Analysis of Ensemble Mean Forecasts: The Blessings of High Dimensionality

BO CHRISTIANSEN

Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 14 June 2018, in final form 30 January 2019)

ABSTRACT

In weather and climate sciences ensemble forecasts have become an acknowledged community standard. It is often found that the ensemble mean not only has a low error relative to the typical error of the ensemble members but also that it outperforms all the individual ensemble members. We analyze ensemble simulations based on a simple statistical model that allows for bias and that has different variances for observations and the model ensemble. Using generic simplifying geometric properties of high-dimensional spaces we obtain analytical results for the error of the ensemble mean. These results include a closed form for the rank of the ensemble mean among the ensemble members and depend on two quantities: the ensemble variance and the bias both normalized with the variance of observations. The analytical results are used to analyze the GEFS reforecast where the variances and bias depend on lead time. For intermediate lead times between 20 and 100 h the two terms are both around 0.5 and the ensemble mean is only slightly better than individual ensemble members. For lead times larger than 240 h the variance term is close to 1 and the bias term is near 0.5. For these lead times the ensemble mean outperforms almost all individual ensemble members and its relative error comes close to -30% . These results are in excellent agreement with the theory. The simplifying properties of high-dimensional spaces can be applied not only to the ensemble mean but also to, for example, the ensemble spread.

1. Introduction

It is a widespread observation in numerical weather prediction and seasonal forecasting that when performing ensemble forecasts the ensemble mean often outperforms most or all of the individual ensemble members (Toth and Kalnay 1997; Hamill and Colucci 1997; Du et al. 1997; Ebert 2001; Hagedorn et al. 2005; Krishnamurti et al. 1999; Casanova and Ahrens 2009; Surcel et al. 2014). A similar behavior is observed in climate modeling when validating an ensemble of different model experiments against observations (Lambert and Boer 2001; Gleckler et al. 2008; Pincus et al. 2008; Knutti et al. 2010; Sillmann et al. 2013; Flato et al. 2013) and when validating air quality models (van Loon et al. 2007; Delle Monache and Stull 2003; McKeen et al. 2005).

In the climate modeling context Christiansen (2018a) recently gave a simple explanation based on nonintuitive

properties of high-dimensional spaces—the curse of dimensionality. It was explained not only why the ensemble mean often outperforms most or all of the individual ensemble members but also why the error of the ensemble mean is almost always $(\sqrt{2} - 1)/\sqrt{2} \approx 30\%$ better than the typical error of individual ensemble members. This factor has frequently been observed in validation of climate models (Gleckler et al. 2008; Sillmann et al. 2013; Flato et al. 2013) and has previously been discussed by Annan and Hargreaves (2011) and Rougier (2016). See also the comment (Rougier 2018) and reply (Christiansen 2018b) exchange to Christiansen (2018a). The superiority of the ensemble mean has previously been explained as a combination of cancellation of errors and nonlinearity of the diagnostics (Hagedorn et al. 2005) or nonlinear filtering (Toth and Kalnay 1997; Surcel et al. 2014). In fact, the first discussion of the factor of $\sqrt{2}$ goes back to Leith (1974) and was also noted by Du et al. (1997). See also the discussion of previous literature in Christiansen (2018a).

The explanation in Christiansen (2018a) is based on general principles and gives a framework from which quantitative and analytical results can be obtained. These principles are valid for all distance measures in

Denotes content that is immediately available upon publication as open access.

Corresponding author: Bo Christiansen, boc@dmu.dk

DOI: 10.1175/MWR-D-18-0211.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

high-dimensional spaces. In the analysis of weather forecasts or climate models the high-dimensional space enters when we consider, for example, the root-mean-square error over all the grid points in an extended spatial region as a measure of the difference between models and observations. We then operate in a space with the dimension given by the number of (independent) grid points. Note that distance measures based on simple spatial averages, such as the difference between the modeled and the observed average Northern Hemisphere temperature, are not of high dimensionality and the results of this paper do therefore not hold for such diagnostics.

The properties of high-dimensional spaces often defy our intuition based on two and three dimensions (Bishop 2007; Cherkassky and Mulier 2007; Blum et al. 2018). The relevant properties of high-dimensional spaces for this study are that independent random vectors are almost always orthogonal and that random vectors drawn from the same distribution have almost the same length. In mathematics these properties are known as “waist concentration” and “concentration of measures” (Talagrand 1995; Donoho 2000; Gorban et al. 2016) and they are a cornerstone of statistical mechanics (see Gorban and Tyukin 2018, for a recent discussion). While the expression “curse of dimensionality” is generally used for the properties of high-dimensional spaces, in the present context these properties turn out to be a blessing as they strongly simplify the analysis and make analytical results possible. For the situation studied in Christiansen (2018a) the vectors representing the ensemble members and observations lie in a thin annulus far from the center and are all mutually orthogonal. The ensemble mean, however, is located near the center. Therefore, the ensemble mean, the observation, and an individual ensemble member form an isosceles right triangle and the error of the ensemble mean will be a factor of $\sqrt{2}$ smaller than the error of the individual ensemble members. This gives a relative error of the ensemble mean of $(1 - \sqrt{2})/\sqrt{2} \sim -30\%$ as often observed. The beneficial properties of high dimensionality are recognized in many areas of machine learning (Kainen 1997; Gorban and Tyukin 2018).

A basic assumption in Christiansen (2018a) is that the ensemble members and the observations are drawn from the same distribution. While this may be a sensible assumption when validating the climatology of climate models it is not necessarily a good assumption in numerical weather prediction. Here, the bias and variance of the model ensemble are expected to depend on the lead time of the forecast. In this paper we analyze the behavior of the ensemble mean in numerical weather prediction, more precisely in the National Oceanic and Atmospheric Administration’s (NOAA) second-generation

global ensemble reforecasts (GEFS) (Hamill et al. 2013). We assume a simple statistical model where observations and ensemble members are drawn from distributions with different variances and where the ensemble may be biased relative to observations. This model was briefly discussed in Christiansen (2018a) regarding the relative error of the ensemble mean. With this model and using the generic geometric properties of high-dimensional spaces we obtain further analytical results for the rank of the ensemble mean, the relation between the ensemble mean and the ensemble spread, and the effect of spatial correlations. These results explain the behavior of the ensemble mean as a function of lead time in the reforecasts. For convenience we have used the model analysis as proxy for the truth but we have tested that only small differences are found using real observations. While the present study focuses on the properties of the ensemble mean, the general simplifying properties of high dimensionality will also be applicable to other (e.g., probabilistic) measures of the forecast skill.

The simple statistical model and the analytic results obtained by using the properties of high-dimensional spaces are discussed in section 2. These results depend on the bias and the variance of both observations and model ensemble. The NOAA global ensemble reforecast dataset is described in section 3. The relation between the simple statistical model and the complex ensemble forecast is discussed in section 4 together with the method to separate bias and the variances in the ensemble forecast. The analytical and numerical results are compared in section 5. The paper finishes with the conclusions in section 6.

