A New Generic Method for Quantifying the Scale Predictability of the Fractal Atmosphere: Applications to Model Verification

XINGQIN FANG AND YING-HWA KUO
University Corporation for Atmospheric Research, and National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 April 2014, in final form 16 December 2014)

ABSTRACT

The authors revisit the issue regarding the predictability of a flow that possesses many scales of motion raised by Lorenz in 1969 and apply the general systems theory developed by Selvam in 1990 to error diagnostics and the predictability of the fractal atmosphere. They then introduce a new generic method to quantify the scale predictability of the fractal atmosphere following the assumptions of the intrinsic inverse power law and the upscale cascade of error. The eddies (of all scales) are extracted against the instant zonal mean, and the ratio of noise (i.e., the domain-averaged square of error amplitudes) to signal (i.e., the domain-averaged square of total eddy amplitudes), referred to as noise-to-signal ratio (NSR), is defined as a measure of forecast skill. The time limit of predictability \( t_{km} \) for any wavenumber \( k_m \) is determined by the criterion \( \text{NSR}(t_{km}) = \phi^{-2m+2} = k_m^{-2} \) or by the criterion \( \log_{\phi}[\text{NSR}(t_{km})] = -2m + 2 = \log_{\phi}(k_m^{-2}) \), where \( \phi = (1 + \sqrt{5})/2 = 1.6180339887... \) is the golden ratio and \( m \) is a scale index. The NSR is flow adaptive, bias aware, and stable in variation (in a logarithm transformation), and it offers unique advantages for model verification, allowing evaluation of different model variables, regimes, and scales in a consistent manner. In particular, an important advantage of this NSR method over the widely used anomaly correlation coefficient (ACC) method is that it could detect the successive scale predictability of different wavenumbers without the need to explicitly perform scale decomposition. As a demonstration, this new NSR method is used to examine the scale predictability of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) 500-hPa geopotential height.

1. Introduction

A fundamental problem in atmospheric sciences is the predictability of atmospheric circulations of specific scales. As reviewed by Anthes (1986), the nonlinear nature of atmospheric processes allows cross-scale energy transfer so that the uncertainty or error in any one scale will eventually contaminate all scales; and the existence of atmospheric instabilities will make any error, no matter how small initially, eventually grow and contaminate even a perfect model’s forecast. Understanding the predictability of atmospheric circulations of different scales and identifying the source and nature of errors are both critical steps toward improving weather and climate models.

An important question for numerical weather prediction (NWP) is how much time will pass before a model forecast becomes useless. In reality, there is no unique answer to this general question, since the level of accuracy required for a forecast to be useful depends very much on the purpose for which the forecast is used (Bengtsson and Simmons 1983). Thus, a more specific question should be asked regarding the time limit of the predictability of a specific scale in the multiscale atmosphere.

On one hand, the error growth in a global model for medium-range forecasts on the synoptic scale is assumed to be associated with baroclinic activities. On the other hand, the moist convective instability may limit the forecast skill on the shorter forecast lead time with an increase of the horizontal resolution and the ability to resolve convective systems. In reality, the development of the high-impact weather systems actually results in multiscale interactions, and it is difficult to separate or isolate the effects from different scales. These facts suggest a careful investigation of the scale predictability with a coherent multiscale continuum of atmosphere in mind.

To quantify the practical predictability of a forecast model in terms of a time limit, two aspects should be considered: one is the proper measure used to assess the...
forecast skill; the other is the criterion used to determine the predictability time limit beyond which a forecast is no longer useful.

In most of the early classical predictability studies, which used either the simplified or ideal models (e.g., Thompson 1957; Lorenz 1963; Lilly 1969; Leith 1971) or the general circulation models (GCMs) (e.g., Charney et al. 1966; Smagorinsky 1969; Jastrow and Halem 1970; Shukla 1981, 1984a), an attempt was made to arrive at a quantitative estimate of the growth rate of an initial error (mostly in terms of the root-mean-square) and to determine the time limit of predictability by means of the error doubling time. This error doubling method was used in a recent study of Hohenegger and Schär (2007) to compare the predictability at synoptic scales versus cloud-resolving scales. However, as shown in Shukla (1985), the error growth rate by itself is not a useful parameter for predictability, partly because it varies significantly for different variables and partly because the ultimate limit of predictability is not only determined by the error growth rate, but also by the saturation value of the error. Neither a larger error, nor a larger error growth rate necessarily means less predictability.

Using a new concept of the ratio of the root-mean-square error to the standard deviation of daily fluctuations, Shukla (1984b, 1985) demonstrated the dependence of predictability on latitude, season, and the variable in question. However, he did not provide an estimate of the predictability in terms of the time limit.

A popular measure of forecast skill for global model evaluations is the anomaly correlation coefficient (ACC; Miyakoda et al. 1972; Hollingsworth et al. 1980; Bengtsson and Simmons 1983; Jolliffe and Stephenson 2012). It calculates the spatial correlation between the forecast and the observed (or analyzed) deviations (i.e., anomalies) from a predefined mean state. The centered version of the ACC also subtracts the spatial mean error. By definition, the ACC is very sensitive to the selected mean state, which can be the 6-h climatology, daily climatology, monthly climatology, seasonal mean, or some running averages. For example, Langland and Maue (2012) calculated the climatology by moving a weighted 21-day-centered mean window at each grid point and for different times. Some empirical thresholds (i.e., 50%, 60%, and 80%) of the ACC are used to quantify the model’s practical predictability time limit (e.g., Hollingsworth et al. 1980). As documented on the ECMWF website, it has been found that the 60% threshold corresponds to the range up to which there is forecast skill for the largest-scale weather patterns; 50% corresponds to forecasts for which the error is the same as that of a forecast based on a climatological average; and 80% would correspond to forecasts for which there is still some skill in large-scale synoptic patterns. Actually, these criteria of the ACC for practical predictability purposes are empirical and can only be applied to larger-scale circulations. As will be shown later, the ACC by itself is not a useful parameter to assess the predictability of smaller scales. If the scale decomposition is applied first, different empirical criteria might be needed to assess the practical predictability of different scales. Furthermore, as argued in Langland and Maue (2012), the ACC is not an optimal metric with which to quantify model forecast skills in all situations, since the ACC can be higher when the large-scale atmospheric flow contains strong anomalies, regardless of whether there is any actual improvement in the model forecast skill. It should be noted that the ACC is strongly modulated by the strength of the anomalies in the forecast and analysis, which may not be related to the model performance.

Another approach to quantify the practical predictability time limit is to compare the forecast error variance with an error criterion derived from the forecasts based on climatology (e.g., Baumhefner 1984; Anthes and Baumhefner 1984; Anthes et al. 1985). As reviewed by Anthes (1986), the limit of atmospheric predictability in global models is considered to be the time required for the difference variance (i.e., the mean square of the difference of some variables) of a pair of solutions that begin with small differences at the initial time to reach the difference variance associated with two randomly chosen atmospheric states. Baumhefner (1984) suggested verifying the forecast by normalizing a conventional measure of skill (such as the root-mean-square error or error variance), as compared with the best case (no error) and/or worst possible case (e.g., the persistence forecast); the limits of predictability for the various scales were estimated by comparing an estimate of the so-called predictability error (i.e., the ensemble average of the root-mean-square difference of the perturbed forecasts) with the root-mean-square climatological forecast error (error from the forecast with climatology) as the error bar. Furthermore, Anthes (1986) suggested estimating the theoretical predictability using twice as much of the climatological forecast error variance as the error variance bar. Obviously, to use this method, scale decomposition is needed to investigate the predictability of different scales.

