Understanding the Distribution of Multimodel Ensembles

BO CHRISTIANSEN

Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 16 March 2020, in final form 17 July 2020)

ABSTRACT

When analyzing multimodel climate ensembles it is often assumed that the ensemble is either truth centered or that models and observations are drawn from the same distribution. Here we analyze CMIP5 ensembles focusing on three measures that separate the two interpretations: the error of the ensemble mean relative to the error of individual models, the decay of the ensemble mean error for increasing ensemble size, and the correlations of the model errors. The measures are analyzed using a simple statistical model that includes the two interpretations as different limits and for which analytical results for the three measures can be obtained in high dimensions. We find that the simple statistical model describes the behavior of the three measures in the CMIP5 ensembles remarkably well. Except for the large-scale means we find that the indistinguishable interpretation is a better assumption than the truth centered interpretation. Furthermore, the indistinguishable interpretation becomes an increasingly better assumption when the errors are based on smaller temporal and spatial scales. Building on this, we present a simple conceptual mechanism for the indistinguishable interpretation based on the assumption that the climate models are calibrated on large-scale features such as, e.g., annual means or global averages.

KEYWORDS: Error analysis; Statistical techniques; Climate models; Ensembles; Model errors

1. Introduction

The analysis of ensembles of climate models has become a standard tool to estimate the effect of emission scenarios or other perturbations on the climate. Community efforts to create and store multimodel ensembles of climate model experiments with standardized forcings and outputs have been undertaken in the CMIP projects (Taylor et al. 2012). When analyzing such ensembles, the mean of the ensemble is often interpreted as the best estimate of the climate and the spread of the ensemble as an estimate of the uncertainty. Increasing the size of the ensemble would therefore be a natural way to obtain more precise estimates. Naively we might argue that the uncertainty on the ensemble mean would decrease as the inverse square root of the ensemble size. This argument is, however, based on two assumptions. The first assumption is that the ensemble is centered around the truth, i.e., the models can be seen as

the truth polluted by noise. This is equivalent to the concept of measurement noise and is called the truth centered or the truth-plus-error interpretation. The second assumption is that the models can be treated as independent. If they are not fulfilled, these assumptions will lead to an underestimation of the spread and therefore to overconfident projections of the climate [see Steinschneider et al. (2015) for a concrete example].

The interest in the potential independence of climate models has increased over the last decade. This includes attempts to estimate the effective number of independent models (Jun et al. 2008; Pennell and Reichler 2011; Leduc et al. 2016), to weight the models according to their amount of independence (Sansom et al. 2013; Knutti et al. 2017; Sanderson et al. 2017), and to select independent subsets (Evans et al. 2013; Herger et al. 2018). See also the review by Abramowitz et al. (2019). There are strong a priori reasons to believe that climate models, such as those compiled in the CMIP archives, are not all independent. Some models from the same institute are closely related variants that only differ in, e.g., resolution. Models from different institutions share code and/or parameterizations and even when such components are developed independently they are often

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

Corresponding author: Bo Christiansen, boc@dmu.dk

built on the same basic physical understanding and the same simplifications (Tebaldi and Knutti 2007). There is, in fact, a close connection between model genealogy and the proximity of the model simulations (Masson and Knutti 2011; Knutti et al. 2013; Boé 2018). The effective number of independent models are often found to be less than half of the nominal ensemble size (Pennell and Reichler 2011; Leduc et al. 2016; Sanderson et al. 2015).

Several methods have been applied to estimate the independence (Pirtle et al. 2010; Abramowitz et al. 2019) but a rigorous definition of independence is often not provided (Annan and Hargreaves 2017). As pointed out in, e.g., Stephenson et al. (2012) we need simplifying, transparent, and defensible statistical frameworks for ensembles of climate models. There are two simple null hypotheses for an ensemble of independent models that have been applied—sometimes tacitly—in previous work. The first is the truth centered interpretation mentioned above. The second is the indistinguishable interpretation in which the observations and the models all are assumed to be drawn from the same distribution. The choice of null hypothesis is important as the two interpretations give different values for typical test statistics used to test if ensemble members are independent. This will be discussed below and in more details in section 2. Note also that the ensemble mean only has a sound physical interpretation in the truth centered interpretation (Christiansen 2018a).