2. Analytical results in high-dimensional spaces

In this section we present the analytical results based on the properties of high-dimensional spaces. We begin by describing the notation, the statistical model, and the simplifying effects of high dimensions in section 2a. In sections 2b and 2c we derive analytical results for the relative error and the rank of the ensemble mean. In section 2d we look at the relation between the ensemble mean and the ensemble spread. Finally, in section 2e we consider the modifications to the analytical results when the elements of the observations and the ensemble members are not independent (corresponding to the presence of spatial correlations in the forecasts) and the effective dimension therefore is smaller than the physical dimension.

a. The statistical model and notation

We first focus on an ensemble forecast for a particular lead time. The forecast has K ensemble members, $\mathbf{x}^k$,

$k = 1, \dots, K$, and the observations are denoted by $\mathbf{o}$. Each $\mathbf{x}^k$ is a vector in N -dimensional space, $\mathbf{x}^k = (x_1^k, x_2^k, \dots, x_N^k)$ and so are the observations, $\mathbf{o} = (o_1, o_2, \dots, o_N)$. We denote the ensemble mean with an overbar: $\bar{\mathbf{x}} = \sum_{k=1}^K \mathbf{x}^k / K$.

Our basic model for the description of the ensemble forecast assumes that the ensemble members are drawn independently from $\mathcal{N}(\mathbf{b}, \sigma_{\text{fc}}^2 \mathbf{I})$ and the observations (representing the truth) from $\mathcal{N}(\mathbf{0}, \sigma_{\text{tr}}^2 \mathbf{I})$, where $\mathbf{b}$ is an N vector, $\mathbf{0}$ is the null vector, and $\mathbf{I}$ is the identity matrix of size N . Thus, we simply assume the ensemble members and the observations are drawn from spherical Gaussian distributions with different variances and the ensemble members may be biased relative to observations with a bias of length $\|\mathbf{b}\| = B\sqrt{N}$. Note that the coordinates of each ensemble member are independently drawn from the same univariate distribution. The same holds for the coordinates of the observations.

For the weather forecasts we expect to be close to the situation where $B = 0$ and $\sigma_{\text{tr}}^2 = 0$ for zero lead time as in this case the ensemble members are centered closely around the observations by construction and we expect the observational errors to be small (in particular if analysis is used as proxy for the truth). For long lead times we might expect—if the model and the real world have identical attractors—that the divergence of trajectories will result in a situation where observations and models are drawn from the same distribution (i.e., $B = 0$ and $\sigma_{\text{tr}}^2 = \sigma_{\text{fc}}^2$). The two limits correspond to what in climate modeling are referred to as the “truth plus error” interpretation and the “indistinguishable” interpretation. In the former the ensemble members are sampled from a distribution centered around the observations and in the latter the ensemble members and observations are all considered exchangeable (Sanderson and Knutti 2012). As mentioned the observational errors are small and, except for the smallest lead times, σ_{tr}^2 therefore reflects the random initial conditions while σ_{fc}^2 reflects both the initial conditions and model deficiencies. A more detailed discussion of how the variances depend on lead times is given in section 4.

For N large the general properties of high-dimensional spaces suggest a situation as shown schematic in Fig. 1. The ensemble members will be situated on a thin annulus with radius $\sigma_{\text{fc}}\sqrt{N}$ and width $\sigma_{\text{fc}}/\sqrt{2}$ (the standard deviation) and the observations on another thin annulus with radius $\sigma_{\text{tr}}\sqrt{N}$ and width $\sigma_{\text{tr}}/\sqrt{2}$. The centers of the two annuli are separated by the distance $B\sqrt{N}$. These results were derived in Christiansen (2018a) by rewriting the distribution of the multivariate spherical Gaussian of $\mathbf{x}$ as a function of $\|\mathbf{x}\|$. The resulting distribution is proportional to

FIG. 1. Schematic diagram showing the situation in high-dimensional space when ensemble members are drawn from $\mathcal{N}(\mathbf{b}, \sigma_{\text{fc}}^2 \mathbf{I})$, $\|\mathbf{b}\| = B\sqrt{N}$, and the observations from $\mathcal{N}(\mathbf{0}, \sigma_{\text{tr}}^2 \mathbf{I})$. The ensemble members are located in a thin annulus with radius $\sigma_{\text{fc}}\sqrt{N}$ and width (standard deviation) $\sigma_{\text{fc}}/\sqrt{2}$. Likewise, the observations are located in a thin annulus with radius $\sigma_{\text{tr}}\sqrt{N}$ and width $\sigma_{\text{tr}}/\sqrt{2}$. The two annuli are separated by the distance $B\sqrt{N}$. The ensemble mean is located near the center of the distribution of the ensemble members. Independent vectors are almost always orthogonal.

$\|\mathbf{x}\|^{N-1} \exp(-\|\mathbf{x}\|^2/2)$ from which it is seen that the peak is centered around $\|\mathbf{x}\| = \sqrt{N-1}$ and that the width (standard deviation) of the peak is $1/\sqrt{2}$ independent of N . The results will, however, together with the analytical results in the rest of this section also hold for other distributions because of the central limit theorem. Furthermore, the ensemble members, the observations, and the bias will all be almost orthogonal.

b. Relative error of the ensemble mean

The situation described above leads to the following relation for the squared error between observations and ensemble members:

$$\|\mathbf{o} - \mathbf{x}^k\|^2 \simeq B^2 N + \sigma_{\text{tr}}^2 N + \sigma_{\text{fc}}^2 N. \quad (1)$$

For the ensemble mean the situation is different. The ensemble mean is special as it for large K will—because of the law of large numbers—be situated near the center of the annulus. Therefore for large K we have

$$\|\mathbf{o} - \bar{\mathbf{x}}\|^2 \simeq B^2 N + \sigma_{\text{tr}}^2 N. \quad (2)$$

Deriving these equations we have used both blessings of high dimensionality; the orthogonality to discard the covariances and the similar lengths to replace $\|\mathbf{x}^k\|^2$ and $\|\mathbf{o}\|^2$ with $\sigma_{\text{fc}}^2 N$ and $\sigma_{\text{tr}}^2 N$ (Christiansen 2018a).

Comparing Eqs. (1) and (2) we see that in this limit, N and K large, the ensemble mean is always closer to the observations than the individual ensemble members.

A measure that is often used when validating ensembles of climate models (Gleckler et al. 2008; Sillmann et al. 2013; Flato et al. 2013) is the error relative to the median error of all ensemble members. If a specific ensemble member has a relative error of, for example, 0.1 it therefore indicates that this ensemble member has an error 10% larger than the median error of the model ensemble. For the relative error of the ensemble mean we then get from Eqs. (1) and (2):

$$\frac{\sqrt{1 + (B/\sigma_{tr})^2} - \sqrt{1 + (B/\sigma_{tr})^2 + (\sigma_{fc}/\sigma_{tr})^2}}{\sqrt{1 + (B/\sigma_{tr})^2 + (\sigma_{fc}/\sigma_{tr})^2}}. \quad (3)$$

This result was found in Christiansen (2018a) and Fig. 2 [similar to Fig. 5 in Christiansen (2018a)] shows the relative error as a function of B/σ_{tr} and σ_{fc}/σ_{tr} . When the observations and ensemble members are from the same distribution (the indistinguishable interpretation, i.e., $B = 0$ and $\sigma_{fc} = \sigma_{tr}$) the relative error reduces to $(1 - \sqrt{2})/\sqrt{2}$, which is close to the value found in many validations of climate models (Gleckler et al. 2008; Sillmann et al. 2013; Flato et al. 2013). Note that in the truth plus error interpretation the relative error is -1 .

c. Rank of the ensemble mean

Equations (1) and (2) show that for large N and K the ensemble mean is always better than all the ensemble members. As mentioned in the introduction it is often observed that the ensemble mean outperforms most or all of the individual ensemble members. To get a more nuanced picture for the rank of the ensemble mean we now need to relax the limit of large N .

