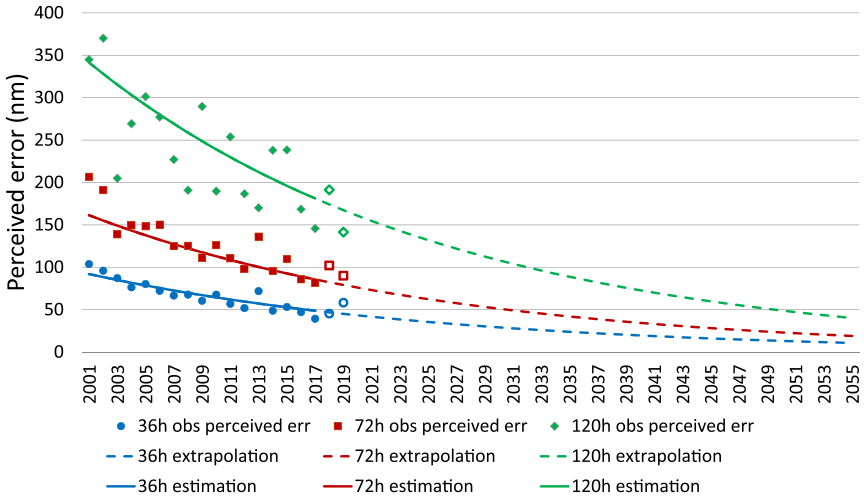

Fig. 4. Measurements (closed and open symbols represent the developmental period of 2001–17 and independent period of 2018–19, respectively) as well as simulated and extrapolated (with best estimates of four-parameter model; continuous and dashed lines, respectively) of seasonal mean perceived error at 120- (green), 72- (red), and 36-h (blue) lead time, as a function of years.

Skill projections below reflect a hypothetical scenario under the assumptions that the rates of daily forecast error growth (assumption P-6) and annual NWP improvements (assumption P-5/1) remain stationary in the future. The projections are also subject to the 0.5% uncertainty in the best estimate of the annual rate of analysis error variance reduction (Table 1). Admittedly, P-6 and P-5/1 are unverifiable but are the simplest assumptions one can make, in the absence of related scientifically based information. The resulting error projection thus offers a reference for future error reduction, from which reality may deviate in case the actual future error growth, and/or forecast system improvement rates deviate from our default, "no change" assumption.

Projection of perceived error. As discussed in the introduction, the predictability of weather and TCs in particular can be assessed in a multitude of ways. As mentioned earlier, LC18 reviewed TC track errors over the past decades to ascertain possible future behavior. In their Fig. 1, seasonal mean perceived error measurements are plotted for two TC basins. In Fig. 4, we examine the same 85 perceived error data points over the 2001–17 period that appear on the Atlantic panel of Fig. 1 in LC18 (except for the omission of tropical depressions, and the use of 36-h lead-time data in place of 24-h lead-time data). Unlike the linear interpolation used by LC18, the simulated perceived error lines in Fig. 4 are based on the four-parameter error model of SAFE-s. With a combination of Eqs. (5), (8), and (9), the evolution of perceived error at 36-h and longer lead times over the years can be simulated as a sum of true forecast error plus best track analysis error:

$$x_{p,i,j}^2 = x_0^2 e^{\beta(j-1)} e^{\alpha(i-1)dt} + \gamma x_0^2 e^{\beta(j-1)}, \tag{11}$$

where $x_{p,i,j}$ is the perceived error at lead time i = 3–7 (36, ..., 120 h) for years j = 1, ..., 17 (2001, ..., 2017 calendar years). As illustrated by Fig. 4, the perceived error formula in Eq. (11) can be linearly interpolated or extrapolated only over shorter periods of time.

Figure 4 offers a visual representation of how just four parameters of the error behavior model can explain the bulk (89%) of the variance in the perceived error measurements. For a subjective evaluation, the independent perceived error data for 2018 and 2019 are also shown with open symbols. Contingent on the validity of assumptions P-6 and P-5/1, the extrapolated parts of the curves in Fig. 4 offer a projection of the expected behavior of *perceived error* into the future. Note that since best track errors are smaller than NWP analysis errors and therefore much smaller than forecast errors, the right-hand side of Eq. (11) is dominated by the forecast error term. It follows that the evolution of perceived error over the years, especially at longer lead times, is well approximated by exponential decay.

