Ensemble Forecasting: A Foray of Dynamics into the Realm of Statistics

Jie Feng[1, 2, 3], Zoltan Toth[4*], Jing Zhang[5], and Malaquias Peña[6]

[1]Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China

[2]Shanghai Key Laboratory of Ocean-land-atmosphere Boundary Dynamics and Climate Change, Shanghai, China

[3]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China

[4]Global Systems Laboratory, NOAA, Boulder, CO

[5]Shanghai Typhoon Institute, China Meteorological Administration, Shanghai, China

[6]Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT

*Corresponding Author: Zoltan Toth, zoltan.toth@noaa.gov

Accepted for publication in:
Quarterly Journal of the Royal Meteorological Society
13 April 2024

ABSTRACT

Uncertain quantities are often described through statistical samples. Can samples for numerical weather forecasts be generated dynamically? At a great expense, they can. With statistically constrained perturbations, a cloud of initial states are created and then integrated forward in time. By now, this technique has become ubiquitous in weather and climate research and operations. Ensembles are widely used, with demonstrated value.

The atmosphere evolves in a multidimensional phase space. Does a cloud of ensemble solutions encompass the evolution of the real atmosphere? Theoretically, random perturbations in high dimensional spaces have negligible projection in any direction, including the error in the best estimate, therefore consistently degrading that. As the bulk of the perturbation variance lies in the null-space of error, samples in multidimensional space do not contain reality.

An evaluation suggests that initial and short-range forecast error and ensemble perturbations are random draws from a high dimensional domain we call the subspace of possible error. Error in any initial condition is partly a result of stochastic observational and assimilation noise, while perturbations explore other, mostly independent directions from the subspace of possible error that may have resulted from other configurations of stochastic noise. What benefits may arise from the deterministic projection of such noise?

Consistent with theoretical expectations, ensemble members consistently degrade the skill of the unperturbed forecast until medium range. The mean and all other products derived from ensembles suffer an 18-hour loss in forecast Information. Since Information is a sufficient statistic, any rational user can benefit more from the unperturbed, than from an ensemble of weather forecasts. Furthermore, case dependent variations in the distribution or spread of ensembles have no impact on commonly used metrics. Can alternative, statistical applications provide comparable, or even higher quality probabilistic and other products, at the fraction of the cost of running an ensemble?

1. INTRODUCTION

Weather forecasting is one of the greatest success stories of natural sciences (Richardson, 1922; Bauer et al., 2015). Drawing on the theory of dynamics and thermodynamics, in an abstract setting, numerical models replicate the larger, resolved-scale dynamics of the atmosphere (Charney, 1949; Kalnay, 2003). In numerical weather prediction (NWP), observations of the atmosphere are collected and fused into an estimate of the initial state, called an analysis. Numerical forecasts initialized from such analyses then attempt to capture the temporal evolution of the atmosphere by exploiting deterministic relationships in nature. Useful forecast skill now extends to 10 days lead time and beyond (Bauer et al., 2015; Zhang et al., 2019) - a feat unimaginable just decades ago.

Despite continual reductions in initial error over the decades, error still amplifies in the forecasts. Eventually errors reach a level comparable with that in states randomly chosen from the climatic distribution, at which point forecasts become useless (Lorenz, 1982). By now it is well understood that the loss of forecast skill is intrinsic to a large class of aperiodic deterministic systems called chaotic dynamical systems (Thompson, 1957; Lorenz, 1963; Li and Chou, 1997; Mu et al., 2004). As it is not due to methodological problems, this loss of skill is unavoidable (Lorenz, 1963). Weather is predictable - but only for a finite period.

Nature unfolds along a unique path in time and 3-dimensional space. NWP forecasts attempt to predict this evolution in a similar form, as a unique sequence of events. Especially at longer lead times a single-value forecast in itself, however, can be rather deceptive. Such forecasts do not indicate how large their error may be, and which part of their variance will match reality. This is a major challenge for weather forecasters and users alike as for optimal decision making the level, and possibly the nature of uncertainty must be known in advance (Leutbecher and Palmer, 2008).

After a brief review of statistical alternatives (Section 2), we introduce the concept of ensemble forecasting, a dynamical approach to assessing forecast uncertainty, along with its current status and presumed benefits (Section 3). Specific methodologies considered in this study, such as forecast system attributes, some metrics of forecast performance, including an analysis of perturbations in multidimensional space, and the sources of forecast error are discussed in Section 4. Long-held assumptions about ensembles are revisited in Section 5, while some theoretical explanation of the experimental results are offered in Section 6. The paper ends with some conclusions and a discussion (Sections 7 and 8, respectively).

2. STATISTICAL METHODS

2.1 Sampling

Statistical tools are available to describe uncertain quantities like weather analyses or forecasts. A sample or a distribution representing the expected error in the best estimate can readily show the range of values a quantity might take. Assuming, as an example, that the error in an analysis follows a normal distribution with known parameters, the black curve centered around reality, whose exact value is unknown indicates the possible position of an analysis. While the blue curve offers an example for a distributional estimate of reality, which if the distribution is statistically reliable (i.e., perturbation variance equals error variance), is identical to the distribution of possible analyses, except translated to center on an arbitrarily selected realization of the analysis (Fig. 1a).

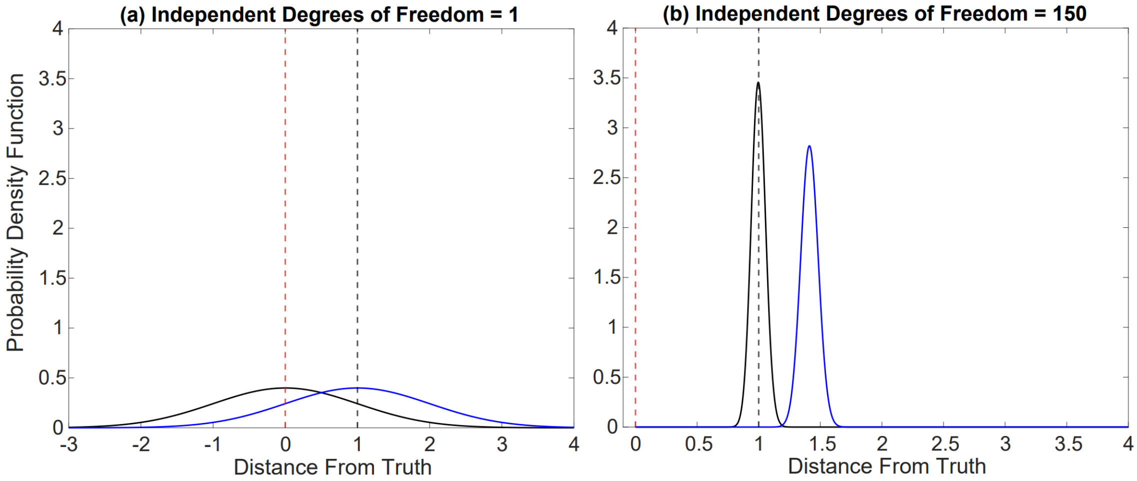

Figure 1. Reality (red dashed vertical line, unknown in practice), the distribution of its best estimate assuming error in it is known and normally distributed (black curve), an example for a best estimate (selected at a distance of the standard deviation from Reality, black dashed vertical line), and the estimated distribution of reality around the best estimate (blue curve) in a 1- (panel a, directional distance) and 150-dimensional space (panel b, absolute distance). For further details, see text here and in Section 6.2.

Error, by definition, is unknown at the time a forecast is made. Error variance, however, may be statistically assessed and used as an indicator of forecast uncertainty, as long as a representative joint forecast-validation sample is available. Error variance in real time NWP guidance (i.e., analysis or forecast) fields at lead time i ($\mathbf{G}_i$), for example, can be estimated based on error variance in a sample of past guidance fields ($\mathbf{S}_i$) similar to $\mathbf{G}_i$ in lead time, location, seasonality, regime, etc. (van den Dool, 1989; Zorita and von Storch, 1999; Hamill and Whitaker, 2006; Li and Ding, 2011):

$$e^2_{\mathbf{G}_i} = \mathrm{E}\left(e^2_{\mathbf{S}_i}\right) , \tag{1}$$

where $\mathrm{E}(\cdot)$ represents the expected value, and $e^2_{\mathbf{S}_i}$ and $e^2_{\mathbf{G}_i}$ are defined as:

$$e^2_{\mathbf{S}_i} = |\mathbf{S}_i - \mathbf{T}|^2 \text{ , } e^2_{\mathbf{G}_i} = |\mathbf{G}_i - \mathbf{T}|^2 , \tag{2}$$

and $\mathbf{T}$ is the corresponding truth or its proxy (e.g., a verifying analysis, see Appendix A).

2.2 Products

As an example, a set of surrogate or perturbed analysis or forecast fields ($\mathbf{P}_i$) can be created by the addition of perturbation fields ($\boldsymbol{\varepsilon}_i$) to the best, unperturbed single-value reference analysis or forecast field (sometimes also called "deterministic", that from here on we call control, $\mathbf{G}_i$):

$$\mathbf{P}_i = \mathbf{G}_i + \boldsymbol{\varepsilon}_i , \tag{3}$$

Conveniently, past error patterns, if available, can serve as perturbations to create a sample of surrogate forecasts (e.g., Delle Monache et al., 2013). If a large enough archive of past error fields are not available, alternative sample generation methods include the addition of random noise (Leith, 1974; Palmer et al., 1990), spatiotemporal shifts of a single forecast (neighborhood methods, e.g., Atger, 2001), or the collection of earlier initialized forecasts valid at the same time (lagged forecasts, Hoffman and Kalnay, 1983).

The mean of a sample is an often used central tendency indicating the expected weather:

$$\mathbf{E}_i = \frac{1}{M_e} \sum_{k=1}^{M_e} \mathbf{P}_{i,k} \quad . \tag{4}$$

where k is the index for perturbations, and M_e is the sample size. If perturbations are centralized before they are added to a reference state:

$$\sum_{k=1}^{M_e} \boldsymbol{\varepsilon}_{0,k} = 0 \quad , \tag{5}$$

the mean will equal the best estimate. In general, the mean captures the common component shared by all members. Typically, by filtering out presumably unpredictable noise, the mean of representative samples lowers forecast error.

To ensure statistical representativeness, the variance or spread of perturbation fields i is set equal to the estimated error variance in the best estimate:

$$V_i = \frac{1}{M_e}\sum_{k=1}^{M_e} |\mathbf{P}_{i,k} \; - \; \mathbf{E}_i|^2 \, . \tag{6}$$

While the mean attempts to capture the predictable forecast signal, the spread (i.e., the standard deviation) measures the residual variability of sample points around their mean. Importantly, statistical sampling of forecast error involves the repeated, mechanistic insertion of perturbations around a single reference (control) forecast at every lead time (Fig. 2a).

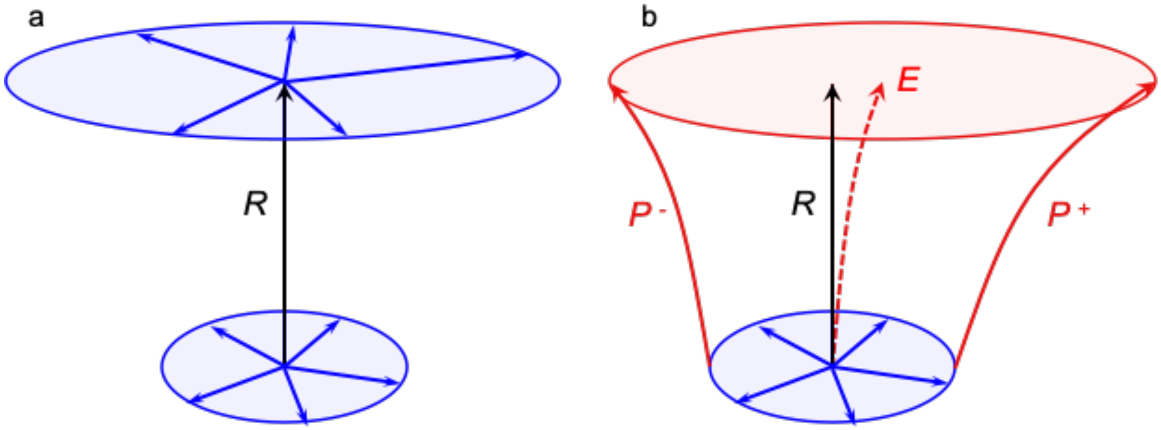

Figure 2. Schematic of statistical (a) vs. dynamical (b) generation of forecast perturbations. In either case, initial perturbations (bottom ellipsoids) are centered on a reference initial condition (R, typically a control analysis and forecast, vertical black line). Forecast perturbations (top ellipsoids) are either statistically added and centered on R (a, blue arrows), or generated via the numerical integration of a dynamical model from perturbed initial conditions (b, red arrows). P^-, P^+ (red solid), and E (red dashed) represent two perturbations initially symmetric around, but later off-center of R, and the mean of the ensemble, respectively. For further explanation, see text.

Using representative samples created around the best (control) single value forecast, a variety of probabilistic and other products can be easily constructed in distributional or categorical (for semi-closed or closed intervals, Anderson, 1996; Ebert, 2001) forms. For decades, statistical post-processing methods have been used to estimate and reduce forecast error, and generate well calibrated forecasts in a variety of probabilistic and other formats (Wilks, 2009; Scheuerer, 2014, Chen et al. 2022). Due to limitations in methodology and the size of forecast archives, statistically generated surrogate forecasts, however, generally lack dynamical balance. Past forecast cases that best match the current forecast at a selected region and lead time, for example, lose such similarity at other locales and lead times. This is due to the high dimensionality of the atmospheric circulation (e.g., van den Dool, 1994). Hence to ensure representativeness, the selection of past forecast cases is often location and lead time dependent (e.g., van den Dool, 1998). Which results in perturbations that lack spatiotemporal and across variable coherence or dynamical balance.

3. DYNAMICAL ALTERNATIVE

3.1 Ensemble Forecasting

Considering the limitations of statistical sampling algorithms and the success of the numerical approach to weather forecasting, a desire for the dynamical sampling of forecast uncertainty arose early on. Instead of the repetitive sampling of individual forecast variables (e.g., weather parameters at selected locations and lead times), why don't we sample the dynamical evolution of the entire atmosphere? In the 1960s an idea about a "glob of points, each of which would follow its own deterministic path" emerged (Edward Epstein, quoted by Lewis, 2005). The basic concept of ensemble forecasting is rather simple. Insert perturbations around the analysis of the atmosphere only once, at the initial time. To represent uncertainty in the analysis (Eq. 1, see also Section 4.3.3), the magnitude of initial perturbations is set equal to that estimated in the analysis. And to retain skill in the mean, the initial sample is typically centered on the best, control estimate of the state (Eq. 5). To create an ensemble, forecast perturbations are then dynamically generated by numerical integrations of the same (or to simulate model related errors, a different, e.g., Houtekamer et al., 2009) numerical model used to make the unperturbed control forecast (Fig. 2b). A collection of such perturbed initial and forecast conditions are hence called an ensemble.