A fundamental characteristic of the error growth in climate and weather predictions is the upscale cascade of error. Lorenz (1969) first demonstrated the slow inverse cascade of errors from small to large scales using a closure model of two-dimensional flow. It was shown that the time taken for the complete uncertainty at wavenumber $2k$ to strongly infect wavenumber $k$ is proportional to the eddy turnover time $\tau(k) = k^{-3/2}[E(k)]^{-1/2}$.
where $E(k)$ denotes the kinetic energy per wavenumber unit at wavenumber $k$. Following this assumption, Palmer (1999) briefly described estimates of predictability arising from the cascade processes and presented a heuristic argument for determining the time it takes for the initial error associated with unresolved wavenumbers higher than $k_e$ to infect the meteorological scales $k_1 \ll k_e$: the time $\Omega(P)$ taken for the uncertainty to propagate from wavenumber $2^p k_1$ to wavenumber $k_1$ is given by $\Omega(P) = \sum_{p=0}^{P-1} \tau(2^p k_1)$, where $P$ is a nonnegative integer. He argued that, in the case of a two-dimensional isotropic homogeneous turbulence in the inertial sub-range between some large (e.g., baroclinic) forcing scale and dissipation scale, $E(k) \sim k^{-5/3}$ and $\tau(k) \sim k^{-2/3}$. In this case, $\Omega(P)$ asymptotes to a finite limit as $P \to \infty$, that is, $\omega(\infty) = 2.77 \tau(k_1)$. This result indicates that the uncertainty in small scales (within the inertial range) can infect the larger meteorological scale and that the predictability time limit for a large-scale system is on the order of an eddy turnover time: a few days. As such, uncertainties on the mesoscale may influence the predictability of large-scale weather systems within a few days.

These theoretical estimates of the scale predictability based on the eddy turnover time are not useful for practical application, in part because the error growth rate in a NWP model is also related to the model physics. Actually, the inherent predictability of the atmosphere and the practical predictability of the atmosphere associated with an NWP model are not the same thing. The latter is both state dependent and model dependent and is thus a subject of interest to the model verification community. Hereafter in this paper, the predictability refers to the practical predictability unless otherwise specified.

Dynamical systems in nature, such as atmospheric flows, stock market indices, and heartbeat patterns, etc., exhibit irregular space–time fluctuations on all scales. The fractal or self-similar nature of space–time fluctuations has been identified since the 1970s (Mandelbrot 1975, 1982). Based on a concept in Townsend (1980) that large eddies are the envelopes enclosing small-scale (turbulent) eddy fluctuations, a general systems theory was developed for fractals (Selvam 1990, 1993, 2005, 2007, 2009, 2011). In this theory, the basic concept involves “visualizing the emergence of successively larger scale fluctuations to result from the space–time integration of enclosed smaller scale fluctuations” (Selvam 2009, p. 333); and the hierarchy of self-similar fluctuations generated is manifested as the observed eddy continuum in power spectral analyses of the fractal fluctuations. As shown in Selvam (2011), one primary assumption deduced from this basic concept is that both the radius ratio and the circulation-speed ratio (or eddy-amplitude ratio) of the two successive large and small eddies are equal to the golden ratio $\varphi = (1 + \sqrt{5})/2 = 1.6180339887 \ldots$. Further derivation of this theory also shows that both the power spectrum and the probability distribution function of the fractal fluctuations follow the same inverse power law, which is close to that of the Gaussian distribution for small-scale fluctuations but gives the observed fat, long tail for large-scale fluctuations (Selvam 2009). Since the power (or variance: i.e., the square of the eddy amplitude) spectrum also represents the probability densities, as in the case of quantum systems such as the electron or photon, fractal atmospheric fluctuations exhibit quantum-like chaos.

In this paper, we apply Selvam’s general systems theory to forecast error diagnostics and the predictability; we then introduce a new generic method to quantify the scale predictability of the fractal atmosphere following the assumption of a gradual upscale cascade of error. The paper is organized as follows. Section 2 introduces a new generic method, designated as the noise-to-signal ratio (NSR) method, for quantifying the scale predictability of the fractal atmosphere. Section 3 describes the data from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) forecasts used in this study. Section 4 applies the new NSR method to investigate the scale predictability of the NCEP GFS 500-hPa geopotential height. Section 5 compares the NSR method with the ACC method. Conclusions and a discussion are presented in section 6.

2. A new generic method for quantifying the scale predictability: The NSR method

a. The one-dimensional dataset

In a numerical model, the fractal atmospheric state within a verification domain (any part of the global domain) can be transferred into a one-dimensional dataset. For example, the $t$-hour forecast, the $t$-hour forecast error, and the analysis at the valid time can be expressed by $F(i)$, $E(i)$, and $A(i)$, respectively, with the location index $i = 1, 2, \ldots, G$, where $G$ is the total number of the Gaussian reduced points within the verification domain.
The value of a variable on each point is associated with the physical location of the point. When the values within a selected domain are transferred into a one-dimensional dataset, the data can be sampled in many different sequential orders. However, no matter which order is selected, the statistics, such as the whole frequency distribution, the mean, and the variance of the dataset, are the same. In other words, when the verification domain in question is verified as a whole (rather than as stratified bins), the actual physical location of each point within this one-dimensional dataset is ignored; from a statistical perspective, each value of the one-dimensional dataset is just one realization of the fractal atmospheric state. The reason to use the Gaussian reduced points is to ensure even sampling over the earth’s surface within the verification domain.

b. The whole-scale eddies: Perturbations against the instant zonal mean

Although the verification is done in limited domains, each point in the selected domain should be taken as a point resident in the entire globe (actually a layer of the Earth atmosphere if taking a terrestrial perspective) so that the eddies should be taken as “whole scale” eddies. The whole-scale eddies at a given time contain all spatial scales; that is, they are composed of eddies of all spatial scales. If we take the perspective that the temporal mean of all eddies averaged over the entire time range is zero, the whole-scale eddies also contain eddies of all temporal scales. To extract the whole-scale atmospheric eddies, the perturbation in the analysis field on an arbitrary point in the verification domain is calculated by the departure from the corresponding instant zonal mean:

$$A'(i) = A(i) - \frac{1}{G_i} \sum_{J_i=1}^{G_i} \text{ALL}_i(J_i)$$  

with the location index $J_i = 1, 2, ..., G_i$, where ALL$_i$(J$_i$) is the one-dimensional dataset of the analysis on all Gaussian reduced points along a full latitude line, where point $i$ belongs and $G_i$ is the total number of those Gaussian reduced points associated with point $i$.

c. A new forecast skill score: The NSR

In the verification domain, let $N(t)$ denote the noise (i.e., the domain-averaged error variance or error energy of the $t$-hour forecast against the corresponding analysis) and $S$ denote the signal (i.e., the domain-averaged total eddy variance or total eddy energy of the analysis). They can be expressed by the following:

$$N(t) = \frac{1}{G} \sum_{i=1}^{G} [E'(i)]^2 = \frac{1}{G} \sum_{i=1}^{G} [F'(i) - A(i)]^2,$$  

and

$$S = \frac{1}{G} \sum_{i=1}^{G} [A'(i)]^2.$$  

A new measure of forecast skill (i.e., the NSR), for the entire verification domain at a specific valid time, as a function of the forecast lead time $t$, could then be defined as

$$\text{NSR}(t) = \frac{N(t)}{S}. $$  

From Eq. (4), we can see that the NSR varies from 0 to a value associated with the worst possible forecast, and the smaller the value, the better the forecast skill.

d. A new method for quantifying the scale predictability

Based on Selvam’s general systems theory, both the radius ratio and the circulation–speed ratio (or the eddy–amplitude ratio) of the two successive large and small eddies are equal to the golden ratio in the fractal atmosphere:

$$\frac{R_m}{R_{m+1}} = \varphi,$$  

and

$$\frac{S_m}{S_{m+1}} = \varphi^2,$$  

where $R_m$ and $R_{m+1}$ are the eddy radii, and $S_m$ and $S_{m+1}$ the corresponding domain-averaged eddy signals (i.e., the average square of the eddy amplitudes) of the two successive large and small eddies with the scale indices $m$ and $m + 1$, respectively. Note that $m$ could be any real number $\geq 1$, with $m = 1$ for the largest scale. However, given the limitation of the data resolution as well as the dynamic diffusion used in the model, a maximum value $M$ could be set as follows: $1 \leq m \leq M$.

Without losing the generality, the eddy radius and the eddy signal for the scale index $m$ can be denoted by

$$R_m = C_R \varphi^{-m+1}$$  

\footnote{The dataset could also be sampled on the icosahedral or any other grids with fairly uniform spacing, with eddies still defined against the instant zonal mean, which could be easily derived from the interpolated data (at a comparable resolution with the original data) along the corresponding full latitude line.}
and
\[ S_m = C_s \varphi^{-2m+2}, \tag{8} \]
where \( C_R \) and \( C_s \) are positive parameters in units of length and signal, respectively.