Projection of true error. Focusing now on the evolution of true error, we note that both forecast error [growth as a function of lead-time days; Eq. (4)] and NWP analysis error [reduction as a function of years; Eq. (9)] follow exponential behavior, with asymptotic values of zero at minus and plus infinity, respectively. The forecast and analysis error evolution curves thus differ only in a single exponent, α and β, respectively, whose numerical values depend on the choice of time units. With a simple horizontal stretching, the two curves can be perfectly overlain, attesting to the special type of similarity among exponential curves with a common asymptotic value. Note that both forecast error growth and analysis error reduction are assumed to be exponential with stationary exponents over the years. It follows then that the evolution of both true analysis and true forecast error, past and future, can be represented with a *single exponential curve*, using two sets of horizontal scales, one for the growth of forecast errors in days, and another for the reduction of analysis errors over years.

Figure 5 is a diagram just like that, based on the best estimate from the four-parameter error model. True analysis (dots) and forecast (solid line) error estimates for years 2001 (red) and 2017 (blue), as examples, are highlighted. For reference, the corresponding 85 seasonal mean 36–120-h perceived forecast error measurements from the developmental sample (black stars), are also displayed. The latter are the same points shown in Fig. 4, except for a horizontal transposition. Perceived error measurements for the two independent seasons of 2018 and 2019 (pink and green stars) are also shown for a subjective evaluation. As seen from Fig. 5, the reduction of analysis error over the 17-yr experimental period is equivalent to the accumulation of forecast error in about 1.7 days. Interestingly, a similar,

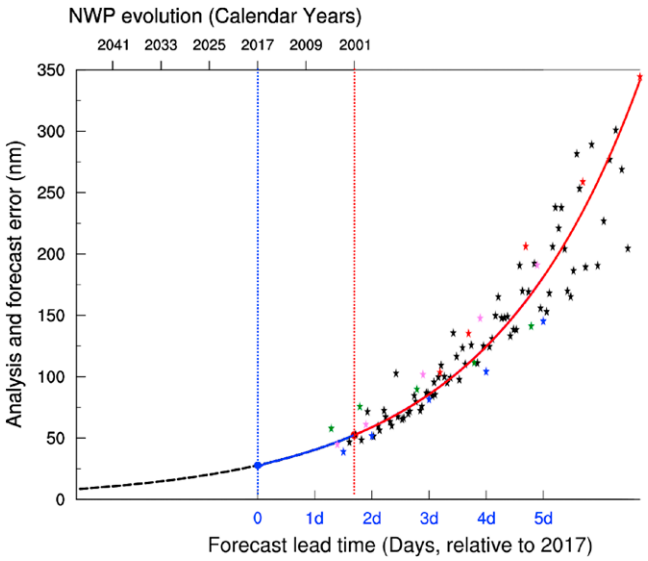

Fig. 5. Exponential growth of true TC positional forecast error and exponential reduction of true analysis error as a function of lead time (days; bottom horizontal axis) and years (top horizontal axis), respectively, estimated for the past (solid line), and extrapolated into the future (dashed line) using the best estimate from the four-parameter error model. The red and blue dots indicate the estimated initial (analysis) error in 2001 and 2017, respectively. A comparison of the bottom and top horizontal scales indicates (cf. red and blue vertical dotted lines) that the lead time of forecasts with any given skill increased by about 1 day decade^{-1} in the past. For reference, seasonal mean perceived positional error measurements are also shown as black stars for the developmental period (red and blue stars for 2001 and 2017, respectively), and pink and green stars for the independent 2018 and 2019 seasons, respectively. For further details, see text.

1-day lead time over a decade gain in forecast horizon is evident on the 500-hPa Northern and Southern Hemisphere extratropics annually averaged pattern anomaly correlation chart of Toth and Buizza (2018; skill in 1987 3-, 5-, and 7-day forecasts extended to 5, 7, and 10 days in 17 July 2007, respectively; see their Fig. 2.1). Whether the 1 day decade^{-1} lead-time gain is also reflective of other large-scale atmospheric processes in general will need to be explored in future studies.

