Using Forecast Temporal Variability to Evaluate Model Behavior

CAROLYN A. REYNOLDS, ELIZABETH A. SATTERFIELD, AND CRAIG H. BISHOP

Naval Research Laboratory, Monterey, California

(Manuscript received 4 March 2015, in final form 28 August 2015)

ABSTRACT

The statistics of model temporal variability ought to be the same as those of the filtered version of reality that the model is designed to represent. Here, simple diagnostics are introduced to quantify temporal variability on different time scales and are then applied to NCEP and CMC global ensemble forecasting systems. These diagnostics enable comparison of temporal variability in forecasts with temporal variability in the initial states from which the forecasts are produced. They also allow for an examination of how day-to-day variability in the forecast model changes as forecast integration time increases. Because the error in subsequent analyses will differ, it is shown that forecast temporal variability should lie between corresponding analysis variability and analysis variability minus 2 times the analysis error variance. This expectation is not always met and possible causes are discussed. The day-to-day variability in NCEP forecasts steadily decreases at a slow rate as forecast time increases. In contrast, temporal variability increases during the first few days in the CMC control forecasts, and then levels off, consistent with a spinup of the forecasts starting from overly smoothed analyses. The diagnostics successfully reflect a reduction in the temporal variability of the CMC perturbed forecasts after a system upgrade. The diagnostics also illustrate a shift in variability maxima from storm-track regions for 1-day variability to blocking regions for 10-day variability. While these patterns are consistent with previous studies examining temporal variability on different time scales, they have the advantage of being obtainable without the need for extended (e.g., multimonth) forecast integrations.

1. Introduction

The ability to assess forecast model skill is a necessity for improving forecast systems. Forecast error has been monitored regularly at operational centers for decades for a variety of metrics, and long records of these performance metrics document steady improvements in the forecasting systems over the years (e.g., Simmons and Hollingsworth 2002; Magnusson and Källén 2013). Forecast error is due to both errors in the forecast model (often in the physical parameterizations) and errors in the initial state, and attributing forecast errors to initial condition errors or model errors is a difficult but worthy goal as it helps guide forecast system development. Lorenz (1982) noted that the differences between forecast error growth rates and growth rates of differences between consecutive forecasts are related to model deficiencies and that, as forecast models improve, these two growth rates should converge. Attempts to identify causes of the nonsystematic component of forecast error include that of Dalcher and Kalnay (1987), building on the work of Leith (1978), in which it is assumed that the exponential part of error growth is due to the self-growth of initial condition errors, while the linear part is due to model deficiencies. Several follow-on studies have applied similar error attribution models to different forecasting systems (e.g., Stroe and Royer 1993; Reynolds et al. 1994; Simmons et al. 1995; Savijarvi 1995; Simmons and Hollingsworth 2002; Magnusson and Källén 2013; Peña and Toth 2014). Other methods developed for error attribution include geometric or shadowing techniques (Judd et al. 2008) and a mapping paradigm technique (Toth and Peña 2007).

A component of forecast error usually attributed to model error is the time-mean forecast error, often referred to as bias. Model bias is examined regularly for weather forecasts, as well as seasonal forecasts and climate integrations (e.g., Klocke and Rodwell 2014). As has been shown for other metrics, weather forecast biases have been decreasing as forecast systems improve (e.g., Jung 2005). Forecast models can also be biased in...
ways that are not necessarily reflected in time-mean errors. Forecast models may exhibit biases in spatial or temporal variability. Skamarock (2004) evaluates aspects of this type of bias by comparing the kinetic energy spectra of forecasts to the $k^{-3}$ spectra observed at larger scales and $k^{-5/3}$ spectra observed on the mesoscale by Nastrom and Gage (1985). Many studies (e.g., Blackmon et al. 1977; Trenberth 1981; Lau and Nath 1987; Cai and Van Den Dool 1991) have evaluated the temporal variability of the atmosphere based on long-term sequences of analyses, using filters to examine variability on different time scales. If long model integrations are available (e.g., AMIP- or CMIP-type studies) then it is possible to compare the atmospheric model temporal variability on different time scales with the analysis variability, as in Lau and Nath (1987). Biases in temporal variability have been interpreted in light of biases in the time-mean state and subsequent waveguide characteristics (e.g., Reynolds and Gelaro 1997). Analyses that separate variability into westward and eastward components have been very useful for diagnosing forecast model deficiencies in the simulation of equatorially trapped waves (e.g., Wheeler and Kiladis 1999; Lin et al. 2006). However, these types of temporal variability diagnostics require a long (usually multimonth to multiyear) forecast integration, something that is often not available from a weather forecast model.

Development of new verification methods, especially those designed to deal with high-resolution forecasts for which traditional verification scores may not be well suited, is a very active area of research. The papers of Gilleland et al. (2009) and Gilleland et al. (2010) provide overviews and intercomparisons of spatial forecast verification methods, including neighborhood, scale-separation, feature-based, and field deformation techniques. While these methods are designed to quantify total forecast error, there is potential for these methods to be used for error attribution as well. In particular, scale-separation techniques, in which forecast error is isolated by scale (e.g., Briggs and Levine 1997; Tustison et al. 2001; Harris et al. 2001; Casati et al. 2010), may provide information on the forecast model’s ability to reproduce observed variability-scale structure, especially if initial condition error can be ruled out as a source of these forecast errors.

In this paper we demonstrate simple diagnostics that can be used to determine if forecast temporal variability accurately captures the variability of the filtered versions of reality that the analyses and forecasts are designed to represent. The diagnostics also allow for an examination of how temporal variability on 1-day time scales in the forecast model can change as the forecast integration time increases. We propose that temporal variability diagnostics may serve as a useful complement to traditional model error diagnostics that examine time-mean errors (e.g., Klocke and Rodwell 2014), spatial scale-separation techniques (e.g., Harris et al. 2001), and new techniques to quantify differences between forecast fields and reality as represented on the scales resolved by the data-assimilation and forecast systems (e.g., Peña and Toth 2014). Diagnostics to assess both spatial and temporal variability will become increasingly important as stochastic techniques, which often have tunable spatial and temporal correlation scales, proliferate. Diagnostics measuring temporal variability (the mean square differences between fields 12 or 24 h apart in a forecast integration) have been used at operational centers such as ECMWF and NCEP to monitor model activity and detect consequences of model changes (A. Persson and Z. Toth 2015, personal communication), but we are not aware of published results concerning these diagnostics. In addition, we derive expected bounds on forecast temporal variability as a function of analysis temporal variability and analysis error variance. We demonstrate the utility of the diagnostics in evaluating the relative fidelity of the forecast model temporal variability for different metrics and different models. We apply these diagnostics to both control and perturbed forecasts produced from the National Centers for Environmental Prediction (NCEP) and Canadian Meteorological Centre (CMC) global ensemble forecasting systems. The diagnostics reflect differences in ensemble design and also reflect upgrades to the CMC ensemble system in a manner that is consistent with the expected impacts of the upgrades. The diagnostics and datasets are described in section 2, results are presented in section 3, and a summary and conclusions are given in section 4.