In the late 1980s and early 1990s, following experiments with models only about half the resolution of operational forecasts at the time, the idea gained momentum. In 1992, related efforts led to the operational implementation of the Global Ensemble Forecast System (GEFS) at the National Centers for Environmental Prediction (NCEP, Toth and Kalnay, 1993). The routine weekend production of ensemble forecasts at the European Center for Medium Range Weather Forecasts (ECMWF) commenced shortly afterward (Molteni et al., 1996). The rest is history (Lewis, 2005).

The dynamical generation of an ensemble, of course, comes at a significant cost. Depending on membership and resolution, in comparison with a single forecast, an order or two more computational resources may be required. Still, today dynamically generated ensembles constitute the main or sole mode of operation at most or all numerical weather and climate prediction centers (Palmer, 2019; Zhou et al., 2019; Chen and Li, 2020). After decades of resistance, operational forecasters and other practitioners from a wide range of application areas (from hydrology, e.g., Schaake et al., 2007, to agriculture, energy, and other sectors, e.g., Alemu et al., 2011; Calanca et al., 2011; Su et al., 2014) and across many timescales (from nowcasting, e.g., Liguori et al., 2012, to multi-seasonal and decadal forecasts, e.g., Krishnamurti et al., 1999; Hou et al., 2018; Liu et al., 2023) have also embraced the practice (e.g., Bougeault et al., 2010). Ensembles and products derived from them, whether they represent the best possible guidance or not, are widely used, with proven value.

3.2 Perceived Benefits

Over the past decades, the potential benefits of ensemble forecasting have been discussed extensively. In this section we offer a brief overview of the perceived benefits. A more detailed analysis follows in Section 5.

3.2.1 Alternative Scenarios

An attractive feature of ensembles is that they offer dynamically consistent alternative scenarios for future weather. Talagrand or Analysis Rank Histograms (Fig. 3, Candille and Talagrand, 2005) demonstrate that the proxy for reality falls with about the same frequency in all intervals defined by an ordered set of ensemble members, an indication that ensemble scenarios are equally likely. A trivial but potentially powerful application is the direct feed of individual ensemble members into decision making algorithms. A cost-benefit analysis in the context of the alternative forecast scenarios allows sophisticated users to optimize their weather dependent course of actions (e.g., Alemu et al., 2011; Khan et al., 2021). A wide variety of probabilistic and other products can also be derived from such samples (Vannitsem et al., 2021) just as easily as from statistical samples generated around single value forecasts.

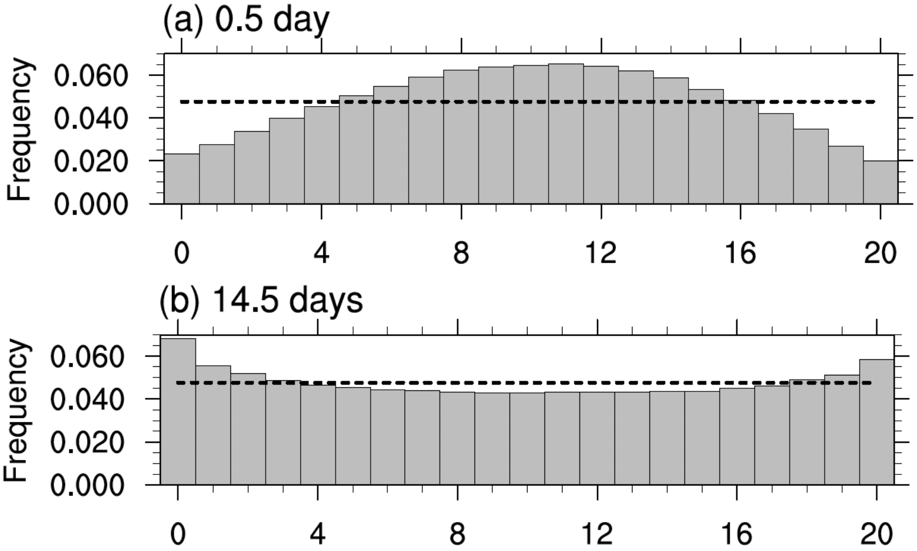

Figure 3. Talagrand (or analysis rank) diagram indicating the frequency of the verifying analysis falling into the intervals defined by the 20 ranked values of 500 hPa geopotential height ensemble member forecasts at individual grid-points, aggregated over the NH extratropics (30º -65ºN) over the 3-month experimental period at 0.5 (a) and 14.5 days lead times (b). A flat distribution (dashed horizontal lines) indicates a perfectly reliable ensemble (where forecast probabilities of events exactly match their observed frequencies).

3.2.2 Error Reduction

Ensembles are well known for the low error in their mean (Eq. 4). As shown in an example from the NCEP ensemble (Appendix A), the error in the mean (red line in Fig. 4) is typically much below that in the control forecast run at the same resolution as the perturbed members (black). This is despite a noticeably higher error in the perturbed forecasts (blue line in Fig. 4). The error reduction in the mean is considered a major benefit of ensembles. A series of studies have suggested that the reduction in forecast error is dynamically conditioned, primarily due to a large projection of initial ensemble perturbations onto the "case-dependent" error in the control analysis (e.g., Toth and Kalnay, 1997, TK97, Ebert, 2001; Wei and Toth, 2003; Buizza et al., 2008; Feng et al., 2019). This presumed effect, often referred to as "nonlinear filtering", is thought to "result in a superior ensemble mean forecast [compared] to a single or even higher-resolution control forecast" (Du, 2007). At the same time, it is maintained that a purely statistical "smoothing effect of [ensemble] averaging partially contributes to this superiority but… in a much less degree… compar[ed] to the nonlinear filtering" (Du, 2007).

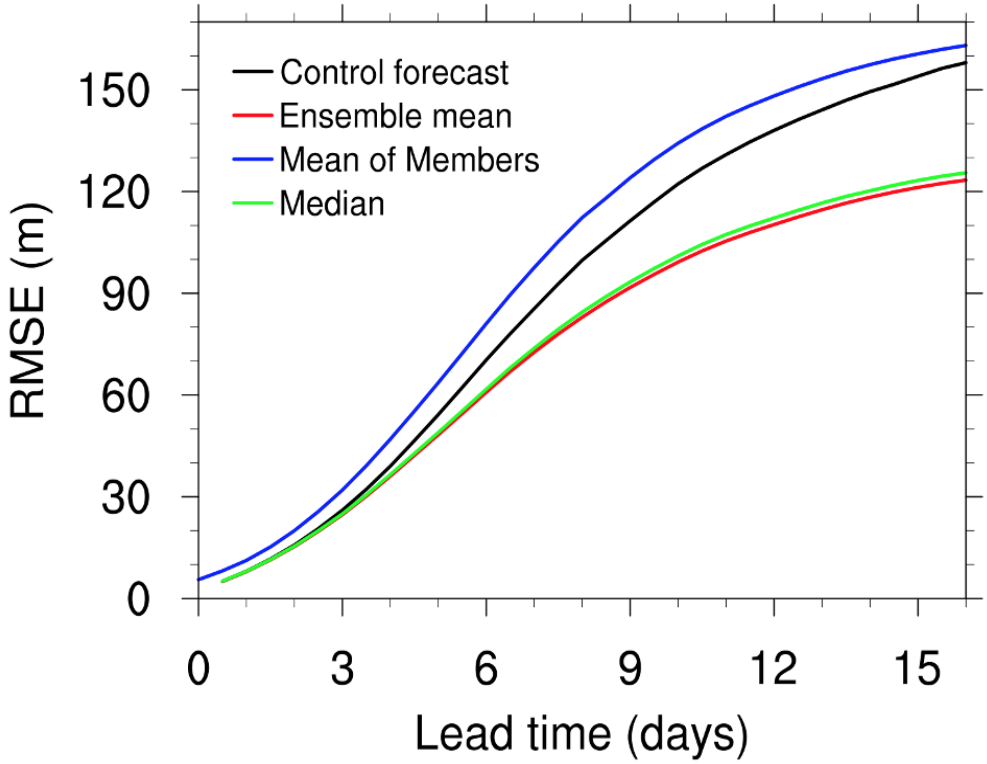

Figure 4. Perceived rms error for the control (black), perturbed (blue), ensemble mean (red), and median (green) NH extratropical 500 hPa height forecasts averaged over the 3-month experimental period.

3.2.3 Spread - Error Relationship

Case-to-case variations in ensemble spread (Eq. 6) are considered an important dynamical indicator of variations in expected forecast error variance (e.g., Murphy, 1988; Buizza, 1997; Goerss, 2000). Many link spatiotemporal variations in spread to fluctuations in atmospheric instabilities, presumably affecting forecast error variance (e.g., Palmer, 2000; Ferranti et al., 2015). For further discussion, see Section 5.4.

3.2.4 Probabilistic Forecasts

A series of related papers (Roulston and Smith 2003, Hagedorn and Smith 2009, Flowerdew et al. 2013, and Christensen et al. 2015) compare verification scores for probabilistic products derived from an ensemble vs. a higher resolution control forecast. Roulston and Smith (2003), for example, find that after applying very similar statistical post-processing methods, 3-10 day lead time ensemble-derived probabilistic forecasts have a much lower Ranked Probability Score (RPS, Murphy, 1969) compared to products derived from a higher resolution unperturbed forecast. Roulston and Smith (2003) and others attribute the favorable score for ensembles to their case-to-case varying distribution that provides “quantitative estimates of the likely forecast accuracy”, concluding that ensemble-based “prediction… is inherently superior to a single “best guess” forecast”.

3.2.5 Bracketing reality

From the beginning, a main objective of ensemble forecasting has been the dynamical sampling of uncertainty in the forecast evolution of the atmosphere. An ensemble brackets or encompasses truth if reality is contained in its range. As is well known, a statistically reliable M_e-member ensemble brackets any single indicator of reality or its proxy in the majority (i.e., (M_e-1)/(M_e+1) fraction) of the cases (Descamps and Talagrand, 2007). As observed for commonly used variables, most of the time the proxy for truth (Appendix A) falls in the range of even somewhat unreliable ensemble forecasts (Fig. 3a). Based on such experience in 1D, the community has assumed that bracketing holds for the multidimensional space of atmospheric dynamics, too. This assumption is reflected in schematics like Fig. 5 (reproduced from Kalnay, 2017), where the evolution of the real atmosphere is contained in, or *dynamically* bracketed by the cloud (i.e., the collection) of ensemble forecast trajectories. This assumption will be evaluated in Section 5.5.

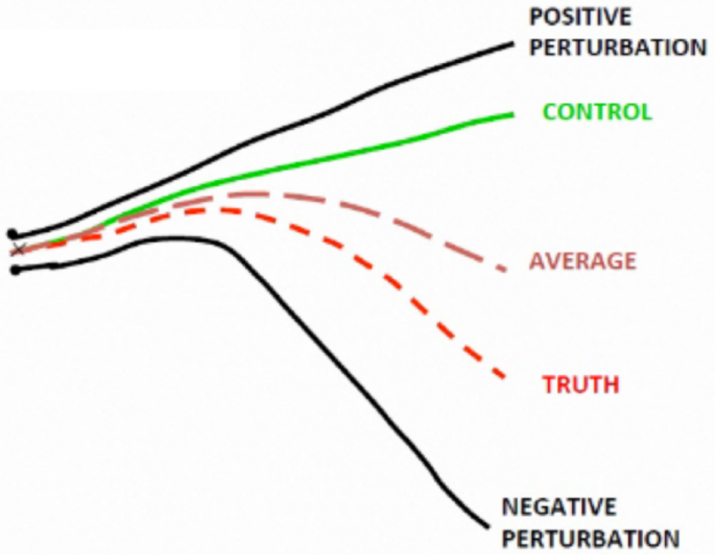

Figure 5. Schematic diagram of ensemble forecast trajectories: the control (green line), perturbed (black), and ensemble mean (long dashed brown) forecasts, and reality (dashed red). Courtesy of E. Kalnay, see text for details.

The introduction of ensembles was partly motivated by the applications, results, and expectations reviewed above. As ensembles proliferated in the weather forecast and user communities, some of the expectations solidified as presumptions. Many of these notions have never been critically examined. Motivated by, and building on the pioneering study of Leith (1974), the rest of this paper revisits some long-held assumptions about ensembles.

4. CONCEPTS AND METHODOLOGY

The assessment of their quality is critical to the optimal use and further development of forecast systems. In this section we review key concepts and tools we consider in the evaluation of ensemble forecasts.

4.1 Forecast Performance Attributes

Reliability. Based on a review of related literature, Toth et al. (2003, 2005) identified two forecast performance attributes: statistical reliability and statistical resolution (e.g., Murphy 1972). Weather forecasts are in the form of abstract “signals”, each of which correspond to a preferably unique weather event or condition in nature. Forecast symbols, as messengers in any communication, are arbitrary. Statistical reliability (e.g., Murphy, 1972) or calibration is one of two main attributes of forecast performance, assessing how truthful the forecast language is to its implied or expressly stated meaning. Specifically, reliability is not concerned about the *sequence* of forecast and observed events, just about their *time average statistics*. For example, is the mean of a sample of forecasts equivalent to the mean of corresponding observations? Naturally, metrics of reliability depend on the form of forecasts (i.e., symbols used, e.g., single value or probabilistic, see Toth et al, 2003). Therefore the reliability of forecasts expressed in different forms is quantitatively not comparable. Statistical reliability is key in the practical use of weather forecasts (Taillardat et al., 2016). Fortunately, just as a text can be corrected for spelling errors without affecting its meaning, forecast bias can be statistically corrected based on past performance (i.e., calibration, e.g., Krzysztofowicz and Kelly, 2000).

Resolution. Weather forecasts attempt to capture the temporal evolution of reality. As such, in contrast to their form, the *sequence of events* foreseen is the *content* of forecasts. Statistical resolution (e.g., Murphy, 1972) concerns how well the dynamical sequence of events in nature is captured by forecast signals indicating such events. In other words, resolution is a system’s ability to foresee the sequence of future weather events, which in a loose sense can also be called the skill of forecast systems. Resolution is independent of the particular form or signals used and is arguably the inherent value, and the most critical attribute of forecast systems. Note that resolution reflects only the similarity in the *sequence* but not in the long-term *statistics* of observed and forecast signals. As reliability is the other way around, the two main attributes of forecast systems are completely independent (Toth et al., 2005).

Importantly, reliability and resolution are the only two attributes of forecast performance based on a comparison of forecasts and observations; other, diagnostic metrics concern only forecasts or observations alone. Therefore, our evaluation will focus on these two attributes since they provide a complete assessment of forecast system performance. Furthermore, as Krzysztofowicz (1992) noted, metrics of resolution are sufficient statistics in a sense that if their output can be calibrated, a forecast system with superior resolution will provide more economic benefit to any user compared to any other forecast system. So in terms of potential benefits, it is enough to compare the statistical resolution of forecast systems.

It follows that reliability and resolution of ensemble forecasts and products derived from them can (and preferably should) be evaluated separately. Most commonly used metrics of forecast performance, however, compound the two attributes with undetermined weights (and possibly include other elements, too). Since our focus is on forecast value, for a comparative evaluation of different forecast systems (such as single value control, and multi-value ensemble forecasts), and for ease of interpretation, we will use a metric of resolution as a primary verification statistic.