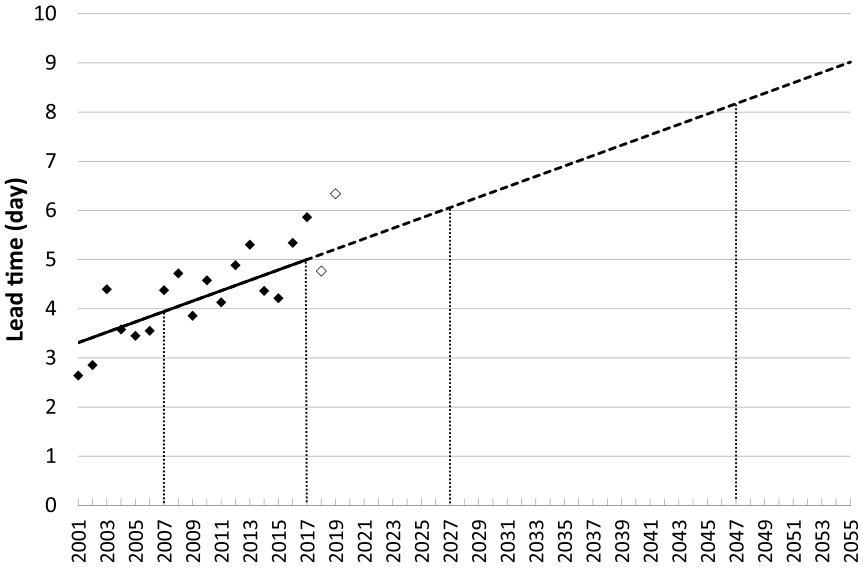

Fig. 6. Lead time until which true forecast error remains below 181 n mi (reached at 5 days in 2017) as simulated (solid line) and extrapolated (dashed line) based on the best estimate from the four-parameter error model. Perceived error measurement-based, linearly interpolated lead times for past years [closed and open diamond for dependent (2001–17) and independent (2018–19) seasons, respectively] are also shown for reference. For further details, see text.

Given the equivalence between a 1-day loss of forecast, and a 10-yr gain of analysis information, Fig. 6 captures a key message of this study: the time horizon of TC position forecasts is expected to expand by 1 day decade^{-1}. In other words, any preselected level of skill that is maintained through day N lead time today is expected to be retained through day N + 1, 2, etc., in one, two, etc., decades into the future. As the future projection here is about relative (i.e., compared to current, and not absolute) skill, its validity is independent of the uncertainty in the estimates of initial (analysis) error variance. Since, as pointed out earlier, uncertainty in the best track position is not relevant here either, the projection depends only on the accuracy of the estimate for the rate of analysis error reduction. Given the ±0.5% (Table 1) uncertainty in the estimate of the rate of analysis error variance reduction, the uncertainty in the projection of TC positional forecast errors can be quantified as a 22–29-h gain of skill per decade. This projection is contingent on the error growth rate (assumption P-6) and the investments into NWP development remaining steady (assumption P-5/1), as observed in past decades.

Conclusions

Tropical cyclones (TCs) are well-identifiable, coherent circulation patterns with potentially large societal impact. The predictability time limit of the atmosphere (i.e., the time period forecast error variance remains below a predefined threshold, given an initial error variance) has been evaluated, conditioned on the presence of such phenomena both in forecasts and observations. Analysis and forecast error variance with respect to reality (i.e., true error) was estimated using a simplified version of the inverse Statistical Analysis and Forecast Error method (SAFE-s). Perceived error variance measurements in the position of TCs in the NHC official forecast as compared to the "best track" analysis was related to true error variance using a series of assumptions, validated by the same experimental measurements. The positional error norm used emphasizes the large-scale circulation of the atmosphere and at initial time reflects not only the error in the initial position of the storm but also in its surroundings affecting the TC's position down to 5 days lead time. Key outcomes are as follows:

- With just four parameters, the error behavior model of SAFE-s explains 89% of the variance in the 36–120-h lead time 2001–17 seasonal mean perceived error measurements. The best estimates for the true initial and best track analysis errors in 2017 are 28 and 11 n mi, respectively, and for the growth of forecast and the reduction of analysis errors from day to day and year to year are 1.45 and 0.96, respectively.
- Simulated perceived error offers a statistically acceptable fit to perceived error measurements, indicating that the assumptions behind SAFE-s are consistent with the experimental measurement data. The results suggest SAFE can be used to estimate true errors not only when perceived errors are evaluated against the same analysis that NWP forecasts are initialized from but also when mostly independent, observationally based analyses are used as reference for truth.
- Perceived error measurements at and beyond 36-h lead time can be fully explained as a sum of exponentially growing initial errors (i.e., true forecast error) and the error in the verifying (best track) analysis, without the need for a model-error-related term. This is an indication that the large-scale dynamics of the atmosphere that affect the propagation of TCs as naturally occurring tracers is well captured by modern NWP models (i.e., no model-related track error; assumption 1).
- The first-ever objective estimate of the true error in the best track position of TCs is about 11 n mi in 2017, or about 38% of the error in the NWP analysis position (assumption 4).
- Initial (analysis) error is found to follow exponential growth in the official NHC forecast (that is based on a subjectively weighted composite of the best-performing NWP forecasts; assumption 2).
- Even at day 5 lead time, forecast error shows no sign of nonlinear saturation. This is related to the chosen norm and condition (i.e., the distance in the position of TCs when a TC is present both in the forecast and observations). With this norm, the steady force arising from atmospheric instabilities that make the observed and forecast storms diverge would remain effective until the distance of the storms approaches the order of magnitude of half the circumference of the latitude belt of TCs (approximately 9.800 n mi at latitude 25°). This is far larger than 5-day track errors even in the early part of the experimental period (~350 n mi).
- As error amplification remains stationary at about 1.45 day^{-1} over the 2001–17 experimental period (assumption 6), a single exponential curve can be used to describe the growth of positional error in all seasons studied.
- The error in the initial, analyzed condition of the official forecast, expressed in terms of TC positional uncertainty, exponentially decreases from year to year at a steady rate of 0.96 (assumption 5). This indicates that the sustained expenditures and talent invested into the development of the NWP observing, data assimilation, and modeling systems provides a steady force, yielding consistent improvements, leading to about a day extension of skill over a period of 10 years.
- It follows that if investments and talent dedicated for NWP improvements remain available as in the past, and the growth of errors also remains unchanged, the skillful range of TC prediction, given the uncertainty in our estimates, is expected to be extended in the future by 22–29 h $decade^{-1}$.

In conclusion, while linear interpolation lines, as noted by LC18, provide noisy estimates unsuitable for extrapolation, the four-parameter error model developed using all available perceived error measurements allows the projection of TC track errors, albeit under hypothetical scenarios, into the future.

Discussion

For how long can forecasts improve? It has been theorized that even if in the distant future analysis errors could be so much reduced that errors would be dominated by inaccuracies in kilometer-scale processes (switch from 2D to 3D turbulence, the realm of convection), or at most by the viscosity of the atmosphere (dissipation effect; Lorenz 1969b), the growth of errors at correspondingly small amplitudes would increase so dramatically that the extra lead-time gain in predictability from improved initial conditions would become negligible. If true, this could be called a hard "bound" on atmospheric predictability. But how distant that point may be in the future? Zhang et al. (2019) found no evidence of a significant increase in the rate of error growth down to about a threefold decrease below today's initial error (as indicated by the quasi-linear growth of error on the logarithmic scale in their Figs. 3 and 4, see blue symbols in the range of 1.5–5-day lead time). Assuming that the current rate of 0.96 annual reduction in initial errors is maintained, the 1-day extension of skillful prediction per decade projection may be valid for at least 25–30 years into the future.

The influence of ENSO and climate change on error growth. As indicated in Table 1, there is only a minimal loss in explained variance as the SAFE-s error behavior model is simplified first with the introduction of assumption 4 (error variance in best track analysis is constant fraction of analysis error variance, 2% loss), and then assumption 5 (exponential reduction of analysis error, an additional 2% loss). In either case, we suspect that most of the unexplained variance may be associated with random sampling noise in the perceived error measurements. When assumption 6 of a constant growth rate is introduced, the loss of explained variance, though still small, is noticeably higher (over 4%) than with the other two assumptions. We expect that part of this unexplained error variance may be associated with seasonally varying levels of instabilities associated with different phases of ENSO (and possibly the Madden–Julian oscillation). Such a possible connection will need to be explored in future studies. On the other hand, faster error growth under possibly more unstable future climatological conditions (manifested, e.g., by possibly stronger wind circulation; Zeng et al. 2019) may shorten the predictability time horizon (i.e., lead time of useful forecasts shortened), and slow the extension of such a limit over the years.