2. Methodology and dataset description

a. Description of diagnostics

The temporal variability of the initial states $a_{i}^{\text{exper}}$ is characterized by taking the root-mean-square difference of a particular variable at a particular location, $i$ days apart; that is,

$$\left( a_{i}^{\text{exper}} \right)^{2} = \frac{1}{n_{i}} \sum_{j=1}^{n_{i}} \left( x_{i,j}^{0} - x_{i,j} \right)^{2}, \quad (1)$$

where $x$ is the model analysis or forecast variable at a particular pressure level and grid point, the subscript indicates the start day of the forecast, and the superscript indicates the forecast time (in days). In this case, $x_{i,j}^{0}$ indicates an initial state valid on day $j$, and $n_{i}$ is the total number of consecutive initial states $x_{i,j}^{0}$ considered. We calculate $a_{i}^{\text{exper}}$ for $i = 1, 10$. The sequence of initial states considered (e.g., the initial states for the control or perturbed ensemble members) are denoted by the superscript exper.
and detailed in Table 1; $a_i$ gives the average temporal variability (measured by root-mean-square difference) in a series of initial states for a time scale of 1 day, while $a_{10}$ gives the average temporal variability over a time scale of 10 days. On average, we expect $a_i$ to increase with increasing $i$, as more modes of variability (associated with longer time scales) contribute to the mean square differences. The calculation may be thought of as analogous to a high-pass filter, in that the full amplitude of modes with periods less than $i$ can contribute to the metric, whereas only a portion of the full amplitude of modes with periods greater than $i$ can contribute to the metric. We also point out that $a_i$ is the mean square error associated with a persistence forecast of length $i$ days.

The temporal variability of the forecasts $f_i^{\text{exper}}$ is characterized in a very similar manner; that is,

$$ (f_i^{\text{exper}})^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_i^{\text{exper}} - x_{j_i}^0)^2, \tag{2} $$

where $x_i^{\text{exper}}$ is the forecast state variable at a particular pressure level and grid point at forecast length $i$ days from the forecast started at day $j$. The prime is used to indicate that the time-mean bias has been removed from the forecast before the square difference is taken. We calculate $f_i^{\text{exper}}$ for $i = 1, 10$. As with $a_i^{\text{exper}}$, initial states and forecasts considered are denoted by the superscript exper and detailed in Table 1. As with $a_i$, we expect $f_i$ to increase with increasing $i$. Differences between $a_i$ and $f_i$ will illustrate how the model forecast variability differs from the variability found in the series of initial states on which the forecasts are based.

In section 2b, below, we derive the expectation that, under certain constraints, $f_i$ should be less than $a_i$ owing to uncorrelated errors in the initial states. There are also other possible reasons as to why $f_i$ and $a_i$ may differ. If, for example, the forecast model becomes less energetic with increasing forecast time, then $f_i$ may steadily diverge from $a_i$. If the forecast model accurately captures phenomena responsible for variability on daily time scales, but fails to capture more slowly varying phenomena, then $f_i$ and $a_i$ may be similar for small values of $i$, but $f_i$ will be smaller than $a_i$ for larger values of $i$.

To investigate potential relationships between these diagnostics and standard forecast error diagnostics, we calculate the mean square forecast error:

$$ \text{MSE} = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_i^{\text{exper}} - x_{j_i}^0)^2. \tag{3} $$
As \( d_i^2 \) is the mean square forecast error of a persistence forecast, our expectation is that mean square error will be considerably smaller than \( d_i^2 \).

We are also interested in how the 1-day variability of the model forecast changes as forecast lead time increases. The 1-day temporal variability as a function of forecast integration time, \( d_i^{\text{exper}} \), is characterized as

\[
(d_i^{\text{exper}})^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_j - x_{j-1})^2. \tag{4}
\]

We calculate \( d_i^{\text{exper}} \) for \( i = 1, 10 \). Changes in \( d_i \) with increasing \( i \) will indicate if the 1-day variability in the forecast model changes with increasing forecast time. As noted above, ECMWF and NCEP use diagnostics similar to (1), (2), and (4) to monitor model activity and detect consequences of model changes (A. Persson and Z. Toth 2015, personal communication), but we are not aware of published results concerning these diagnostics. The choice of notation, \( a_i \) for “analysis” temporal variability, \( f_i \) for “forecast” temporal variability, and \( d_i \) for “daily” or “day-to-day” (forecast) temporal variability, hopefully provides an intuitive relationship between the notation and its meaning.