With $\mathbf{z}^k = \mathbf{x}^k - \mathbf{b}$ we get without any approximations

$$\|\mathbf{o} - \mathbf{x}^k\|^2 = \|\mathbf{o} - \mathbf{b}\|^2 + \|\mathbf{z}^k\|^2 + 2(\mathbf{o} - \mathbf{b}) \cdot \mathbf{z}^k \quad (4)$$

and

$$\|\mathbf{o} - \bar{\mathbf{x}}\|^2 = \|\mathbf{o} - \mathbf{b}\|^2 + \|\bar{\mathbf{z}}\|^2 + 2(\mathbf{o} - \mathbf{b}) \cdot \bar{\mathbf{z}}. \quad (5)$$

For independent Gaussian distributed vectors, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{I})$ and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathbf{I})$, the dot product $\mathbf{u} \cdot \mathbf{v}$ has zero mean and a standard deviation of $\sigma_u \sigma_v \sqrt{N}$ for large N . Likewise, if $\mathbf{c}$ is a constant vector then $\mathbf{c} \cdot \mathbf{u}$ has zero mean and a standard deviation of $\|\mathbf{c}\| \sigma_u \sqrt{N}$. The last term of Eq. (4), therefore, has a standard deviation of

$$\Delta = 2\sigma_{fc} \sqrt{\sigma_{tr}^2 + B^2} \sqrt{N}, \quad (6)$$

FIG. 2. The relative error of the ensemble mean as a function of normalized model variance σ_{fc}/σ_{tr} and normalized model bias B/σ_{tr} . Theoretical results calculated from Eq. (3). Similar to Fig. 5 in Christiansen (2018a).

as $\mathbf{o} \cdot \mathbf{z}^k$ and $\mathbf{b} \cdot \mathbf{z}^k$ are independent as $\mathbf{z}^k$ have constant length and $\mathbf{o}$ and $\mathbf{z}^k$ are orthogonal for large N . For large N the mean and standard deviation of $\|\mathbf{z}^k\|^2$ is $\sigma_{fc}^2 N$ and $\sqrt{2}\sigma_{fc}^2 \sqrt{N}$. The former comes from the constant lengths and the latter comes from the mean of $2\|\mathbf{z}^k\|$ (which is $2\sigma_{fc}\sqrt{N}$) multiplied by the standard deviation of $\|\mathbf{z}^k\|$, which is $\sigma_{fc}/\sqrt{2}$ (Fig. 1). For $\|\bar{\mathbf{z}}\|^2$ the mean and standard deviation likewise are $\sigma_{fc}^2 N/K$ and $\sqrt{2}\sigma_{fc}^2 \sqrt{N}/K$.

In Eq. (5) we now approximate $\|\bar{\mathbf{z}}\|^2$ with $\sigma_{fc}^2 N/K$ and ignore its variability and the cross-term $(\mathbf{o} - \mathbf{b}) \cdot \bar{\mathbf{z}}$ as they are a factor $1/K$ smaller than the similar terms in Eq. (4). We also note that $\|\bar{\mathbf{z}}\|^2$ and $\|\mathbf{z}^k\|^2$ are uncorrelated when K is large.

In Eq. (4) we approximate $\|\mathbf{z}^k\|^2$ with $\sigma_{fc}^2 N$. The variability of this term is ignored as its distribution is positively skewed and as this skewness only decays slowly with N . Therefore, the approximation without this term is superior to the approximation including only the symmetric part except for the largest N .

Thus, to leading order the probability of $\|\mathbf{o} - \mathbf{x}^k\|^2$ being smaller than $\|\mathbf{o} - \bar{\mathbf{x}}\|^2$ is equal to the probability of $\mathcal{N}(\sigma_{fc}^2 N, \Delta^2)$ [from the two last terms in Eq. (4)] being smaller than $\sigma_{fc}^2 N/K$ [from the second term in Eq. (5)]. So the fraction of the ensemble members closer to the observations than the ensemble mean becomes

$$\begin{aligned} & \frac{1}{\sqrt{2\pi}\Delta} \int_{-\infty}^{\sigma_{fc}^2 N/K} \exp\left(-\frac{(z - \sigma_{fc}^2 N)^2}{2\Delta^2}\right) dz \\ &= \Phi\left[\frac{\frac{\sqrt{N}}{2} \frac{K-1}{K} \frac{\sigma_{fc}}{\sigma_{tr}}}{\sqrt{1 + \left(\frac{B}{\sigma_{tr}}\right)^2}}\right], \end{aligned} \quad (7)$$

where Φ is the cumulative probability density of the Gaussian.

The fraction of the ensemble members better than the ensemble mean calculated from this equation is shown in Fig. 3 for $N = 30$ and $N = 100$ for large K . We note from Eq. (7) that the fraction decreases with N and becomes zero for $N \rightarrow \infty$ in agreement with the arguments earlier in this section. We see from Fig. 3 that even for $N = 30$ the fraction is quite low when $\sigma_{fc} \geq \sigma_{tr}$ and the bias is moderate. For the truth plus error interpretation the fraction is identically zero while it is $\Phi(-\sqrt{N}/2)$ for the indistinguishable interpretation.

We have confirmed numerically that Eq. (7) is a very good approximation even for $N = 10$. We have also confirmed numerically that this approximation also holds for large N for other distributions such as the uniform distribution and the t distribution as should be expected from the central limit theorem.

d. Ensemble spread and ensemble mean error

Here we take brief look at the relation between the error of the ensemble mean and the ensemble spread, which has attracted interest in previous studies (Whitaker and Loughe 1998; Buizza and Palmer 1998; Marzban et al. 2011; Wilks 2011; Fortin et al. 2014). If such a robust relation exists then the ensemble itself contains information about its forecast error. The squared ensemble spread δ^2 defined as the ensemble variance of $\|\mathbf{x}^k - \bar{\mathbf{x}}\|$ is $\delta^2 = \sigma_{fc}^2/2$ in the limit of large K and N . Keeping an extra term in Eq. (2) the squared error of the ensemble mean becomes

$$\begin{aligned} \|\mathbf{o} - \bar{\mathbf{x}}\|^2 &\simeq NB^2 + N\sigma_{tr}^2 + N\sigma_{fc}^2/K \\ &= NB^2 + N\sigma_{tr}^2 + 2\delta^2 N/K. \end{aligned} \quad (8)$$

The factor $1/K$ in the last term will contribute to make the dependence on δ^2 weak for observed values of bias and variances (see section 5). Note, that also after bias correction will the relationship between the squared error and the ensemble spread be weak.

e. The effective dimension

In the considerations above we assumed that elements $(x_1^k, x_2^k, \dots, x_N^k)$ of each ensemble members were drawn independently. For weather and climate modeling this corresponds to the unrealistic situation where there are no spatial correlations. When we consider the GEFS forecast the horizontal fields will be given at a $1^\circ \times 1^\circ$ longitude–latitude grid. So for the region north of 20°N we have around 25 000 grid points. However, because of the spatial correlations the effective dimension will be much lower.