4.2 Forecast Information and Noise

As noted above, since they depend on the form of forecasts, reliability scores are quantitatively not comparable across different forecast systems. On the other hand, irrespective of their form, all forecast systems attempt to predict the sequence of future events; they differ only in what signals they use for communicating this. Unlike reliability, resolution therefore can be measured by common metrics, each assessing correlation between forecast and observed anomalies (Krzysztofowicz 1992, Krzysztofowicz and Evans 2008).

Information. In verification, we compare forecast quantities with the observed state, described here with its case-dependent anomaly from the climatic mean[1]. Let us consider forecast anomalies from the climatic mean of a model with realistic variability[2], standardized by the climatic

[1] Differences between points in phase space are independent of the choice of the reference point (or origin) used in defining possible coordinate systems. Here we adopt a convenient and often used representation of atmospheric states through their anomalies from the climatic mean (e.g., Chen and Li, 2021).

[2] We note that forecasts from operational prediction systems have a realistic level of variability, i.e., the overall variance in forecast (**F**–**C**) and verifying proxy for truth (i.e., analysis) anomaly fields (**T**–**C**) are near equal (see, e.g., less than 10% deviation between the solid and dashed black curves in Fig. 8a, introduced later).

variance (Fig. 6). Further, we consider an orthogonal decomposition of forecast anomalies along, and orthogonal to the observed anomaly. Predictive capability or statistical resolution is measured here by the variance of the projection of forecast anomaly onto the observed anomaly, defined with respect to the climatic mean of nature:

$$I_i = |\mathbf{F}_i^{\mathrm{o}} - \mathbf{C}|^2 / |\mathbf{T} - \mathbf{C}|^2, \tag{7}$$

which we call forecast Information (I). $\mathbf{F}_i$ and $\mathbf{T}$ are an i lead time forecast and the corresponding truth or its proxy (e.g., a verifying analysis), respectively, $\mathbf{C}$ is the climatic mean, $\mathbf{F}_i^{o}$ is the orthogonal projection of $\mathbf{F}_i$ on the observed anomaly $\mathbf{T} - \mathbf{C}$, and $|\bullet|$ is the Euclidean norm. In the rest of the manuscript, Information refers to I defined above. Information is the variance of the observed anomaly explained by a forecast, a direct measure of predictive capability. In other words, Information is the anomaly variance shared between reality and a forecast.

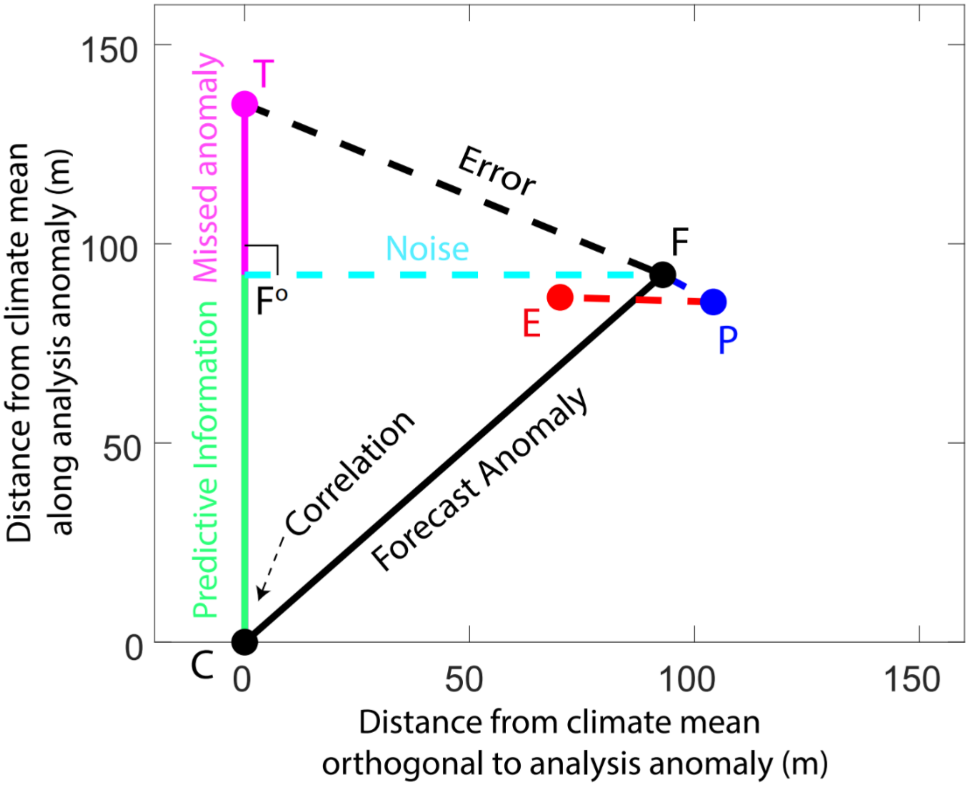
T
Missed anomaly
Error
Noise
F
F°
E
P
Predictive Information
Correlation
Forecast Anomaly
C
150
100
50
0
0
50
100
150
Distance from climate mean along analysis anomaly (m)
Distance from climate mean orthogonal to analysis anomaly (m)

Figure 6. Schematic representing the phase space position of the seasonally and diurnally varying climatic mean (**C**), an unperturbed (**F**), perturbed (**P**), and ensemble mean forecast (**E**), and the corresponding truth or its proxy (**T**) on the information (along the verifying analysis anomaly, vertical axis) and noise (orthogonal to the verifying analysis anomaly, horizontal axis) plane. **P** and **E** are rotated into the **C**-**T**-**F** plane. Key performance metrics used in this study include the rms error (**F**-**T**, or its square, the variance error, black dashed line); variances of forecast Information (**F**°-**C**, solid green) and Noise (**F**- **F**°, dashed cyan); the analysis anomaly missed by the forecast (**T**-**F**°, solid pink); and pattern anomaly correlation (cosine of the angle at **C**). The

position of the points indicates the performance of twice-daily 8-day lead time NCEP GEFS NH extratropical (30º -65ºN) 500 hPa height forecasts averaged over the Dec 1, 2017 - Feb 28, 2018 experimental period. For further details, see text.

Noise. In contrast, next we define variance in a forecast's anomaly that is orthogonal to the observed anomaly as Noise:

$$N_i = |\mathbf{F}_i - \mathbf{F}_i^o|^2 / |\mathbf{T} - \mathbf{C}|^2. \tag{8}$$

Noise is an indicator for the level of divergence between a forecast and reality. Since Information that is identical to, and Noise that is unrelated to the observed anomaly constitute an orthogonal decomposition, for forecast systems with a realistic level of variance they are not independent quantities:

$$I_i + N_i = 1. \tag{9}$$

Information and Noise are therefore positively and negatively oriented, alternative and interchangeable metrics of forecast performance, respectively. Information / Noise variances standardized by the climatic variance range between 1 / 0 (perfect knowledge about nature) and 0 / 1 (no knowledge), respectively. Though related, Information and Noise defined above are different from "Information entropy" (Shannon 1948) or noise used in signal processing (e.g., Tuzlukov 2010, see Appendix B).

Error. Error variance (Eq. 1) is one of the most often used metrics of forecast performance. Error measures the difference between a model forecast and reality. Theoretically, the initially quasi-exponential, then saturating growth of forecast error can be described by a logistic curve (Lorenz, 1984, see Appendix C). As seen from Fig. 6 (dashed black line), error can be decomposed into Noise contained in (dashed cyan line, Eq. 8), and Information missed by a forecast (continuous pink line, Eq. C2).

Information Density. Pattern anomaly correlation (PAC or r_i, Jolliffe and Stephenson, 2003) is another commonly used performance metric, an inverse measure of the angle between forecast and verifying analysis anomalies taken from the climatic mean ($\mathbf{F}_i - \mathbf{C}$ and $\mathbf{T} - \mathbf{C}$, respectively in Fig. 6). The square of PAC is interpreted here as Information Density (I_i^d, see Fig. 8d):

$$I_i^d = r_i^2 = \frac{I_i}{I_i + N_i} \ . \tag{10}$$

Note that for forecasts with the same, and only with the same anomaly variance, Information and Information Density are interchangeable.

4.3 Divergence of Trajectory Segments

At initial time, data assimilation systems capture partial Information about the state of nature, which numerical models then project into the future. Forecast error can be interpreted as the difference between segments of trajectories of dynamical systems. The difference between the evolution of two initially close segments on the trajectory of one, or two similar dynamical systems may be due to a number of factors.

4.3.1 Difference in Model Dynamics

An important difference between the real atmosphere and its numerical models, beyond the latter being an abstract representation of reality, is that models explicitly consider only the larger-scale circulation. Following Leith (1974) and Zhou and Toth (2020), we assume that on larger scales well resolved, numerical models replicate atmospheric dynamics in the extratropics near perfectly. Hence our study uses Northern Hemisphere extratropical 500 hPa height as a primary dataset.

4.3.2 Difference between Equilibria

Though numerical models capture the dynamics of large-scale extratropical circulation well, their equilibrium (i.e., climatic mean) state differs from that of the real atmosphere. In other words, the attractor of numerical models is displaced from that of reality. When a model is initialized with a state close to that observed, it gradually drifts toward the model's own climatology. By definition, this process is governed by the stable dynamics of the model. As climatic drift in Northern Hemisphere extratropical 500 hPa height is negligible, differences between the climatic mean of forecast and reanalysis fields are not considered in the definition of anomalies (Eqs. 7 and 8). Considering also Section 4.3.1, in the rest of this study we disregard the effect of model imperfections on forecast performance.

4.3.3 Off-Trajectory States

Though ideally the best (i.e., the control) and perturbed analyses of the atmosphere should all be in dynamical balance, in reality, both lie off the model trajectory (which approximates the trajectory of the large-scale motions of reality). Analysis fields contain random noise originating from both observational error and assimilation methods, while perturbations reflect intentionally imposed constraints (e.g., Tribbia and Baumhefner, 2004; Molteni et al., 1996; TK97; Houtekamer and Mitchell, 1998). When a numerical model is applied to such imbalanced atmospheric states, over a relatively short (i.e., shorter than two-day) period, the stable part of dynamics pulls the evolving states close to the model trajectory. Once a forecast asymptotes the trajectory, the initial imbalance has no further effect on the divergence of trajectory segments. This is consistent with the findings that initial perturbations alter forecast performance just over a relatively short time period, after which only perturbation amplitude matters (Buizza et al., 2005, Magnusson et al., 2009, Raynaud and Bouttier, 2016). Therefore, imbalances in the evolution of error and perturbations are not considered explicitly in this study.

4.3.4 Difference in the Position on the Trajectory

The divergence in the evolution of two points on the trajectory of a system that are originally close in phase space (but distant in time) is primarily driven by unstable dynamics (Lorenz E. N, 1982; Buizza et al., 1993; Mu et al., 2003; Feng et al., 2018). In the absence of model error, forecast uncertainty and the loss of predictability that ensembles aim to quantify arises due to such divergence. Assuming the variance distance between trajectory segments (i.e., error variance) follows a logistic evolution (see Error under Section 4.2), both the growth of Noise and the loss of Information can be described analytically. As seen in Appendix C, due to the effect of unstable dynamics, with increasing lead time, Information is gradually converted into Noise variance, until all skill is lost. As the effects due to imbalances, or differences in the dynamics and equilibrium of systems are all negligible, error and perturbation behavior studied in this paper are ascribed to the effect of unstable dynamics alone.

4.4 Perfect Model - Perfect Ensemble Setup

Common verification practice also followed in this study involves the evaluation of forecasts such as the 20-member NCEP ensemble used in this study against verifying analysis fields. To eliminate the possible effect of specific data assimilation, modeling, and ensemble generation methods on evaluation results, in this study verification statistics will also be recalculated for 19 remaining members of the NCEP ensemble, replacing the verifying analysis with a randomly chosen ensemble member as truth. Reality and error in this simulated environment are generated by the same techniques as forecasts and perturbations, thus eliminating any influence from imperfect NWP methodologies. Following a long tradition established with the use of the term “perfect model” in observing system and other simulated experiments, we refer to this as a “perfect ensemble” setup. Note that the word “perfect” here does not imply an ultimate or ideal ensemble, but rather, a simulated environment where the ensemble forecast system uses a numerical model and perturbation generation method that are identical to those used in simulating reality and the error in its best estimate.

5 EXPERIMENTAL RESULTS

Though ensembles are of multi-value form, just like a single control forecast, their members cover zero in probability space. Hence probabilistic and related products, just as from a single forecast, must be derived via statistical inter- and extrapolation (Vannitsem et al., 2021). And whether single value- (e.g., van den Dool, 1989; Hamill and Whitaker, 2006; Delle Monache et al., 2013) or ensemble-derived (e.g.,Taillardat et al., 2016), the reliability of probabilistic and other products can only be assessed and enforced by statistical methods, using a sample of past cases (Krzysztofowicz and Kelly, 2000).

Unfortunately, approximations in complex numerical models introduce biases into both single value and ensemble forecasts. Difficulties in the estimation of the magnitude of initial, and in the representation of model related errors also render the spread and distribution of ensemble forecasts unreliable (Vannitsem et al, 2021). Hence in terms of one of the two major forecast performance attributes, statistical reliability, ensembles offer no benefit compared to single value NWP forecasts. Products from both need to be statistically formulated, assessed and calibrated before their use. Next we evaluate what benefits ensembles may bring in terms of the second major forecast performance attribute, statistical resolution or forecast skill, or other unique aspects listed in Sections 3.2.1-3.2.5.

5.1 Forecast Quality

Diagrams like Fig. 3b (Section 3.2.1) attest that members of ensembles offer equally likely scenarios. But are those scenarios also equally likely with the forecast started from the best, unperturbed control analysis[3]? An abundance of evidence indicates that they are not. Assuming perturbations are random draws from the distribution of initial error (Sections 2.1 and 3.1), based on simple statistical considerations the addition of perturbations to the best control analysis

[3] Whether a control analysis (and forecast) is produced by a data assimilation system in practice or not is immaterial. Whether the primary estimate of the state of nature is in single value form around which an ensemble is introduced a posteriori, or an ensemble, the mean of which necessarily has a smaller error (Leith 1974, first full par. in the left column of p. 411), a state with a superior estimate either exists, or can be identified, from which deviations of other estimates can be considered "perturbations".

doubles their error variance compared to the control (Palmer et al., 2006). This is born out in results from operational systems like the NCEP GEFS, where initial and short range perturbed forecast error variance is about double that in the control forecast, negatively affecting performance at all ranges (cf. rms error for the perturbed (blue) and control forecasts (black) in Fig. 4). Moreover, we find that during the first few days, error variance in perturbed forecasts is higher not only in an expected sense, but also for each individual member. Shown in Fig. 7a and 7b is the distribution of error in operational and perfect ensemble (see Section 4.4) perturbed forecasts, standardized separately in each case and for each lead time by the error in the control forecast. Apparently, shifts in phase space location introduced by ensembles induce a degradation in forecast quality similar to that due to spatiotemporal or other shifts made in statistical sample generation.