**b. Derivation of bounds on expected temporal variability**

Given certain assumptions, we can derive an expected relationship for the relative magnitudes of \( a_i \) and \( f_i \). Note that since each initial state is equal to the true filtered state \( \bar{f}_j \) plus an analysis error \( e_j \), (1) can be rewritten in the form

\[
(a_i^{\text{exper}})^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} [(f_{j+i} + e_{j+i}) - (f_j + e_j)]^2
= \frac{1}{n_i} \sum_{j=1}^{n_i} [(f_{j+i} - f_j) + (e_{j+i} - e_j)]^2. \tag{5}
\]

Now, since analysis error is closely tied to observation error and since observation bias correction attempts to remove the dependence of observation error on the true state, it is of interest to assume that \( 1/n_i \sum_{j=1}^{n_i} (f_{j+i} - f_j) (e_{j+i} - e_j) = 0 \) provided \( n_i \) is large. With this plausible assumption, (5) becomes

\[
(a_i^{\text{exper}})^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (f_{j+i} - f_j)^2 + \frac{1}{n_i} \sum_{j=1}^{n_i} (e_{j+i} - e_j)^2. \tag{6}
\]

For a global model, the assumption that \( \langle (e_j^a)^2 \rangle = \langle (e_{j+1}^a)^2 \rangle \) is accurate (where the angle brackets indicate the average over an infinite time series). With this assumption the temporal correlation of analysis errors \( p \) satisfies the equation

\[
p = \langle (e_{j+1}^a + e_j^a) \rangle / \sqrt{\langle (e_{j+1}^a)^2 \rangle} = \langle e_{j+1}^a e_j^a \rangle / \langle (e_j^a)^2 \rangle.
\]

Hence, without loss of generality, we can statistically model the relationship between \( e_j^a \) and \( e_{j+1}^a \) by the equation \( e_{j+1}^a = p e_j^a + \eta \), where \( \eta \) is a random stochastic variable that is uncorrelated with \( e_j^a \) and has a mean of zero. By squaring both sides of \( e_{j+1}^a = p e_j^a + \eta \) and taking expectations, one can readily show that the condition \( \langle (e_j^a)^2 \rangle = \langle (e_{j+1}^a)^2 \rangle \) is met, provided \( \langle \eta^2 \rangle = (1 - p^2) \langle (e_j^a)^2 \rangle \). The equation \( e_{j+1}^a = p e_j^a + \eta \) implies that

\[
\langle (e_{j+1}^a - e_j^a)^2 \rangle = \langle (e_{j+1}^a)^2 \rangle + \langle (e_j^a)^2 \rangle - 2p \langle (e_j^a)^2 \rangle
= 2(1 - p^2) \langle (e_j^a)^2 \rangle. \tag{7}
\]

Taking the expected value of (6) over an infinite number of independent sets of \( n_i \) data points and then using (7) yields

\[
\langle (a_i^{\text{exper}})^2 \rangle_{n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} (f_{j+i} - f_j)^2 + \frac{1}{n_i} \sum_{j=1}^{n_i} (e_{j+i} - e_j)^2
= \langle (f_{j+i} - f_j)^2 \rangle + 2(1 - p) \langle (e_j^a)^2 \rangle, \tag{8}
\]

where the angle brackets \( \langle \rangle_{n_i} \) indicate that the expectation is over independent sets of \( n_i \) realizations. In regions of few observations or regions with significant model bias, it is highly likely that \( p \langle (e_j^a)^2 \rangle = \langle e_{j+1}^a e_j^a \rangle > 0 \) and extremely unlikely that \( p \langle (e_j^a)^2 \rangle = \langle e_{j+1}^a e_j^a \rangle < 0 \). In addition, because \( p \) is a correlation, it is impossible for \( p \) to be greater than unity. Thus, in general, \( p \) is bounded by 0 \( \leq p < 1 \) and hence, from (8), \( \langle (a_i^{\text{exper}})^2 \rangle \) is bounded according to

\[
\frac{1}{n_i} \sum_{j=1}^{n_i} (f_{j+i} - f_j)^2 < \langle (a_i^{\text{exper}})^2 \rangle_{n_i}
\leq \frac{1}{n_i} \sum_{j=1}^{n_i} (f_{j+i} - f_j)^2 + 2\langle (e_j^a)^2 \rangle. \tag{9}
\]

The purpose of model diagnostics is to identify those situations in which the system behavior differs from that of a perfect system; hence, it is of interest to note that if we assume the initial and forecast states are drawn from the same distribution as the true states and that the model is perfect then (2) and (4) can be rewritten in the form

\[
\langle (e_j^a)^2 \rangle = \langle (e_{j+1}^a)^2 \rangle \]

\[
\langle (a_i^{\text{exper}})^2 \rangle = \frac{1}{n_i} \sum_{j=1}^{n_i} (f_{j+i} - f_j)^2 + \frac{1}{n_i} \sum_{j=1}^{n_i} (e_{j+i} - e_j)^2.
\]
\[
\langle (f_i^\text{exp})^2 \rangle_{n_i} = \left\langle \frac{1}{n_i} \sum_{j=1}^{n_i} [M_i(t_j^0) - \bar{f}_i]^2 \right\rangle_{n_i} = \left\langle \frac{1}{n_i} \sum_{j=1}^{n_i} [f_j^0 - \bar{f}_j]^2 \right\rangle_{n_i},
\]
\[
\langle (d_i^\text{exp})^2 \rangle_{n_i} = \left\langle \frac{1}{n_i} \sum_{j=1}^{n_i} [M_i(t_j^0) - \bar{f}_j^0]^2 \right\rangle_{n_i} = \left\langle \frac{1}{n_i} \sum_{j=1}^{n_i} (f_j^0 - \bar{f}_j^0)^2 \right\rangle_{n_i} = \langle (f_1^\text{exp})^2 \rangle_{n_i},
\]

where \( M \) represents the forecast model and the subscript represents the integration length in days. Using (9) in (10) then gives

\[
\langle (a_i^\text{exp})^2 \rangle_{n_i} > \langle (f_i^\text{exp})^2 \rangle_{n_i} \geq \langle (a_i^\text{exp})^2 \rangle_{n_i} - 2 \langle (a_i^\text{err})^2 \rangle_{n_i} \quad \text{and} \quad \langle (a_1^\text{exp})^2 \rangle_{n_i} \geq \langle (a_1^\text{exp})^2 \rangle_{n_i} - 2 \langle (a_1^\text{err})^2 \rangle_{n_i},
\]

(11)

where \( \langle (a_i^\text{err})^2 \rangle \) is the analysis error variance. Note that the lower bound given in (11) pertains to the extraordinary case in which the correlation between consecutive analysis errors is 0. Accurate estimates of the analysis error variance are generally harder to obtain than accurate estimates of the error variance \( \langle (e_i^\text{exp})^2 \rangle \) of the forecast used as the "first guess" in the data-assimilation scheme (e.g., Peña and Toth 2014). In useful data-assimilation schemes the analysis error variance is smaller than the forecast error variance. Hence, (11) also implies the bound

\[
\langle (a_i^\text{exp})^2 \rangle_{n_i} > \langle (f_i^\text{exp})^2 \rangle_{n_i} > \langle (a_i^\text{exp})^2 \rangle_{n_i} - 2 \langle (a_i^\text{err})^2 \rangle_{n_i}, \quad \text{and} \quad \langle (a_1^\text{exp})^2 \rangle_{n_i} \geq \langle (a_1^\text{exp})^2 \rangle_{n_i} - 2 \langle (a_1^\text{err})^2 \rangle_{n_i}.
\]

