Ensemble Averaging and the Curse of Dimensionality

BO CHRISTIANSEN

Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 28 March 2017, in final form 10 November 2017)

ABSTRACT

When comparing climate models to observations, it is often observed that the mean over many models has smaller errors than most or all of the individual models. This paper will show that a general consequence of the nonintuitive geometric properties of high-dimensional spaces is that the ensemble mean often outperforms the individual ensemble members. This also explains why the ensemble mean often has an error that is 30% smaller than the median error of the individual ensemble members. The only assumption that needs to be made is that the observations and the models are independently drawn from the same distribution. An important and relevant property of high-dimensional spaces is that independent random vectors are almost always orthogonal. Furthermore, while the lengths of random vectors are large and almost equal, the ensemble mean is special, as it is located near the otherwise vacant center. The theory is first explained by an analysis of Gaussian- and uniformly distributed vectors in high-dimensional spaces. A subset of 17 models from the CMIP5 multimodel ensemble is then used to demonstrate the validity and robustness of the theory in realistic settings.

1. Introduction

It has become a key component in climate modeling to compare an ensemble of different model experiments with observations. In such studies, the model mean (also known as the ensemble mean) has often turned out to be closer to the observations than all or most of the individual models for a wide range of climate variables, model systems, and error measures (Lambert and Boer 2001; Gleckler et al. 2008; Pincus et al. 2008; Knutti et al. 2010; Sillmann et al. 2013; Flato et al. 2013). Similar behavior has been described for numerical weather prediction and seasonal forecasts (Toth and Kalnay 1997; Hamill and Colucci 1997; Du et al. 1997; Hagedorn et al. 2005; Krishnamurti et al. 1999; Casanova and Ahrens 2009), as well as for air quality modeling (van Loon et al. 2007; Delle Monache and Stull 2003; McKeen et al. 2005). In the climate modeling context, explanations have been suggested by Annan and Hargreaves (2011) and Rougier (2016). Rougier (2016) noted that the mean squared error of the model mean is less than the average of the mean squared errors of the individual models. However, this does not hold for all metrics, particularly not for the root-mean-square error (RMSE). It also fails in explaining why all or most of the individual model members have larger errors than the model mean. Noting this, Rougier (2016) then suggested an explanation involving systematic biases. However, the previous explanations do not seem compelling. As the discussed behavior is ubiquitous and found across a range of climate variables and metrics, it requires a simple and general explanation. We will here suggest such a general explanation based on the geometric behavior of high-dimensional spaces. This explanation assumes that the observations and the models are drawn from the same distribution, and it also correctly predicts how much better the model mean will be. Furthermore, it reveals why the superiority of the model mean will be more common as the dimension increases.

The curse of dimensionality refers to the complexities of high-dimensional spaces that often defy our intuition based on two and three dimensions (Bishop 2007; Cherkassky and Mulier 2007). Apart from the well-known fact that the number of samples needed to obtain a given density grows exponentially with dimension, there are other less appreciated features of high-dimensional spaces (Blum et al. 2017). For a sphere or cube, the volume increasingly concentrates near the surface when the dimension increases. Independent random vectors in high-dimensional spaces are almost always orthogonal, and almost every point is an outlier in its own projection. Below, we will show that another

Corresponding author: Bo Christiansen, boc@dmi.dk

DOI: 10.1175/JCLI-D-17-0197.1

consequence is that for a sample of random points in a high-dimensional space, the typical distance between two points is significantly larger than the typical distance between one point and the mean of the sample.

If one considers a spatial field of, for example, the surface temperature, rainfall, or the 500-hPa geopotential height, and uses, for example, the root-mean-square error over the grid points as a measure of the difference between models and observations, then one operates in a space with the dimension given by the number of grid points. The number of grid points may be very large, though in many cases, nearby grid points are not independent, and the correct dimension should reflect the effective number of independent points. This number of spatial degrees of freedom is difficult to estimate (Christiansen 2015) and depends on the field, the time scale, and the region under consideration (Jones and Briffa 1996; North et al. 2011). It is on the order of 50–100 for monthly surface temperatures in the Northern Hemisphere (Wang and Shen 1999; Bretherton et al. 1999) and probably an order of magnitude larger for daily means. For daily precipitation, the number of spatial degrees of freedom is much larger (Moron et al. 2006; Benestad 2013).

The relevant effect of the high-dimensional space is easy to demonstrate in any programming language. Draw 30 independent vectors with 100 elements, and calculate the mean vector. Draw an additional vector, also with 100 elements, to represent the observations. Calculate the errors with respect to the observations for the 30 individual vectors and for the mean vector. Observe that the error of the mean vector is (almost always) smaller than all the errors of the individual vectors.

The theory is explained and illustrated by simple examples in section 2. In section 3, we extend the simple examples and compare our study with previous work. Section 4 demonstrates that the predictions of the theory hold when a subset of phase 5 of the Coupled Model Intercomparison Project (CMIP5) multimodel ensemble is considered.

2. Simple analysis and theory

Consider K vectors in N-dimensional space $\mathbf{x}^k$, $k = 1, 2, \ldots K$, where $\mathbf{x}^k = (x_1^k, x_2^k, \ldots x_N^k)$. These vectors represent K models, and $\mathbf{x}^k$ could, for example, be a spatial field, with N then being the number of grid points. The model mean $\overline{\mathbf{x}} = \sum_{i=1}^{K} \mathbf{x}^k/K$ is then a vector of dimension N. The observations are also given by a vector of dimension N, $\mathbf{z} = (z_1, z_2, \ldots z_N)$. We now make the simple assumption that the observations and the models are drawn independently from the same distribution. This describes the situation when the only differences between the individual models, and between the observations and the individual models, are due to internal variability. Thus, the models are assumed to be without bias and with a realistic variability.