No model-related error in TC track forecasts? An interesting, and possibly controversial outcome of this study is that numerical models themselves may not impart any noticeable error into TC track forecasts. Correspondingly, track errors may originate exclusively from initial error. In a series of SAFE error estimation studies, PT14 and Feng et al. (2017, 2020) found no indication for model related errors in extratropical circulation forecasts either.[6] How is this possible, given the countless studies demonstrating a range of problems with NWP models?

We first note that having no model related component in track or large-scale circulation errors is not an indication that NWP models are perfect. No model-related track error only implies that the simulation of *large-scale dynamics* responsible for the transposition of TCs, or for the evolution of synoptic-scale flow may be near perfect. At the same time, many finer-scale processes may not be represented well or at all in global NWP models. We hypothesize that due to their size, many of these finer-scale processes would not have a material effect on how realistically the large-scale dynamics is captured in models. In other words, it is only the statistics of finer-scale processes that directly affect the dynamics of the large-scale circulation. This notion is born out by the success of NWP models that, by necessity, parameterize the effects of all motions with scales finer than the effective spatial resolution of the models. Finer-scale atmospheric processes, though critical for the successful

[6] Some slow processes in the tropics are reported to be missing from an NCEP model by PT14. If these are in the deep tropics, however, they would have no effect on TC translation governed by subtropical and then midlatitude flow.

short-range prediction or climatological simulation of important characteristics of weather (e.g., convective bands, tornadoes), may not directly affect, for example. the movement of a TC as a whole.

Nevertheless, the earlier and current SAFE results are in apparent conflict with a long series of studies concluding that beyond the dynamically "internal" divergence of model solutions, forecasts are also affected by an "external" error presumably due to model deficiencies (Leith 1978, 1982). In particular, the growth of forecast error, especially at short ranges, has been found consistently and significantly higher when forecasts are compared against verifying analysis fields (this error is referred to as "perceived error" in the present study), as opposed to another forecast [started from an earlier or later, or a perturbed initial condition—referred to as internal error (e.g., Lorenz 1984; Savijarvi 1995)]. The faster growth of perceived, compared to internal error is commonly attributed to the emergence of model related errors, usually without identifying any specific mechanism, and is generally considered as a telltale sign of model imperfection.

In their discussion, Feng et al. (2020) offer an alternative explanation for the superexponential growth of perceived, versus exponential growth of internal (or in their parlance, true) short range forecast error. They demonstrate an initially strong, then decaying correlation between error in forecast and corresponding verifying analysis fields [cf. Eq. (3) in present paper]. This results in biases when short range true error is estimated by perceived error. In particular, true error is underestimated, while its growth is overestimated. A serious overestimation of error growth based on perceived error behavior is indeed found both in operational and simulated [observing system simulation experiment (OSSE)] forecast environments [see Figs. 7 and 3 in Feng et al. (2020), respectively]. This finding may question the need for, and the validity of the prevalent "external" model error hypothesis for explaining the superexponential growth of perceived error as compared to the exponential growth of "internal" differences between lagged or perturbed ensemble forecasts, or true error. The topic of model related error certainly deserves more exploration in future studies.

Comparison with previous estimates. To our knowledge, of the four unknown parameters in Table 1, the annual decrease and the size of the true initial position error have not been estimated before. Subjective positional error estimates for TCs and hurricanes combined for the 2000–09 (Torn and Snyder 2012; real time operational position) and 2010 (Landsea and Franklin 2013; best track position) seasons are around 26 and 20 n mi, respectively. These estimates appear to be somewhat larger than the 15–20 and 14–19 n mi range estimates from the present study, for the respective periods (cf. best track error and its uncertainty estimates in Table 1).

As for NH TC track forecasts, with a comparison of lagged forecasts from ECMWF, Met Office, and MeteoFrance, Plu (2011) estimates that error doubling is in the range of 30–50 h. The estimate with the best performing, ECMWF system is 42 h, which equals to an amplification of 1.49 over a 24-h period, identical to Aberson's (1998) estimate using naturally occurring analogs, and relatively close to the 1.454 estimate in Table 1 of the present study. Note that Plu's numbers refer to all NH basins and are admittedly influenced by serial correlations among the intercompared forecasts, leading to a possible overestimation of growth rate, thus precluding any precise comparison with results reported in Table 1.