(12)

Equations (11) and (12) place bounds on the relationship between forecast variability and analysis variability for a data assimilation and forecasting system that is functioning the way it should. That is, forecast temporal variability should lie between corresponding analysis variability and analysis variability minus 2 times the analysis error variance. Hence, violation of the bounds given in (11) and (12) is indicative of a time variability error in the forecasting and analysis system. Klocke and Rodwell (2014) have discussed error diagnostics based on the mean of short-term forecast change that reveal mean model error associated with fast processes. Equations (11) and (12) complement the Klocke and Rodwell diagnostic by providing bounds on the degree of variability that the model should support.

While these assumptions are not met by real systems, this derivation does point out that \( f_i < a_i \) or \( d_i < a_i \) are not necessarily indicative of model error. On the other hand, \( f_i > a_i \) or \( d_i > a_i \) are not expected in a perfect system, and are therefore indicative of either model error or inconsistencies between the model and analysis (e.g., the forecast model “builds in” realistic variability that is filtered out of the initial states).

c. Forecast set descriptions

We considered both NCEP and CMC global ensemble forecasts taken from the THORPEX Interactive Grand Global Ensemble (TIGGE) archive (Bougeault et al. 2010), using both the control ensemble member and the first perturbed ensemble member. This provides the opportunity to look at differences in the diagnostics between different forecast centers and between control and perturbed ensemble members. The NCEP Global Ensemble Forecast System produces forecasts from the NCEP Global Forecast System (GFS; Han and Pan 2011). The ensemble transform method with rescaling is used to produce the initial perturbations (Wei et al. 2008), and stochastic total tendency perturbations (STTP; Hou et al. 2006, 2008) are used to account for model uncertainty in the perturbed members. As the NCEP control does not include STTP, the control and perturbed members differ by initial conditions and stochastic forcing.

The Canadian Global Ensemble Prediction System (Charron et al. 2010; Gagnon et al. 2013; Houtekamer et al. 2014) produces forecasts from the Canadian Global Environmental Multiscale model (GEM; Côté et al. 1998a,b; Girard et al. 2014), using an ensemble Kalman filter (EnKF) for initial perturbations. Model uncertainty is incorporated in the perturbed members through both stochastic forcing [physics tendency perturbations (PTP) and stochastic kinetic energy backscatter (SKEB)] as well as parameterization differences. Therefore, the control and perturbed member of the Canadian ensemble differ through PTP and SKEB (included in the perturbed forecast but not the control), as well as differences in parameters in the parameterization schemes relating to gravity wave drag, mixing length, and orographic blocking.

The diagnostics were first calculated for the period 1 January–31 March 2013. However, the Canadian global ensemble prediction system underwent a significant upgrade at 1200 UTC 13 February, as described in Gagnon et al. (2013). This upgrade included a new version of the GEM model (from GEM 4.2.5 to GEM
4.4.1 with improved turbulent mixing and orographic blocking schemes and improved treatment of topography). The EnKF was also upgraded, including increased resolution (from approximately 100 to 66 km), increased volume of assimilated AMSU radiance observations, and improved observation bias corrections. The incorporation of model uncertainty was also upgraded, including changes to PTP and SKEB. Of particular interest here was the change made to the PTP scheme. Gagnon et al. (2013) note that very high quantities of precipitation in a 24-h period were forecast at a few grid points in the tropics in the older implementation—a problem traced to PTP and convection. Modifications to the PTP such that convective tendencies are not perturbed when there is a positive convective available potential energy (CAPE) reduced this problem. To see how this change is reflected in the diagnostics, we calculate the diagnostics for two time periods: before and after the upgrade (1 January–13 February and 14 February–31 March, respectively). This is done for both the NCEP and CMC forecasts to see if changes are apparent in the CMC ensemble that are not apparent in the NCEP ensemble.

Diagnostics were calculated for the initial states (and forecasts started from these initial states) every 0000 and 1200 UTC. The mean square differences were only calculated at 24-h increments so as to exclude diurnal variations. While the initial states and forecasts were produced at different resolutions, all of the diagnostics were performed on $1^\circ \times 1^\circ$ gridded data. Configuration details for the different experiments considered are provided in Table 1. Diagnostics were calculated for temperature, zonal wind, meridional wind, geopotential height, and specific humidity at 850, 500, and 200 hPa. To be concise, we only present results for the 500-hPa level. Results for fields at the 850- and 200-hPa levels were qualitatively similar.

3. Results

Results are first presented for area averages in order to give a concise overview of the diagnostic results. This is followed by plan-view examinations of the diagnostics in order to provide more detailed information on spatial differences.

a. Area-averaged diagnostics

In Fig. 1, we show the four diagnostic quantities defined in (1)–(4): $a_i^2$ (black) representing variability in the analyses, $f_i^2$ (red) representing variability in the forecasts, $d_i^2$ (blue) representing the day-to-day forecast variability, and mean square error (green) for $i = 1, 10$, for the control member forecasts from NCEP (top row) and CMC (bottom row) calculated for 500-hPa geopotential height. These quantities are averaged for three regions: the southern midlatitudes (70$^\circ$–30$^\circ$S, left panels), the tropics (20$^\circ$S–20$^\circ$N, middle panels), and northern midlatitudes (30$^\circ$–70$^\circ$N, right panels). Note that $f_i^2$ and $d_i^2$ are defined to be identical for $i = 1$. The differences between $a_i^2$ and $f_i^2$ are indicated by the magenta curves with 95% confidence intervals, and the differences between $a_i^2$ and $d_i^2$ are indicated by the cyan curves with 95% confidence intervals. The 95% confidence intervals were calculated using a z score on a time series of paired differences of area average values following $\mu_d = \bar{d} \pm 1.96(\sigma_d/\sqrt{n'})$, where $\bar{d}$ and $\sigma_d$ represent the mean and standard deviation of the distribution of differences. We accounted for temporal correlation by reducing the number of independent samples following $n' \cong n(1 - \rho_1)/(1 + \rho_1)$, where $\rho_1$ is the lag-1 autocorrelation coefficient. If the confidence intervals do not cross the zero axis, we can conclude that the quantities are significantly different at 95% confidence.