Let \( k_m \) represent the wavenumber and \( L_m \) represent the wavelength associated with the scale index \( m \). For the largest-scale eddy, we have \( m = 1 \), \( k_1 = 1 \), and \( L_1 = 2R_1 = 2C_R \). Since \( L_m k_m = L_{m+1} k_{m+1} \), from Eq. (7) we then derive that \( k_m \) is related to \( m \) by
\[ k_m = \varphi^{m-1}, \tag{9} \]
and \( L_m \) is related to \( m \) and \( k_m \) by
\[ L_m = 2C_R \varphi^{-m+1}. \tag{10} \]

Note that the wavelengths for the same wavenumber vary with latitude (corresponding to a different \( C_R \) value). For example, the wavelengths corresponding to wavenumbers \( k = 1 \sim 10 \) are \( L_1 = 2C_R \), \( L_2 = 2C_R \), \ldots, and \( L_{10} = C_R/5 \). At the equator, the wavelength for wavenumber 1 (5) is about 40000 km (8000 km); while at the 60° latitude, the wavelength for wavenumber 1 (5) is about 20000 km (4000 km).

Interestingly, from Eqs. (8) and (9), which are based on Selvam’s general systems theory, we further derive the following spectral energy relationship:
\[ E_{k_m} = S_{k_m} / k_m = C_s k_m^{-3}, \tag{11} \]
where \( E_{k_m} \) is the spectral energy per wavenumber unit at wavenumber \( k_m \). This relationship actually indicates the well-known −3 spectral decay law in the atmosphere.

Figure 1 shows the golden rectangles and golden spirals, illustrating the beautiful fractal with self-similarity. In Fig. 1, analog to the golden ratio, for the fractal atmosphere that with scale index 1 ≤ \( m \) ≤ \( M \), the scale radius \( R_m \) and the scale signal \( S_m \) can be represented by the side length and area, respectively, of the \( m \)th golden square. The total eddy signal \( S = \sum_{m=1}^{M} S_m \) can be represented by the largest golden rectangle; thus, we have
\[ S_m = S \varphi^{-2m+1}, \tag{12} \]
where the following property of the golden ratio, that its successive powers obey the Fibonacci recurrence, is used:

![Figure 1. The golden rectangles and golden spirals. The side lengths of each rectangle are in the golden ratio \( \varphi \), as is the side length of a larger square and the side length of the next smaller square. The red spiral is a golden spiral, a special type of logarithmic spiral for which the growth factor is the golden ratio \( \varphi \) while the green spiral made from the quarter circles tangent to the interior of each square is a good approximation of the golden spiral, with the overlapping portions appearing in yellow. For the fractal atmosphere with scale index 1 ≤ \( m \) ≤ \( M \), the scale radius \( R_m \) and the scale signal \( S_m \) can be represented by the side length and area, respectively, of the \( m \)th golden square. The value \( N_m \) is the noise component that is supposed to significantly contaminate the scale signal \( S_m \).](image-url)
\[ \varphi^n = \varphi^{n-1} + \varphi^{n-2}. \quad (13) \]

Note that \( n \) in Eq. (13) can be any real number and is not required to be an integer.

Following the error cascade assumption, the model errors grow from the errors on the smallest scale; and with the increase of forecast lead time, the ubiquitous error growth on smaller scales can be visualized as the errors on the successive larger scales, and, finally, the error growth on smaller scales can be visualized as the required to be an integer.

As illustrated in Fig. 1, it is reasonable to assume that, for \( 1 \leq m \leq M \), the scale signal \( S_m \) is significantly contaminated when the noise component that is supposed to contaminate the signal on that scale, \( N_m \), is equal to the signal itself, which is the area of the \( m \)th golden square:

\[ N_m = S_m. \quad (14) \]

Accordingly, the noise at forecast time \( t_m \) has significantly contaminated the signal \( S_m \) when

\[ \frac{N(t_m)}{S_m} = \varphi \quad (15) \]

or

\[ \frac{N(t_m)}{S} = \varphi^{-2m+2}, \quad (16) \]

where \( N(t_m), S_m, \) and \( S \) are equal to the area of the \( m \)th golden rectangle, the \( m \)th golden square, and the largest golden rectangle, respectively, in Fig. 1.

The predictability range \( t_k \) for wavenumber \( k_m \) in the whole-scale atmospheric flow can be determined by the following criterion:

\[ \text{NSR}(t_k) = \varphi^{-2m+2} = k_m^{-2}. \quad (17) \]

For example, the NSR criteria for the predictability for the scale index \( m = 1, 2, \ldots, M \) are \( \varphi^0, \varphi^{-2}, \ldots, \varphi^{-2M+2} \), with the 10 leading values rounded to 1.00000, 0.38197, 0.14590, 0.05573, 0.02129, 0.00813, 0.00311, 0.00119, 0.00045, and 0.00017. In other words, the NSR criteria for the predictability for the wavenumber \( k = 1, 2, \ldots, K \) are \( 1, 2^{-2}, \ldots, K^{-2} \) (\( K \) is the maximum wavenumber), with the 10 leading values of 1, 2, 3 \(-2, 4 \), 5, 6 \(-2, 7, 8 \), 9 \(-2, 10 \).

For practical applications, especially for introducing more linearity and stronger stationarity in the statistical analysis, we propose a logarithm function with base golden ratio \( \varphi \): \( \log_\varphi(x) \). Applying the logarithm with base \( \varphi \) to Eq. (17), we have

\[ \log_\varphi[\text{NSR}(t_k)] = -2m + 2 = \log_\varphi(k_m^{-2}), \quad (18) \]

and the \( \log_\varphi(\text{NSR}) \) criteria for the predictability of the 10 leading scale indices (corresponding to wavenumber \( k = 1, \varphi, \ldots, \varphi^9 \)), are exactly 0, –2, –4, –6, –8, –10, –12, –14, –16, and –18. In other words, the \( \log_\varphi(\text{NSR}) \) criteria for the predictability of the 10 leading wavenumbers are 0, –2, –log2.0, –log2.5, –log2.0, –log2.5, –log2.6, –log2.7, –log2.8, –log2.9, and –2 log2.10.

As will be shown later, the \( \log_\varphi(\text{NSR}) \) criteria with respect to the scale index or the wavenumber (in the logarithm form with base golden ratio) is especially convenient for revealing the predictability time limit of different scales for a forecast model.

3. Data

We now use the 12-hourly NCEP GFS 500-hPa geopotential height 0–192-h forecasts for the Northern Hemisphere initialized four times daily from 1 August to 31 October 2012 to demonstrate the practical applications of the new NSR method for quantifying the scale predictability of the fractal atmosphere in a forecast model. Note that the data in a horizontal resolution of 0.5° is transferred onto a comparable N160 Gaussian reduced grid (https://software.ecmwf.int/wiki/display/EMOS/Reduced+Gaussian+Grids).

As we know, with the continued improvements in model dynamics and physics, data assimilation methods, and observing systems, the quality of global atmospheric analysis has been significantly improved. Although not perfect by any means, the global analysis provides a good estimate of the true atmospheric state for verification purposes where the analysis error can be omitted. For the sake of simplicity, the NCEP GFS analysis is treated as the true atmospheric state for forecast error diagnostics.

4. Scale predictability of the NCEP GFS 500-hPa geopotential height derived by the new NSR method

a. Examples of the one-dimensional dataset

We perform the verification on three different domains: 1) region A [−(15°–35°N and 30°–150°W); 2) region B [−(25°–50°N and 30°–150°W)]; and 3) the Northern Hemisphere [−(NH: 0°–90°N and 0°–360°W)]. Figure 2 shows the one-dimensional datasets of the analysis eddy amplitude \( |A'(i)| \), and the forecast error amplitude \( |E_{12h}(i)|, |E_{24h}(i)|, \) and \( |E_{72h}(i)| \) of the NCEP GFS 500-hPa geopotential height in the NH valid at 0000 UTC 31 August and 0000 UTC 31 October 2012, respectively. Figure 3 shows the one-dimensional datasets of the analysis eddy amplitude \( |A'(i)| \), and the forecast error amplitude \( |E_{12h}(i)|, |E_{24h}(i)|, |E_{72h}(i)|, |E_{144h}(i)|, \) and \( |E_{168h}(i)| \) for region A valid at 0000 UTC 31 August
31 August 2012 and for region B valid at 0000 UTC 31 October 2012, respectively.