We will therefore have to distinguish between the physical dimension N (the number of grid points) and the effective dimension N^* . The effective dimension or

FIG. 3. The rank of the ensemble mean (fraction of ensemble members less than ensemble mean) as a function of variance and bias for $N = 30$ and $N = 100$. Theoretical results calculated from Eq. (7).

the effective number of spatial degrees of freedom is an important property of spatiotemporal fields. The effective dimension depends on the considered time scale and is often used when evaluating the statistical significance of a spatial signal. The effective dimension can be estimated by several methods (see, e.g., Bretherton et al. 1999; Wang and Shen 1999).

We will here consider how the situation where $N > N^* \gg 1$ will affect the results above (where the assumption was that $N = N^* \gg 1$). The following results can easily be obtained if we consider the idealized situation where the N elements of the ensemble members come in identical groups of length L (i.e., $x_1^k = x_1^k, \dots, x_L^k, x_{L+1}^k = x_{L+2}^k, \dots, x_{2L}^k$, etc.). The effective dimension is then N/L .

We now repeat the rewriting of the multivariate spherical Gaussian as a function of radius (Christiansen 2018a) but now we take the reduced number of independent elements into consideration. We find that the lengths of the vectors in Fig. 1 still depend on the physical dimension N [e.g., in Eqs. (1) and (2)].

FIG. 4. The root-mean-square error of the GEFS reforecast of 2-m temperature as a function of lead time for 15 Jan in two different years [(top) 1985, (bottom) 1986]. Thin black curves are the 11 ensemble members, thick black curve the ensemble mean. Full blue curve is the estimate from Eq. (2) and dashed blue curve is the estimate from Eq. (1) with observed values of B , σ_{ic} , and σ_{tr} from Figs. 10 and 11. (left) For the region north of 20° N, and (right) for the region defined by 40°–45°N, 10°–40°E.

However, the widths (standard deviations) will be $\sigma_{fc}\sqrt{N/N^*}/\sqrt{2}$ and $\sigma_{tr}\sqrt{N/N^*}/\sqrt{2}$.

For the dot product of two ensemble members we then get the standard deviation $\sigma_{fc}^2\sqrt{N^2/N^*}$ and for the standard deviation of $\|\mathbf{z}^k\|^2$ we get $\sqrt{2}\sigma_{fc}^2\sqrt{N^2/N^*}$.

Using these relations we find that for the fraction of the ensemble members better than the ensemble mean, Eq. (7), N should be replaced with N^* . In the relation between the error of the ensemble mean and the ensemble spread, Eq. (8), the first two terms are unchanged while the last term becomes $N\delta^2/2/N^*$ as δ^2 now is $N\sigma_{fc}^2/2/N^*$. The result for the relative error of the ensemble mean, Eq. (3), does not depend on the dimension and will be unchanged.

3. The global ensemble reforecast

In this section we describe the ensemble reforecast data and show how the error and rank of the ensemble mean change as functions of lead time. In section 5 we will use the theoretical results of section 2 to relate the behavior of the ensemble mean to the variances and bias.

For comparison with the analytical results we use data from the NOAA's second-generation global ensemble reforecast system (GEFS; Hamill et al. 2013).

The reforecasts were downloaded from <https://www.esrl.noaa.gov/psd/forecasts/reforecast2/download.html>.

The ensemble forecasts consist of one control forecast and 10 perturbed ensemble members with lead times up to 384 h (16 days). The forecasts are available for every day in the period 1985 to the present. The internal resolution of the model is T254L42 the first 8 days and T190L42 thereafter.

We have downloaded the forecasts with 3 h intervals for lead times up to 72 h and with 6-h intervals up to 384 h for the period 1985–2017 (33 years). The downloaded horizontal fields are at a $1^\circ \times 1^\circ$ longitude–latitude resolution. Here we mainly consider the 33 forecasts of 2-m temperature for 15 January. Observations are taken as the unperturbed (control) ensemble member at lead time 0. Using the analysis as a proxy for truth is common practice [but not without problems (Bowler et al. 2015)]. However, we have tested that almost identical results are obtained when using either an independent model-based reanalysis or a station observation-based gridded dataset instead of the analysis. The main difference is that in these cases the root-mean-square errors do not vanish for the smallest lead times.

Figure 4 shows the root-mean-square errors of the 2-m temperature for the different ensemble members as a

FIG. 5. Relative error of the GEFS reforecast as a function of lead time for the same two forecasts as in Fig. 4. Thin black curves are the 11 ensemble members, thick black curve the ensemble mean. Blue curve is the estimate from Eq. (3) with observed values of $\sigma_{\text{ic}}/\sigma_{\text{tr}}$ and B/σ_{tr} from Figs. 10 and 11. Red line is $(1 - \sqrt{2})/\sqrt{2}$. (top), (bottom) Years 1985 and 1986 as in Fig. 4. (left) For the region north of 20°N , and (right) for the region defined by $40^\circ\text{--}45^\circ\text{N}$, $10^\circ\text{--}40^\circ\text{E}$.

function of lead time for two different 15 January dates (top 1985, bottom 1986). The root-mean-square errors are calculated for the extratropics, $20^\circ\text{--}90^\circ\text{N}$ (left), and for a smaller area, $40^\circ\text{--}45^\circ\text{N}$, $10^\circ\text{--}40^\circ\text{E}$ (right). As expected the errors increase fast in the first few time steps where after the increase slows down. Saturation is reached for a lead time approximately around 100 h for the extratropics and around 50 h for the smaller area. Also note that the spread of the ensemble members increases with lead time and is larger for the smaller region than for the extratropics. Figure 4 also shows the root-mean-square errors of the ensemble mean (thick black curves). We observe that the error of the model mean is smaller than almost all errors of the individual ensemble members for large lead times. This effect is more clear for the extratropics than for the smaller region. Close inspection shows that also for the smallest lead times the error of the ensemble mean is smaller than the errors of the individual ensemble members. There is a daily cycle in the forecasts, with peaks for lead times 12 h, 36 h, etc., which is most pronounced for the smaller region.

Figure 5 shows—for the same forecasts as in Fig. 4—the errors relative to the median error of all ensemble members. Note that the relative error of the ensemble mean is always smaller than zero. For the smallest lead

times it is close to -1 , but it increases fast and is close to zero for lead times around 30–40 h. Then it slowly decreases and there is a tendency for convergence toward $(1 - \sqrt{2})/\sqrt{2}$ for large lead times.