Aberson's 1.49 24-h growth-rate estimate, however, reflects the behavior of a large number of Atlantic TCs *observed* over the 1956–95 period, warranting a more careful comparison. We first note that though Aberson's 1.49 amplification factor is close to, it is clearly out of the 1.432–1.476 range estimated in the present study (cf. the growth rate and uncertainty range in Table 1). Model-based predictability results are often conceived as too optimistic—How would a purely observationally based study find a predictability horizon beyond that found

in the present SAFE-s study, using model-based NHC official forecasts? Why would the official forecasts exhibit slower error growth than Aberson's (1998) analog method? Part of the 3% difference between 24-h amplification rates in Aberson's (1998) and the present studies may be explained by the sampling of different circulation regimes in the two nonoverlapping datasets used. The faster growth rate between TC trajectories reported by Aberson (1998), however, may be more due to the methodologies used. Aberson (1998) intends to assess the growth rate of errors developing between the tracks of initially close-by storms. As his 40-yr observational period is way too short for a storm to emerge in close proximity to another storm (Lorenz 1969a) and Aberson (1998) applies a technique suggested by Fraedrich and Leslie (1989), where all storm trajectories are transposed in space so that their initial positions are exactly collocated. The observed storms of course develop and move under conditions in their original position, scattered across the Atlantic basin, and not under the conditions to where they are collocated. The divergence rate of their initially collocated tracks therefore reflects not only the internal unstable dynamics of the atmosphere, but also the diversity of climatologically prevailing conditions over different parts of the Atlantic basin from where initial storm positions are collocated (e.g., sun elevation, proximity to landmasses, large-scale environmental flow), which would be near identical without the artificially relocation of the storms.

In the parlance of NWP modeling, the relocation amounts to a misspecification of the large-scale environmental conditions, triggering an extra level of divergence between TC trajectories, beyond what is due to naturally occurring instabilities at any locale. Though based on observational data only, the transposition of TCs in Aberson (1998; and in the original study of Fraedrich and Leslie 1989) amounts to the introduction of a methodological approximation akin to "model error" in NWP. We speculate that this methodological approximation in the analog study may lead to an error growth faster than what would be observed with a perfect model. Note that even a more advanced analog approach was found to produce forecasts with significantly faster error growth than a state-of-the-art NWP system over the 1991–2000 time period (see Fig. 3[7] in Fraedrich et al. 2003). Paradoxically, in this case limited resolution, thus necessarily imperfect NWP models may offer a better representation of the large-scale dynamics of TC movements than a model based on atmospheric analogs (produced by nature itself).

[7] We attribute the higher NWP errors (most visible before 48-h lead time) to biases present in NWP track forecasts from the 1980s and 1990s that are virtually nonexistent in statistical forecasts.

Acknowledgments. The first author expresses her appreciation for the hospitality she enjoyed as a Visiting Scientist at the Global Systems Laboratory of NOAA/OAR/ESRL in Boulder, Colorado, where most of this study was carried out. The tropical cyclone forecast error database was kindly provided by Chris Landsea and John Cangialosi of the National Hurricane Center. The authors are also grateful to Jie Feng (Oklahoma University) and Tim Marchok (Geophysical Fluid Dynamics Laboratory) for their valuable comments and suggestions. Discussions with several colleagues at the National Hurricane Center of NOAA are gratefully acknowledged. Expert comments by Chris Snyder, Nathan Hardin, and two anonymous reviewers greatly improved the presentation of the material. The research was jointly supported by the National Key Research and Development Program of China (Grants 2017YFC1501601), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences.

Data availability statement. The perceived NHC official track forecast error data used in this study were obtained for the 2001–16 seasons on 25 August 2018, and for the 2017 season on 29 May 2019, from www.nhc.noaa.gov/verification/errors/1970-present_OFCL_v_BCD5_ind_ATL_TI_errors_noTDs.txt. Preliminary perceived track error data for the 2018–19 seasons were kindly provided by Dr. John Cangialosi of NHC on 22 May 2020. As the NHC datasets referenced above undergo occasional corrections (at which time the corrected data files are overwritten), the data files used in the present study are posted under the same file name at www.escience.cn/people/zhoufeifan/index.html.