We first assume that all components are drawn from a standard Gaussian distribution (zero mean and unit variance). We begin by considering the Euclidean metric $\|\mathbf{w} - \mathbf{v}\| = \sqrt{\sum_{n=1}^{N}(w_n - v_n)^2}$. In one dimension, the distance becomes simply the absolute value of the difference $|w_1 - v_1|$. The length of the kth model is $\|\mathbf{x}^k\|$, and its error is $\|\mathbf{x}^k - \mathbf{z}\|$. Likewise, the error of the model mean is $\|\overline{\mathbf{x}} - \mathbf{z}\|$. Note that this error differs by a factor of $\sqrt{N}$ from the usual root-mean-square error. This is a convenient choice, as the error then equals the distance from the model to the observations. This choice does not influence the comparison of errors of the individual models with errors of the model mean.

Figure 1 shows the situation in the one-dimensional case ($N = 1$) with 30 models ($K = 30$). Here, 30 independent random numbers are drawn to represent the models, and one additional number is drawn to represent the observations. In the one-dimensional case, the distances are simply the absolute values, and the errors are the absolute values of the differences. The model mean is closer to zero than most of the individual models, but the histogram of the lengths of the individual models peaks at zero (left panel). Considering the errors (right panel), the model mean therefore does not stick drastically out from the individual models. However, for $N = 1$, there is already a tendency for the model mean to be better than most individual models. This can be understood from the assumption that both models and observations are drawn from the same distribution. Let us assume that the observation is positive. For large K, the model mean z will be close to zero. If the distribution is symmetric, then around half of the individual models will be negative and have larger errors than the model mean. In fact, all models larger than $2z$ will also have larger errors than the model mean. Analytical integrations show that on average, for large K, the model mean will be better than around 65% [5/8 and $1 - \arctan(2)/\pi$ for uniformly and Gaussian-distributed numbers] of the individual ensemble members.

For higher dimensions, the situation changes drastically (Fig. 2). Here, 30 independent random vectors of 50 dimensions are drawn representing the models, and one additional 50-dimensional vector is drawn to represent the observations. Now the distribution of the lengths of the individual models does not peak at zero, but at a finite value. This distribution is quite narrow, and the model mean is closer to zero and totally separated from the individual models (Fig. 2, left panel). In the right panel of Fig. 2, we see the desired result: the

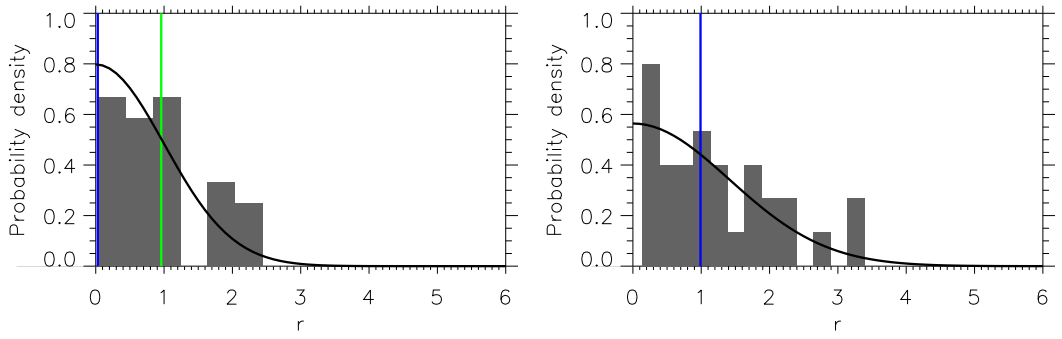

FIG. 1. (left) The histogram of lengths of 30 independent random numbers drawn from a standard Gaussian distribution. The length of the mean is shown as the blue vertical line, and the length of the observations is the green vertical line. The full curve is the distribution $P(r)$ from Eq. (1). (right) Errors of the 30 numbers with respect to the observations (an additional random number). The full curve is the distribution $P(r/\sqrt{2})/\sqrt{2}$. The error of the mean with respect to the observations is shown as the blue vertical line.

error of the model mean is smaller than the errors of the individual models. This is due to two properties. The first is that the model mean is near zero, so the error of the model mean has the value of a typical length of an individual model. The second property is that, in contrast, the errors of the individual models are larger than the typical lengths of the individual models.

The dependence on the dimension N is shown in Fig. 3. For each N, $K = 30$ independent random N-dimensional vectors are drawn representing K models. One additional random N-dimensional vector is drawn to represent the observations. The errors of the individual models are calculated, and the mean and standard deviation are shown with the cyan curves. The error of the model mean is also calculated (red curves). As above, this error is comparable to the lengths of the individual models, as the model mean is situated near zero (green curve), and the individual models and the observations are drawn from the same distribution. For N larger than about 15, the errors of the individual models are significantly larger than the error of the model mean. As the dimension N increases, both the errors of the individual models and the error of the model mean increase. However, the errors of individual models increase fastest, and the standard deviations remain almost constant. Thus, when the dimension increases, the model mean becomes an increasingly better estimate of the observations than the individual models.