There is some support for the truth centered interpretation (Knutti et al. 2010; Sanderson and Knutti 2012, and references therein) and it has often been assumed more or less explicitly (see, e.g., the discussion in Jun et al. 2008). However, there is an increasing amount of evidence pointing toward the indistinguishable interpretation. The distance between the ensemble mean and the observations do not converge toward zero as the number of models increases (van Loon et al. 2007; Potemski and Galmarini 2009; Knutti et al. 2010; Bishop and Abramowitz 2013). The correlation between model errors are often found to be positive and near 0.5 (Jun et al. 2008; Annan and Hargreaves 2010; Pennell and Reichler 2011; Bishop and Abramowitz 2013; Herger et al. 2018; Abramowitz et al. 2019), which is the value predicted from the indistinguishable interpretation. Additionally, the error of the ensemble mean is often 30% smaller than the typical error of individual models, which again is the value predicted by the indistinguishable interpretation. For details see, e.g., Palmer et al. (2006) and the discussions in Christiansen (2018a, 2019).

The rationale for the truth centered interpretation is that it is the situation that would be expected after calibration of statistical models. This is a basic idea behind

ensemble regression methods (Bishop 2007, chapter 14). Climate models are based on physical knowledge and are obviously not calibrated or trained like simple statistical models. However, in the process of developing the models, the observations over the last 100 years have served as a guideline. If, e.g., a model's global mean temperature has deviated significantly from the observed, then efforts have been made to correct the problem, perhaps by improving relevant processes or perhaps by changing the forcings. Before a new version of a model with updated or new parameterizations is used it is often tuned to ensure, e.g., that the mean outgoing long wave radiation at the top of the atmosphere matches that of the observations (Mauritsen et al. 2012; Hourdin et al. 2017). Other often used targets for the tuning are the global mean surface temperature, the global mean outgoing radiation, El Niño–Southern Oscillation, the twentieth century warming, and the meridional overturning circulation (Hourdin et al. 2017). Thus, the calibration of the climate models is often focused on the largest scales (Flato et al. 2013), although different model centers may consider different aspects of the climate system. Perfect tuning is not possible due to structural limitations in the climate models.

A main difficulty for the indistinguishable interpretation is the lack of an obvious mechanism to produce such a distribution. In ensemble weather forecasting the model is initialized from a number of conditions representing the uncertainty in the observed initial state. When the model is integrated forward in time, the width of the distribution broadens because of the chaotic divergence of the trajectories. This model distribution should be compared to the distribution of the truth, i.e., a perfect model initialized from the same conditions. Thus, in weather forecasts the indistinguishable interpretation appears naturally as the target for a good model (Ehrendorfer 1994; Palmer 2000). However, initial condition ensembles with a single climate model are in general much narrower than the multimodel ensembles (Haughton et al. 2014), although this to some extent depends on the variable and scale under consideration (Kumar et al. 2016; Swart et al. 2015) (see also the analysis in section 4). The multimodel ensemble is therefore not mainly determined by initial conditions but more by the model deficiencies, which might be different in different models. Thus, the multimodel ensemble is basically a representation of our incomplete knowledge of the climate system.

We note that the real climate and the climate models are systems of very high dimensionality. When comparing different models or when comparing models with

observations, the differences are often calculated as the root-mean-square error or the correlation over a large number of grid cells or times. We have previously used the properties of high dimensional spaces to analyze the errors of the ensemble mean (Christiansen 2018a, 2019). Here, we will argue that the indistinguishable interpretation arises for high dimensional systems when the calibration is performed on large-scale quantities. Note that we see this as a conceptual mechanism and that we do not expect climate models to follow the indistinguishable interpretation perfectly.