For the NCEP control forecasts (top row), $a_i^2$ is larger than $f_i^2$ at longer lead times for all three regions, although the differences are not statistically significant. As noted in section 2b, this is not necessarily an indication of model error. The 1-day variability (blue curves) decreases steadily with increasing forecast time in all three regions, although $d_i^2$ is not statistically significantly different from $a_i^2$ for any $i$. In contrast, for CMC forecasts, $a_i^2$ is smaller than $f_i^2$ in the tropics, indicating that the CMC control tropical forecasts actually have more temporal variability in the height field than the CMC control initial states, although these differences are relatively small and not statistically significant. As this is not expected in the perfect model scenario, this finding raises the possibility that the temporal variability of the CMC model may be unrealistically large in the tropics. Another possibility is that the CMC analyses are unrealistically smooth, and the forecast model ‘builds in’ realistic variability. This may be due to the impact of the digital filter, recently replaced by the four-dimensional incremental analysis update (4DIAU) technique for the global deterministic system and under development for the global ensemble system. As described in Buehner et al. (2015), replacement of the digital filter with 4DIAU resulted in increased skill in short-term forecasts of tropical heights, specifically in the representation of semidiurnal tides. The digital filtering of the full fields had resulted in a degradation of the tidal signal. In addition, Buehner et al. show that the digital filter depletes the mesoscale portion of the kinetic energy spectrum, resulting in a substantial spinup of KE in these scales over the first few days of integration. Replacement of the digital filter with 4DIAU retains these
scales of motion in the analyses. Diagnostics for the 500-hPa temperature and zonal wind (not shown) indicate similar relationships to those shown for the geopotential height.

Figure 2 is the same as Fig. 1 except the results are shown for 500-hPa specific humidity instead of 500-hPa geopotential height. While the analysis and forecast temporal variability in geopotential height (Fig. 1) were almost identical in the NCEP and CMC systems, larger differences between the two systems are apparent for specific humidity, especially in the tropics. In general, higher spatiotemporal resolution will yield higher temporal variability. Hence, the difference in resolution between NCEP and CMC may account for part of the higher variability of the NCEP system.

For the NCEP system, the analysis variability $a_i^2$ is greater than the forecast variability $f_i^2$, and these differences are statistically significant in the tropics for all $i$. The forecast 1-day temporal variability $d_i^2$ decreases with forecast time, and $d_i^2$ is statistically significantly smaller than $a_i^2$ for all $i$ in the tropics and for several values of $i$ in the midlatitudes. The behavior in the CMC system is quite different. With the CMC system, $a_i^2$ is smaller than $f_i^2$ for all three regions, and this difference is statistically significant for most time intervals for the southern midlatitudes and tropics. In addition, the 1-day variability in the forecast increases with forecast time ($d_i^2 > a_i^2$), and this difference is also statistically significant for all $i$ in the southern midlatitudes and tropics. As noted above in regards to the geopotential height field, this is not expected in a perfect system and is consistent with the observations.
with the excessive smoothing of the control forecast initial state due to the digital filter in the CMC system and subsequent spinup over the first few days of forecast integration.

The green mean square forecast error curves for both geopotential height and specific humidity show that, as expected, forecast error is substantially smaller than what would be obtained from a persistence forecast (black curves). The forecast errors from the two systems are quite similar, with smaller forecast errors for NCEP in the extratropical height fields and smaller forecast errors for the CMC system in the tropical specific humidity fields.

For both NCEP and CMC systems, both $a_i$ and $f_i$ appear to saturate between days 5 and 8 for midlatitude height and a bit earlier for midlatitude specific humidity. In contrast, in the tropics, $a_i$ and $f_i$ are still increasing even at 10 days. The relative increase of $a_{10}$ over $a_1$ varies by region and variable. For geopotential height, the increase is about 70% in the southern and northern midlatitudes and 85% in the tropics. The differences between the midlatitudes and tropics are far more pronounced for specific humidity than for geopotential height. This means that, on average, changes to the tropical specific humidity field and the tropical and midlatitude height field over a 10-day period are considerably larger than changes over a 24-h period, indicative of substantial variability on time scales longer than a day. In contrast, day-to-day differences in midlatitude specific humidity are almost as large as differences for fields that are 10 days apart, indicative of higher-frequency events dominating temporal variability. We explore these regional differences in more detail in section 3b.

To quantify the trends in the 1-day variability in the forecasts with increasing integration time, Fig. 3 shows...

---

**Fig. 2.** As in Fig. 1, but for specific humidity (g $^2$ kg $^{-2}$).
For $i = 1, 10$ for 500-hPa geopotential height, specific humidity, temperature, and zonal wind (black, blue, red, and green curves, respectively) for the NCEP control forecasts (thick curves) and CMC control forecasts (thin curves with marks). In a perfect system, from (12) we expect $d_i \leq a_i$. This is true for most of the NCEP fields (with the exception of geopotential height in the tropics for $i = 1, 2$); $d_i$ decreases with increasing $i$, indicating that temporal variability in the NCEP forecasts decrease as integration times increase. The uniform rate of change in the midlatitudes indicates that this decrease is not clearly tied to either an initial adjustment process or the decrease in resolution after day 8. In the tropics, the decrease in $d_i$ is largest during the first few days, which may be indicative of some adjustment process occurring early in the integration time. For some CMC fields, most notably specific humidity, $d_i < a_i$, indicating that the forecast model “builds in” temporal variability beyond what is found in the sequence of initial states. In the tropics, particularly for specific humidity and temperature, $d_i$ increases between $i = 1$ and $i = 2$, leveling off after that. This behavior is indicative of some initial adjustment or inconsistency between the analyses and the forecast model, with fairly constant forecast temporal variability after that. It is consistent with the spinup attributed to the digital filtering of the analysis fields [since replaced in the deterministic system by 4DIAU as discussed in Buehner et al. (2015), and under development in the ensemble system].