Figures 6 and 7 summarize the situation for all 33 forecasts for 15 January as a function of lead time. Figure 6 shows the relative error of the ensemble mean for all 33 ensemble forecasts together with its average over the 33 forecasts. The behavior we saw for the two individual forecasts is confirmed. Figure 7 shows the rank of the model mean for the 33 ensemble forecasts. The rank is 0 if the error of the ensemble mean is lower than the error of all 11 individual ensemble members. Likewise, the rank is 11 if the error of the ensemble mean is higher than the error of all ensemble members. For the smallest and largest lead times the ensemble mean is almost always closer to observations than all the ensemble members. For lead times between 20–100 h the ensemble mean is more often outperformed by one or more ensemble members.

We can directly test if the blessings of high dimensionality are fulfilled for the model ensemble. For each lead time and for each of the 33 forecasts we calculate the angles between pairs of ensemble members. The angles $\theta_{k,l}$ are given by $\mathbf{x}^k \cdot \mathbf{x}^l = \|\mathbf{x}^k\| \|\mathbf{x}^l\| \cos \theta_{k,l}$.

FIG. 6. The relative error (RMS) of the ensemble mean as a function of lead time for all the 33 forecasts of the 2-m temperature for 15 Jan. Thin black curves are the 33 individual forecasts. Thick black curve is mean over all 33 forecasts. Blue curve is the theoretical relative error from Eq. (3) with observed values of σ_{ic}/σ_{tr} and B/σ_{tr} from Figs. 10 and 11. Red line is $(1 - \sqrt{2})/\sqrt{2}$. (top) For the region north of 20°N, and (bottom) for the region defined by 40°–45°N, 10°–40°E.

These angles are distributed narrowly about $\pi/2$ as shown in the top panel of Fig. 8 indicating that the ensemble members are almost orthogonal. This holds for all lead times although the distributions are slightly wider for the smallest lead times. The distribution of the squared lengths, $\|\mathbf{x}^k\|^2$, and of the cross-term $\mathbf{x}^k \cdot \mathbf{x}^l$ are shown in the bottom panel of Fig. 8. The squared lengths are narrowly distributed around a mean that increases with lead time. For all lead times the cross terms are distributed around zero and are well separated from the squared lengths. This indicates that the cross terms can be ignored in the expansion $\|\mathbf{x}^k - \mathbf{x}^l\|^2 = \|\mathbf{x}^k\|^2 + \|\mathbf{x}^l\|^2 - 2\mathbf{x}^k \cdot \mathbf{x}^l$ and that $\|\mathbf{x}^k\|^2$ and that $\|\mathbf{x}^l\|^2$ can be replaced by $N\sigma_{ic}^2$.

4. Relation between the simple statistical model and ensemble forecasts

The analytical expressions in section 2 were derived for a simple statistical model including model bias and

FIG. 7. The rank of the ensemble mean as a function of lead time based on all the 33 forecasts of the 2-m temperature for 15 Jan. Thin curves are the 33 individual forecasts. Thick black curve is mean over all 33 forecasts and dashed black curves the mean plus/minus one standard deviation. The blue curves are analytic results [Eq. (7)] for $N^* = 5$ (top curve), 10, 30, and 100 (bottom curves) with observed values of σ_{ic}/σ_{tr} and B/σ_{tr} from Figs. 10 and 11. (top) For the region north of 20°N, and (bottom) for the region defined by 40°–45°N, 10°–40°E.

different variances for observations and the model ensemble. In this section we first, in section 4a, discuss how to relate the simple statistical model to the complex ensemble forecast. In section 2 we saw that in the simple statistical model the relative error and the rank of the ensemble mean both depend on the variances of the forecasts and observations and on the bias. In section 4b we discuss how to estimate these quantities to be able to compare the observed ensemble mean to the theory of section 2. The results in this section do not depend on the properties of high-dimensional spaces.

a. The dynamical perspective

There are two sources of errors in ensemble forecasts. The first is related to uncertainties in the initial conditions due to, for example, measurements errors and

FIG. 8. (top) The distribution of angles θ_{ij} between pairs of ensemble members, $\cos\theta_{ij} = \mathbf{x}^k \cdot \mathbf{x}^l / \|\mathbf{x}^k\| \|\mathbf{x}^l\|$, as a function of lead time. (bottom) The distribution of squared lengths $\|\mathbf{x}^k\|^2/N$ (black curves) and the cross terms $\mathbf{x}^k \cdot \mathbf{x}^l/N$ (blue curves) as a function of lead time. For each lead time the distribution of lengths is calculated from all ensemble members and all the 33 different forecasts. The angles and the cross terms are calculated over all $K(K-1)/2$ different pairs of ensemble members and the 33 different forecasts. The ensemble members have for each forecast been centered to their common mean before the calculation of lengths and angles. Distributions are shown as the mean (thick solid curve), mean plus/minus one standard deviation (thin solid curves), and the 5% and 95% quantiles (dashed curves). The calculation is based on the region north of 20°N . Full black curve in bottom panel is identical to σ_{fc}^2 in top panel of Fig. 10.

scarceness of observations. The second is related to model deficiencies originating, for example, from unresolved scales and physical parameterizations. The effect of these errors and attempts to separate them have been studied by Tribbia and Baumhefner (1988), Stensrud et al. (2000), and Nicolis et al. (2009). See also the review by Palmer (2000). The situation in phase-space is illustrated schematically in Fig. 9. The gray cloud illustrates the probability distribution of the initial conditions. The blue clouds indicate the modeled probability distributions at two later times.

In the real atmosphere there is a single true initial condition and a single trajectory in phase space and consequently a single atmospheric state at each later time. However, we can imagine the evolution of the distribution of the initial conditions in the real atmosphere (or in a perfect model). This is illustrated as the red clouds in the figure. It is these distributions—the blue and red clouds—that should be compared for a given lead time when validating the ensemble forecast.

FIG. 9. Schematic illustration of the time evolution in phase space. The probability distribution of the initial conditions (gray cloud) evolves differently in the model (blue clouds) and the real atmosphere (red clouds) because of model deficiencies. The curves illustrate individual trajectories. The widths of the clouds indicate the variances of the model σ_{fc}^2 and the real atmosphere σ_{tr}^2 . The distance between the clouds indicates the model bias B . Note that these quantities all depend on time.

The widths of the clouds (in the subspace or projection under consideration; here the near-surface temperature) indicate the variances of the model, σ_{fc}^2 , and the real atmosphere, σ_{tr}^2 , and their mean distance indicates the model bias B . It is these parameters, which all are time dependent, that enter the formulas in the previous sections.

The temporal development of the distributions are given by the Liouville equation (Ehrendorfer 1994, 2006). However, in ensemble forecasts the distributions are estimated from a finite number of ensemble members due to the complexities of the atmosphere (e.g., Palmer 2000). In Fig. 9 the initial conditions of the ensemble are all located inside the small gray cloud. The blue curves show how the ensemble members develop over time. The figure indicates a strong divergence in the beginning which soon saturates and is followed by a slower divergence. For the real atmosphere the trajectories are illustrated by the red curves.