The truth centered and the indistinguishable interpretation are both very idealized viewpoints. To get more freedom in our analysis we use a simple statistical model where observations and models are drawn from distributions with different variances and which include a bias. The two interpretations can be found as limits of this more comprehensive model. We use the simplifying properties of high dimensional space (Bishop 2007; Christiansen 2018a) to obtain analytical results for the error correlations and for the error of the ensemble mean under this statistical model. We then analyze the near-surface temperature and the 500-hPa geopotential height fields from a multimodel CMIP5 ensemble of historical experiments. We use the results from the simple statistical model as a tool for interpreting the climate model ensemble. We show that there is a transition from an approximately truth centered situation when global annual means are considered to an indistinguishable interpretation when monthly gridpoint anomalies are considered. Based on this observation we present a simple conceptual mechanism for the indistinguishable interpretation.

In section 2 we introduce the simple statistical model that allows the models and observations to be drawn from different distributions. We use the properties of high dimensional space to obtain analytical results for the ensemble mean and the error correlations. In section 3 we analyze the CMIP5 ensemble using the simple statistical model for interpretation. In section 4 we consider how the indistinguishable interpretation may arise from calibration of the large-scale quantities.

2. The simple statistical model and the blessings of dimensionality

In section 2a we introduce the simple statistical model that was also used in Christiansen (2018a, 2019). We then describe the relevant properties of high dimensional space in section 2b. Then, in section 2c we discuss analytical results for the ensemble mean and the error correlations based on these properties. The derivation of the analytical results is given in appendix A.

a. The simple statistical model

We use the same notation as in Christiansen (2018a, 2019). The K ensemble members are denoted by $\mathbf{x}^k$, $k = 1, \dots, K$, 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: $\bar{\mathbf{x}} = \sum_{k=1}^K \mathbf{x}^k / K$. The N -dimensional space can represent both spatial and temporal fields.

We now consider a model where the ensemble members are drawn from $\mathcal{N}(\mathbf{b}, \sigma_{\text{mod}}^2 \mathbf{I})$ and the observations from $\mathcal{N}(\mathbf{0}, \sigma_{\text{obs}}^2 \mathbf{I})$. Here $\mathbf{b}$ is an N vector of length $\|\mathbf{b}\| = B\sqrt{N}$ and $\mathbf{I}$ is the identity matrix of size N . Thus, the ensemble members and the observations are drawn from Gaussian distributions with different variances, and the ensemble members may be biased relatively to observations. The center, representing the expectation value of the observations, has been set to $\mathbf{0}$ without lack of generality as we only consider differences in our analysis. Note, that while σ_{obs}^2 represents internal variability in the observations, σ_{mod}^2 includes both internal variability and structural differences between models. The Gaussians are spherical so that the coordinates of all the ensemble members are independently drawn from the same univariate distribution. The same holds for the coordinates of the observations. This is not a serious limitation as will be explained in the next subsection. The situation $B = 0$ and $\sigma_{\text{obs}}^2 = 0$ corresponds to the truth centered interpretation and the situation $B = 0$ and $\sigma_{\text{obs}} = \sigma_{\text{mod}}$ corresponds to the indistinguishable interpretation (Sanderson and Knutti 2012). We note that the analytical results in the next subsection include B and σ_{obs}^2 only in the combination $B^2 + \sigma_{\text{obs}}^2$.

b. The blessings of high dimension

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). As a simple example, a unit cube in N dimensions will have 2^N vertices each with a length of $\sqrt{N}/2$. The volume of the cube within a distance ε to the edge is $1 - (1 - \varepsilon)^N$. For $N = 100$ there are more than 10^{30} vertices and more that 99% of the volume is within a distance 0.05 to the edge. Thus, the volume increasingly concentrates near the surface when the dimension increases.

The properties of high dimensional spaces are sometimes called the “curse of dimensionality” and sometimes the “blessing of dimensionality” depending on the considered problem. The relevant properties of high dimensional spaces for our purpose are in mathematics known as “waist concentration” and “concentration of measure” (Talagrand 1995; Donoho 2000; Hall et al.