Diagnostics based on the control and perturbed ensemble member initial states and forecasts may differ owing to the way in which the ensembles are constructed, where $a_i$ may be larger for the perturbed member than the control member, depending on the characteristics of the initial perturbations and $f_i$ may differ owing to differences in the forecast model for the control and perturbed members. In addition, the behavior of the CMC ensemble may change after the upgrade to the system on 13 February 2013. To explore these possibilities, we examine and compare the diagnostics for the control and perturbed members for NCEP and CMC for period 1 (1 January–13 February 2013) and for period 2 (14 February–31 March 2013). Examination of the midlatitude and tropical regions indicates that the largest differences are observed in the tropics, so we focus on that area here. We also focus on the first 5 days of integration, as we expect differences due to ensemble design to be most apparent during the first part of the forecast.

Diagnostics calculated for the tropics for 500-hPa geopotential height, specific humidity, and temperature are shown in Figs. 4–6, respectively. As in Figs. 1 and 2, the analysis variability $a_i^2$, the forecast variability $f_i^2$, 1-day forecast variability $d_i^2$, and the mean square error are indicated by the black, red, blue, and green curves, respectively. Thick curves are for the perturbed members and thin curves are for the control members with $a_i^2 - f_i^2$ for the perturbed members shown in magenta with 95% confidence intervals and $a_i^2 - d_i^2$ for the perturbed members shown in cyan with 95% confidence intervals. For the NCEP forecasts (Figs. 4–6, top panels), comparison of the curves in the left and right panels indicates that while temporal variability is higher during period 1 than period 2, the relationships of the curves to each other, and general trends, remain the same. During both time periods, for all three variables, $a_i^2$ is
comparable to or greater than $f_{2i}$ for both the control and perturbed forecasts, where $a_{2i}$ and $f_{2i}$ for the perturbed forecasts are higher than $a_{2i}$ and $f_{2i}$ for the control forecasts. These differences are consistent with the facts that the ensemble initial states have perturbation components that are uncorrelated in time and that the perturbed forecasts include stochastic forcing, while the control forecasts do not. For height, specific humidity, and temperature, $d_{2i}$ decreases with increasing forecast time, although the differences between $a_{2i}$ and $d_{2i}$ are only statistically significant for specific humidity and temperature.
The CMC diagnostics (bottom panels of Figs. 4–6) share some characteristics with NCEP but also exhibit differences. As with the NCEP system, $a_i^2$ and $f_i^2$ are larger during the first period than during the second period. The perturbed initial states and forecasts have higher values of both $a_i^2$ and $f_i^2$ than the control initial states and forecasts, although the differences between the perturbed and control forecasts are larger in the CMC case than in the NCEP case. These results indicate that the CMC system, with perturbed model formulation, and two types of stochastic forcing, has larger differences between the control and perturbed forecasts than are found in the NCEP system, with one type of stochastic forcing. For the CMC perturbed forecasts, for height and temperature, $f_i^2$ is statistically significantly larger than $a_i^2$ during the first period (for $i > 1$), and $a_i^2$ is

![Fig. 5. As in Fig. 4, but for 500-hPa specific humidity (g$^2$kg$^{-2}$) instead of 500-hPa geopotential height.](image-url)
statistically significantly larger than \( f_i^2 \) during the second period. For height and specific humidity, \( d_i^2 \) remains approximately constant with increasing forecast time during the first period (and increases for temperature), while during the second period \( d_i^2 \) decreases substantially between \( i = 1 \) and \( i = 2 \) for all three fields and is statistically significantly smaller than \( a_i^2 \) for all \( i \). In contrast to the perturbed member, the CMC diagnostics for the control member are quite similar for the first and second periods. These diagnostics clearly show the impact of the upgrade in the CMC system on the temporal variability of the perturbed forecasts.

Figure 7 shows \((d_i - a_i)/a_1\) for \( i = 1, 5 \) for the 500-hPa geopotential height, specific humidity, temperature, and zonal wind (black, blue, red, and green curves, respectively) for the NCEP perturbed member (thick curves) and the CMC perturbed member (thin curves with marks) for the period before (left panel) and after
The results for the NCEP perturbed member are similar to those shown for the NCEP control member (Fig. 3), both before and after the upgrade, with $d_i < a_i$ (except for 500-hPa height for $i = 1, 2$ during period 1) and a decreasing trend in temporal variability for temperature, winds, and specific humidity. The behavior of the CMC perturbed member changes markedly, with $d_i$ either increasing or changing little with time before the upgrade, and $d_i$ decreasing with time after the upgrade.

The fact that the qualitative characteristics of the diagnostics do not change between period 1 and period 2 for the NCEP members, while they do change for the CMC members, is consistent with the fact that the NCEP system configuration remained constant during the full time period considered, while the CMC system underwent a major upgrade. The fact that these changes in the diagnostics are clear for the perturbed member but not the control member indicate that the system upgrades affecting the perturbed member only, such as the changes to PTP and SKEB, had a larger impact on these particular diagnostics than changes affecting both perturbed and control members, such as the deterministic forecast model upgrades. These changes (the smaller values of $f_j$ and decreasing trend in $d_i$ during the second period) are consistent with the upgrades made to the PTP scheme in which tendency perturbations are not made when there is positive CAPE as computed by the Kain–Fritsch convection scheme. As Gagnon et al. (2013) note, this resulted in a significant decrease in spurious amounts of precipitation, particularly in the tropics, and is consistent with the decrease in tropical temporal variability as illustrated by these diagnostics. The situation in which $f_j > a_i$, as is seen in the first period for the geopotential height, temperature, and wind for the perturbed forecasts, would not be expected in a perfect model scenario, and this condition is no longer observed in the perturbed forecasts after the upgrade. There is, however, a trend of decreasing temporal variability during the second period. Examination of the mean-square error (green) curves in Figs. 4–6 clearly illustrates the improvements obtained in the perturbed forecasts after the upgrade. Relative reductions in the CMC perturbed member forecast error after the upgrade are 2–3 times larger than those seen for the CMC control forecasts or for the NCEP control or perturbed forecasts.