For each forecast and each lead time we have a model ensemble from which we can estimate the distributional quantities. However, for each forecast and each lead time we only have one realization of the real atmosphere. To estimate the distribution related to the real atmosphere (the red clouds) we need to combine information from many forecasts just as is done in verifications of ensemble forecasts (Toth et al. 2003).

b. Variance-bias decomposition of the error

We now focus on the forecasts for a specific lead time. We are interested only in the bias and variance related

to the divergence from the initial conditions and not in the bias and variance related to the climatological year-to-year variations. However, without further assumptions it is not possible to estimate both bias and variance of the observations as we have only one observation for each forecast. We, therefore, consider all forecasts with start day at the same calendar day, here 15 January. We combine these forecasts by removing for every start day the ensemble mean for that day from both the forecast ensemble and observation. This amounts to assuming a model where the different days only differ by a shift that is the same for models and observations. In other words, we consider anomalies with respect to the ensemble mean (which also depends on lead time). Thus, ensemble members are drawn from $\mathcal{N}(\mathbf{b} + \mathbf{c}_d, \sigma_{fc}^2 \mathbf{I})$ and the observations from $\mathcal{N}(\mathbf{c}_d, \sigma_{tr}^2 \mathbf{I})$, where $\mathbf{c}_d$ represents the year-to-year variability in climate.

To be more precise we assume we have forecasts for D start days and each forecast has K ensemble members. We extend the previous notation so $\mathbf{x}_d^k$ are the forecasts and $\mathbf{o}_d$ the observations for day d , $d = 1, \dots, D$. Each $\mathbf{x}_d^k$ and $\mathbf{o}_d$ are N vectors as previously. Defining $\mathbf{X}_d^k = \mathbf{x}_d^k - \bar{\mathbf{x}}_d$ and $\mathbf{O}_d = \mathbf{o}_d - \bar{\mathbf{o}}_d$ with $\bar{\mathbf{x}}_d = \sum_k \mathbf{x}_d^k / K$ we get

$$\begin{aligned} \sum_{k,d} \|\mathbf{x}_d^k - \mathbf{o}_d\|^2 / (KD) &= \sum_{k,d} \|\mathbf{x}_d^k - \mathbf{O}_d\|^2 / (KD) = \\ \sum_{k,d} \|\mathbf{X}_d^k - \langle \bar{\mathbf{X}} \rangle + \langle \bar{\mathbf{X}} \rangle - \langle \mathbf{O} \rangle + \langle \mathbf{O} \rangle - \mathbf{O}_d\|^2 / (KD) &= \\ \sum_{k,d} \|\mathbf{X}_d^k - \langle \bar{\mathbf{X}} \rangle\|^2 / (KD) + \sum_d \|\mathbf{O}_d - \langle \mathbf{O} \rangle\|^2 / D &+ \|\langle \bar{\mathbf{X}} \rangle - \langle \mathbf{O} \rangle\|^2. \end{aligned} \quad (9)$$

Here angle brackets denote the average over the D start days (e.g., $\langle \mathbf{O} \rangle = \sum_d \mathbf{O}_d / D$). The last equality is obtained noting that the cross terms, $\sum_{k,d} (\mathbf{X}_d^k - \langle \bar{\mathbf{X}} \rangle) \cdot (\mathbf{O}_d - \langle \mathbf{O} \rangle) / (KD)$, $\sum_{k,d} (\mathbf{X}_d^k - \langle \bar{\mathbf{X}} \rangle) \cdot (\langle \bar{\mathbf{X}} \rangle - \langle \mathbf{O} \rangle) / (KD)$, and $\sum_{k,d} (\mathbf{O}_d - \langle \mathbf{O} \rangle) \cdot (\langle \bar{\mathbf{X}} \rangle - \langle \mathbf{O} \rangle) / (KD)$, disappear identically. The first cross term disappears because $\sum_k \mathbf{X}_d^k = \mathbf{0}$ by definition. The two other cross terms disappear because $\langle \bar{\mathbf{X}} \rangle - \langle \mathbf{O} \rangle$ is a constant so these terms are just means over anomalies. Referring back to section 2a we recognize the terms on the right-hand side as $N\sigma_{fc}^2$, $N\sigma_{tr}^2$, and the squared bias NB^2 .

The resulting decomposition is shown in the top panels of Figs. 10 and 11. The total error grows fast for the first 24 h. For lead times smaller than 100 h σ_{tr}^2 dominates the error while σ_{fc}^2 grows slowly and the two terms become comparable at the largest lead times. The bias is relatively small for all lead times. The daily cycle is apparent in the total error and the bias but not in the variances. Using an observation-based gridded dataset

FIG. 10. Decomposition of the mean square error into variance and bias shown as a function of lead time. For every lead time and year the ensemble mean has been removed from both the ensemble and the observation [see Eq. (9)]. (top) σ_{fc}^2 , σ_{tr}^2 , and squared bias B^2 . (bottom) The ratios σ_{fc}/σ_{tr} and B/σ_{tr} which enter Eqs. (3) and (7). The region north of 20°N.

instead of analysis as proxies for the truth gives similar results except for the smallest lead times where the bias and σ_{tr}^2 now are finite.

Figures 10 and 11 (bottom panels) show the ratios σ_{fc}/σ_{tr} and B/σ_{tr} , which are the quantities relevant for Eqs. (3) and (7). The ratio of variances σ_{fc}/σ_{tr} dominates over B/σ_{tr} for the first few lead times and for lead times larger than approximately 100 h. The model is underdispersed ($\sigma_{fc}/\sigma_{tr} < 1$) except for the first few lead times. For the largest lead times σ_{fc}/σ_{tr} approaches 1 and B/σ_{tr} varies around 0.5 with a daily cycle that is strongest for the smallest region and that reflects the daily cycle in the bias (top panels of Figs. 10 and 11). Thus, even for the longest lead times the requirements for the indistinguishable interpretation are not entirely fulfilled. For lead times between 10 and 100 h the ratio of variances is comparable to B/σ_{tr} for the extratropical region. For the smaller region B/σ_{tr} dominates for these lead times.

5. Relating the properties of the ensemble mean to variance and bias

We now compare the mean errors and ranks from the GEFS reforecasts—described in section 3—with the analytical expressions obtained in section 2. These expressions depend on the bias and the variances of the

FIG. 11. As in Fig. 10, but for the smaller region defined by 40°–45°N, 10°–40°E.

ensemble and observations and they therefore also depend on the lead time. The bias and variances were estimated in section 4.

We first consider the errors of the two individual forecasts shown in Fig. 4. Full and dashed blue curves in this figure are estimates of the error of the ensemble mean [Eq. (2)] and the error of an individual ensemble member [Eq. (1)] using B , σ_{fc} , and σ_{tr} from the GEFS reforecasts (shown in top panels of Figs. 10 and 11). We find a good agreement between the theoretical values and those calculated in section 3 regarding the magnitude and general dependence of lead time.

When values of the ratio of variances σ_{fc}/σ_{tr} and the normalized bias B/σ_{tr} from the GEFS reforecasts (shown in bottom panels in Figs. 10 and 11) are inserted into Eq. (3) we get the values for the relative error of the ensemble mean shown with full blue curves in Figs. 5 and 6.

For individual winters (Fig. 5) we again see a good agreement with a strong increase in the relative error for the first lead times, a maximum around 30 h, and a decrease toward $(1 - \sqrt{2})/\sqrt{2}$ thereafter. For the mean over all 33 forecasts (Fig. 6) the same general picture prevails and now the agreement is excellent. The small relative error around 30 h is due to a small value of the ratio of variances (cf. Fig. 2).