2005; Gorban et al. 2016, 2020). Here “concentration of measure” means that random vectors drawn from the same distribution in high dimensions have almost always the same lengths. More precisely, the variance of the lengths relative to the expectation of the lengths converges toward zero. “Waist concentration” denotes the fact that independent vectors in high dimensions are almost always orthogonal. More precisely, when the dimension increases the variance of the angle between independent vectors converges toward $\pi/2$. Thus, in high dimensions we can set dot products to zero and substitute the length of a random vector with its expectation value.

The simple statistical model is obviously unrealistic. It assumes that σ_{obs}^2 and σ_{mod}^2 are the same for all coordinates. This is of course not the case when climate fields are considered, e.g., the near-surface temperature has larger variability near the poles than over the equator. Also, the assumption of Gaussianity of the individual components x_n^k may be unrealistic. Perhaps even more importantly, the simple model assumes independent components while in many cases nearby grid points are strongly dependent. However, the concentration of measure—just as the central limit theorem (e.g., Clusel and Bertin 2008)—does not only hold for independent and identically distributed variables but has been extended to various classes of nonidentical and dependent variables (Kontorovich and Ramanan 2008; Chazottes 2015). So generally, quoting Chazottes (2015), “A random variable that smoothly depends on the influence of many weakly dependent random variables is, on an appropriate scale, very close to a constant.” It is only this constant—which in our case is an average variance σ_{mod}^2 —that enters the simple statistical model and $\|\mathbf{x}^k\|^2/N$ can therefore be substituted with σ_{mod}^2 .

In the present paper high dimensional space enters when we consider a spatial field and calculate, e.g., the mean-square error over the grid points. The number of grid points may be very large but in many cases nearby grid points are not independent. The “many” in the quotation above is related to the amount of dependence among the grid points: with strong dependence we would need more grid points. A relevant measure here is the effective number of independent points (or degrees of freedom) as it is related to how fast the correlations decay with distance. Some algorithms directly use the average decorrelation length to estimate the effective number of degrees of freedom. In general this number is rather difficult to estimate (Christiansen 2015) and depends both on the field, the time scale, and the region under consideration (Jones and Briffa 1996; North et al. 2011). However, it is still large and on the order of 50–100 for monthly surface temperatures in the Northern Hemisphere (Wang and Shen 1999; Bretherton et al. 1999).

In appendix B we illustrate the concentration of measure with numerical experiments including dependent variables. In particular, we demonstrate the convergence with the increasing number of effective degrees of freedom.

c. Some analytical results

Using the simplifying properties of high dimensional spaces described in the previous subsection we can obtain analytical results for the error correlation and the ensemble mean. The results are discussed here but the detailed derivations are given in appendix A.

Considering the mean-square error between observations and ensemble mean we get

$$\|\mathbf{o} - \bar{\mathbf{x}}\|^2/N = B^2 + \sigma_{\text{obs}}^2 + \sigma_{\text{mod}}^2/K. \quad (1)$$

Here we have used that the cross terms vanish (the orthogonality) and that $\|\mathbf{o}\|^2/N = \sigma_{\text{obs}}^2$ and $\|\mathbf{x}^k - \mathbf{b}\|^2/N = \sigma_{\text{mod}}^2$ for all k (same lengths). Thus, the mean-square error converges for large K toward $B^2 + \sigma_{\text{obs}}^2$ as $1/K$. Only in the truth centered interpretation ($B = 0$ and $\sigma_{\text{obs}}^2 = 0$) will the mean-square error go to zero for large N and in this situation the root-mean-square error is $\sigma_{\text{mod}}/K^{1/2}$. In the indistinguishable interpretation ($B = 0$ and $\sigma_{\text{obs}} = \sigma_{\text{mod}}$) we get $\sigma_{\text{mod}}\sqrt{1 + 1/K}$. This result was also found in van Loon et al. (2007), Potemski and Galmarini (2009), and Bishop and Abramowitz (2013) based on expectation values that are valid for all N but only as the mean over many realizations. The finite error in the limit of large K has been observed in multimodel ensembles by Knutti et al. (2010) and Bishop and Abramowitz (2013). Dividing Eq. (1) by σ_{obs}^2 we get the dimensionless expression for the mean-square error:

$$\|\mathbf{o} - \bar{\mathbf{x}}\|^2/N/\sigma_{\text{mod}}^2 = (B/\sigma_{\text{mod}})^2 + (\sigma_{\text{obs}}/\sigma_{\text{mod}})^2 + 1/K. \quad (2)$$

Defining the error of the k th model as $\mathbf{e}^k = \mathbf{x}^k - \mathbf{o}$ we get for the correlation between two model errors:

$$\begin{aligned} \text{corr}(\mathbf{e}^k, \mathbf{e}^m) &= \frac{\sigma_{\text{obs}}^2 + B^2}{\sigma_{\text{mod}}^2 + \sigma_{\text{obs}}^2 + B^2} \\ &= \frac{(\sigma_{\text{obs}}/\sigma_{\text{mod}})^2 + (B/\sigma_{\text{mod}})^2}{1 + (\sigma_{\text{obs}}/\sigma_{\text{mod}})^2 + (B/\sigma_{\text{mod}})^2}. \end{aligned} \quad (3)$$

Here we have again used the properties of high dimensional spaces. Similar analytical results for $B = 0$ were reported in Annan and Hargreaves (2011). In the truth centered interpretation the correlation becomes zero. However, in the indistinguishable interpretation it

FIG. 1. The error correlation, the relative error of the ensemble mean, and the root-mean-square error of the ensemble mean in the limit of large K in the simple statistical model where observations are drawn from $\mathcal{N}(\mathbf{0}, \sigma_{\text{obs}}^2 \mathbf{I})$ and models from $\mathcal{N}(\mathbf{b}, \sigma_{\text{mod}}^2 \mathbf{I})$. The three measures are shown as function of normalized model variance $\sigma_{\text{obs}}/\sigma_{\text{mod}}$ and normalized model bias B/σ_{mod} . Calculated from Eqs. (3), (4), and (2). The relative error of the ensemble mean was shown with other scaling in Fig. 5 in Christiansen (2018a). Cyan and red dots indicate the truth centered and the indistinguishable interpretation. Black (TAS) and orange (500-hPa heights) symbols indicate the fields from Tables 2 and 3. Triangles and full circles are based monthly and annual values, respectively. The size of the symbols indicates the spatial averaging of the fields. From largest to smallest the symbols indicate global means, climatology, means, and anomalies.

becomes 0.5 as also noted by Bishop and Abramowitz (2013). Such error correlations around 0.5 have been reported by Jun et al. (2008), Annan and Hargreaves (2010), Pennell and Reichler (2011), Herger et al. (2018), Bishop and Abramowitz (2013), and Abramowitz et al. (2019).

A measure that is often used when validating ensembles of climate models is the error relative to the median error of all ensemble members (Gleckler et al. 2008; Sillmann et al. 2013; Flato et al. 2013). If a specific ensemble member has a relative error of, e.g., 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 have

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

This result, which requires that both K and N are large, was found in Christiansen (2018a) and discussed further in Christiansen (2019). In the indistinguishable interpretation the relative error reduces to $(1 - \sqrt{2})/\sqrt{2} = -0.29$, 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). In the truth centered interpretation the relative error is -1 .

The three measures from Eqs. (3), (4), and (2) are shown as functions of $\sigma_{\text{obs}}/\sigma_{\text{mod}}$ and B/σ_{mod} in Fig. 1.

FIG. 2. The root-mean-square error of the ensemble mean in the simple statistical model as function of the number of ensemble members. The error is calculated from Eq. (2), which is valid for large N . The root-mean-square error is shown for the same parameters as the histograms in Fig. 3. Orange and brown curves are on top of each other.