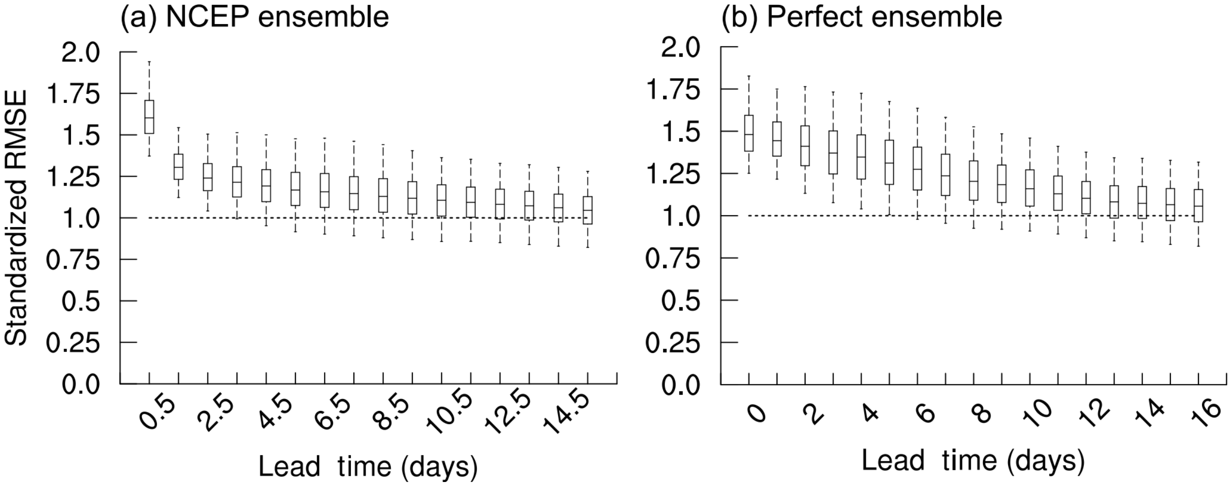

Figure 7. NH (30º -65ºN) 500 hPa height perturbed forecast rms error evaluated against the verifying analysis (a) and a randomly selected member (b), standardized by the error in 0.5 - 15.5 (panel a) and 0 - 16 day (panel b) control forecasts, ranked from lowest to highest, and averaged over all 180 cases. The top and bottom of whiskers and boxes represent the average of the extreme sample point and 25 / 75% quantile values of the 20 and 19 ranked perturbed forecast error values in panels (a) and (b), respectively.

Likewise, perturbed forecasts (blue line in Fig. 8b) have lower Information compared to the control forecast (black line), reflecting an 18-hour loss in skill, equivalent to about an 8-year setback in NWP developments (Zhou and Toth, 2020). Significantly, the mean of the ensemble (red) shows a similar loss of Information[4]. The addition of random initial perturbations, like noise

[4] A degradation in performance was first pointed out by Leith (1974) in the case of an ideal ensemble formed around reality, in comparison with a perfect control forecast.

acquired in signal propagation, reduces Information in all members (not shown). One may argue then that unless other sources of Information are also considered, all products derived from ensemble forecasts will have Information lower than that in the control forecast, which is a key conclusion of this study. This is because new Information about nature cannot be created by taking a function of constituent members all characterized by lower quality. This situation is exemplified by the lower level of Information in the median of the ensemble (green curve in Fig. 8b). We recall that Information is a measure of statistical resolution, or the inherent value in forecasts. Since Information is a sufficient statistic (Section 4.1), the results here indicate that any user may derive more benefit from a control forecast than from an ensemble. For optimal decision making, one must use the control forecast, possibly with an added, statistically derived estimate of uncertainty. Is there some other value present in ensembles that may be missed by either of the two main forecast performance attributes, reliability or resolution?

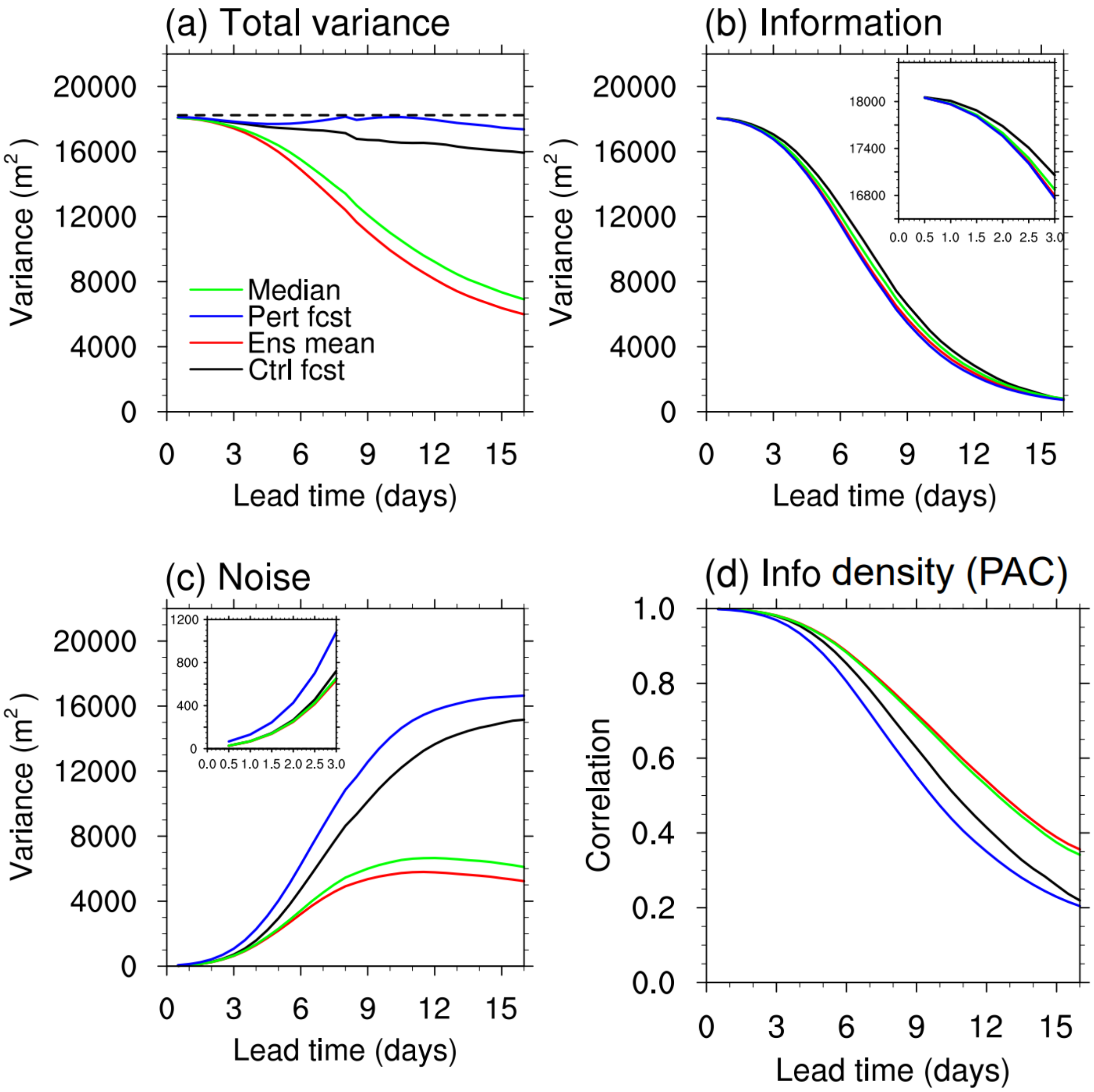

Figure. 8. Sample mean non-standardized (a) total variance, (b) Information variance, (c) Noise variance, and (d) Information Density (or pattern anomaly correlation) of 500-hPa geopotential height forecasts in the NH extratropics (30º - 65ºN) over the 90-day experimental period (Appendix A). The dashed line in panel (a) indicates the climatic variance present in the analysis.

5.2 Error Reduction

The lower error in the mean of an ensemble (cf. red and black lines in Fig. 4) suggests yes, ensembles may have other benefits (Section 3.2.2). But how do we reconcile the reduction of error in the mean (Fig. 4), a negatively oriented performance metric, with a concurrent decrease in Information (Fig. 8b), a positively oriented metric? According to Eq. C3 (Appendix C), error can be reduced either by increasing Information, or decreasing Noise. As revealed by Fig. 8c, the moderate reduction in Information is more than compensated with the large reduction of Noise in the mean. This is also apparent in the evaluation of 8-day forecasts in Fig. 6. Apparently, the mean of an ensemble is a very efficient Noise filter. This is also reflected in the well-known, much smoother character of the mean as compared to single value forecasts (Ancell 2013), which is reflected in a significant reduction of overall variance in the mean (Fig. 8a). Therefore, contrary to commonly-held expectations (Section 3.2.2), error in the mean is reduced not because of a gain, but despite a loss of forecast Information, due to an effective reduction of unpredictable Noise.

5.3 Probabilistic Forecasts

Here we revisit the reason behind the lower error metrics found for ensemble- vs. control-based probabilistic forecasts (Section 3.2.4). It turns out that unlike assumed by many, commonly used probabilistic scores like Continuous Ranked Probability Score (CRPS, and its categorical equivalent, RPS) are not affected by variations in the shape or spread of forecast distributions (Hersbach 2000). They depend only on the average of the spread of forecast distributions over the verification period. If not "case-dependent" variations in the shape of distributions, as suggested in the literature, then what explains the lower RPS error for ensemble-derived probabilistic forecasts? As Hersbach (2000) points out, CRPS (and hence RPS) is analogous to mean absolute error (MAE, which itself is closely related to error defined by Eq. C1). The significantly lower RPS and other scores reported in Roulston and Smith (2003) and other studies for probabilistic forecasts derived from an ensemble vs a single control forecast is then a result of, just as in case of the error in the mean (Section 5.2), the reduced level of Noise in the position of ensemble distributions (i.e., their median) as compared to single value forecasts (cf. green and black curves in Fig. 8c).

5.4 Spread - Error Relationship

An indication of the magnitude of forecast error (or error variance, Eq. C1) by spatiotemporal fluctuations in ensemble standard deviation (or spread, Eq. 6) is another perceived benefit of ensembles (Section 3.2.3). The correlation between the two quantities, however, is rather low, explaining only about 10% in the day-to-day variability of the error magnitude (see, e.g., Fig. 5 of Hopson 2014). Perhaps not surprisingly, we found no anecdotal or documented evidence for the practical use of this relationship. As we saw in Section 5.3, fluctuations in spread certainly do not enhance forecast Information. What may then explain the correlation between spread and error? By-and-large, the realizations of the atmosphere follow a multinormal distribution (Toth, 1995). In such a space, distances between states, just as in a univariate normal distribution, depend on a state's anomaly from the climatic mean (Li et al., 2018). As forecast error measures the distance between trajectory segments of dynamical systems in multidimensional space (Section 4.3), it must also depend on the anomalies of the forecast and observed states. Evidence of this relationship for different forecast systems is presented by Toth (1991a, 1991b) and Kleeman (2011). We hypothesize that the weak relationship between spread and error may at least partially be explained by the dependence of both quantities on the climatic anomaly of the control forecast.

5.5 Bracketing in Multidimensional Space

In one dimension, all perturbed states necessarily lie in the direction defined by reality and its forecast. The concept of bracketing, or encompassing truth is straightforward: reality must fall within the range of perturbed states (Section 3.2.5). Assuming a well-behaved unimodal distribution, this is possible only if some perturbed members have an error lower than the unperturbed estimate of reality, which is what we observe for all variables in today's ensembles. Does bracketing in any selected single direction guarantee bracketing in the multidimensional space of dynamics?

First we generalize the intuitive concept of bracketing into multidimensional spaces like that occupied by the dynamics of the atmosphere. There, just as in 1D, bracketing is considered satisfied if reality falls in the range of perturbed states in the direction defined by reality and its forecast. This is the case dependent direction of error in the unperturbed control, out of many independent degrees of freedom. Reduced error hence is still a necessary (but not sufficient) condition for bracketing in multiple dimensions. For bracketing to work in this space, perturbations must have a strong projection on, or be congruent with the case specific direction of error in the control. Bracketing case specific error patterns in a multidimensional space is a much harder challenge than bracketing single 1D variables.

Experimental results in Fig. 7 show that even for a subset of the atmosphere (500 hPa height variable over the NH extratropics) the necessary condition for bracketing of reduced error in the perturbed states is violated. Until day 3.5 and 5 days, all members of the operational and perfect ensembles, respectively, have an error larger than that in the control forecast. An alternative interpretation of Fig. 7 is that the time evolution of reality (or its proxy), the control forecast, and the range of perturbed forecasts are shown by the Y=0 line, the Y=1 line, and the boxplots, respectively. Fig. 7 thus can be considered as a factual alternative to popular schematics like Fig. 5 circulating in the community about ensemble forecasting. Clearly, in the case-dependent direction of error in the control forecast (and also in the control analysis in Fig.

7b), reality or its proxy is far removed from the range of initial and short-range ensemble members. The widely-held assumption that the evolution of the atmosphere is contained in dynamically generated ensembles (Section 3.2.5) is untenable.

6. THEORETICAL CONSIDERATIONS

6.1 Simulation

In search of an explanation for the universal loss of skill, and the failure of dynamical bracketing demonstrated in Fig. 7, we now turn our attention to the nature of high dimensional spaces. For a quantitative assessment, we hypothesize that (i) unstable atmospheric dynamics responsible for the divergence of forecast and observed trajectory segments (cf. Section 4.3.4), as suggested by Toth (1991b, 1993) and Palmer et al. (2006), evolve in a multinormal space with a large number of independent and identically distributed (iid) variables[5] (M_d), and that (ii) error and ensemble perturbations, after a short period of transitionary behavior (see Section 4.3.3) are random draws from this domain we call the subspace of possible error. If these assumptions are valid, some basic features of forecast error and ensemble perturbation behavior should be statistically reproducible.

Our aim here is to compare the error in the initial unperturbed and perturbed states. While this can be accomplished for the perfect ensemble described in Section 4.4, we will use 12-hour forecasts instead as an indicator for error in the operational system. Plotted in Fig. 9 are 12-hour lead time operational (20 dots, panel a) and perfect initial ensemble members (19 dots, panel b) for 180 cases along with the proxy for reality (vertical bar), as a function of distance from the control 12-hour forecast (panel a) or control analysis (panel b, both plotted at point 0,0) in the direction of error in the control (X axis, directional distance) and in the subspace orthogonal to it (Y axis, absolute distance in the null-space of error in the control), on a scale standardized by the sample mean error in the control. Note that the distance of the bars and points from the control point (0,0) measures the size of error in, and perturbation around the control.

To validate the hypotheses above, we proceed with the generation of 20 random points from a distribution with a varying number of iid standardized normal variables (i.e., dof). Just like

[5] Whether an orthogonal basis describing such a space can be determined in practice or not is irrelevant for our study; we are concerned only about the number of independent normal iid variates (dof) of this space.

in the perfect ensemble experiment reported in Fig. 9b, one randomly chosen point is considered reality, while the remaining 19 the perturbed states. And just as is the case with the perfect ensemble, the simulation experiment is repeated 180 times. We find that the distribution of the error from the perfect and simulated ensembles are statistically indistinguishable at the 5% significance level for samples with a dof in the range of 28-38, with dof=33 yielding the best fit (Appendix D), for which the results are plotted in Fig. 9c.