Inserting the values of the ratio of variances σ_{fc}/σ_{tr} and the normalized bias B/σ_{tr} from the GEFS reforecasts into Eq. (7) we get the ranks of the ensemble

mean that are shown with blue curves in Fig. 7. As the analytical expression includes the unknown effective dimension N^* we plot it here for $N^* = 5, 10, 30$, and 100. For the extratropical region a very good agreement is found for $N^* = 100$. For the smaller region N^* around 10 gives the best agreement. These values are in agreement with the effective dimension found with the methods of Wang and Shen (1999).

The daily cycle in both the relative error and the rank of the ensemble mean that is most clearly seen for the smallest region is a consequence of the daily cycle in the bias.

In general we see for the smallest lead times a situation corresponding to truth plus error interpretation where model ensemble members are sampled from a distribution centered about the observations. This is obviously a simple consequence of the initialization and corresponds in our simple model to $B = 0$ and $\sigma_{tr} = 0$ for which Eq. (3) gives a relative error of -1 and a rank of 0 for the ensemble mean. The latter indicates that the ensemble mean is better than all individual ensemble members. For the largest lead times the situation corresponds approximately to the situation found for the climatology of climate models in Christiansen (2018a). As mentioned before this is sometimes called the indistinguishable interpretation (Sanderson and Knutti 2012) where model ensembles and observations all are considered exchangeable corresponding to $B = 0$ and $\sigma_{tr} = \sigma_{fc}$. In this limit the rank of the ensemble mean is again 0 for high dimensions ($\simeq \Theta(-\sqrt{N^*}/2)$) but the relative error is $(1 - \sqrt{2})/\sqrt{2}$.

We note that the very good agreement between the analytical and empirical results is not due to a spurious cancellation of errors as it is also found when the bias and variance are calculated from one-half of the starting days and the error of the ensemble mean from the other half.

Finally, we consider the relation between the error of the ensemble mean and the ensemble spread. We note that the last term in Eq. (8) is smaller than the other terms both because of the factor $1/K$ but also because σ_{fc}^2 is considerably smaller than σ_{tr}^2 for most lead times (Figs. 10 and 11). Under these circumstances it is difficult to observe a relationship between the ensemble mean error and the ensemble spread. This has been confirmed with simple numerical experiments where the relationship only shows up after averaging over many realizations and is in agreement with the weak connection reported in the literature (see, e.g., Kolczynski et al. 2011). The relationship $\delta^2 = N\sigma_{fc}^2/2/N^*$ between the ensemble variance δ^2 and σ_{fc}^2 gives us an estimator of the effective dimension

N^* . Using this estimator we find values broadly in agreement with those used above.

6. Conclusions

In ensemble forecasts of weather or climate the ensemble mean often outperforms most or all of the individual ensemble members. This holds when, for example, the root-mean-square error over many grid points is used as the measure of the difference between models and observations. It was demonstrated by Christiansen (2018a) that this is a simple consequence of the nonintuitive geometric properties of high-dimensional spaces assuming that the ensemble members and observations are drawn from the same distribution. While this assumption that ensemble members and observations are drawn from the same distribution might hold for the climatology of climate models it does not hold for short-term weather forecasts where bias and variance depend strongly on lead time.

In this paper we first considered the properties of the ensemble mean in a simple theoretical model. We then studied the behavior of the ensemble mean as a function of lead times in the GEFS ensemble reforecast system and compared the results to the simple theoretical model. In this simple model the ensemble members and observations are drawn from distributions with different variances and separated by a bias (Fig. 1). Based on this model we extended the work of Christiansen (2018a) and derived an analytical expression for the rank of the ensemble mean relative to the ensemble members. We also found an analytical relation between the ensemble spread and the ensemble mean error. Finally, we derived analytical expressions for the modifications needed when spatial correlations reduce the effective dimension. All these derivations rely heavily on the blessings of high-dimensional spaces, which simplify the calculations as independent vectors are almost always orthogonal and vectors drawn from the same distribution have almost the same length. The analytical expressions depend on the ratio of the bias to the variance of the observations as well on the ratio between the variances of the ensemble and the observations.

For the GEFS reforecast we found considering the extratropical 2-m temperature that the relative error of the ensemble mean increases rapidly for the first lead times to reach a value of 0 around lead times of 30–40 h (Fig. 6). At the time of initialization the relative error is -1 for large ensembles as the ensemble members are just observations polluted with noise. For larger lead times the relative error decreases again and approaches the theoretical value of $(1 - \sqrt{2})/\sqrt{2}$, although it does not fully reach it even for the largest lead times of 384 h.

Similarly, the rank of the ensemble mean is 0 for the lowest lead times, reaches a maximum around lead times of 30–40 h and decreases for larger lead times (Fig. 7). We directly showed that the blessings of high dimensionality are fulfilled for the model ensemble for all lead times: the ensemble members are orthogonal and their lengths are almost identical (Fig. 8).

For comparison with the simple theoretical model we needed the bias and the variances of both observations and models. These quantities were calculated by assuming that the distribution of model ensembles and observations for the same day every year only differ by an offset of the mean. This offset can be different for different years but is the same for model ensembles and observations. Using the variances and bias estimated in this way we found an excellent agreement between the simple theoretical model and the GEFS forecasts regarding the relative error and rank of the ensemble mean as functions of lead time. We also noted a weak relationship between ensemble spread and the ensemble mean error as predicted by the simple theoretical model.

The results above were obtained when forecasting the 2-m temperature on 15 January but we have confirmed that we get similar results for other fields such as the 500-hPa geopotential height and the total cloud cover and for other dates such as 15 July. Focusing on smaller regions—and therefore fewer dimensions—we found that the ensemble mean is still in general better than most individual ensemble members but that the effect as expected is less pronounced (Fig. 7).

In this paper we analyzed the properties of the ensemble mean by applying a simple statistical model. However, the blessings of high dimensionality may also be applied when analyzing more advanced statistical models or other measures of the forecast skill. The blessings of high dimensionality hold for all distance measures, the only requirement is that the diagnostic is of high dimensionality. Such diagnostics include the root-mean-square error and correlations between model and observations when these are calculated over an extended spatial region. Diagnostics which are not of high dimensionality include those based on simple spatial averages.

Acknowledgments. This work is supported by the NordForsk-funded Nordic Centre of Excellence project (Award 76654) Arctic Climate Predictions: Pathways to Resilient, Sustainable Societies (ARCPATH) and by the project European Climate Prediction System (EUCP) funded by the European Union under Horizon 2020 (Grant 776613). NOAA's second-generation global ensemble reforecast dataset was downloaded from <https://www.esrl.noaa.gov/psd/forecasts/reforecast2/download.html>.