b. Spatial characteristics

While area-averaged diagnostics are useful for gaining a broad understanding of the general behavior of the systems, important spatial variability can be masked in the averaging. The plan-view diagnostics shown in this section will allow for a more detailed understanding of the spatial differences in these diagnostics. Figure 8 shows $a_{i \text{nceptl}}$ and $f_{i \text{nceptl}}$ (top row), $a_{10 \text{nceptl}}$ and $f_{10 \text{nceptl}}$ (middle row), and the percent differences between the two (bottom row) for the NCEP control
Fig. 8. Diagnostics calculated for 500-hPa geopotential height (m): (top left) $a_{1}^{\text{spec}}$, (top right) $f_{1}^{\text{spec}}$, (middle left) $a_{10}^{\text{spec}}$, (middle right) $f_{10}^{\text{spec}}$, (bottom left) percent differences between $a_{1}^{\text{spec}}$ and $a_{1}^{\text{spec}}$, and (bottom right) percent difference between $f_{10}^{\text{spec}}$ and $f_{1}^{\text{spec}}$. 
member forecasts for 500-hPa geopotential height. Comparison of the left and right top figures indicates that for 500-hPa height, the NCEP 1-day variability in the forecast is very similar to that seen in the initial state sequence. The 1-day variability has maxima in the midlatitude jet regions in the North Pacific and North Atlantic, as well as over the Southern Ocean. The 10-day variability (middle row) also has relatively large values in these regions; however, the maxima are shifted and/or extended downstream. Specifically, in the northern Pacific, the maximum shifts from the central part to the eastern part of the basin. In the northern Atlantic, the region of high values extends farther east to include northern Europe. These results are consistent with studies examining temporally filtered variances of analyses, such as Blackmon et al. (1977) and Lau and Nath (1987), who find the bandpass (2.5–6 days) filtered variance maxima in the storm tracks, and the low-pass (>10 days) filtered variance maxima downstream of the storm tracks, in preferred blocking regions. Specifically, the 1-day variability maxima shown in the top panels of Fig. 8 are quite similar to bandpass 500-hPa variance maxima shown in Fig. 1 of Lau and Nath’s study, with a maximum stretching across the North Pacific (from 140°W to 140°E) and a maximum in the North Atlantic centered at 60°W. Likewise, the 10-day variability maxima (middle panels of Fig. 8) share characteristics with low-pass variance results, with a North Pacific maximum centered at 150°W and maxima over the North Atlantic and northern Europe (at 60°E). The spatial patterns in Fig. 8 are also quite similar to the high-pass (less than 7 days) and bandpass (7–90 day) filtered variances shown in Fig. 1 of Cai and Van Den Dool (1991).

In the Southern Hemisphere, the maximum shifts from the southern Indian Ocean at $i = 1$ to the southern Pacific Ocean at $i = 10$. As in the Northern Hemisphere, these patterns are consistent with previous studies of temporally filtered variances, such as Fig. 17 in Trenberth (1981), who likewise find a shift in the maximum from the southern Indian Ocean in 2–8-day bandpass filtered variance, to the southern Pacific Ocean for the 8–64-day bandpass filtered variance. While $a_{10}$ and $f_{10}$ (left and right middle panels) are similar, close examination reveals some differences, with the maximum over the Gulf of Alaska (Atlantic) being stronger (weaker) in $f_{10}$ than in $a_{10}$.

The similarities between the spatial patterns produced using the current diagnostics and the spatial patterns available from previous studies using low-pass and bandpass filters is highlighted to stress the potential of this technique for diagnosing the temporal forecast model characteristics on different time scales without the need for multimonth forecast integrations. Blackmon et al. (1977), Lau and Nath (1987), and Cai and Van Den Dool (1991) examined analysis time series of 3–4 months for 9–18 winter seasons [Trenberth (1981) used 8 yr of continuous data]. For their comparison of analysis and forecast temporal variability, Lau and Nath (1987) used 12 winter periods from a 15-yr-long model integration. The 10-day forecasts considered here are not sufficiently long to apply the low-pass (>10 day) or bandpass (7–90, 8–64 day) filters used in the previous studies.

The percent differences of 10-day variability versus 1-day variability (Fig. 8, bottom panels) show considerable spatial inhomogeneity, with large values (over 150%) in some areas of the tropics/subtropics and polar regions and relatively small values in the storm-track regions. In fact, the values in the central North Pacific and southern Indian Ocean are close to zero in the forecasts, indicating that high-frequency (1 day) variability accounts for most of the changes seen over a 10-day time period. The results for the CMC system (not shown) are similar. The large values in the subtropics and polar regions may indicate seasonal trends in temperature and height that will be more prominent in 10-day differences than in 1-day differences. Similar figures for specific humidity (not shown) indicate that throughout much of the midlatitudes, 10-day variability is no larger than 1-day variability. In contrast, there are subtropical–tropical regions where the 10-day variability is over 60% larger than the 1-day variability, perhaps reflecting convectively coupled equatorial waves or seasonal trends.

To highlight differences between analysis and forecast temporal variability, Fig. 9 shows the percent difference between $f_i$ and $a_i$ [i.e., $100 \times (f_i - a_i)/a_i$] for 500-hPa height for the NCEP control member (top left) and CMC control member (top right). The percent differences are fairly small for both NCEP and CMC, with differences greater than 10% confined to the tropics and subtropics. These positive values, indicating tropical regions where differences between the analysis and 1-day forecast are larger than differences between subsequent analyses, are consistent with adjustment processes that occur during the first forecast day. The bottom panels of Fig. 9 show the percent difference between $d_{10}$ and $a_1$ for the NCEP control member (bottom left) and CMC control member (bottom right). For NCEP forecasts, the negative values over most areas (with regional exceptions in the Arctic, southern Indian Ocean, and Western Australia) indicate a fairly pervasive decrease in temporal variability in the forecast model with increasing forecast time. The mixed signal in CMC indicates a less spatially systematic change in temporal forecast behavior than is seen in the NCEP system and may also reflect issues associated with the suppression of variability in the analyses due to the digital filter.
The same plots shown for 500-hPa geopotential height in Fig. 9 are shown for 500-hPa specific humidity in Fig. 10. These figures indicate that the area-averaged increasing trend for CMC and decreasing trend for NCEP are not spatially uniform. In particular, while the temporal variability decreases over most of the deep tropics in the NCEP forecasts, there are areas in the northern subtropics, notably the western Atlantic, where values increase. The CMC system shows substantial increases in temporal variability over much of the globe, and as noted before, may be related to the smoothing effect of the digital filter on the analysis and subsequent spinup (Buehner et al. 2015). Examination of the moisture biases for these forecasts (not shown) does not reveal a straightforward correspondence between higher temporal variability and a high specific humidity bias. It is interesting to note that some of the areas of enhanced variability, such as the Caribbean Sea and the western and central North Pacific subtropics, are common to both the NCEP and CMC systems, indicating that both forecast models have a tendency to overestimate temporal variability in these regions when compared to the initial states.