Appendix: The choice of forecast error metric

It is well understood that due to the chaotic nature of the atmosphere, errors in weather forecasts grow until all forecast skill is lost (e.g., Lorenz 1969b). This is a consequence of the divergence of initially close-by trajectory segments in the phase space of the atmosphere and its models. The rate of divergence, hence the quality of forecasts can be evaluated by a vast array of metrics (e.g., Jolliff and Stephenson 2003). Depending on the distance norm, variable, and domain chosen, each metric emphasizes different aspects of the same observed (and forecast) circulation. Consequently, the apparent performance of forecast systems is generally metric dependent.

It is important to recognize that the subjective choice of metric does not influence, nor fully reflect the objective evolution of the observed and forecast circulations, or their difference (from which various error metrics are derived). From the perspective of nature, the choice of metric is always arbitrary, driven by study objectives and practical considerations. As we understand from adjoint and other sensitivity studies (e.g., Wu et al. 2007) that the track of TCs (and associated forecast errors), for example, depend not only on initial position (or error) of TCs, but also on their intensity and broader environment, which the TC track error metric does not reflect.

The choice of TC track error as a metric in the present study was motivated by its long term use and the availability of historical records. It follows that in SAFE-s, where initial error is estimated from forecast error measurements using an inverse method, initial error (x_0) is affected by the quality of not only the TC center position, but also of other features in NWP analyses (feeding into the official forecast) that have an effect on the quality of ensuing TC position forecasts. All these effects are reflected in the metric of positional error of the center of the TC, in the units of nautical miles. This is consistent with the common knowledge [also discussed by Plu (2011) and Landsea and Franklin (2013)] that forecast TC position (and its error) depends more on the environment, and less on the position of a TC at initial time.

References

Aberson, S. D., 1998: Five-day tropical cyclone track forecasts in the North Atlantic basin. *Wea. Forecasting*, **13**, 1005–1015, https://doi.org/10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Birgin, E. G., J. E. Martinez, and R. Marcos, 2001: Algorithm 813: SPG—Software for convex-constrained optimization. *ACM Trans. Math. Software*, **27**, 340–349, https://doi.org/10.1145/502800.502803.

Cangialosi, J. P., 2019: 2018 Hurricane season. National Hurricane Center Forecast Verification Rep., 73 pp., www.nhc.noaa.gov/verification/verify3.shtml.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474–491, https://doi.org/10.3402/tellusa.v39i5.11774.

Feng, J., Z. Toth, M. Pena, 2017: Spatially extended estimates of analysis and short-range forecast error variances. *Tellus*, **69A**, 1325301, https://doi.org/10.1080/16000870.2017.1325301.

——, ——, ——, and J. Zhang, 2020: Partition of analysis and forecast error variance into growing and decaying components. *Quart. J. Roy. Meteor. Soc.*, **146**, 1302–1321, https://doi.org/10.1002/qj.3738.

Fraedrich, K., and L. M. Leslie, 1989: Estimates of cyclone track predictability. I: Tropical cyclones in the Australian region. *Quart. J. Roy. Meteor. Soc.*, **115**, 79–92, https://doi.org/10.1002/qj.49711548505.

——, C. C. Raible, and F. Sielmann, 2003: Analog ensemble forecasts of tropical cyclone tracks in the Australian region. *Wea. Forecasting*, **18**, 3–11, https://doi.org/10.1175/1520-0434(2003)018<0003:AEFOTC>2.0.CO;2.

Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.*, **58**, 3525–3539, https://doi.org/10.1175/1520-0469(2001)058<3525:LRDIHA>2.0.CO;2.

Hurricane Research Division, 2019: Tropical cyclone records. NOAA, www.aoml.noaa.gov/hrd/tcfaq/E6.html.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.

Komaromi, W. A., and J. D. Doyle, 2017: Tropical cyclone outflow and warm core structure as revealed by HS3 dropsonde data. *Mon. Wea. Rev.*, **145**, 1339–1359, https://doi.org/10.1175/MWR-D-16-0172.1.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

——, and J. P. Cangialosi, 2018: Have we reached the limits of predictability for tropical cyclone track forecasting? *Bull. Amer. Meteor. Soc.*, **99**, 2237–2243, https://doi.org/10.1175/BAMS-D-17-0136.1.

Lawrence, M. B., 1990: National Hurricane Center verification. NOAA, www.nhc.noaa.gov/verification/verify3.shtml.