REFERENCES

- Annan, J. D., and J. C. Hargreaves, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate*, **24**, 4529–4538, <https://doi.org/10.1175/2011JCLI3873.1>.
- Bishop, C., 2007: *Pattern Recognition and Machine Learning*. 2nd ed. Springer-Verlag, 740 pp.
- Blum, A., J. Hopcroft, and R. Kannan, 2018: *Foundations of Data Science*. Cornell University, 479 pp., <https://www.cs.cornell.edu/jeh/book.pdf>.
- Bowler, N. E., M. J. P. Cullen, and C. Piccolo, 2015: Verification against perturbed analyses and observations. *Nonlinear Processes Geophys.*, **22**, 403–411, <https://doi.org/10.5194/npg-22-403-2015>.
- Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990–2009, [https://doi.org/10.1175/1520-0442\(1999\)012<1990:TENOSD>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2).
- Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, [https://doi.org/10.1175/1520-0493\(1998\)126<2503:IOESOE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2).
- Casanova, S., and B. Ahrens, 2009: On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Mon. Wea. Rev.*, **137**, 3811–3822, <https://doi.org/10.1175/2009MWR2893.1>.
- Cherkassky, V. S., and F. Mulier, 2007: *Learning from Data: Concepts, Theory, and Methods*. 2nd ed. John Wiley & Sons, 560 pp.
- Christiansen, B., 2018a: Ensemble averaging and the curse of dimensionality. *J. Climate*, **31**, 1587–1596, <https://doi.org/10.1175/JCLI-D-17-0197.1>.
- , 2018b: Reply to “Comments on ‘Ensemble averaging and the curse of dimensionality.’” *J. Climate*, **31**, 9017–9019, <https://doi.org/10.1175/JCLI-D-18-0416.1>.
- Delle Monache, L., and R. B. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode. *Atmos. Environ.*, **37**, 3469–3474, [https://doi.org/10.1016/S1352-2310\(03\)00475-8](https://doi.org/10.1016/S1352-2310(03)00475-8).
- Donoho, D. L., 2000: High-dimensional data analysis: The curses and blessings of dimensionality. *AMS' Conf. on Mathematical Challenges of the 21st Century*, Los Angeles, CA, American Mathematical Society, 1–32.
- Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, [https://doi.org/10.1175/1520-0493\(1997\)125<2427:SREFOQ>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2).
- Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, [https://doi.org/10.1175/1520-0493\(2001\)129<2461:AOAPMS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2).
- Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.*, **122**, 703–713, [https://doi.org/10.1175/1520-0493\(1994\)122<0703:TLEAIP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<0703:TLEAIP>2.0.CO;2).
- , 2006: The Liouville equation and atmospheric predictability. *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 59–98, <https://doi.org/10.1017/CBO9780511617652.005>.
- Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866, <https://doi.org/10.1017/CBO9781107415324.020>.
- Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, <https://doi.org/10.1175/JHM-D-14-0008.1>.
- Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, <https://doi.org/10.1029/2007JD008972>.
- Gorban, A. N., and I. Y. Tyukin, 2018: Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philos. Trans. Royal Soc. London*, **376A**, <https://doi.org/10.1098/rsta.2017.0237>.
- , —, and I. Romanenko, 2016: The blessing of dimensionality: Separation theorems in the thermodynamic limit. *IFAC-PapersOnLine*, **49**, 64–69, <https://doi.org/10.1016/j.ifacol.2016.10.755>.
- Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, <https://doi.org/10.1111/j.1600-0870.2005.00103.x>.
- Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, [https://doi.org/10.1175/1520-0493\(1997\)125<1312:VOERSR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2).
- , G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, <https://doi.org/10.1175/BAMS-D-12-00014.1>.
- Kainen, P. C., 1997: Utilizing geometric anomalies of high dimension: When complexity makes computation easier. *Computer Intensive Methods in Control and Signal Processing*, M. Kárný and K. Warwick, Eds., Birkhäuser, 283–294, https://doi.org/10.1007/978-1-4612-1996-5_18.
- Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, <https://doi.org/10.1175/2009JCLI3361.1>.
- Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, N. S. Altman, and A. Deng, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.*, **139**, 3954–3963, <https://doi.org/10.1175/MWR-D-10-05081.1>.
- Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, <https://doi.org/10.1126/science.285.5433.1548>.
- Lambert, S. J., and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dyn.*, **17**, 83–106, <https://doi.org/10.1007/PL00013736>.
- Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, [https://doi.org/10.1175/1520-0493\(1974\)102<0409:TSOMCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2).
- Marzban, C., R. Wang, F. Kong, and S. Leyton, 2011: On the effect of correlations on rank histograms: Reliability of temperature and wind speed forecasts from finescale ensemble reforecasts. *Mon. Wea. Rev.*, **139**, 295–310, <https://doi.org/10.1175/2010MWR3129.1>.
- McKeen, S., and Coauthors, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *J. Geophys. Res.*, **110**, D21307, <https://doi.org/10.1029/2005JD005858>.
- Nicolis, C., R. A. P. Perdigao, and S. Vannitsem, 2009: Dynamics of prediction errors under the combined effect of initial

- condition and model errors. *J. Atmos. Sci.*, **66**, 766–778, <https://doi.org/10.1175/2008JAS2781.1>.
- Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**, 71, <https://doi.org/10.1088/0034-4885/63/2/201>.
- Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.*, **113**, D14209, <https://doi.org/10.1029/2007JD009334>.
- Rougier, J., 2016: Ensemble averaging and mean squared error. *J. Climate*, **29**, 8865–8870, <https://doi.org/10.1175/JCLI-D-16-0012.1>.
- , 2018: Comments on “Ensemble averaging and the curse of dimensionality.” *J. Climate*, **31**, 9015–9016, <https://doi.org/10.1175/JCLI-D-18-0274.1>.
- Sanderson, B. M., and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, **39**, L16708, <https://doi.org/10.1029/2012GL052665>.
- Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res. Atmos.*, **118**, 1716–1733, <https://doi.org/10.1002/jgrd.50203>.
- Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, [https://doi.org/10.1175/1520-0493\(2000\)128<2077:UICAMP>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2).
- Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142**, 1093–1105, <https://doi.org/10.1175/MWR-D-13-00134.1>.
- Talagrand, M., 1995: Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, **81**, 73–205, <https://doi.org/10.1007/BF02699376>.
- Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, [https://doi.org/10.1175/1520-0493\(1997\)125<3297:EFANAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2).
- , O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley & Sons, 137–163.
- Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Mon. Wea. Rev.*, **116**, 2276–2288, [https://doi.org/10.1175/1520-0493\(1988\)116<2276:TROIID>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2276:TROIID>2.0.CO;2).
- van Loon, M., and Coauthors, 2007: Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmos. Environ.*, **41**, 2083–2097, <https://doi.org/10.1016/j.atmosenv.2006.10.073>.
- Wang, X., and S. S. Shen, 1999: Estimation of spatial degrees of freedom of a climate field. *J. Climate*, **12**, 1280–1291, [https://doi.org/10.1175/1520-0442\(1999\)012<1280:EOSDOF>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<1280:EOSDOF>2.0.CO;2).
- Whitaker, J. S., and A. F. Lough, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302, [https://doi.org/10.1175/1520-0493\(1998\)126<3292:TRBESA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2).
- Wilks, D. S., 2011: On the reliability of the rank histogram. *Mon. Wea. Rev.*, **139**, 311–316, <https://doi.org/10.1175/2010MWR3446.1>.