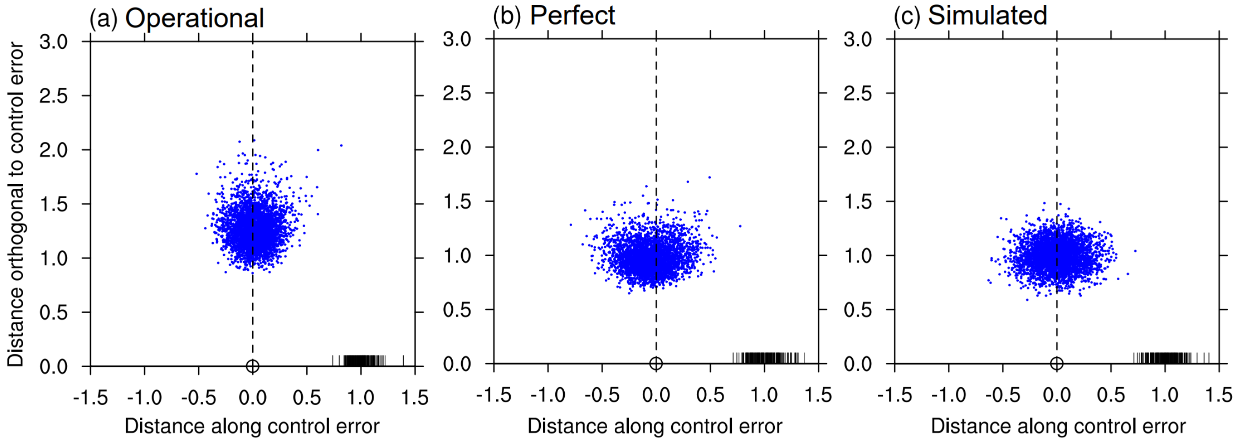

Figure. 9. Perturbed NH 500 hPa height (a) operational 12-hr forecasts, and (b) perfect initial ensemble members (3,600 and 3420 individual blue dots from 20 and 19 members from each of 180 cases for panels a and b, respectively), plotted along (horizontal axis) and orthogonal (vertical axis) to the error in the unperturbed control 12-hour forecast (panel a, open circle at 0,0) or control analysis field (panel b) on a scale standardized by the sample mean error in the control, and the corresponding proxy for truth (black bars). Panel (c) is a statistical simulation of panel (b) with a 33 dof standardized multi-normal distribution. For further details, see text.

Notable on all panels in Fig. 9 is the small projection of perturbations introduced around the control forecast or analysis (0,0) onto the realization of error in the control (i.e., absolute value of X of perturbed points). This is in contrast with the magnitude of perturbations in the null-space of error (i.e., Y value of perturbed points), which is comparable to the magnitude of error (i.e., distance of black bars from the control at 0,0). Consequently, error variance for most members is almost doubled compared to the control (cf. the distance between reality and the control vs. the perturbed states, consistent with rms error at 12-hour lead time in Fig. 4). As error in all members is increased compared to the control, their cloud forms further away from reality. Consistent with Fig. 7, reality or its proxy is not encompassed by either the operational or perfect ensembles. In all cases, the simulated ensemble also fails to bracket truth. Unlike in 1D, statistical reliability (i.e., perturbation variance matching error variance) apparently does not imply bracketing in multidimensional space.

The remarkable visual and statistical similarity of the simulated (Fig. 9c) to the perfect ensemble (Fig. 9b), and a lesser, but still strong similarity to the operational ensemble data (Fig. 9a) indicate that the experimental results are consistent with the hypotheses that (i) perturbation and error dynamics captured by NWP analyses and forecasts evolve in a multinormal space with a large number of iid variables, and that (ii) ensemble perturbations and error are indeed random samples from such a space. The space of resolved-scale error and perturbation dynamics is contingent on Information captured in an analysis or forecast. The similarity of panels *a* and *b* in both Figs. 7 and 9 also indicates that the problematic behavior observed in the operational ensemble, including their low skill and failure in bracketing cannot be addressed by perfecting data assimilation, modeling, or perturbation methodologies used.

6.2 Interpretation

For an interpretation of error and perturbation results in Fig. 9, we consider the orthogonal decomposition of anomaly variance into Information and Noise (Section 4.2). Depending on available observations and data assimilation techniques, NWP analyses and forecasts capture a certain amount of Information about the evolution of the larger scale condition of the atmosphere, often considered *deterministic*. According to our hypotheses (Section 6.1), two states of the atmosphere given at the resolution of today's operational systems can differ in M_d independent ways, which for the NH extratropical height is estimated at 33. Given *stochastic* observational and methodological noise, error in any analysis is then just one random realization from this finer scale "subspace of possible error". And perturbations which we assume are random draws from the same space simulate alternative realizations of analysis error that could have happened under different realizations of stochastic observational and methodological errors. Importantly, both Information about nature, and Noise contaminating a forecast are carried forward by the same model dynamics, albeit at different scales, used in numerical models.

In 1D, reality and its best and perturbed estimates all occupy a single, common direction. Statistically reliable perturbations along this single direction bracket reality (Fig. 1a). In 1D space, statistical reliability is analogous with bracketing (Fig. 3a). The dynamical evolution of the atmosphere, on the other hand, manifests in high dimensional space. As the independent degrees of freedom (dof, M_d) increases, random draws from such a space spread out across more directions, lowering their expected projection on any single direction to $1/M_d$, including that of the error in unperturbed (control) estimates. Such behavior is often referred to as the "curse of dimensionality" (e.g., Bellman, 1961), which appears to be the fundamental cause of the failure of any sample, whether statistically or dynamically generated, in matching the level of Information in unperturbed estimates, or encompassing reality. As the bulk of perturbation variance projects into the null-space of control error, error in each perturbed member is necessarily increased, failing to meet a necessary condition for bracketing.

A contributing factor to the loss of Information in, and the lack of bracketing by the perturbed members is the reduction of variability in the magnitude of both error and perturbations, which results in an even sharper separation of reality and its samples. Fluctuations in the magnitude of error (about 0.12 standardized units along X of the black bars in Figs. 9b and 9c) and perturbations (0.12 along Y and 0.18 along X of the blue points) are greatly reduced compared to the standard deviation of 1 in 1D. This behavior is due to the Law of Large Numbers (e.g., Rose and Smith, 2002). If error in the best estimate (or control analysis) of a state is assumed to follow an iid normal distribution then theoretically, the distance of such guesses from reality follows a chi-square distribution (black curve in Fig. 1b), and the distance of perturbed states around any analysis from reality a non-central chi-square distribution (blue curve in Fig. 1b). The higher the dof, the narrower both of these distributions become. The demonstration in Fig. 1b is for M_d=150 dof. Unlike in 1D (Fig. 1a), all perturbed states in high dimensional spaces are further displaced from reality. Fig. 1b hence indicates that the failure of operational, perfect, and simulated ensemble members to match the skill of unperturbed estimates, or to bracket reality is due to the peculiar geometry of high dimensional spaces.

Finally, we contrast the time evolution of perturbations that are aligned, or congruent with, vs. orthogonal to the error in the control. Initial perturbations congruent with the control error uniformly reduce or increase error in the perturbed state over the entire domain. Resulting perturbed states lie on a line defined by reality and the control initial condition. If error in each initial condition is assumed to grow logistically, the relative differences between smaller and larger initial errors will be retained in the forecast phase. Consequently, trajectories started with initial perturbations congruent with the control error will remain dynamically congruent in the forecast phase. Such forecast trajectories necessarily lie on a 2D surface defined by the trajectories of reality and the control forecast, ever diverging from, and never crossing each other (as suggested in Fig. 5).

Such orderly error behavior is never observed with real life ensembles. On the opposite, error curves for perturbed members evaluated over any subdomain display an incongruent, crisscrossing nature. This is evident in Fig. 10, where the member that is best / worst over the NH extratropics in the 12-24 hour range (solid blue / red lines), for example, performs the worst / best a few days later (60-120 hour lead time range), respectively (or at other locales, not shown). This behavior can be explained by the random nature of initial perturbations in a high dimensional space. With negligible projection on the actual error in the control, such perturbations improve / degrade the control initial condition in a random fashion over different parts of the domain. Model dynamics transposes the random initial spatial variations in skill into the time domain. The random fluctuations in forecast skill seen in Fig. 10 hence arise as the influence of improvements and degradations in initial condition from different parts of the domain reach the verification area. Clearly, ensemble perturbations behave like random (albeit spatiotemorally correlated) noise. And paradoxically, these random fluctuations provide statistical bracketing for single observed variables (Fig. 3), while fail to dynamically bracket the full state or its evolution (Figs. 7 and 9).

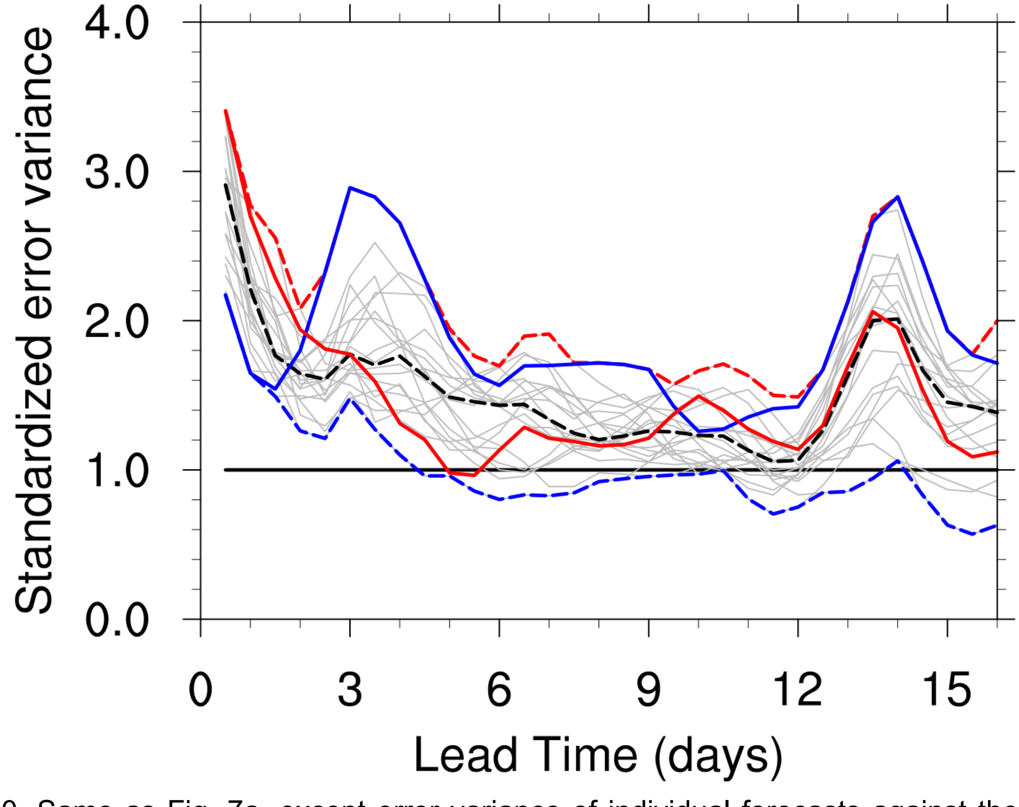

Figure 10. Same as Fig. 7a, except error variance of individual forecasts against the verifying analysis for the single case initialized at 12 UTC, 30 Dec 2017. The three dashed curves represent the error in the best (bottom, blue), median (middle, black), and worst member (top, red) at each lead time separately. The blue and red solid curves show the error variance in the members best and worst at the 12-hr lead time, respectively. Light gray curves show the error variance of individual members.

6.3 Nonlinear Effects

To avoid the introduction of sampling error, initial perturbations are symmetrically arranged around the control analysis (Eq. 5), setting the mean of operational ensembles equal to the control analysis. What explains the moderate and large reduction of Information and Noise in the ensemble mean compared to the control forecast, respectively? As noted by Gilmour et al. (2001), perturbations with amplitudes small relative to their saturation value develop quasi-linearly, leaving the mean mostly unaffected. This is reflected in the overlap of the black (control) and red (ensemble mean) curves in Fig. 8b (Information), and especially in Fig. 8c (Noise) in the 0-1 day lead time range).

Noticeable nonlinearities first emerge on the smallest scales due to the asymmetric evolution of the amplitude and position of affected features (Ancell 2013). This results in a deviation of the mean from the control forecast. As nonlinear mixing in the 1-2 days lead time range is low, Noise removal is minimal; the difference between the mean and the control forecasts is dominated by the loss of Information. This is evidenced by the noticeably larger difference between the black (control) and red (ensemble mean) curves for forecast Information (Fig. 8b) n, compared to Noise Fig. 8c).

With increasing lead time, the phase and amplitude of perturbations on the smallest scales become fully randomized. At this stage of full nonlinear mixing, the mean of a typically sized ensemble removes a significant part of Noise present in the control forecast on scales with such fully saturated perturbation amplitudes. Simultaneously, error first in the control, then in the perturbed members also saturates, at which point all forecast Information on these small scales is lost. Due to the upscale propagation of energy, the same perturbation dynamics is repeated on successively larger scales. On scales with newly randomized perturbations, Information in the mean compared to the control forecast is temporarily reduced, after which a large part of Noise on such scales is removed. This succession of temporary reduction of Information and the additive removal of Noise on ever larger scales explains the increasing - steady - decreasing reduction of Information in the mean (compare black and red curves in Fig. 8b), and the cumulative removal of Noise compared to the control forecast (compare black and red curves in Fig. 8c), as a function of increasing lead time.

As perturbation energy moves upscale, the growth, as well as the overall variance of perturbations shifts to ever larger scales (cf. Fig. 1 of Prive and Errico, 2015). This results in a general reduction of the independent degrees of freedom in perturbation dynamics. Which explains the increase in the likelihood that the skill in some members over a limited domain rises above that of the control forecast, as noted earlier in Fig. 7 for longer leads times. Ensembles, however, fail to bracket reality or its proxy at initial and short lead times even over a subdomain of the 500 hPa height over the extratropical NH (with an estimated dof of 33). So bracketing observed at later lead times is only statistical, not dynamical in nature.

The dof of the full dynamics of short-range perturbations resolved by today's NWP systems is estimated to be in the range of 150-200 (see Appendix D). Bracketing in that space, as demonstrated for dof=150 in Fig. 1b, is even more challenging. Could the addition of more members help? The ratio of bracketing, more formally defined in Appendix E is a function of dof and ensemble membership. In 1D, the bracketing ratio with typical membership is sufficiently close to 1 ((M_e-1)/(M_e+1)), ensuring that most of the time statistically reliable samples encompass a proxy of reality. In the high dimensional space of the full resolved-scale dynamics of atmospheric circulation the chance of even large size randomly generated ensembles encompassing reality, however, is astronomically low (see Fig. E1).

7 CONCLUSIONS

We exploit an orthogonal decomposition of forecast anomaly from the climatic mean into Information identical, and Noise orthogonal to the observed anomaly. Generally, Information about the state of natural systems is limited. Information is further reduced as forecast variance in chaotic systems like the atmosphere is gradually converted into Noise (Fig. 6). For decades, statistical sampling has been successfully applied to assess uncertainty in weather forecasts (Fig. 2a). Could forecast samples be generated dynamically, asked forerunners of ensemble forecasting. The practice of ensemble forecasting matured in the 1990s. Initial perturbations are added to the best estimate of the state, from which alternative scenarios are dynamically projected into the future (Fig. 2b). After statistical calibration, probabilistic and other products derived from ensembles are widely used today, with demonstrated value.