The latter measure is shown for large K so that the term $1/K$ disappears. The truth centered and the indistinguishable interpretation correspond to the points $(0, 0)$ and $(1, 0)$, respectively. The measures all only depend on $(B/\sigma_{\text{mod}})^2 + (\sigma_{\text{obs}}/\sigma_{\text{mod}})^2$. This means, e.g., that a value of 0.5 for the error correlations can be obtained both for the indistinguishable interpretation and also for values of $\sigma_{\text{obs}}/\sigma_{\text{mod}}$ smaller than one if a bias is present. Note that as the analytical equations are found as a consequence of operating in a high dimensional space, they hold even for a single realization in contrast to results found as the expectation over many realizations.

The decay with K of the ensemble mean error, Eq. (2), is illustrated in Fig. 2 for different values of the parameters. Figure 3 shows the rank histograms for the same parameters. The rank histograms are calculated numerically for $K = 45$ and $N = 1000$ and the bias is chosen equal for all components, $\mathbf{b} = B/\sqrt{N}(1, 1, \dots)$. A rank of zero means that the observation is smaller than all models and a rank of K means the observation is larger than all models. Such rank histograms measure the degree to which the model distributions are consistent with the distribution of observations (Talagrand et al. 1998; Hamill 2001) and has previously been used for the analysis of multimodel ensembles by Annan and Hargreaves (2010) and Haughton et al. (2014).

As expected, the ensemble mean error converges toward zero for large K when we are close to the truth centered interpretation (Fig. 2). In this situation the decay of the ensemble mean error is also most pronounced as measured by the difference between small

and large K . Furthermore, the rank histogram (Fig. 3) is peaked around the center ($K/2$). When we are closer to the indistinguishable interpretation the decay of the ensemble mean error is slower and saturates at a finite value. Now the rank histogram is almost flat. Note that the rank histograms do depend independently on both σ_{obs} and B , as a finite B will show up as an asymmetry in the rank histograms.

3. The CMIP5 multimodel ensemble

Here we investigate the errors in a subset of 45 climate models from the CMIP5 (Taylor et al. 2012) multimodel ensemble. The models that are chosen from the major modeling centers are briefly identified in Table 1 and details can be found in Table 9.A.1 of Flato et al. (2013). We use monthly means from the historical experiments and from each model we use only one ensemble member (r1i1p1). As observations we use NCEP–NCAR data (Kalnay et al. 1996). We focus on the near-surface temperature (TAS) and the geopotential height at 500 hPa. The model data, which come in different horizontal resolutions, are interpolated to the horizontal NCEP resolution (144 longitudes, 73 latitudes) using a simple nearest neighbor procedure. The monthly and annual mean climatologies are calculated for all grid points using the period 1980–2005. Monthly and annual anomalies are calculated by subtracting the climatology. 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 climatologies

FIG. 3. Rank histograms for the simple statistical model. In each panel the parameters are given in the left legend, and the values of the three measures are in the right legend. The number of ensemble members is $K = 45$ and the dimension is $N = 1000$. The histograms are from numerical calculations while the cyan curve is from the analytical expression Eq. (5). The rank histograms are shown for the same parameters as in Fig. 2.

the root-mean-square errors are calculated over the 144×73 grid points and the 12 calendar months and for annual means the errors are calculated over the 144×73 grid points and the 26 years. The results in this section are all robust to changes in the included period and to, e.g., detrending of the data. We also note that we get similar results using the MERRA-2 (Gelaro et al. 2017) or ERA5 (Hersbach et al. 2019) reanalyses as observations instead of NCEP–NCAR.