We can understand this analytically. The surface area of a hypersphere with radius r in N dimensions is $S_{N-1} = 2\pi^{N/2} r^{N-1}/\Gamma(N/2)$.[1] The standard Gaussian probability distribution is $P(\mathbf{x}) = (2\pi)^{-N/2} \exp(-\sum_{n=1}^{N} x_n^2/2)$. So as a function of r,

$$P(r) = S_{N-1} P(\mathbf{x}) = \frac{2^{1-N/2}}{\Gamma(N/2)} r^{N-1} \exp(-r^2/2). \quad (1)$$

The maximum of $P(r)$ is reached for $r = \sqrt{N-1}$, and the width (standard deviation) of the peak is $\sqrt{2}/2$, independent of N. Thus, for increasing N, the models lie in an annulus with radius r in $\sqrt{N-1} \pm \sqrt{2}/2$. The mean distance between the center and the observations is therefore $\sqrt{N-1}$. As the model mean will be situated near the center (see next paragraph), this is also the mean distance between the model mean and the observations (the model mean error).

The mean distance between two points on the surface of the unit hypersphere (mean chord length) is $\sqrt{2}$ in the limit of large N. This is a special case of the general fact that in high-dimensional space, two independent random vectors are almost always orthogonal. Thus, the mean distance between individual models is $\sqrt{2}\sqrt{N-1}$, which is also the mean distance between the observations and the individual model minus the individual model errors. Because the mean radius of the annulus increases while the width stays the same for increasing N, the model mean will become more separated from the individual models.

Let us also consider how the length of the model mean $\bar{\mathbf{x}}$ approaches zero. We have $\|\bar{\mathbf{x}}\|^2 = \sum_{n=1}^{N} \bar{x}_n^2$. Each $\bar{x}_n$ is Gaussian-distributed with variance $1/K$, so $\sqrt{K}\|\bar{\mathbf{x}}\|$ is χ distributed with N degrees of freedom. As the mean and variance of a χ distribution with N degrees of freedom are $\mu = \sqrt{2}\Gamma[(N+1)/2]/\Gamma(N/2)$, and $N - \mu^2$, it is seen that $\|\bar{\mathbf{x}}\|$ has a mean of $\mu/\sqrt{K}$ and a standard deviation of $\sqrt{(N-\mu^2)/K}$. Note that $\mu \approx \sqrt{N}$ for large N, so $\|\bar{\mathbf{x}}\| \approx \sqrt{N/K}$. We see that the required number of models K for $\|\bar{\mathbf{x}}\|$ to be small

[1] $\Gamma(N/2) = (N-1)!!\sqrt{\pi}/2^{(N-1)/2}$, where $m!! = m(m-2)\ldots 2$ for m even and $m!! = m(m-2)\ldots 1$ for m odd.

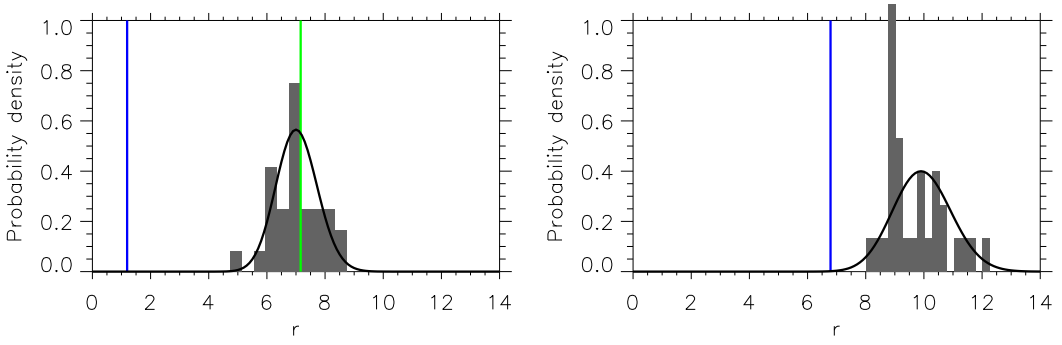

FIG. 2. As in Fig. 1, but for $N = 50$, that is, now based on 30 independent random 50-dimensional vectors to represent the models and an additional random 50-dimensional vector to represent the observations. Each x_i^k is drawn from a standard Gaussian.

relative to $r = \sqrt{N-1}$, for which $P(r)$ peaks, is independent of N.

The results are not particular for the Gaussian distribution, as we have confirmed by repeating the numerical experiments with t-distributed and uniformly distributed numbers. This can also be argued from the central limit theorem, which implies that the squared lengths of random vectors will be Gaussian-distributed for large N independent of the distributions of the individual components. In the non-Gaussian cases, further analytic results are difficult to obtain, but note that a cube in N dimensions will have 2^N vertices. For a unit hypercube, the volume is one, but when N increases, the volume will be increasingly concentrated in a thin layer near the surface. The vertices will increase their distance from the center as $\sqrt{N}/2$, while the volume of the central N-dimensional sphere inscribed in the cube will decrease very fast as 2^{-N}. Thus, the hypercube will be very spiky and can be perceived, when projected on a three-dimensional space, to have the form of a sea urchin or porcupine (Hecht-Nielsen 1990). The model mean is therefore special, as it is located near the center, while the individual models are located in the spikes.