Leith, C. E., 1978: Objective methods for weather prediction. *Annu. Rev. Fluid Mech.*, **10**, 107–128, https://doi.org/10.1146/annurev.fl.10.010178.000543.

——, 1982: Statistical methods for the verification of long and short range forecast. *Seminar on Problems and Prospects in Long and Medium Range Weather Forecasts*, Reading, ECMWF, 313–333, www.ecmwf.int/en/elibrary/10708-statistical-methods-verification-long-and-short-range-forecasts.

Leung, T. Y., M. Leutbecher, S. Reich, and T. G. Shepherd, 2019: Atmospheric predictability: Revisiting the inherent finite-time barrier. *J. Atmos. Sci.*, **76**, 3883–3892, https://doi.org/10.1175/JAS-D-19-0057.1.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

——, 1969a: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

——, 1969b: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

——, 1984: Estimates of atmospheric predictability at medium range. *Predictability of Fluid Motions*, G. Holloway and B. West, Eds., American Institute of Physics, 133–139, http://eaps4.mit.edu/research/Lorenz/Estimates_of_Atmospheric_Predictability_1984.pdf.

Madden, R., and P. Julian, 1971: Detection of a 40-50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702–708, https://doi.org/10.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2.

NHC, 2019: Forecast verification procedures. NOAA, www.nhc.noaa.gov/verification/verify2.shtml.

Peña, M., and Z. Toth, 2014: Estimation of analysis and forecast error variances. *Tellus*, **66A**, 21767, https://doi.org/10.3402/tellusa.v66.21767.

Plu, M., 2011: A new assessment of the predictability of tropical cyclone tracks. *Mon. Wea. Rev.*, **139**, 3600–3608, https://doi.org/10.1175/2011MWR3627.1.

Rotunno, R., and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063–1076, https://doi.org/10.1175/2007JAS2449.1.

Savijarvi, H., 1995: Error growth in a large numerical forecast system. *Mon. Wea. Rev.*, **123**, 212–221, https://doi.org/10.1175/1520-0493(1995)123<0212:EGIALN>2.0.CO;2.

Schenkel, B. A., and R. E. Hart, 2012: An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets. *J. Climate*, **25**, 3453–3475, https://doi.org/10.1175/2011JCLI4208.1.

Schubert, S. D., and M. Suarez, 1989: Dynamical predictability in a simple general circulation model: Average error growth. *J. Atmos. Sci.*, **46**, 353–370, https://doi.org/10.1175/1520-0469(1989)046<0353:DPIASG>2.0.CO;2.

Torn, R. D., and C. Snyder, 2012: Uncertainty of tropical cyclone best-track information. *Wea. Forecasting*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

Toth, Z., and R. Buizza, 2018: Weather forecasting: What sets the forecast skill horizon? *The Gap between Weather and Climate Forecasting: Subseasonal to Seasonal Prediction*, A. Robinson and F. Vitard, Eds., Elsevier, 17–45.

Vigh, J. L., and W. H. Schubert, 2009: Rapid development of the tropical cyclone warm core. *J. Atmos. Sci.*, **66**, 3335–3350, https://doi.org/10.1175/2009JAS3092.1.

Wang, C., C. Deser, J. Yu, P. N. DiNezio, and A. C. Clement, 2017: El Nino and Southern Oscillation (ENSO): A review. *Coral Reefs of the Eastern Tropical Pacific*, P. W. Glynn, D. Manzello, and I. Enochs, Eds., Coral Reefs of the World, Vol. 8, Springer, 85–106, https://doi.org/10.1007/978-94-017-7499-4_4.

Wu, C. C., J. H. Chen, P. H. Lin, and K. H. Chou, 2007: Targeted observations of tropical cyclone movement based on the adjoint-derived sensitivity steering vector. *J. Atmos. Sci.*, **64**, 2611–2626, https://doi.org/10.1175/JAS3974.1.

Yamaguchi, M., and S. J. Majumdar, 2010: Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts. *Mon. Wea. Rev.*, **138**, 3634–3655, https://doi.org/10.1175/2010MWR3176.1.

Zeng, Z., and Coauthors, 2019: A reversal in global terrestrial stilling and its implications for wind energy production. *Nat. Climate Change*, **9**, 979–985, https://doi.org/10.1038/s41558-019-0622-6.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S. J. Lin, J. H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.