Although the physical location of each point in each dataset is ignored (as indicated in section 2) when the selected domain is verified as a whole, the indices in Figs. 2 and 3 run from south to north and from west to east starting at the date line or the west edge of each latitude. Therefore, the points in the datasets of Fig. 3a (Fig. 3b) correspond to segments of the points with indices from about 17 000 (28 000) to about 39 000 (52 000) in Fig. 2a (Fig. 2b). In addition, the one-dimensional datasets with point indices running in this order actually carry some structural information on the spatial variations.

**Fig. 2.** The one-dimensional datasets of the analysis eddy amplitude $|A'(i)|$ (blue) and the forecast error amplitude $|E_i^{12h}(i)|$ (green), $|E_i^{24h}(i)|$ (cyan), and $|E_i^{72h}(i)|$ (red) of the NCEP GFS 500-hPa geopotential height in the NH $[-(0°-90°N$ and $0°-360°)]$ valid at (a) 0000 UTC 31 Aug 2012 and (b) 0000 UTC 31 Oct 2012. Note that the forecast error amplitude datasets are shifted upward by 200, 300, and 400 m, respectively. The corresponding signal, noise, and NSR values are listed.
FIG. 3. The one-dimensional datasets of the analysis eddy amplitude $|A'(i)|$ (blue) and the forecast error amplitude $|E^{16h}(i)|$ (green), $|E^{24h}(i)|$ (cyan), $|E^{72h}(i)|$ (red), $|E^{96h}(i)|$ (black), $|E^{120h}(i)|$ (orange), $|E^{144h}(i)|$ (purple), and $|E^{168h}(i)|$ (brown) for (a) region A $[15^\circ–35^\circ N$ and $30^\circ–150^\circ W]$ valid at 0000 UTC 31 Aug 2012 and (b) region B $[25^\circ–50^\circ N$ and $30^\circ–150^\circ W]$ valid at 0000 UTC 31 Oct 2012. Note that the forecast error amplitude datasets are shifted upward by 200, 300, 400, 600, 800, 1000, and 1200 m, respectively. The corresponding signal, noise, and NSR values are listed. The points in the datasets of (a) and (b) correspond to segments of the points with indices from about 17000 to about 39000 in Fig. 2a and from about 28000 to about 52000 in Fig. 2b, respectively.
As shown by the $|A'(i)|$ datasets (in blue) in Fig. 2, the 500-hPa geopotential height analysis eddy amplitudes in the NH have significant meridional and temporal variations. The analysis eddy amplitudes on a fall day of 31 October (see the blue dataset in Fig. 2b) are, in general, much larger than those on a summer day of 31 August (see the blue dataset in Fig. 2a), with a signal of $7674.8 \text{ m}^2$ versus $2858.19 \text{ m}^2$, mostly attributed to the
much larger analysis eddy amplitudes in the middle and high latitudes (on the right side of the plot). Obviously, the two $|A'(t)|$ datasets (in blue) in Fig. 2 for the NH, which cover the whole Northern Hemisphere, have more diversities than those of the two much smaller regional subdomains (region A and region B) in Fig. 3, which cover only about ten percent of the Northern Hemisphere.

For the derivation of the scale predictability based on the NSR method, these variations and diversities are not explicitly considered. The only information derived from the one-dimensional analysis eddy amplitude dataset is a statistical measure: the signal, or the domain-averaged square of the total eddy amplitudes. Therefore, from the statistical perspective, the specific verification domain should have as little diversity as possible in order to derive scale predictability features with better representativeness. For example, the mesoscale circulation systems over the tropics may behave quite differently from those over the polar regions; therefore, the predictability of the mesoscale circulations over the tropics may be quite different from that of the circulation systems with the same horizontal scale over the polar regions. This is a general requirement for any overall spatial verification.

The forecast error amplitude datasets in Figs. 2 and 3 are, in general, analogous to the corresponding analysis eddy amplitude datasets; the larger the analysis eddy amplitude, the larger the forecast error amplitude. As a result, the NSR values for the same lead time are generally of comparable order.

Since the atmosphere is constantly changing, the forecasts valid at a specific time with different lead times have different initial conditions, while the forecasts with different lead times integrated from the same initial conditions should be verified against different evolving atmospheric states. The forecast error amplitudes and the NSR values valid at a specific time do not necessarily increase with the forecast lead time, since they are integrated from different initial conditions. For example, as shown in Fig. 3a, the NSR value of the 168-h forecast is smaller than that of the 144-h forecast. In addition, as shown by the forecast error datasets in Figs. 2 and 3, the error structures for forecasts produced from different initial conditions also differ. However, it is expected that, from the statistical perspective, the NSR values should increase with the forecast lead time until totally saturated. This will be illustrated in the following monthly verification results.

b. Examples of scale predictability derived by the NSR method

The examples shown in Fig. 3 are used to demonstrate the application of the NSR method to derive the predictability time limit of different wavenumbers. As derived in section 2, the NSR or $\log_{10}(\text{NSR})$ values valid at the same time but with different lead times can be used to quantify the scale predictability when bounded by the corresponding criteria listed at the end of section 2.

Table 1 lists the 500-hPa geopotential height NSR $[\log_{10}(\text{NSR})]$ values of the NCEP GFS forecasts for two individual examples, as shown in Fig. 3: region A valid at 0000 UTC 31 August 2012 and region B valid at 0000 UTC 31 October 2012. With reference to the criteria listed at the end of section 2, for one example, bounded by the $\log_{10}(\text{NSR})$ criterion of $-8$ for the scale index $m = 5$ (corresponding to wavenumber $k_{m=5} \approx 6.85$), it is derived that the predictability limit time of the 500-hPa geopotential height of wavenumber $k_{m=5} \approx 6.85$ is shorter in region A valid at 0000 UTC 31 August 2012 (between 12 and 24 h) than in region B valid at 0000 UTC 31 October 2012 (between 24 and 72 h); for another example, bounded by the $\log_{10}(\text{NSR})$ criterion of $-2\log_{10}5(\sim -2.26)$ for the wavenumber $k = 5$ (corresponding to the scale index $m \approx 2.13$), it is derived that the predictability time limit of wavenumber $k = 5$ can be longer than 120 h or even 168 h in region A valid at 0000 UTC 31 August 2012 but is a little shorter than 120 h in region B valid at 0000 UTC 31 October 2012. It can also be derived from Table 1 that the predictability limit time of the 500-hPa geopotential height of wavenumber $k_{m=6} \approx 11.09$ [with the $\log_{10}(\text{NSR})$ criterion of $-10$] is shorter than 12 h in region A valid at 0000 UTC 31 August 2012 but is as long as about 24 h in region B valid at 0000 UTC 31 October 2012. It is noted that the temporal resolution of the detected predictability time limits will depend on the output frequency of the selected forecasts.

Since the predictability time limit from an individual case perspective, either for an individual valid time or an individual initial time, is not representative, it is desirable to extract meaningful predictability time limits from a statistical perspective. The following two subsections illustrate the time evolution of the $\log_{10}(\text{NSR})$, the mean value of the $\log_{10}(\text{NSR})$, and the statistical features of the predictability time limits.

c. The time evolution of $\log_{10}(\text{NSR})$ from 1 August to 31 October 2012

Figure 4 shows the time evolution of $\log_{10}(\text{NSR})$ of the NCEP GFS 500-hPa geopotential height forecasts with lead time values of 12, 24, 72, 120, and 168 h, and the square root of signal $\sqrt{S}$ every 6 h from 1 August to 31 October 2012 for the three verification domains.

We find in Fig. 4 that $\log_{10}(\text{NSR})$ and $\sqrt{S}$ generally correlate inversely; the larger the signal, the smaller the $\log_{10}(\text{NSR})$, especially for the shorter-range (12 and 24 h) forecasts. That is, the model generally has better short-range
forecast skill for a stronger signal, even if the absolute forecast error is larger (see the example datasets in Figs. 2 and 3). This is because the NSR by definition is automatically and dynamically adaptive to flow, with its value representing the consistent model performance. Since the NSR is adaptive to flow and it also directly penalizes the systematic forecast bias and pattern errors, it should be a useful generic measure that is advantageous for consistent model comparison among different variables, different atmospheric regimes, and different scales, etc. On the contrary, many other widely used measures, such as the mean square error, the root-mean-square error, and the ACC, may either be strongly modulated by the absolute strength of the large-scale atmospheric variability or be significantly contaminated by the systematic bias and could not offer consistent comparison. Therefore, the new NSR method would be an important addition to existing model evaluation measures.