Above, we have shown that the model mean often outperforms the individual ensemble members, but the results are also valid when considering the median of the models instead. Figure 4 shows results when all components are drawn from a uniform distribution (over $[-1/2, 1/2]$) and the median of the models is considered instead of the mean.

The results are also not sensitive to the distance function used (and hence, not to the error metric) and do not, in particular, depend on the convexity of the distance function. This has been confirmed by repeating the numerical experiments with the distance functions $\sum_{n=1}^{N}|x_n - y_n|$ and $[\sum_{n=1}^{N}(x_n - y_n)^2]^{1/4}$. Also, the spatial correlations could be used.

In the validation of model ensembles (Gleckler et al. 2008; Sillmann et al. 2013; Flato et al. 2013), the errors of the ensemble members are often given relative to their median error. Thus, a value of 0.1 means that the specific ensemble member has an error 10% larger than the median error of the model ensemble. Remarkably, the relative error for the model mean is consistently around -0.3 for a large range of fields and measures. This is clearly seen in Figs. 3 and 7 in Gleckler et al. (2008), Fig. 10 in Sillmann et al. (2013), and Figs. 9.7 and 9.37 in Flato et al. (2013). Thus, the model mean is very often 30% better than the typical individual model.

While this might seem surprising, it is a simple consequence of the theory above. In high-dimensional space, two independent random points are almost always orthogonal, and they are almost at the same distance from the center where the model mean is situated. Thus, the observations, any ensemble member, and the model mean form an isosceles (two equal sides) right triangle, and the relative error of the model mean is, therefore, $(1 - \sqrt{2})/\sqrt{2} \approx -0.29$.

Figure 5 (top) shows the relative error of Gaussian-distributed values as a function of the dimension N. Here we have chosen the number of models K to be 17 to be consistent with the next session. As the dimension N increases, the range of individual model errors decreases. However, the error of the model mean is close to -0.30 for N larger than 5–10. As a result, the error of the model mean is lower than the errors of all individual models for N larger than approximately 40. The relative errors do not depend on the width of the distribution, and very similar results are obtained for uniformly distributed values (Fig. 5, bottom).

Thus, the curse of dimensionality does not only predict that the model mean is better, but also how much better.

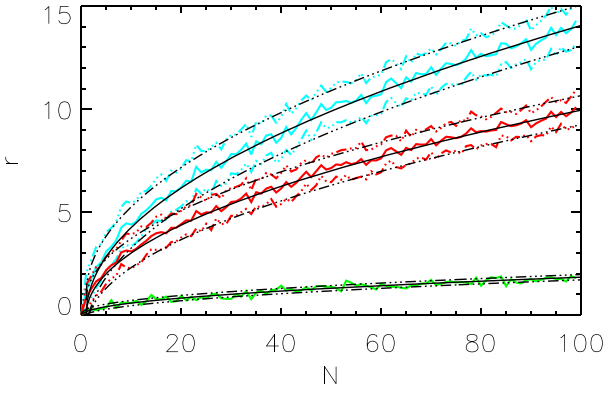

FIG. 3. The situation as a function of N. For each N, 31 random Gaussian-distributed N-dimensional vectors are drawn to represent the $K = 30$ models and the observations. The errors of the individual models (full cyan curve) and of the model mean (full red curve) are shown as a function of the dimension N. The dashed curves show $\pm$ one standard deviation. The green curve is the distance from the model mean to zero. Black curves are analytic results as described in the text. From top: $\sqrt{2(N-1)} \pm 1$, $\sqrt{N-1} \pm \sqrt{2}/2$, and $\mu/\sqrt{K} \pm \sqrt{(N-\mu^2)/K}$, with $\mu = \sqrt{2}\Gamma[(N+1/2)]/\Gamma(N/2)$.

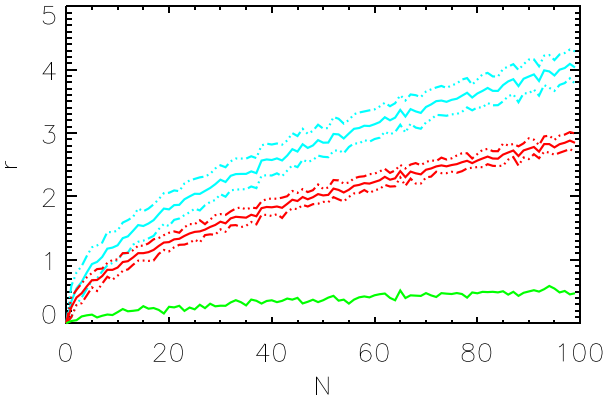

FIG. 4. As in Fig. 3, but for uniformly distributed data and using the model median instead of model mean. No analytic results are obtained in this case.

3. Extending the model and previous work

In the previous section, both the ensemble members and the observations were independently drawn from the same Gaussian distribution, $N(\mathbf{0}, \sigma\mathbf{I})$, where $\mathbf{0}$ and $\mathbf{I}$ are the N-dimensional zero vector and the N-dimensional identity matrix, respectively. Obviously, the choice of zero means does not change the generality of the results. Choosing different variances for the different dimensions also does not change the results, as long as the total variance is not dominated by a few dimensions.