As noted previously, the CMC ensemble system underwent a significant upgrade in February 2013, after which the area-averaged diagnostic characteristics for the perturbed ensemble member changed considerably (Figs. 4–7). Figure 11 shows the percent difference between $d_5$ and $a_1$ for the 500-hPa temperature for the NCEP perturbed member (top row) and CMC perturbed member (bottom row) for the periods prior to (left panels) and after (right panels) the CMC system upgrade. As expected, the results for the NCEP forecasts are similar for the two periods, while the CMC relative difference is larger than 50% over much of the tropical Pacific before the upgrade but is actually negative over most of the tropics after the upgrade. Similar figures for the 500-hPa specific humidity (not shown)
indicate similar behavior, though the patterns are much noisier and of finer spatial scale. These results show that the diagnostics are successful in illustrating the impact of the CMC upgrade in reducing spuriously large tropical precipitation events (Gagnon et al. 2013). The diagnostics quantify the impact of these changes on temporal variability, complementing other assessment diagnostics such as those focused on spatial variability.

4. Summary and conclusions

We use simple diagnostics to quantify the temporal variability in analyses, $a_i$, and forecasts, $f_i$, for increasing time differences from $i = 1, 10$ days. In addition, we introduce a diagnostic that reflects the day-to-day variability of the forecasts, $d_i$, for increasing forecast lead time $i$. We show that, in a perfect system, we expect $a_i < f_i$ and $a_i < d_i$ owing to the presence of uncorrelated analysis errors. We formalize the expected bounds on $f_i$ in (11) and (12), showing that forecast temporal variability should lie between corresponding analysis variability and analysis variability minus 2 times the analysis error variance. We apply these diagnostics to control and perturbed ensemble initial states and forecasts from the NCEP and CMC global ensemble forecasting systems to demonstrate the utility of the diagnostics in quantifying aspects of forecast performance related to temporal variability. We relate the results to ensemble design and, in the case of CMC, a system upgrade.

While $a_i > f_i$ and $a_i > d_i$ for most NCEP fields, which is expected in a perfect system, $a_i < f_i$ and $a_i < d_i$ for several CMC fields, indicating that the CMC system may have excessive temporal variability as compared to the analyses. This is probably due to the excessive smoothing of the CMC analyses through the application of a digital filter [since replaced by 4DIAU in the deterministic global system as discussed in Buehner et al. (2015), and under development in the ensemble system].

![Figure 10](image-url) As in Fig. 9, but for 500-hPa specific humidity instead of 500-hPa height.
Comparison of the diagnostics for the control and perturbed ensemble members (Figs. 4–6) highlights larger differences for the CMC system than for the NCEP system, consistent with the fact that the CMC perturbed and control members differ in both model formulation and two types of stochastic forcing, while the NCEP perturbed and control members differ by one type of stochastic forcing. Trends in $d_i$ illustrate how both the control and perturbed NCEP forecasts show a small but steady decrease in day-to-day temporal variability with increasing forecast time. In contrast, the CMC control forecasts show increasing temporal variability for temperature and humidity during the first few days, illustrating the spinup of the system after the initial excessive digital filter smoothing.

The diagnostics also clearly reflect the upgrade in the CMC system on 13 February 2013. Before the upgrade, $f_i$ was greater than $a_i$ in the tropics for the CMC perturbed ensemble member for height, winds and temperature, which, as shown in section 2b, is not expected in a perfect system. After the upgrade, $f_i$ was less than $a_i$ in the tropics for the perturbed member. The trends in $d_i$ for the perturbed member also change, remaining fairly constant or increasing before the upgrade and decreasing after the upgrade. These differences are consistent with changes made to the stochastic physics perturbations in order to reduce excessive precipitation (Gagnon et al. 2013).

An advantage of these diagnostics is the ability to assess forecast temporal variability on different time scales without the need for very long forecast integrations. For example, the locations of the maxima in height field variability (Fig. 8) shift or extend from the North Atlantic and North Pacific jet regions for $i = 1$ downstream to northern Europe and the eastern North Pacific for $i = 10$. These shifts are consistent with patterns found in temporal filtering diagnostics of analyses time series that differentiate between regions of synoptic variability and blocking (e.g., Blackmon et al. 1977; Lau and Nath 1987).
Cai and Van Den Dool 1991) using low-pass (>10 day) and bandpass (7–90 and 8–64 day) filters that could not be applied to the 10-day forecast integrations considered here.

Diagnostics measuring temporal variability are complementary to other diagnostics, such as those that focus on time-mean quantities or model bias (e.g., Klocke and Rodwell 2014), spatial scale-separation techniques (e.g., Harris et al. 2001), and techniques to quantify differences between forecast fields and reality as represented on the scales resolved by the data-assimilation and forecast systems (e.g., Peña and Toth 2014). Using diagnostics to assess the accuracy of both temporal and spatial variability will become increasingly important as stochastic techniques to account for model uncertainty proliferate in ensemble forecasting systems, as both spatial and temporal correlations are often parameters in these schemes that need to be tuned. Potential future work includes consideration of other forecast systems, as well as an extension to a comparison with observations.

Acknowledgments. This research is supported by the Chief of Naval Research through the NRL Base Program, PE 0601153N. The forecasts were obtained from the THORPEX Interactive Grand Global Ensemble (TIGGE) data portal at ECMWF. We thank three anonymous reviewers for very conscientious critiques of this paper. We thank the ECMWF for permission to use its forecast data.

REFERENCES