We show in Fig. 4 that all of the time series of log$_u$(NSR) have complicated multiscale intraseasonal oscillations. The diurnal cycle is evident in all the verification domains. In the NH (see Fig. 4c), there is a significant quasi-biweekly oscillation, which might be associated with the boreal summer monsoon systems (Kikuchi and Wang 2009); while in the two regional subdomains (see Figs. 4a,b), there are more short-term oscillations. This might be the result of a weaker smoothing effect on both the noise and the signal in the smaller regional subdomains that have less spatial diversity. This further supports that, from a statistical perspective, it is desirable for the specific verification domain to have as little diversity as possible in order to ensure better spatial homogeneity and better representativeness of the signal. Although there are no issues in calculating the forecast skill measures, such as the log$_u$(NSR) or the ACC, over the entire globe or the entire hemisphere, the calculation of the measures over a specific verification domain with greater homogeneity would allow a more representative estimate of the scale predictability for specific domains with specific features. We also argue that the verification could be performed in a small region of interest as long as the region as a whole has some meaningful representativeness.

Another important feature of the log$_u$(NSR) time series as shown in Fig. 4 is that the log$_u$(NSR) values for the forecasts with the same lead time generally have comparable magnitudes and very stable variations with time even though the total seasonal variational range may spread up to six. In addition, the log$_u$(NSR) values for forecasts with different ranges have consistent variability while the magnitudes increase with the lead time: that is, the log$_u$(NSR) time series have strong stationarity.

### Table 1. Individual examples of 500-hPa geopotential height NSR and log$_u$(NSR) (in parentheses) values of the NCEP GFS forecasts.

<table>
<thead>
<tr>
<th>Region and valid time</th>
<th>12 h</th>
<th>24 h</th>
<th>48 h</th>
<th>72 h</th>
<th>96 h</th>
<th>120 h</th>
<th>144 h</th>
<th>168 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region A 0000 UTC 31 Aug 2012</td>
<td>0.0120 ($-0.19$)</td>
<td>0.0287 ($-0.41$)</td>
<td>0.0760 ($-0.93$)</td>
<td>0.0906 ($-0.93$)</td>
<td>0.0002 ($-1.30$)</td>
<td>0.0084 ($-1.20$)</td>
<td>0.0012 ($-0.93$)</td>
<td>0.0012 ($-0.93$)</td>
</tr>
<tr>
<td>Region A 0000 UTC 31 Oct 2012</td>
<td>0.0120 ($-0.19$)</td>
<td>0.0287 ($-0.41$)</td>
<td>0.0760 ($-0.93$)</td>
<td>0.0906 ($-0.93$)</td>
<td>0.0002 ($-1.30$)</td>
<td>0.0084 ($-1.20$)</td>
<td>0.0012 ($-0.93$)</td>
<td>0.0012 ($-0.93$)</td>
</tr>
<tr>
<td>Region B 0000 UTC 31 Aug 2012</td>
<td>0.0331 ($-0.41$)</td>
<td>0.0674 ($-0.93$)</td>
<td>0.1512 ($-0.93$)</td>
<td>0.1841 ($-0.93$)</td>
<td>0.0313 ($-1.30$)</td>
<td>0.0391 ($-1.20$)</td>
<td>0.0412 ($-0.93$)</td>
<td>0.0412 ($-0.93$)</td>
</tr>
<tr>
<td>Region B 0000 UTC 31 Oct 2012</td>
<td>0.0331 ($-0.41$)</td>
<td>0.0674 ($-0.93$)</td>
<td>0.1512 ($-0.93$)</td>
<td>0.1841 ($-0.93$)</td>
<td>0.0313 ($-1.30$)</td>
<td>0.0391 ($-1.20$)</td>
<td>0.0412 ($-0.93$)</td>
<td>0.0412 ($-0.93$)</td>
</tr>
</tbody>
</table>
In statistics, a stationary process is one in which the mean and variance do not change over time. Therefore, the stable log$_u$(NSR) time series is useful for investigating the consistent long-term mean feature of the model performance in both the short-range, small-scale forecasts and the long-range, large-scale forecasts. This ability to compare short-range small-scale forecasts in a consistent way with long-range large-scale forecasts is quite attractive,
particularly for centers using a unified modeling approach (e.g., the Met Office).

An interesting feature of the log$_\phi$(NSR) in region A (see Fig. 4a) is that the growth rate of the log$_\phi$(NSR) between 24 and 168 h during August is much smaller than that of October. Although the short-range forecasts in August have worse skill than those in October, the longer-range forecast skills in these two periods are comparable. Derived by means of the NSR method, the scale predictability of wavenumber $k_{m=5} \approx 6.85$ (corresponding to the scale index $m = 5$) in August is, in general, shorter than 24 h, while it is generally longer than 24 h in October; however, the scale predictability of wavenumber $k_{m=2} \approx 1.62$ (corresponding to the scale index $m = 2$) in both periods is close to 168 h. This illustrates the complex variability of the predictability as a function of time (season) and scale. The comparison between Figs. 4a and 4b otherwise presents the variation of the predictability with latitude. Also, along with latitude there could be additional factors contributing to the predictability difference, such as fractional coverage of land versus ocean.

d. Mean features of the scale predictability derived by the NSR method

To obtain the mean features of the scale predictability over a period of time, we need to extract the temporal average of the NSR or log$_\phi$(NSR). We use the arithmetic mean for the log$_\phi$(NSR), which is equivalent to using the geometric mean for the NSR:

$$\log_{\phi}(\text{NSR}) = \frac{1}{T} \sum_{it=1}^{T} \log_{\phi} [\text{NSR}(it)]$$

$$= \log_{\phi} \left[ \prod_{it=1}^{T} \text{NSR}(it) \right]^{1/T}, \quad (19)$$

with it = 1, 2, ..., T, where T is the total number of times verified during the period.

Studied in this paper are three different periods in 2012: August, September, and October. Figure 5 shows the temporal average of the log$_\phi$(NSR) of the NCEP GFS 500-hPa geopotential height forecasts with different lead times every 12 h from 12 to 192 h during August, September, and October for the three verification domains: region A, region B, and the NH.

From Fig. 5, we can see that the log$_\phi$(NSR) generally follows the piecewise linear growth with lead time in all verification regions and each verification period. Region A, which is located in a low latitude belt, has a shorter predictability time limit than region B, which is located in the middle latitudes; October has better predictability than August and September for all three regions; and the predictability in the regional subdomains is very different from that in the whole Northern Hemisphere.

Table 2 summarizes some of the mean features of the scale predictability of the 500-hPa geopotential height in the NCEP GFS in 2012 derived by the NSR method. A coarse linear interpolation or extrapolation is performed if necessary to retrieve the steady growth of the NSR for calculating the time limits of the predictability of the different wavenumbers.

Figure 6 is designed to effectively present the same information in Table 2, with the derived predictability time limit (a key quantity of interest for a forecaster, as it has a very clear meaning) on the y axis as a function of wavenumber (size of meteorological feature) on the x axis. The wavenumber itself is a function of the scale index (on the y axis).

It should be noted that more accurate predictability time limits could be obtained using forecast outputs at higher temporal resolution, especially for the earlier
stage forecasts. Users of the NSR should select proper output frequency of the forecasts for their verification purposes with desired accuracy.

The results described above are the features of the practical predictability of the NCEP GFS model for the specific verification variable, time, and space. Different modeling systems should have different practical predictability limits, just as they have different forecast skills. It goes without saying that the NSR method provides a consistent approach for model comparison.

5. Comparison of the NSR method with the ACC method

a. The sensitivity of the ACC to the definition of the mean

The centered ACC that has the spatial mean error removed is defined by:

$$\text{ACC} = \frac{\left[ (F - C) - (F - C) \right] \left[ (A - C) - (A - C) \right]}{\sqrt{\left[ (F - C) - (F - C) \right]^2 \left[ (A - C) - (A - C) \right]^2}},$$

(20)

where $F$ is the one-dimensional forecast dataset, $A$ the one-dimensional analysis dataset at the valid time, and $C$ the one-dimensional mean dataset against which the anomaly is calculated.