We now more systematically extend the model to allow for a common bias between models and observations and for models and observations to have different variances. Thus, the ensemble members are drawn from $N(\mathbf{b}, \sigma^2_{\text{mod}}\mathbf{I})$ and the observations from $N(\mathbf{0}, \sigma^2_{\text{obs}}\mathbf{I})$. With $\|\mathbf{b}\| = B\sqrt{N}$, and using the general results about the length and orthogonality for large N and K, we get

$$\|\mathbf{z} - \overline{\mathbf{x}}\|^2 = B^2N + \sigma^2_{\text{obs}}(N-1), \tag{2}$$

and

$$\|\mathbf{z} - \mathbf{x}^k\|^2 = B^2N + \sigma^2_{\text{obs}}(N-1) + \sigma^2_{\text{mod}}(N-1). \tag{3}$$

For the relative error of the model mean, we get

$$\frac{\sqrt{1 + \left(B/\sigma_{\text{obs}}\right)^2} - \sqrt{1 + \left(B/\sigma_{\text{obs}}\right)^2 + \left(\sigma_{\text{mod}}/\sigma_{\text{obs}}\right)^2}}{\sqrt{1 + \left(B/\sigma_{\text{obs}}\right)^2 + \left(\sigma_{\text{mod}}/\sigma_{\text{obs}}\right)^2}}. \tag{4}$$

The relative error is shown as a function of $\sigma_{\text{mod}}/\sigma_{\text{obs}}$ and B/σ_{obs} in Fig. 6. The relative error changes only slowly with both parameters: for example, for $B/\sigma_{\text{obs}} = 1$ and $\sigma_{\text{mod}}/\sigma_{\text{obs}} = 1$, the relative error is still -20%. When the bias grows, the relative error increases (approaches zero), making the error of the ensemble mean closer to the errors of the individual ensemble members. The relative error also increases when $\sigma_{\text{mod}}/\sigma_{\text{obs}}$ decreases, as could be the case if the models are imperfectly tuned. However, it is important to note that as for the simpler model in the last section, the ensemble mean will still outperform all ensemble members as follows from comparing Eqs. (2) and (3). This happens because the distribution of $\|\mathbf{x}^k\|$ is an annulus with fixed width but with a radius that increases with N.

Van Loon et al. (2007) observed that when observations and ensemble members are drawn from a one-dimensional distribution $N(0, \sigma)$, the expected value of the mean squared error of the ensemble mean is σ^2, and the expected value of the mean squared error of an individual ensemble member is $2\sigma^2$. This corresponds to the one-dimensional model considered in the beginning of section 2. However, as shown there, one cannot conclude from this result that the ensemble mean is closer to the observations than any individual ensemble member.

Annan and Hargreaves (2011) assumed, as in the present paper, a model where observations and ensemble members are drawn from the same distribution. Based on numerical analysis, they found that for high dimensions, the errors of the individual ensemble members are a factor of $\sqrt{2}$ larger than the error of the ensemble mean. However, they only provided heuristic arguments, failing to realize that random vectors in high dimensions are almost always orthogonal. The factor of $\sqrt{2}$ was previously also noted by Du et al. (1997), based on theoretical arguments by Leith (1974).

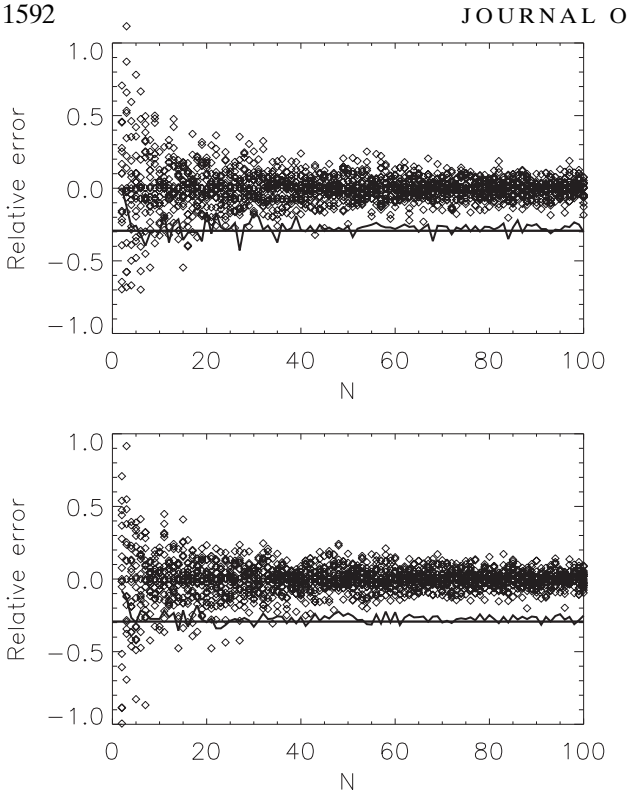

FIG. 5. The error relative to the median error for $K = 17$ as a function of dimension N. For each N, 18 random (top) Gaussian- or (bottom) uniformly distributed N-dimensional vectors are drawn to represent the $K = 17$ models and the observations. Open squares are the 17 individual ensemble members. Black curve is the model mean.