Ensembles are assumed to (i) encompass the evolution of the real atmosphere (Fig. 5), (ii) capture case-dependent variations in forecast error, and (iii) provide higher quality single value (ensemble mean, Fig. 4) and (iv) probabilistic guidance. With a combination of theoretical and experimental approaches, these assumptions have been revisited. Using a statistical analysis, first we found that the divergence of segments of observed and/or forecast trajectory segments, and hence error and perturbation dynamics reside in a high dimensional (150-200 independent degrees of freedom, Appendix D) domain we call the subspace of possible error. This subspace is contingent on the larger scale condition of the deterministically evolving atmosphere, which one may associate with a “case”. Theoretically, sample points from high dimensional spaces have negligible projection in any preselected direction, including the error in any initial state. Consequently, unlike in 1D (Fig. 1a), sample points in high dimensions consistently degrade the quality of the best estimate, and also miss to encompass reality (Fig. 1b).

Information captured by an analysis is determined by the sophistication of the observing, data assimilation, and modeling systems. Experimental results suggest that error and perturbations are random draws from the high-dimensional subspace of possible error (Fig. D1). Error in initial conditions results from specific realizations of stochastic noise in observations and data assimilation procedures, while perturbations represent alternative realizations of possible error that may have realized under different configurations of stochastic noise. As in real time Information and Noise are inseparable, numerical forecasts project their sum, the total initial variance into the future. What value may the dynamical generation of forecast samples (i.e., ensembles) via the deterministic projection of alternative Noise realizations may bring?

An analysis of an operational and a perfect ensemble reveals that as theoretically expected, but contrary to assumption (i) above, initial perturbations and ensemble forecasts do not contain the state and evolution of the atmosphere (Figs. 9 and 7, respectively). Also as expected, out to medium range, all members of the operational and perfect ensembles have larger error and less Information than that in the unperturbed control forecast. Unlike in 1D (Fig. 1a), ensemble members do not provide any scenario that is closer to reality than the control; ironically, they explore instead different ways that the control can be degraded (Fig. 1b). And importantly, contrary to assumption (iii), the mean and arguably (iv) all probabilistic and other products derived from ensembles have less Information than that in the control forecast (Fig. 8b).

Incidentally, an analysis by Hersbach (2000) shows that variations in the distribution of ensembles do not even have an effect on commonly used verification metrics. While other studies, as an alternative to assumption (ii), suggest that the low-level correlation found between case-to-case variations in spread and error may be explained by each being influenced by the amplitude of forecast anomalies. In any case, how could random draws from the subspace of possible error have any predictive information about the specific realization of error that is driven by stochastic observational and data assimilation processes? After all, it is only the subspace of possible error from which perturbations are also drawn, but not any specific, stochastically driven realization from it that is “case” dependent (on the well-known large-scale conditions).

Our diagnosis indicates that the smaller error in the mean, a well-known benefit of ensembles, is due to an efficient filtering of Noise (Fig. 8c) compared to individual forecasts. The smoother nature of the median of ensemble distributions is also what explains the lower scores found in probabilistic forecasts derived from an ensemble vs. a control. Unfortunately, nonlinear filtering removes not only Noise, but some forecast Information as well (Fig. 8b). Interestingly, Information is preserved in the mean only during the early, linear phase of the evolution of perturbations where their initial symmetry is still preserved and where ensembles are generally considered useless. Later, the loss of Information in the mean and other products amounts to an about 18-hour loss of lead time in warning about future weather events, or an 8-year setback in international NWP developments. The significance of this is that since Information is a sufficient verification statistic, any rationally acting user benefits more from an unperturbed control than from an ensemble of forecasts.

Importantly, all behavior observed in operational ensembles is reproduced with a perfect ensemble. This confirms that their failure to meet expectations is not due to methodological shortcomings but lies rather in the multidimensional and nonlinear nature of atmospheric dynamics. At a great computational expense, ensembles recreate the same Information present in the control forecast M_e times, albeit at a lower level, while with painstaking accuracy generate M_e alternative realizations of dynamically balanced error of a somewhat larger magnitude. Ensembles lack statistical reliability or any discernible benefit from case-dependent variations, and have demonstrably less Information. Should the use of statistical alternatives be reconsidered? Filtering applications may reduce Noise in the best estimate while preventing or flexibly controlling the loss of forecast Information. With developing machine learning applications like recent data-driven weather modeling (Bi et al., 2023; Chen et al., 2023), spatiotemporal and cross-variable covariances may also be induced into statistically generated perturbations. All the while calibrated probabilistic and other products of interest can be derived from statistical samples of error in past control forecasts, instead of dynamically generated ensembles.

8 DISCUSSION

Applicability. Though real-life results in this work are presented only with a single configuration, the NCEP GEFS, they arise out of general system characteristics. Specifically, (a) the phase space of all complex systems is high dimensional, in which (b) nonlinear saturation randomizes perturbations on increasingly larger scales, reducing Information in the entire distribution. Therefore, the main conclusions of this study may in general be applicable to numerically created ensembles of high dimensional multiscale dynamical systems. As an example, finer scale processes resolved by increased resolution models are accompanied by higher degrees of freedom where compared to synoptic scales, saturation of error happens faster, resulting in an even earlier onset of nonlinear perturbation behavior compared to what is found with the NCEP ensemble.

Continuous Approach. Whether forecast samples are represented in a quantized form of a finite sample (i.e., ensembles), or by a continuous function (e.g., the Liouville Equations – Ehrendorfer, 2006), the underlying problems highlighted above remain the same. The loss of Information and the lack of bracketing therefore may equally affect continuous or quantized dynamical estimates of forecast uncertainty. In light of the availability of viable statistical alternatives, Leith's (1974) early assessment about ensembles may be applicable to continuous dynamical approaches as well: "sample sizes M_e>1 will have to be justified on the basis of the detailed knowledge obtained…".

Stochastic Perturbations. Traditionally, the effect of finer scale processes on motions explicitly resolved in numerical models is parameterized deterministically, conditioned on the resolved scales. More recently, with the intent of producing stochastic perturbations, random processes are inserted into some parameterization schemes. When such perturbations are added to forecast states during model integration, ensembles may become more reliable (e.g., Buizza et al., 1999; Berner et al., 2009), with reduced error in their mean (e.g., Sardeshmukh et al. 2023). Just like initial perturbations, these random perturbations, however, also increase forecast Noise in individual members, and reduce Information both in the members and their mean. From a forecast Information perspective, one might consider stochastic Noise as adding insult to injury sustained from the introduction of initial perturbations first.

Data Assimilation. Ensemble-derived products used in data assimilation describe covariances in the behavior of short-range forecast error. These products are based on ensemble forecasts issued at an earlier time and valid at the time of the analysis. As such, covariances have no forecast Information about the future state of the atmosphere; rather, they help find the best estimate of reality, given available observational data. Key limitations of ensembles such as the loss of Information or the lack of bracketing therefore do not affect data assimilation applications. Interestingly, the high dimensionality of the space of error discussed in Sections 5.5 and 6 has long been recognized in the context of ensemble-based data assimilation (e.g., “localization” algorithm of Szunyogh et al, 2008).

New Elements. The degraded performance of all short-range perturbed forecasts compared to the control forecast evaluated over large areas is not a new finding (Palmer et al. 2006). Important implications such as the loss of Information in all derived products, and the failure of dynamically generated ensemble forecasts to encompass reality, however, have not been previously recognized. Neither have the lower error in the mean or in probabilistic forecasts derived from ensembles been attributed exclusively to Noise filtering, nor has the random nature of ensemble perturbations in the high-dimensional subspace of possible error, or the significance of the stochastic nature of error been recognized. Correspondingly, some basic characteristics of ensembles and the potential viability of alternative statistical sampling methods have for many remained elusive.

Acknowledgements. We sincerely thank Drs. Jian-Wen Bao and Jeff Beck of the Physical Sciences Laboratory and the Global Systems Laboratory (GSL) of NOAA, respectively for their valuable comments on earlier versions of this report. The constructive comments of two anonymous Reviewers greatly improved the presentation of the material. The first author (ZT) is grateful for many invigorating discussions he had over the years with members of the ensemble forecast and user communities. Discussions with Drs. Feifan Zhou (Institute of Atmospheric Physics), Duane Rosenberg (GSL), and the encouragement and continued support of Jennifer Mahoney and Kevin Kelleher, current and former Directors, and Curtis Alexander and DaNa Carlis, current and former Deputy Directors of GSL are also gratefully acknowledged. We also thank Dr. Jun Du of the Environmental Modeling Center, NCEP for his help with some of the references. The second author was partially supported by the National Natural Science Foundation of China (Grant No. 42288101 and No. 42105054).

Data Availability. Numerical analysis and forecast data used in this study were provided by Dr. Xiaqiong Zhou at the Environmental Modeling Center (EMC) of NCEP. The standard deviation and climate mean datasets used in some calculations were kindly provided by Drs. Bo Yang and Suranjana Saha (EMC/NCEP). The data can be downloaded from NCEP’s operational products inventory under GFS Ensemble Forecast System (GEFS) at https://www.nco.ncep.noaa.gov/pmb/products/gens/.

References

Alemu, E. T., R. N. Palmer, A. Polebitski, and B. Meaker, 2011: Decision support system for optimizing reservoir operations using ensemble streamflow predictions. J. Water Resour. Plann. Manage., 137(1), 72–82.

Ancell, B. C., 2013: Nonlinear Characteristics of Ensemble Perturbation Evolution and Their Application to Forecasting High-Impact Events. Weather and Forecasting, 28, 6, 1353-1365, available from: < https://doi.org/10.1175/WAF-D-12-00090.1>

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. Journal of Climate, 9(7), 1518–1530.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. Nonlinear Processes in Geophysics, European Geosciences Union (EGU), 8 (6), pp.401- 417, hal-00331059.

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

Bellman, R. E., 1961: Adaptive Control Processes. Princeton University Press, Princeton, NJ.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. Nature, https://doi.org/10.1038/s41586-023-06185-3.

Bougeault, P., and coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. Bull. Amer. Meteor. Soc., 91, 1059–1072, doi:10.1175/2010BAMS2853.1.

Buizza, R., J. Tribbia, F. Molteni, and T. Palmer, 1993: Computation of optimal unstable structures for a numerical weather prediction model. Tellus, 45A, 388–407.

Buizza, R. and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. J. Atmos. Sci., 52, 1434–1456.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Mon. Wea. Rev., 125, 99–119

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. Monthly Weather Review, 126(9), 2503–2518.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q.J.R. Meteorol. Soc., 125: 2887-2908. https://doi.org/10.1002/qj.49712556006

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Z, Wei, and Y. J. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon. Wea. Rev.,133, 1076–1097.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. Q. J. R. Meteorol. Soc., 134, 2051-2066. doi:10.1002/qj.346

Buizza, R., & M. Leutbecher, 2015: The forecast skill horizon. Quarterly Journal of the Royal Meteorological Society, 141(693), 3366–3382. https://doi.org/10.1002/qj.2619

Calanca, P., D. Bolius, A. P. Weigel, and M. A. Liniger, 2011: Application of long-range weather forecasts to agricultural decision problems in Europe, J. Agric. Sci., 149, 15–22.

Candille, G. and Talagrand, O., 2005: Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society, 131(609), 2131–2150.

Charney, J., 1949: On a physical basis for numerical prediction of large-scale motions in the atmosphere. J. Meteor. 6, 371-385.

Charney, J. G., R. Fjoertoft, and J. V. Neumann, 1950: Numerical integration of the barotropic vorticity equation. Tellus 2, 237–254.

Chasalow, K.E., and K. E. Levy, 2021: Representativeness in Statistics, Politics, and Machine Learning. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Chen, J. and X. Li, 2020: The Review of 10 Years Development of the GRAPES Global/Regional Ensemble Prediction. Advances in Meteorological Science and Technology, 10(2), 9-29, DOI：10.3969/j.issn.2095-1973.2020.02.003.

Chen, L., X. Zhong, F. Zhang, and coauthors, 2023: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. npj Clim Atmos Sci, 6, 190. https://doi.org/10.1038/s41612-023-00512-1

Chen, T.-C., S. G. Penny, J. S. Whitaker, S. Frolov, R. Pincus, and S. Tulich, 2022: Correcting systematic and state-dependent errors in the NOAA FV3-GFS using neural networks. Journal of Advances in Modeling Earth Systems, 14. https://doi.org/10.1029/2022MS003309

Chen, X., and T. Li, 2021: An improved method for defining short-term climate anomalies. J. Meteor. Res., 35(6), 1012–1022, doi: 10.1007/s13351-021-1139-2.

Christensen, H. M., 2015: Decomposition of a new proper score for verification of ensemble forecasts. Monthly Weather Review, 143(5), 1517–1532. https://doi.org/10.1175/MWR-D-14-00150.1

Delle Monache, L., F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight, 2013: Probabilistic weather prediction with an analog ensemble. Mon. Weather Rev., 141 (10), 3498-3516.

Descamps, L., and O. Talagrand, 2007: On some aspects of the definition of initial conditions for ensemble prediction. Mon. Wea. Rev., 135, 3260–3272.

Du, J., 2007: How to evaluate the quality of an EPS

and its forecasts? Part VI in Uncertainty and ensemble forecast, p.19-24, Science and Technology Infusion Lecture Series, NOAA's National Weather Service, Office of Science and Technology. Available online from: https://www.nws.noaa.gov/ost/climate/STIP/uncertainty.htm, DOI: 10.25923/vpje-w924

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Monthly Weather Review, 129(10), 2461–2480. https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2

Ehrendorfer, M., 2006. The Liouville equation in atmospheric predictability. In: Palmer, T., Hagedorn, R. (Eds.), Predictability of Weather and Climate. Cambridge University Press, pp. 59–98.

Errico, R. and D. Baumhefner, 1987: Predictability experiments using a high-resolution limited area model. Mon. Wea. Rev., 115, 488-504.

Feng, J., R. Q. Ding, J. P. Li, and Z. Toth, 2018: Comparison of nonlinear local Lyapunov vectors and bred vectors in estimating the spatial distribution of error growth. J. Atmos. Sci., 75, 1073–1087.

Feng, J., J. P. Li, J. Zhang, D. Q. Liu, and R. Q. Ding, 2019: The relationship between deterministic and ensemble mean forecast errors revealed by global and local attractor radii. Adv. Atmos. Sci., 36(3), 271–278.

Feng, J., Z. Toth, and M. Peña, 2020: Partition of Analysis and Forecast Error Variance into Growing and Decaying Components. Quart. J. Roy. Meteor. Soc., 146(728), 1302-1321.

Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. Quarterly Journal of the Royal Meteorological Society, 141(688), 916–924. https://doi.org/10.1002/qj.2411

Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. Tellus, Series A: Dynamic Meteorology and Oceanography, 66(1). https://doi.org/10.3402/tellusa.v66.22662

Gilmour, I., and L. A. Smith, 1997: Enlightenment in Shadows. In: Applied nonlinear dynamics and stochastic systems near the millennium. Eds. J. B Kadtke and A. Bulsara, AIP, New York, pp. 335-340.

Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear regime duration: is 24 hours a long time in synoptic weather forecasting? Journal of the Atmospheric Sciences, 58, 3525–3539.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 69(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. Monthly Weather Review, 128(4), 1187–1193. https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2

Hagedorn, R. and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. Meteorological Applications, 16(2), 143–155. https://doi.org/10.1002/met.92

Hamill, T., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Mon. Wea. Rev., 134, 3209–3229.

Hamill, T. M., Whitaker, J. S., & Snyder, C., 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Monthly Weather Review, 129(11), 2776–2790. https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2

Held, I., 2015: Small Earth, deep atmosphere, and hypohydrostatic models. Retrieved 4 Nov 2022 from GFDL Blog site: https://www.gfdl.noaa.gov/blog_held/65-small-earth-deep-atmosphere-and-hypohydrostatic-models/

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5), 559–570.

Hoffman, R.N. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus A, 35A: 100-118. https://doi.org/10.1111/j.1600-0870.1983.tb00189.x

Hopson, T. M., 2014: Assessing the Ensemble Spread–Error Relationship, Monthly Weather Review, 142(3), 1125-1142. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/142/3/mwr-d-12-00111.1.xml

Hou, Z., J. P. Li, R. Q. Ding and J. Feng, 2018: The application of nonlinear local Lyapunov vectors to the Zebiak–Cane model and their performance in ensemble prediction. Clim. Dyn., 51, 283–304.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. Mon. Wea. Rev., 126: 796–811.

Houtekamer, P., H. Mitchell, X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. Mon. Weather Rev., 137, pp. 2126-2143, 10.1175/2008MWR2737.1.

Jolliffe, I. T. and Stephenson, D. B. (eds), 2003: Forecast Verification: a Practitioner's Guide in Atmospheric Science. Chichester: Wiley.

Kalnay, E., 2003: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge: Cambridge University Press.

Kalnay, E., 2017: Historical perspective: earlier ensembles and forecasting forecast skill. Presentation at the ECMWF Annual Seminar 2017: Ensemble prediction: past, present and future. 11-14 Sep. 2017, Reading, England. https://www.ecmwf.int/sites/default/files/medialibrary/2017-03/AS2017_Programme.pdf

Khan, A. N., N. Iqbal, A. Rizwan, R. Ahmad, and D. H. Kim, 2021: An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings. Energies, 14(11). https://doi.org/10.3390/en14113020

Kleeman, R., 2011: Information Theory and Dynamical System Predictability. Entropy 13, no. 3, 612-649. https://doi.org/10.3390/e13030612

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, , S. Gadgil, and S. Surendan, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. Science, 285, 1548–1550.

Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, Water Resources Research, 35 (9), pp. 2739-2750

Krzysztofowicz, R., and K. S. Kelly, 2000: Bayesian improver of a distribution. Stochastic Environmental Research and Risk Assessment, 14(6), 449–470. https://doi.org/10.1007/PL00009785

Krzysztofowicz, R., & Evans, W. B. (2008). Probabilistic forecasts from the national digital forecasts database. Weather and Forecasting, 23(2), 270–289. https://doi.org/10.1175/2007WAF2007029.1

Le Cam, L., 1986: The Central Limit Theorem around 1935. Stat. Sci. 1, 78–96.

Lewis, M., 2005: Roots of Ensemble Forecasting. Mon. Wea. Rev., 133(7), 1865–1885.

Liguori, S., M. Rico-Ramirez, A. Schellart, A. Saul, 2012: Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. Atmos. Res., 103, 80-95.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. Mon. Wea. Rev., 102, 409–418.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. Journal of Computational Physics, 227(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Li, J. P. and J. Chou, 1997: The existence of the atmosphere attractor. Sci. China Ser. D, 40, 215–224.

Li, J. P., and R. Q. Ding, 2011: Temporal–spatial distribution of atmospheric predictability limit by local dynamical analogues. Mon. Wea. Rev., 139, 3265–3283.

Li, J. P., J. Feng, and R. Q. Ding 2018: Attractor Radius and Global Attractor Radius and their Application to the Quantification of Predictability Limits. Clim. Dyn., 51, 2359–2374, https://doi.org/10.1007/s00382-017-4017-y.

Liu, T., Y. Gao, X. Song, et al., 2023: A multi-model prediction system for ENSO. Sci. China Earth Sci. 66, 1231–1240. https://doi.org/10.1007/s11430-022-1094-0

Lorenz E. N, 1963: Deterministic nonperiodic flow. J. Atmos. Sci., 20, 130–141.

Lorenz E. N., 1969: The predictability of a flow which possesses many scales of motion. Tellus 21: 289–307.

Lorenz E. N, 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34: 505–513.

Lumen, 2022: Symbols and language. In: Module 2 (Culture and Society) of Sociology, an open educational resource. Lumen Learning, available from: https://courses.lumenlearning.com/alamo-sociology/chapter/reading-elements-of-culture/

Magnusson, L., J. Nycander, & E. Källén, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble prediction. Tellus, Series A: Dynamic Meteorology and Oceanography, 61(2), 194–209.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. Quart. J. Roy. Meteor. Soc.,122,73–119.

Mu, M., W. S. Duan, J. F. Chou, 2004: Recent advances in predictability studies in China (1999–2002). Adv. Atmos. Sci., 21, 437–443. https://doi.org/10.1007/BF02915570

Mu, M., W. S. Duan, B. Wang, 2003: Conditional nonlinear optimal perturbation and its applications. Nonlin Processes Geophys. 10, 493–501.

Murphy, A. H., 1972: Scalar and vector partitions of the probability score: part I. Two-state situation. J. Appl. Meteor., 11, 273-282.

Murphy, A. H., 1969: On the “ranked probability score.” J. Appl. Meteor., 8, 988–989.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. Quart. J. Roy. Meteor. Soc., 114, 463–493

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith, 2006: Ensemble prediction: a pedagogical perspective. ECMWF newsletter, 106(106), pp.10-17.

Palmer, T. N., R. Mureau, and F. Molteni, 1990: The Monte Carlo forecast. Weather, 45: 198−207.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. Reports on Progress in Physics, 63(2), 71–116. https://doi.org/10.1088/0034-4885/63/2/201

Palmer, T., 2019: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. Quarterly Journal of the Royal Meteorological Society, 145, 12–24, https://doi.org/10.1002/qj.3383.

Peña, M., and Z. Toth, 2014: Estimation of analysis and forecast error variances. Tellus A: Dynamic Meteorology and Oceanography, 66:1, DOI: 10.3402/tellusa.v66.21767

Privé, N. C., and R. M. Errico. 2015: Spectral analysis of forecast error investigated with an observing system simulation experiment. Tellus A, 67: 25977

Raynaud, L., and F. Bouttier, 2016: Comparison of initial perturbation methods for ensemble prediction at convective scale. Quart. J. Roy. Meteor. Soc., 142, 854–866, https://doi.org/10.1002/qj.2686.

Richardson, L. F., 1922: Weather Prediction by Numerical Process (Cambridge Univ. Press).

Roulston, M. S., & L. A. Smith, 2003: Combining dynamical and statistical ensembles. Tellus, Series A: Dynamic Meteorology and Oceanography, 55(1), 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x

Rose, C., and M. D. Smith, 2002: Mathematical Statistics with Mathematica. Springer-Verlag, 311 pp.

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

Sardeshmukh, P. D., J. A. Wang, G. P. Compo, and C. Penland, 2023: Improving Atmospheric Models by Accounting for Chaotic Physics. J. Climate, 36, 5569–5585, https://doi.org/10.1175/JCLI-D-22-0880.1.

Schaake J. C., T. H. Hamill, R. Buizza, M. Clark, 2007: HEPEX – the hydrological ensemble prediction experiment Bulletin of the American Meteorological Society, 88 (10).

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics. Quarterly Journal of the Royal Meteorological Society, 140(680), 1086–1096. https://doi.org/10.1002/qj.2183

Su, X., H. Yuan, Y. Zhu, Y. Luo, and Y. Wang, 2014: Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. J. Geophys. Res. Atmos., 119, 7292–7310, doi:10.1002/2014JD021733.

Szunyogh, I., Kostelich, E. J., Gyarmati, G., Kalnay, E., Hunt, B. R., Ott, E., … Yorke, J. A. (2008). A local ensemble transform Kalman filter data assimilation system for the NCEP global model. Tellus, Series A: Dynamic Meteorology and Oceanography, 60 A(1), 113–130. https://doi.org/10.1111/j.1600-0870.2007.00274.x

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Monthly Weather Review, 144(6), 2375–2393. https://doi.org/10.1175/MWR-D-15-0260.1

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings, ECMWF Workshop on Predictability, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.].

Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of the large scale atmospheric pattern. Tellus, 9, 275- 295.

Toth, Z., 1991a: Estimation of Atmospheric Predictability by Circulation Analogs, Monthly Weather Review, 119(1), 65-72. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/119/1/1520-0493_1991_119_0065_eoapbc_2_0_co_2.xml

Toth, Z., 1991b: Circulation Patterns in Phase Space: A Multinormal Distribution?, Monthly Weather Review, 119(7), 1501-1511. Retrieved Mar 14, 2022, from https://journals.ametsoc.org/view/journals/mwre/119/7/1520-0493_1991_119_1501_cpipsa_2_0_co_2.xml

Toth, Z., 1993: Preferred and Unpreferred Circulation Types in the Northern Hemisphere Wintertime Phase space, Journal of Atmospheric Sciences, 50(17), 2868-2888. Retrieved Mar 21, 2022, from https://journals.ametsoc.org/view/journals/atsc/50/17/1520-0469_1993_050_2868_paucti_2_0_co_2.xml

Toth, Z., and E. Kalnay, 1993: Ensemble Forecasting at the NMC: The generation of perturbations. Bull. Amer. Meteorol. Soc., 74, 2317-2330.

Toth, Z., 1995: Degrees of freedom in Northern Hemisphere circulation data. Tellus, 47A, 457–472.

Toth, Z., and E. Kalnay, 1997: Ensemble Forecasting at NCEP: the breeding method. Mon. Wea. Rev., 125, 3297–3318.

Toth, Z., O. Talagrand, and Y. Zhu, 2005: The attributes of forecast systems: A framework for the evaluation and calibration of weather forecasts. Predictability of Weather and Climate, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 584—595.

Toth, Z., Y. Zhu, and T. Marchok, 2001: On the ability of ensembles to distinguish between forecasts with small and large uncertainty. Weather and Forecasting,16, 436–477.

Tribbia, J. J., and D. P. Baumhefner, 2004: Scale Interactions and Atmospheric Predictability: An Updated Perspective. Mon. Wea. Rev., 132, 703–713, https://doi.org/10.1175/1520-0493(2004)132<0703:SIAAPA>2.0.CO;2.

Tuzlukov, V., 2010: Signal Processing Noise, Electrical Engineering and Applied Signal Processing Series, CRC Press. 688 pages. ISBN 9781420041118

Unger, D. A., 1985: A method to estimate the continuous ranked probability score. Preprints, Ninth Conf. on Probability and Statistics in Atmospheric Sciences, Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

van den Dool, H. M., 1989: A new look at weather forecast through analogs. Mon. Wea. Rev., 117, 2230–2247.

van den Dool, H.M., 1994: Searching for analogues, how long must we wait? Tellus A, 46: 314-324. https://doi.org/10.1034/j.1600-0870.1994.t01-2-00006.x

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., … Ylhaisi, J. (2021). Statistical postprocessing for weather forecasts review, challenges, and avenues in a big data world. Bulletin of the American Meteorological Society. American Meteorological Society. https://doi.org/10.1175/BAMS-D-19-0308.1

Wang, X., C. H. Bishop, and S. J. Julier, 2004: Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? Mon. Wea. Rev., 132, 1590–1605.

Wang, X., Barker, D. M., Snyder, C., & Hamill, T. M., 2008: A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part II: Real observing experiments. Monthly Weather Review, 136(12), 5132–5147. https://doi.org/10.1175/2008MWR2445.1

Wei, M. and Z. Toth, 2003: A new measure of ensemble performance: perturbations versus Error Correlation Analysis (PECA). Mon. Wea. Rev., 131, 1549–1565.

Whitaker, J. S., & Loughe, A. F., 1998: The relationship between ensemble spread and ensemble mean skill. Monthly Weather Review, 126(12), 3292–3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications, 16(3), 361–368. https://doi.org/10.1002/met.134

World Meteorological Organization (WMO), 2021, Guidelines on Ensemble Prediction System Postprocessing, ISBN 978-92-63-11254-5.

Yang, S. C., M. Cai, E. Kalnay, M. Reinecker, G. Yuan, and Z. Toth, 2006: ENSO bred vectors in coupled ocean-atmosphere general circulation models. Journal of Climate, 19(8), 1422–1436. https://doi.org/10.1175/JCLI3696.1

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

Zhou, F., and Z. Toth, 2020: On the Prospects for Improved Tropical Cyclone Track Forecasts. Bull. Amer. Meteor. Soc. 1–55, https://doi.org/10.1175/BAMS-D-19-0166.1

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the NCEP Global Ensemble Forecast System in a parallel experiment. Wea. Forecasting, 32, 1989–2004.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. J. Climate, 12, 2474–2489.

APPENDIX A: Experimental Data

Experimental results in this study are based on operational analysis and forecast data from the NCEP Global Ensemble Forecast System (GEFS), initialized twice a day (00Z and 12Z) from the period Dec 1, 2017 - Feb 28, 2018, for a total of 180 cases on a 1º x 1º latitude-longitude grid, out to 16 days lead time at 12-hour output frequency (Zhou et al., 2017). Note that the unperturbed control forecast is run at the same resolution as the perturbed forecasts. Most statistics are computed over the Northern Hemisphere extratropics in the 30º-65º latitude band. The perturbation methods and numerical model used to generate the NCEP ensemble are typical of those used at many other centers.

As reality (or truth) is unknown, true error cannot be measured in practice. In this study, we use NWP analysis fields as a proxy for truth. The difference between a forecast and this proxy can be called "perceived" error. With some assumptions, true error can be estimated based on perceived error measurements (Pena and Toth, 2014). Despite quantitative differences at short lead times, the qualitative behavior of true and perceived error are similar (Feng et al., 2020). Beyond 2 days lead time, the bias in perceived forecast error induced by error in the verifying analysis field used as a proxy for truth is relatively small.

APPENDIX B: Information and Noise in Signal Processing vs. Weather Forecasting

Here we discuss what is common in and different between Information (I) as defined by Eq 7 and “Information entropy” or Shannon entropy (SE, Shannon 1948) as used in information theory, and Noise as defined in Eq. 8 compared with its use in signal processing. Both Information and SE provide a measure of uncertainty in our knowledge of a particular event out of all of its possible outcomes. While SE was introduced in the context of communication, Information is designed to quantify knowledge captured in analyzed or forecast states of a natural system like the atmosphere. Conveniently, I in its standardized form captures the fraction of forecast variance identical to the real state.