We tested three different definitions of mean to examine the sensitivity of the ACC to the selection of the mean: the 92-day seasonal mean of the NCEP GFS analysis four times daily from 1 August to 31 October 2012; the 10-day running mean of the NCEP GFS analysis four times daily centered at the valid time; and the instant zonal mean of the NCEP GFS as defined in Eq. (1) of section 2. Figure 7 shows the time evolution of the ACC of the NCEP GFS 500-hPa geopotential height forecasts with lead time of 12, 24, 72, 120, and 168 h from 1 August to 31 October 2012 for region A with the anomalies calculated against these three different definitions of the mean. It should be noted that, because of data availability and also for the sake of simplicity, we do not test against the common climatology set by the WMO or the one operationally used by the NCEP.

As shown in Fig. 7, both the ACC calculated against the seasonal mean and against the instant zonal mean

![Fig. 6. The predictability time limits of the NCEP GFS 500-hPa geopotential height forecasts on different scales during August (solid line with + signs), September (solid line with open circles), and October (solid line with black dots) for the three different verification domains: (a) region A [−(15°–35°N and 30°–150°W)], (b) region B [−(25°–50°N and 30°–150°W)], and (c) NH [−(0°–90°N and 0°–360°)].](image-url)
FIG. 7. The time evolution of the ACC of the NCEP GFS 500-hPa geopotential height forecasts with lead time of 12 (blue), 24 (brown), 72 (red), 120 (green), and 168 h (purple) every 6 h from 1 Aug to 31 Oct 2012 for region A \([-15^\circ-35^\circ N \text{ and } 30^\circ-150^\circ W]\) with the anomalies calculated against: (a) the seasonal mean, (b) the 10-day running mean, and (c) the instant zonal mean.
are somehow similar, while their time variation details sometimes differ and the latter is, in general, a little smaller. On the contrary, the ACC calculated against the 10-day running mean is very different, in terms of amplitudes, phases, or peaks; it is much smaller than the ACC calculated against the other two kinds of mean. As expected, the ACC is sensitive to the definition of the mean.

b. The weak stationarity of the ACC

Totally different from the time evolution of the log_u(NSR) shown in Fig. 4a, the ACC values (in either Fig. 7a or Fig. 7c) of the same lead time have many sharp variations in time, and the variation ranges are particularly large for the long-range forecasts. More importantly, the ACC values of the different forecast ranges have widely varying magnitudes, incoherent variability, and strong nonlinearity in growth rate. Therefore, contrary to the log_u(NSR) time series, the ACC time series have very weak stationarity, which makes it difficult to calculate meaningful long-term mean statistics. It is also very difficult to investigate the model performance in the short-range forecasts of small-scale weather systems because the ACC values are always very close to 1.0 for short-range forecasts.

c. The mean features of the model predictive skill derived by the ACC method

Figure 8 shows the temporal average of the ACC of the NCEP GFS 500-hPa geopotential height with different forecast lead times every 12 h from 12 to 192 h during August (solid line with + signs), September (solid line with open circles), and October (solid line with black dots) for region A [−(15°–35°N and 30°–150°W)] with the anomalies calculated against (a) the seasonal mean, (b) the 10-day running mean, and (c) the instant zonal mean.

Although similar qualitative conclusions can be drawn on the GFS model forecast skills from Fig. 8 and Fig. 4a such that the GFS has better forecast skill in October than in August and September and the model has the worst skill in September for all forecast ranges in region A, it is not possible to derive an objective and quantitative evaluation of the scale predictability from the ACC values in Fig. 8.

In addition, the strong nonlinearity of the ACC’s variation among the lead times is quite evident, especially in shorter-range forecasts. The left side of Fig. 4a and the solid line with crosses in Fig. 5a clearly illustrate that there is a larger rate of increase in the log_u(NSR) with lead time during shorter-range forecasts (from 24 to 72 h) than that during the longer-range forecasts (from 120 to 168 h) in region A in August. This feature might be related to the model’s capability in predicting the convections. It is natural that the model performance should vary with both time and space, which might be occupied by different weather regimes. However, this apparent error growth in short-range forecast cannot be easily identified in Fig. 8 by the ACC, since the rate of decrease in the ACC with the lead time in shorter-range forecasts is very small. The fundamental issue is that the ACC does not allow a clear separation of error growth for weather systems at different scales. Consequently, the model performance in the short-range forecasts, which should be dominated by error in smaller-scale systems, is overpowered by the larger-scale weather systems. This actually reflects the fact that the log_u(NSR) values are much more sensitive than the ACC values to the short-range, small-scale forecast errors and are thus more useful for detecting the varying practical predictability of the model, especially during the early stage of forecasts.
6. Conclusions and discussion

The ACC is one of the standard methods for the evaluation and comparison of the performance of global NWP models. The correlation coefficient of the forecast anomaly field with the analysis anomaly field both valid at the same time is used as a measure of model forecast skill. The calculation of the ACC is known to be sensitive to the definition of the mean state from which the anomaly is extracted. Although the WMO has established a standard definition of the climate mean state for the global NWP community, different operational centers still use different climatologies to calculate the ACC for their own verifications. In addition, the ACC time series have very weak stationarity, which makes it difficult to calculate meaningful long-term mean statistics, and sometimes certain running mean time series have to be derived as smoother replacements but with less representativeness. Since the ACC does not penalize systematic bias and errors in patterns, it is often used together with the bias and RMSE to assess model performance. Another significant shortcoming is that the ACC is strongly modulated by the strength of the anomalies in the forecast and analysis, which may not be related to the model performance. Because the entire anomaly fields are used for the calculation, it is not possible to differentiate the performance of the model on different scales without explicitly performing scale decomposition. For example, the short-range (12–24 h) forecast of a global model should be dominated by errors on the smaller scales. Yet, the ACC is generally very close to one in the short-range (12–24 h) forecasts, because the ACC is dominated by the performance of the model on the larger scale of motions (with large amplitude and high correlation). In other words, the ACC is not sensitive to short-range, small-scale errors. As a result, it is impossible to evaluate the performance of the model on small scales based on the ACC by itself.

In this paper, we propose a new measure of model forecast skill, which we name the NSR. The basic concept is to calculate the ratio of the noise (defined as the average square of the model error amplitudes) to the signal (defined as the average square of total eddy amplitudes) across all scales. On the surface, this is not significantly different from that of the ACC. However, by taking advantage of the intrinsic properties of the self-similarity and the inverse power law of the fractal atmosphere based on the general systems theory of Selvam (1990), and by following the assumption of the upscale cascade of error, we show that the time limit of predictability \( t_{km} \) for any wavenumber \( k_m = \varphi^{m-1} \) of the fractal atmosphere could be determined by the criterion \( \text{NSR} \left( t_{km} \right) = \varphi^{-2m+2} = k_m^{-2} \) or by the criterion \( \log_m[\text{NSR} \left( t_{km} \right)] = -2m + 2 = \log_m(k_m^{-2}) \), where \( \varphi = (1 + \sqrt{5})/2 = 1.6180339887… \) is the golden ratio and \( m \) is a scale index, with the real number \( m \in [1, M] \) and real number \( k_m \in [1, K] \) \( (K \) and \( M \) are the largest wavenumber and the maximum scale index, respectively, associated with the data resolution as well as the dynamic diffusion used in the model). In this way, the predictability of different spatial scales can be quantitatively detected without explicitly performing scale decomposition. This is an important advantage of the NSR method over the ACC method.