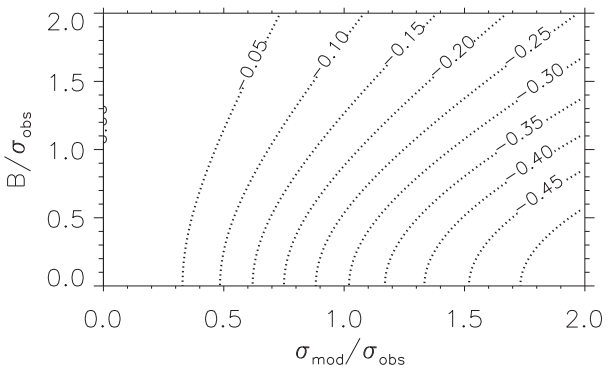

FIG. 6. The relative error as a function of the relative standard variation $\sigma_{\text{mod}}/\sigma_{\text{obs}}$ and the relative bias B/σ_{obs} ($B = \|\mathbf{b}\|/\sqrt{N}$) when observations are drawn from $N(\mathbf{0}, \sigma^2_{\text{obs}}\mathbf{I})$ and ensemble members from $N(\mathbf{b}, \sigma^2_{\text{mod}}\mathbf{I})$.

Rougier (2016) studied the asymptotic situation for large N. He considered the model where the ensemble members are drawn from $N(\mathbf{z} + \mathbf{b}_k, \sigma_{\text{mod}}\mathbf{I})$, where $\mathbf{b}_k = b_k\mathbf{I}$, that is, the ensemble members have biases along the same direction but of different lengths. If $b_k = 0$ for all k, this reduces to the situation where the ensemble members are scattered randomly around the observation. In this situation, the error of the ensemble mean will trivially be smaller than the errors of the individual ensemble members for large K (the relative error will be -1). This limit is also found for large $\sigma_{\text{mod}}/\sigma_{\text{obs}}$ in the model above [Eq. (4)]. Rougier (2016) studied the conditions when the b_k values are small enough for the error of the ensemble mean to still be smaller than the errors of all the individual ensemble members. For large K and N, one such condition is $|b_k| < \sigma_{\text{mod}}$ for all k. In the spirit of the present paper, we have in the limit of large K that $\overline{\mathbf{x}} = \overline{b} + \mathbf{z}$ and $\|\overline{\mathbf{x}} - \mathbf{z}\|^2 = \overline{b^2}N$ (overline denoting ensemble mean). And in the limit of large N (because of the orthogonality), $\|\mathbf{x}^k - \mathbf{z}\|^2 = N\sigma^2_{\text{mod}} + b_k^2 N$. From this follows Rougier's "Result 2" in the limit of large K. We also find that the relative error of the ensemble mean is $\overline{b} - \sqrt{\sigma^2_{\text{mod}} + \overline{b^2}}/\sqrt{\sigma^2_{\text{mod}} + \overline{b^2}}$. Note that in contradiction with the model suggested in the present paper, this does not predict that the relative error of the ensemble mean is $(1 - \sqrt{2})/\sqrt{2}$, as found in the analyses of climate models.

We close this section with the related question considered by, for example, Hagedorn et al. (2005): How can a poor model add skill? The curse of dimensionality might help explain this. As we have seen in the limit of large N, the center is a special point that is vacant of both observations and individual ensemble members. Therefore, if the new model drags the ensemble mean closer to zero, then it might add skill to the ensemble mean. Assume we already have K models drawn from $N(\mathbf{0}, \sigma^2_{\text{mod}}\mathbf{I})$. The mean is $\overline{\mathbf{x}} = \sum_k \mathbf{x}^k/K$. In high dimensions, all $\mathbf{x}^k$ have the length r and are orthogonal, so $\|\overline{\mathbf{x}}\|^2 = r^2/K$. Adding a model $\mathbf{x}_{\text{new}}$ with length r_{new} gives the mean $\sum_k(\mathbf{x}^k + \mathbf{x}_{\text{new}})/(K+1)K$ with squared length $(Kr^2 + r^2_{\text{new}})/(K+1)^2$. The new length is less than the old when $r_{\text{new}}/r < \sqrt{(2K+1)/K}$. Thus, a model that is poor in the sense that its error is larger than the errors of the other models can still add skill to the mean model, as long as its error does not exceed the errors of the other models with a factor of more than $\sqrt{2}$.

4. The CMIP5 multimodel ensemble

Here, we investigate the errors in a subset of 17 climate models from the CMIP5 (Taylor et al. 2012). We will show how the superiority of the model mean depends on the degrees of freedom. We will also show that there is symmetry between models and observations; it is not some particularity of either the observations (observational errors) or the models (biases) that make the model mean superior.

TABLE 1. The CMIP5 models included in this study. Details can be found in Flato et al. (2013).

Modeling center/group	ID	Model name
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	CSIRO-BOM	ACCESS1.0
Beijing Climate Center, China Meteorological Administration, China	BCC	BCC_CSM1.1
College of Global Change and Earth System Science, Beijing Normal University, China	GCESS	BNU-ESM
Canadian Centre for Climate Modelling and Analysis, Canada	CCCma	CanCM4
Centro Euro-Mediterraneo per I Cambiamenti Climatici, Italy	CMCC	CMCC-CESM
EC-EARTH consortium	EC-EARTH	EC-EARTH
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences, and Center for Earth System Science, Tsinghua University, China	LASG-CESS	FGOALS-g2
First Institute of Oceanography, State Oceanic Administration, China	FIO	FIO-ESM
NOAA Geophysical Fluid Dynamics Laboratory, United States	NOAA GFDL	GFDL CM3
NASA Goddard Institute for Space Studies, United States	NASA GISS	GISS-E2-R-CC
Met Office Hadley Centre, United Kingdom	MOHC	HadGEM2-ES
Institute of Numerical Mathematics, Russia	INM	INM-CM4.0
L'Institut Pierre-Simon Laplace, France	IPSL	IPSL-CM5A-LR
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies, Japan	MIROC	MIROC-ESM
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	MIROC	MIROC5
Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology), Germany	MPI-M	MPI-ESM-MR
Norwegian Climate Centre, Norway	NCC	NorESM1-M