Noise, either defined by Eq. 8 (N) or as used in signal processing, refers to impediments to accessing information. “In signal processing, noise is a general term for unwanted (and, in general, unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion” (Tuzlukov, 2010). Meanwhile, Noise in the context of forecast states of dynamical systems refers to dynamically constrained forecast variance that is unrelated to reality (see Section 4.3.4).

APPENDIX C: Error, Noise, and Information

Following Lorenz (1984), we assume that the divergence of initially nearby segments of a chaotic dynamical system's trajectory, and in the absence of model error, true forecast error (i.e., the difference between a forecast and reality) follows a logistic curve:

$$d_i^2 = R \cdot c/(e^{-\alpha\cdot i\cdot \Delta t} + c), \tag{C1}$$

where $c = d_0^2/(R - d_0^2)$, d_0^2 is the variance of initial error, R is the range between the lower and upper saturation values (that is double the climatic variance, Leith, 1974), is the exponential growth rate, and t is the time increment.

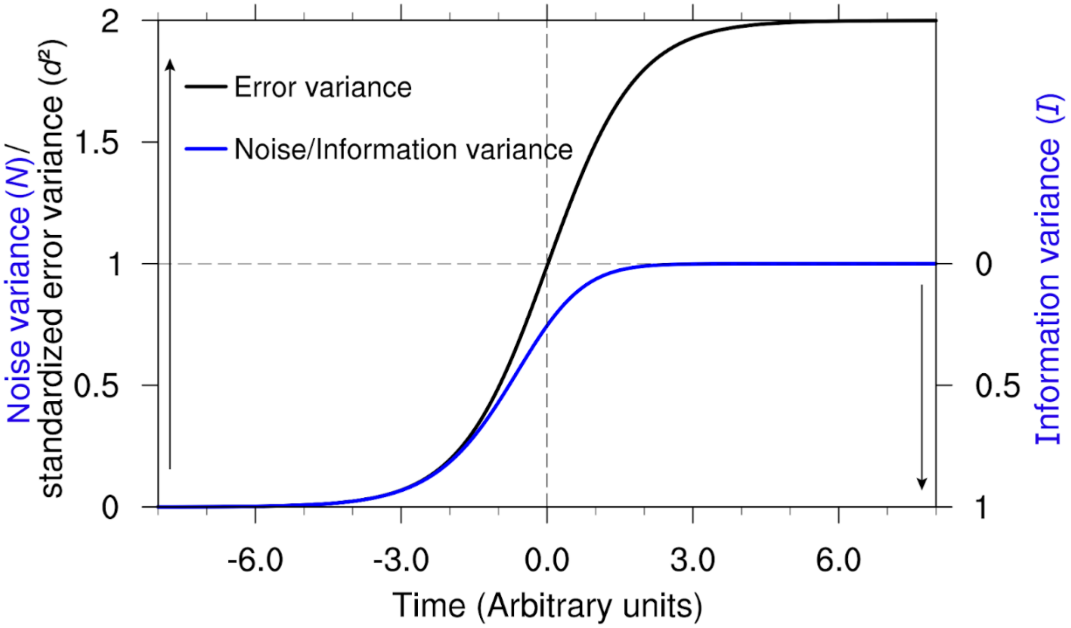

Figure C1. Schematic depicting the growth of noise (blue line, left axis) and the decrease of information variance (blue line, right axis) in a forecast characterized by logistically growing standardized error (black line). For further details, see text.

Error variance (d_i^2) can also be expressed as a function of Information I_i, i.e., the variance of truth *missed by*, and noise variance (N_i) that is *included in* a forecast (see the top right-angled triangle in Fig. 6):

$$d_i^2 = |\mathbf{F}_i - \mathbf{T}|^2/\,|\mathbf{T} - \mathbf{C}|^2 = N_i + (1 - \sqrt{I_i})^2\ . \tag{C2}$$

For forecast systems with realistic variability, exploiting Eq. 9, error variance can be written as a function of either Noise (not shown) or Information variance only:

$$d_i^2 = 2(1 - \sqrt{I_i})\ . \tag{C3}$$

Considering also Eqs. C1 and 9, a rearrangement of Eq. C3 defines the time evolution of Noise (not shown) and Information (see blue line in Fig. C1) as:

$$I_i = (2 - d_i^2)^2/4 \ . \tag{C4}$$

APPENDIX D: Degrees of Freedom

The experiment reported in Fig. 9c is repeated with different values for dof and the frequency distribution of error amplitudes in perturbed states in the perfect (Fig. 9b) and simulated ensembles are compared using the Kolmogorov-Smirnov 2-sample test (Chakravart et al., 1967). Error amplitudes in both the perfect and simulated ensembles are standardized by the sample-mean rms error of the control analysis. The best fit is found at M_d = 33 (used in the construction of Fig. 9c, see Fig. D1a), with a range of values between 28 and 38 still acceptable at the 5% statistical significance level. To reduce noise, the test statistic is processed with a 5-point triangular filter before it is plotted as a function of dof (Fig. D1b). The results indicate that the experimental data in Fig. 9b are consistent with the hypothesis that the global ensemble perturbations form a random sample in a high dimensional phase space.

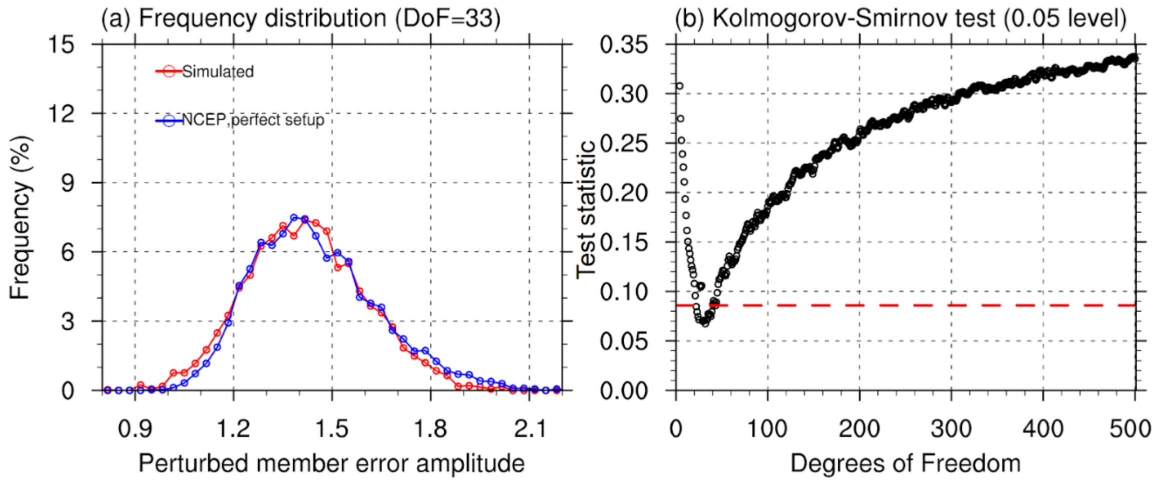

Figure D1. (a) The frequency of error in perturbed initial conditions from the NCEP (perfect setup, blue) and simulated (dof=33, red) ensembles. (b) Test statistic for the two-sample Kolmogorov-Smirnov test showing the maximum absolute difference (black open circles) between the empirical perturbed state error distribution functions from the perfect and simulated ensembles like those in panel (a). Values below the red dashed line indicate dof values where the actual and simulated distributions are statistically indistinguishable at the 0.05 significance level.

M_d = 33 degrees of freedom (dof) estimated for the NH 500 hPa extratropical height, of course, assess only a small part of the full space of atmospheric dynamics at the resolution of today's models. Using the statistical evaluation described above, the best estimate for the dof of global 500 hPa height variability is found to be 50. Though global extratropical 500 hPa height covers a large subspace of atmospheric dynamics, it does not reflect independent variations across the entire planetary circulation (Palmer et al., 2006, see their Appendix). Due to strong dynamical connections across variables, conservatively we expect a factor of less than 2 increase in dof if all independent model variables are considered. And due to the low aspect ratio of the atmospheric fluid at today's resolution of global models (e.g., Held, 2015), we anticipate a similarly low, less than a factor of 2 increase in dof were all levels included. Such considerations suggest that the dof of the subspace of Noise (i.e., initial error and short-range perturbation dynamics resolved by today's operational forecast systems) may be 3-4 times higher than that of the global 500 hPa height field, in the range of $M_d^{overall}$= 150-200.

APPENDIX E: Bracketing Ratio in Multiple Dimensions

Bracketing ratio F_{M_d,M_e} is a positively oriented metric, defined here for multidimensional applications as the relative frequency of reality (or its proxy) falling within (or bracketed by) the ensemble cloud in the direction *congruent with the error in the control* (see Appendix D). The bracketing ratio is a function of the degrees of freedom (M_d) and the number of ensemble members (M_e). Note that bracketing ratio F_{M_d,M_e} (Section 6.3) is an inverse measure of the ensemble outlier statistic (e.g., Buizza and Palmer, 1998), generalized for multidimensional applications, as well as a generalization of the probability of an ensemble member having an error lower than that in the control, shown in the table of the Appendix in Palmer et al (2006).

For the illustration below, F_{M_d,M_e} is calculated as follows. Missed Information in the control and initial ensemble perturbation vectors d_0 and ε_0 are given by ($d_{0,1}, d_{0,2}, \ldots, d_{0,M_d}$) and ($\varepsilon_{0,1}, \varepsilon_{0,2}, \ldots, \varepsilon_{0,M_d}$), respectively. Since $\varepsilon_{0,i}$ is a random sample of $d_{0,i}$, and following the standardization introduced in Section 5.5, we assume the elements $\varepsilon_{0,i}$ and $d_{0,i}$ both follow independent and identical standard Gaussian distributions N(0,1). Therefore, the distribution of projection of the ensemble perturbation ε_0 on the analysis error d_0 has an expected value of zero and a variance of 1, also conforming to a Gaussian distribution N(0,1).

We consider an ensemble with M_e members. The projection of the members onto the direction congruent with the Missed Information divide the probability space of N(0,1) into M_e+1 intervals. We mark the threshold designating the upper percentile of 1/(M_e+1) as S. The truth is bracketed if the Euclidean norm of d_0 is smaller than S. The Euclidean norm of d_0 is calculated as $\sqrt{\sum_{i=1}^{M_d} d_{0,i}^2}$, where $\sum_{i=1}^{M_d} d_{0,i}^2$ follows the chi-square distribution $\chi(M_d)$. Therefore, the general form of the formula for the bracketing ratio illustrated in Fig. E1 is:

$$F_{M_d,M_e} = P(x < S^2)\,, \tag{E1}$$

where $x = \sum_{i=1}^{M_d} d_{0,i}^2 \sim \chi(M_d)$ and $P(\cdot)$ stands for the probability of x being smaller than S^2. For a single variable (M_d=1), Eq. E1 recovers the inverse of the formula for the often used ensemble outlier statistic:

$$F_{1,M_e} = 1 - 2 / (M_e + 1)\,. \tag{E2}$$

As an illustration, Fig. E1b displays the expected value of F_{M_d,M_e} as a function of the degrees of freedom (M_d) and the number of ensemble members (M_e). Highlighted are marginal values for M_d = 33, the estimated dof of the NH extratropical 500 hPa height field (E1a), and M_e = 20, the membership of the NCEP ensemble over the experimental period (E1c). In sharp contrast with realistic-size ensembles in low dimensions ($F_{M_d,20} \sim 1$, Fig. E1c), even for large ensembles (e.g., M_e = 200) and for a limited domain like the NH extratropical 500 hPa height (M_d = 33), truth is bracketed only in 1 out of about 500 million cases (Fig. E1a). This answers a question Gilmuor and Smith's (1997) posed in a broader context: ensembles can capture reality "only in", but not "even in" low dimensional systems. In the full space of resolved atmospheric dynamics ($M_d^{overall}$ ~ 175), truth would be encompassed at an astronomical rate so low that is computationally directly inaccessible.

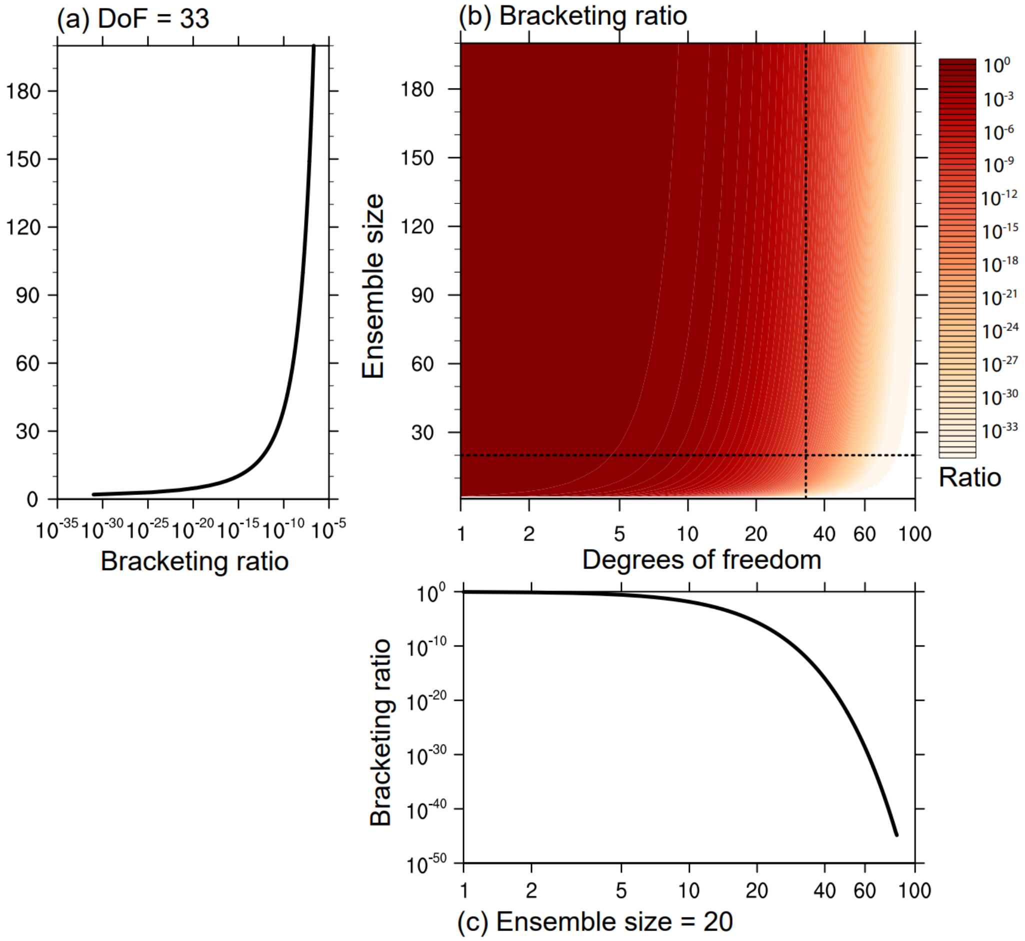

Figure E1. Ratio of cases a simulated ensemble of varying size brackets reality, as a function of the independent degrees of freedom (panel b). Bracketing ratio for dof = 33 and a 20-member ensemble is highlighted in panels (a) and (c), respectively.