There are several other unique advantages of the NSR method. First, the mean (from which the eddies are calculated) is defined by the instant zonal mean. Therefore, by definition, the NSR is adaptive to flow. Second, the time series of the logarithm with base \( \varphi \) of the NSR, \( \log_{\varphi}(\text{NSR}) \), have comparable, coherent, and stable variations in different ranges of forecasts, implying strong stationarity, which is advantageous for calculating meaningful mean statistics on model performance in both short-range, small-scale and long-range, large-scale forecasts. Third, the variation of the \( \log_{\varphi}(\text{NSR}) \) from shorter- to longer-range forecasts as a function of lead time provides an assessment of the model error growth between scales (e.g., how fast errors on smaller scales propagate and contaminate motions of larger scales). Since the NSR directly penalizes the systematic bias and pattern errors, it can effectively detect where and when the model loses skill on different scales because of the forecast error growth. Fourth, because the calculation is based on the instantaneous forecast and analysis datasets, the NSR method can be consistently applied to any time, space, variables, ensemble members, and models; it is a generic method. For example, the meaningful mean (e.g., monthly mean statistics) can be calculated and compared among the different forecast models. This will provide useful insight on the relative performance of the different models or of the different physics in the same model, or of the same model run at different resolutions.

As a demonstration of the advantages of the NSR method, we evaluate the performance of the NCEP GFS model forecast of the 500-hPa geopotential height from August to October in 2012 and compare that derived with the ACC method. The relative performance of the NCEP GFS in different months is clearly evident in both measures of the NSR and the ACC. However, the strong sensitivity of the ACC to the definition of the mean is shown to be quite substantial. Also, there can be huge variations in the ACC in time for different forecast lead times; the NSR does not have these issues. Moreover, the NSR method provides additional insights on the predictability time limits of different scales, how they differ in different regions dominated by different
weather systems, and how they change with time as the season progresses, without the need to explicitly perform the scale decomposition.

With the ability to reveal the predictive skills on different scales as well as the model error growth between scales, the NSR method can provide useful insights on the relative performance of different global models (e.g., the NCEP GFS vs the ECMWF) or the relative performance of the same model at different resolutions (e.g., the NCEP GFS vs the NCEP CFS) on different scales. For example, questions such as “which model performs better on smaller scales (e.g., which may be driven by convection)?” and “which model has a faster error growth rate between scales?” can be addressed.

The analysis of the NSR results of a forecast model may provide useful guidance on future model improvement. The NSR method can be applied over subdomains of a global model, which may shed light on model challenges over a specific region. With the aid of a global analysis, the NSR method can also be used to verify a regional model. Exciting developments in NWP are being made in high-resolution convection-permitting models, and in extended-range coupled prediction. It would be interesting to use the NSR method to measure predictability for subdaily forecasts and for monthly forecasts from coupled models. These applications of the NSR method deserve to be pursued in future studies.

There are two important attributes of the new NSR method: 1) to define the NSR as a normalized measure to assess the short-range, small-scale forecast skills and the long-range, large-scale forecast skills in a consistent way; and 2) to identify the golden ratio associated criteria as the thresholds to quantitatively detect the successive scale predictability time limits given the NSR values. The second attribute works only if Selvam’s general systems theory is applicable for the Earth atmosphere. In Selvam’s general systems theory, the atmosphere is assumed fractal, and the fractal dimension is the golden ratio. It is under this theory in addition to the assumption of upscale error cascade that the NSR values could be further interpreted and used to derive the successive scale predictability time limits. We find that Selvam’s theory actually implies a $-3$ spectral decay law, which has already been demonstrated by many previous theoretical, numerical, or observational studies to be valid for the Earth atmosphere at least down to scales of a few hundreds of kilometers. Many insightful studies (see the appendix) actually suggest that the $-3$ power law should extend to a much smaller scale than the transition mesoscales identified by the shallower slope of $-5/3$ in the Nastrom–Gage spectrum (Nastrom and Gage 1985) before reaching the dissipation range.

In a nutshell, our new NSR method for assessing the short-range, small-scale forecast skills and the long-range, large-scale forecast skills in a consistent way and for quantifying the practical predictability of atmospheric circulations of different scales without the need of explicit scale decomposition could be generic, robust, and useful for model verification.

Acknowledgments. We thank the reviewers and Christopher A. Davis for their valuable comments, which substantially improved the presentation of this paper. This work is jointly supported by the National Aeronautics and Space Administration and by the National Science Foundation under NSF Award AGS-1033112 and by the NOAA Hurricane Forecast Improvement Project through the Developmental Testbed Center under Award NA08NWS4670046.

APPENDIX

Atmospheric Spectrum Structure

In their seminal work, Nastrom and Gage (1985) presented an analysis of the flight observations of wind and temperature, showing a distinct transition from a steep spectral slope of $-3$ at synoptic scales to a shallower slope of $-5/3$ at mesoscales. The spectrum agrees well with Charney’s (1971) theory of geostrophic turbulence at large scales but deviates from the theory at scales where it shallows. As reviewed in Tulloch and Smith (2006), understanding the source and structure of the Nastrom–Gage spectrum has posed a puzzle in atmospheric science over the past 20 years. In particular, the fact that the small scale has a slope of $-5/3$ has invited multiple explanations.

Explanations in the literature for the shallower mesoscale spectrum fall into three general categories: 1) an inverse cascade of small-scale energy, produced perhaps by convection; 2) production of gravity waves by divergent flows; or 3) a direct cascade of energy from the large scales. More recent observations and analyses present new facts debating the old explanations of the first two types, and only theories of the third type are left as plausible universal explanations of the Nastrom–Gage spectrum. Therefore, most of the different theories proposed are based on the hypothesis of an energy cascade. Tung and Orlando (2003) proposed a theory that the subdominant $-5/3$ cascade reveals itself at the wavenumbers where the direct cascading energy spectrum exceeds that of the enstrophy cascade. They supported their theory using a standard two-layer quasigeostrophic numerical model forced by baroclinic instability and reproduce the mesoscale shallow spectrum. As noted by Smith (2004), because
the forward energy cascade rate depends on the dissipation scale, the transition scale of Tung and Orlando (2003) will always coincide with the effective Kolmogorov scale of the dissipation mechanism and so changes with filter strength and grid resolution. He argued that if the theory of Tung and Orlando (2003) is correct, then, the atmosphere must possess a mechanism that selectively dissipates the forward cascade at some fixed $O(1)$ km scale, independent of energy flux. Tran and Shepherd (2002) also indicated that the choice of forcing mechanisms and dissipation operators has implications for the spectral distribution of energy and enstrophy dissipation and, thus, for the possible existence of energy and enstrophy cascades. Furthermore, the choice of dissipation operators has implications for permissible scaling of the energy spectrum. They remarked that, in choosing forcing mechanisms and dissipation operators for numerical reasons, one should be mindful of these constraints. Tulloch and Smith (2006) proposed a finite-depth surface-quasigeostrophic model that highlights the transition between quasi-two-dimensional barotropic flow and three-dimensional baroclinic flow. They demonstrated how turbulent motions at the synoptic scale can produce a balanced, forward cascade of temperature, resulting in an upper-tropospheric spectrum with a break at a scale that is a function of fundamental background parameters. Later, Tulloch and Smith (2009) showed that a quasigeostrophic model driven by baroclinic instability exhibits such a transition near its upper boundary (analogous to the tropopause) when surface temperature advection at that boundary is properly resolved and forced. The transition wavenumber depends linearly on the ratio of the interior potential vorticity gradient to the surface temperature gradient. Obviously, the transition scales as derived in these numerical models are tunable under different numerical or dynamical assumption details. There are still many limitations and uncertainties in the related numerical studies.