The models that are chosen from the major modeling centers are briefly identified in Table 1. Details can be found in Flato et al. (2013; their Table 9.A.1). We use monthly means from the historical experiments, and from each model, we use one ensemble member (r1i1p1). As observations, we use NCEP–NCAR data (Kalnay et al. 1996). We focus on the near-surface temperature (TAS). The model data, which come in different horizontal resolutions, are interpolated to the horizontal NCEP resolution (2.5° × 2.5° longitude–latitude grid) using a simple nearest-neighbor procedure.

The monthly and annual mean climatologies of the SAT are calculated using the period 1980–2005. The climatologies are calculated for all grid points, for the zonal means, and for the Northern Hemisphere (NH) mean. The errors are then calculated as the root-mean-square error over the included space and time, as in Gleckler et al. (2008). Thus, for monthly zonal mean climatologies, the root-mean-square errors are calculated over the 73 latitudes and the 12 calendar months.

The resulting relative errors (Fig. 7, top five rows) show the expected behavior. When most degrees of freedom are included, as for monthly climatologies for all grid points, the model mean has a relative error around −0.3 and smaller errors than all the individual models. When the number of degrees of freedom is reduced, as for the annual and monthly climatologies for the NH mean, several individual models are better than the model mean. It is difficult to estimate the number of degrees of freedom in the monthly climatology, but we note from Fig. 5 that the full separation between model mean and individual models can be expected for N around 35. Also note that the scatter of the model mean between −0.4 and −0.2 found in Fig. 7 is consistent with Fig. 5 for N less than 40. Figure 7 (sixth row) also shows the root-mean-square error of monthly anomalies of the TAS over the period 1980–2005 and over all grid points. This example has a huge number of degrees of freedom, and the model mean has a relative error near −0.3 and is well separated from the individual models. Note that this happens even though the models are initialized a long time before 1980 and any forecast skill is lost.

The errors in the upper part of Fig. 7 were based on models that had been bias corrected by adding in each grid point the difference between the long-term means of the model and the observations. We have confirmed that the bias correction had no effect on the results regarding the difference between the model mean and the individual models. The two rows in the middle of Fig. 7 demonstrate this.

The basic assumption that separates the high-dimensional explanation from explanations involving model biases is that the observation and the models are drawn from the same distribution. We test this symmetry by exchanging the observations with one of the 17 models and calculating the new model mean and the new set of errors. We can do this swapping for each of the 17 models, therefore giving us 17 test cases. Thus, in the ith case, the observations and the

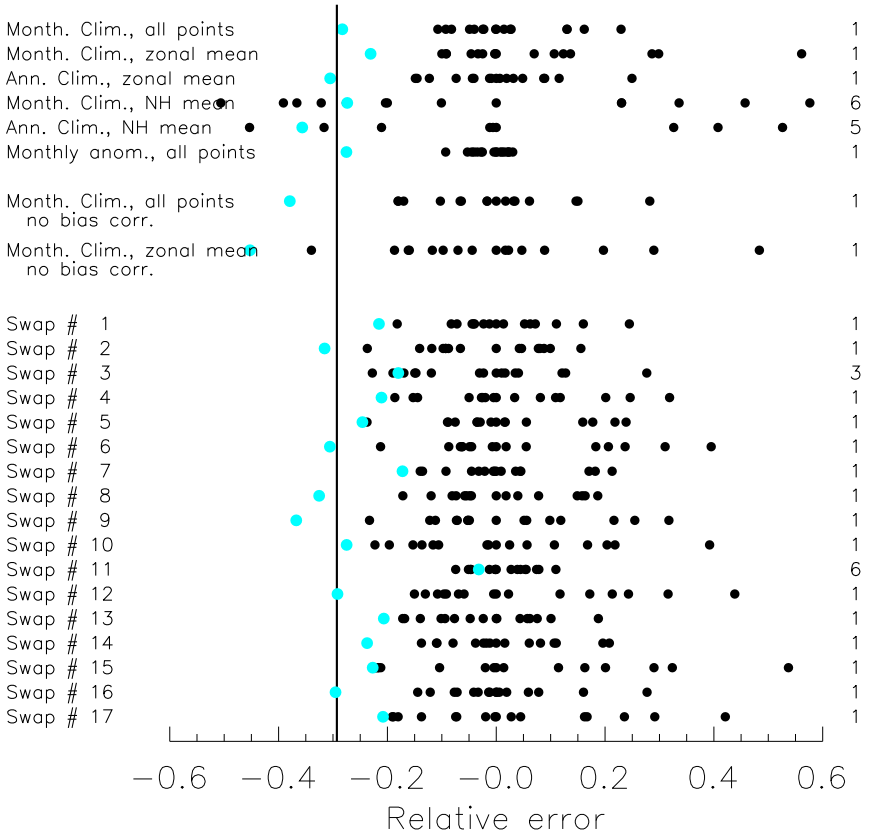

FIG. 7. Relative error of 17 CMIP5 models for different diagnostics of the near-surface temperature. Black dots are the 17 individual ensemble members. Cyan dots are the model mean. Numbers to the right of each row are the rank of the model mean. The first five rows show the RMSE of the seasonal and annual climatology over all grid points, the zonal mean, and the NH mean. The sixth row shows the errors of monthly anomalies over the period 1980–2015 over all grid points. Rows 7 and 8 are equivalent to the two first rows, but now no bias correction has been applied. The last 17 rows show results of swapping each of the individual models with observations. The diagnostic is the same as in the first row: the RMSE of seasonal climatology over all grid points.