Recently, the spectrum analysis and the structure function analysis are used on high-resolution observations or model simulations to revisit the atmospheric spectrum issue. As reviewed in Augier and Lindborg (2013), while the $k^{-5/3}$ power law and the corresponding $r^{2/3}$ second-order structure function can be obtained from observations (e.g., Frehlich and Sharman 2010) or from the model simulations by some general circulation models (e.g., Koshyk and Hamilton 2001; Hamilton et al. 2008; Augier and Lindborg 2013) and regional mesoscale numerical weather prediction models (e.g., Skamarock 2004; Ricard et al. 2013), other global models, such as the ECMWF’s weather prediction model Integrated Forecast System (IFS), produce mesoscale spectrum significantly steeper than that analyzed from the observations, even at relatively high resolution (e.g., Shutts 2005; Burgess et al. 2013; Augier and Lindborg 2013). The shallower mesoscale spectrum was not seen in earlier deterministic versions of the ECMWF forecast model (Palmer 2001; Palmer et al. 2005). Palmer (2001) and Palmer et al. (2005) noted that the tropospheric kinetic energy spectrum produced by this model steepened in the mesoscale, rather than shallowing. With the addition of stochastic backscatter, a shallower mesoscale spectrum appeared (Shutts 2005; Palmer et al. 2005). The stochastic backscatter actually compensates for a model deficiency by injecting energy at the smallest resolvable scales. In the ECMWF T799 analysis, the power law characteristic of the two-dimensional turbulent enstrophy-cascading subrange, extends into the mesoscale to spatial wavelengths of 400 km or less (see Fig. 1 in Burgess et al. 2013); at 250 hPa, the spectral slope is close to $-3$, and the spectrum break is still not that distinctive above 250 hPa; its mesoscale spectrum has a power law between $-2.0$ and $-2.5$, which is steeper than the well-known $-5/3$ mesoscale shallow spectrum. The spectral break moves upscale from about $k = 60$ to $k = 20$ between 230 and 100 hPa, and remains fairly stable at higher altitudes. Their study suggested that this transition layer, and the range of wavenumbers at which the spectral break occurs, is model dependent; in particular, for a higher-resolution analysis with a shallower mesoscale spectrum, the break might appear at lower altitudes and higher $k$ than their results. As expected, the mesoscale shallowing emerges as a result of the divergent component exceeding the rotational component of the flow in the kinetic energy spectra from forecasts produced with a more recent, higher-resolution (T1279) version of the ECMWF forecast system (see Fig. 6 in Burgess et al. 2013), the mesoscale spectrum is slightly weaker with a somewhat shallower slope than that in the T799 analysis (but both are steeper than the well-known marked $-5/3$ mesoscale shallow spectrum) and extends to higher wavenumbers (with the dissipation range beginning somewhere between $k = 200$ and $k = 300$). It is interesting that the well-known $-5/3$ mesoscale shallow spectrum does not find a counterpart in the ECMWF model, one of the best numerical weather prediction systems in the world.

Kolmogorov (1941) assumed the presence of anomalous dissipation in developing his power law for 3D isotropic and homogeneous turbulence. His argument for the $-5/3$ spectrum is essentially as follows: for isotropic homogeneous turbulence, $E(k)$ should depend only on the local wavenumber $k$ and the energy flux of the cascade $\varepsilon$ through that wavenumber. Then using dimensional arguments, $E(k)$ must have the form of the $-5/3$ power. Here $\varepsilon$ is the energy flux from low wavenumbers to high wavenumbers and should be a constant in the inertial subrange (where there is no forcing or
dissipation). Charney’s (1971) theory argued that at the large scale, rotation and stratification conspire to make the atmosphere quasi-two-dimensional. Stirring by baroclinic instability (or any planetary mechanism) will induce a forward cascade of potential enstrophy, reflected in a $-3$ kinetic energy spectrum below the stirring scale. The forward enstrophy cascade in this theory should proceed down to scales at which rotation becomes less important, where unbalanced motions and instabilities might efficiently lead to dissipation. He suggested that, in the atmosphere, a reasonable estimate puts this scale an order of magnitude smaller than the observed transition scale. In addition, in a recent work of Yushkov (2012) on the structure function analysis of sound speed and local entropy in the turbulent atmosphere, behavior of structure functions also indicates significant deviation from the asymptotic Kolmogorov–Obukhov (Kolmogorov 1941) law on the inertial subrange scales, which significantly exceed the Taylor microscale and are not strongly affected by viscosity. These studies suggest that the $-3$ power law should extend to a much smaller scale than the transition mesoscales identified by the shallower slope of $-5/3$ in the Nastrom–Gage spectrum (Nastrom and Gage 1985) before reaching the dissipation range. That is, the realistic atmospheric spectrum should break and transition at much smaller scales rather than at mesoscales if it does break and transition.

Despite many attempts over the last 30 yr, the atmospheric turbulence spectrum analyzed from the observations has not been fully explained, and the real physical mechanisms that account for the universality of the atmospheric spectrum structure are still unclear. The review presented above raises an interesting and bold question: is the atmospheric turbulence spectrum analyzed from the observations that have high horizontal resolution and instabilities might efficiently lead to dissipation. He suggested that, in the atmosphere, a reasonable estimate puts this scale an order of magnitude smaller than the observed transition scale. In addition, in a recent work of Yushkov (2012) on the structure function analysis of sound speed and local entropy in the turbulent atmosphere, behavior of structure functions also indicates significant deviation from the asymptotic Kolmogorov–Obukhov (Kolmogorov 1941) law on the inertial subrange scales, which significantly exceed the Taylor microscale and are not strongly affected by viscosity. These studies suggest that the $-3$ power law should extend to a much smaller scale than the transition mesoscales identified by the shallower slope of $-5/3$ in the Nastrom–Gage spectrum (Nastrom and Gage 1985) before reaching the dissipation range. That is, the realistic atmospheric spectrum should break and transition at much smaller scales rather than at mesoscales if it does break and transition.

Despite many attempts over the last 30 yr, the atmospheric turbulence spectrum analyzed from the observations has not been fully explained, and the real physical mechanisms that account for the universality of the atmospheric spectrum structure are still unclear. The review presented above raises an interesting and bold question: is the atmospheric turbulence spectrum analyzed from the observations that have high horizontal resolution and instabilities might efficiently lead to dissipation. He suggested that, in the atmosphere, a reasonable estimate puts this scale an order of magnitude smaller than the observed transition scale. In addition, in a recent work of Yushkov (2012) on the structure function analysis of sound speed and local entropy in the turbulent atmosphere, behavior of structure functions also indicates significant deviation from the asymptotic Kolmogorov–Obukhov (Kolmogorov 1941) law on the inertial subrange scales, which significantly exceed the Taylor microscale and are not strongly affected by viscosity. These studies suggest that the $-3$ power law should extend to a much smaller scale than the transition mesoscales identified by the shallower slope of $-5/3$ in the Nastrom–Gage spectrum (Nastrom and Gage 1985) before reaching the dissipation range. That is, the realistic atmospheric spectrum should break and transition at much smaller scales rather than at mesoscales if it does break and transition.

We can imagine that the overestimation of the energy at larger scales and/or the underestimation of the energy at smaller scales would distort the energy spectrum to a deeper slope than the real one; on the contrary, the underestimation of the energy in the larger scales and/or the overestimation of the energy in the smaller scales would distort the energy spectrum to a shallower slope. Unfortunately, this usually happens in the practice of spectrum analysis or structure function analysis when the dataset (either from observations or model simulations) is incomplete, not representative, or is overdiffused or overfiltered. Almost all of the observational datasets used to calculate the spectrum structure are either 1D regional segments or 2D regional boxes; that is, they represent only the regional limited sampling of the real atmosphere that actually possesses a full spectrum of scale, ranging from wavenumber 1 to the maximum resolved wavenumber. The structure function analysis should be very sensitive to the length of the data segments; and without the presence of larger scales, the regressed slope for the investigated spectrum band would tend to be shallower. In addition, some techniques, such as detrending, filtering, and combining, etc., introduced in the data preprocessing may distort the energy spectrum results as well. For example, as recognized in Skamarock (2004), the spectrum computed using the full length $L$ of an ideal dataset differs much from that using the $3L/4$ of the data. In practice, datasets much shorter than this ratio are used for the calculation of the mesoscale spectrum.

It is desirable that a realistic observational atmospheric spectrum is used to check the spectrum behaviors in the numerical model. As shown in Kahn et al. (2011), higher horizontal spatial resolution observations over the entire globe are necessary to observe the global characteristics of small-scale “turbulence,” although this is not an easy undertaking. The limited spatial resolution places a constraint on earlier studies of global atmospheric analyses, restricting the extent to which two-dimensional turbulence theory describes the atmospheric circulation. In addition, the distribution of energy among scales in model simulations may suffer from the deficiencies in model physics and dynamics. With the improvement of meteorological analyses that have much higher spatial resolution and include a well-resolved stratosphere and with the availability of observations that have high horizontal resolution and global coverage, it is expected that the mutual corroboration among theories, observations, and model simulations will help us understand the universality of the atmospheric spectrum structure, determine the intrinsic and practical predictability of atmospheric circulations of different scales, identify the source and nature of errors, and finally improve weather and climate models.

REFERENCES

Baumhefner, D. P., 1984: The relationship between present largescale forecast skill and new estimates of predictability error