*i*th model are swapped so the model ensemble now includes 16 real models and the observations. The model mean is now calculated from this ensemble, and the errors are calculated relative to the *i*th model. As in the upper row of Fig. 7, the root-mean-square errors are calculated over the monthly climatology and all grid points. The results are shown in the lower part of Fig. 7. We see that in almost all cases (15 out of 17), the model mean has lower errors than the individual models. We also see that the relative errors are close to −0.29, except for one case. This demonstrates that the superiority of the model mean is an inherent property of the averaging process and is not connected to any model bias or imperfect observations.

Note that the spread of the relative errors of the individual models in the swap experiments is larger than for the original data (first row in Fig. 7). This suggests that deviations from the hypothesis—observations and models are drawn from the same distribution—do exist. We find that this hypothesis is sufficient to explain the superiority of the model mean, but the swap experiment should not be seen as a general test of the interchangeability of models and observations, as it probably lacks statistical power. Adding an additional model closely related to one of the existing models (such as an r2i1p1 ensemble member or a model from the same family as an already included model) will not change the general results about the model mean. However, when one of these closely related models is swapped with the observations, its twin will have a much smaller relative error than the rest.

5. Conclusions

We have given a general explanation for the observation that the ensemble mean often outperforms all the

individual ensemble members. This explanation does not need any assumptions beyond the assumption that observations and models are drawn from the same distribution. The explanation holds even in the simplest examples of Gaussian- or uniformly distributed data. The explanation is based on the nonintuitive properties of high-dimensional spaces. In such spaces, randomly chosen vectors are almost always orthogonal. Another important property is that the largest part of the volume is concentrated in a thin layer, with a distance to the center that increases with the dimension. On the other hand, the ensemble mean is exceptional, as it is located close to the center. This explanation also predicts the observation that the error of the ensemble mean often is 30% lower than the median error of the individual ensemble members.

Studying a subset of the CMIP5 multimodel ensemble, we confirmed previously published results that the model mean is, in general, superior to individual models and that the relative error of the model mean is around −30%. Here, we considered monthly and annual climatologies, as well as monthly anomalies of the near-surface temperatures. Similar results are found for other fields, such as zonal wind and geopotential height at different pressure levels.

We found, as expected from the theory, that the superiority of the ensemble mean is positively correlated with the number of degrees of freedom. We also found that the superiority of the ensemble mean does not depend on the application of simple bias correction. The validity—in the present context—of the assumption that observations and models are drawn from the same distribution was investigated by swapping individual models with the observations and repeating the calculations of the model mean and the relative errors. Again, we found that the new model mean has lower errors then the individual models and that the relative errors are close to −0.29. This demonstrates that the lower errors of the model mean are due to the averaging process itself and are not connected to imperfections in the models or observations.

More meaningful comparisons of individual models and model means could be done after applying dimension-reduction techniques such as principal component analysis.

Acknowledgments. The NCEP–NCAR reanalysis data were provided by the NOAA/CIRES Climate Diagnostics Center (Boulder, Colorado) from their website (http://www.cdc.noaa.gov/). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1) for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

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.

Benestad, R. E., 2013: Association between trends in daily rainfall percentiles and the global mean temperature. *J. Geophys. Res. Atmos.*, **118**, 10 802–10 810, https://doi.org/10.1002/jgrd.50814.

Bishop, C., 2007: *Pattern Recognition and Machine Learning*. 2nd ed. Information Science and Statistics Series, Springer, 738 pp.

Blum, A., J. Hopcroft, and R. Kannan, 2017: Foundations of data science. Cornell University, 454 pp., https://www.cs.cornell.edu/jeh/book.pdf.

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.

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, 538 pp.

Christiansen, B., 2015: The role of the selection problem and non-Gaussianity in attribution of single events to climate change. *J. Climate*, **28**, 9873–9891, https://doi.org/10.1175/JCLI-D-15-0318.1.

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.

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.

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.

Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

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.

Hecht-Nielsen, R., 1990: *Neurocomputing*. Addison-Wesley, 433 pp.

Jones, P. D., and K. R. Briffa, 1996: What can the instrumental record tell us about longer timescale paleoclimatic reconstructions?

Climatic Variations and Forcing Mechanisms of the Last 2000 Years, P. D. Jones, R. S. Bradley, and J. Jouzel, Eds., NATO ASI Series, Vol. 41, Springer, 625–644, https://doi.org/10.1007/978-3-642-61113-1_30.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

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.

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.

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.

Moron, V., A. W. Robertson, and M. N. Ward, 2006: Seasonal predictability and spatial coherence of rainfall characteristics in the tropical setting of Senegal. *Mon. Wea. Rev.*, **134**, 3248–3262, https://doi.org/10.1175/MWR3252.1.

North, G. R., J. Wang, and M. G. Genton, 2011: Correlation models for temperature fields. *J. Climate*, **24**, 5850–5862, https://doi.org/10.1175/2011JCLI4199.1.

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.

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.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

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.

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.