In Christiansen (2018b) we presented a direct test of the validity of the simple statistical model for a smaller version of the multimodel ensemble than used in the present paper. A similar test was presented in Christiansen (2019) for an ensemble of weather forecasts. These tests demonstrated that the lengths $\|\mathbf{x}^k - \bar{\mathbf{x}}\|$ of the different models were narrowly distributed (concentration of measure) and that the angles between models were close to $\pi/2$ (waist concentration). We have repeated the test for the present ensemble. Calculating the $\|\mathbf{x}^k - \bar{\mathbf{x}}\|/\sqrt{N}$ for all models we find a mean length of $2.57 K$ with a standard

deviation of $0.46 K$ for the monthly climatology of TAS. For the angles ϕ between models [from $(\mathbf{x}^i - \bar{\mathbf{x}}) \cdot (\mathbf{x}^j - \bar{\mathbf{x}}) = \|\mathbf{x}^i - \bar{\mathbf{x}}\| \|\mathbf{x}^j - \bar{\mathbf{x}}\| \cos(\phi)$] we find a mean of 1.59 with a standard deviation of 0.28 . Similar results are obtained for the other fields.

We now consider the rank histograms of the observations among the $K = 45$ models. Then we will use the rank histograms to fit the simple statistical model. The rank histograms for TAS are shown in Fig. 4. For the global annual means the distribution of these 26 ranks (one for each year, Fig. 4a) is rather closely centered around $K/2$ as should be expected for a truth centered ensemble. For the monthly climatology the distribution—calculated over all grid points and calendar months—of the ranks is much flatter indicating a less truth centered ensemble (Fig. 4b). The deviation from flatness shows that the model ensemble is overdispersed. For monthly means (Fig. 4c) the distribution—calculated over all grid points and all months—is even flatter. For the monthly anomalies—again calculated over all grid points and all

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

Modeling center (or group)	Institute ID	Model name
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	CSIRO-BOM	ACCESS1.0, ACCESS1.3
Beijing Climate Center, China Meteorological Administration	BCC	BCC-CSM1.1, BCC-CSM1.1(m)
Canadian Centre for Climate Modeling and Analysis	CCCMA	CanESM2, CanCM4
National Center for Atmospheric Research	NCAR	CCSM4
Community Earth System Model Contributors	NSF-DOE-NCAR	CESM1(BGC), CESM1(CAM5), CESM1(FASTCHEM), CESM1(WACCM)
Centro Euro-Mediterraneo sui Cambiamenti Climatici	CMCC	CMCC-CESM, CMCC-CM, CMCC-CMS
Centre National de Recherches Météorologiques–Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	CNRM-CERFACS	CNRM-CM5, CNRM-CM5-2
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-QCCCE	CSIRO-Mk3.6.0
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS, Tsinghua University	LASG-CESS	FGOALS-g2
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	LASG IAP	FGOALS-s2
The First Institute of Oceanography, SOA, China	FIO	FIO-ESM
NOAA Geophysical Fluid Dynamics Laboratory	NOAA GFDL	GFDL-CM2.1, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M
NASA Goddard Institute for Space Studies	NASA GISS	GISS-E2-H, GISS-E2-H-CC, GISS-E2-R, GISS-E2-R-CC
National Institute of Meteorological Research–Korea Meteorological Administration	NIMR-KMA	HadGEM2-AO
Met Office Hadley Centre (additional HadGEM2-ES realizations contributed by Instituto Nacional de Pesquisas Espaciais)	MOHC	HadCM3
Institute for Numerical Mathematics	INM	INM-CM4
Institut Pierre-Simon Laplace	IPSL	IPSL-CM5A-LR, IPSL-CM5A-MR, IPSL-CM5B-LR
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	MIROC	MIROC-ESM, MIROC-ESM-CHEM
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC	MIROC5, MIROC4h
Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	MPI-M	MPI-ESM-MR, MPI-ESM-LR, MPI-ESM-P
Meteorological Research Institute	MRI	MRI-CGCM3, MRI-ESM1
Norwegian Climate Centre	NCC	NorESM1-M, NorESM1-ME

FIG. 4. Rank histograms showing the distribution of the rank of observations among the 45 CMIP5 models. Red curves are from the fitted simple statistical model. (a) Annual global means, (b) monthly climatology, (c) monthly means, and (d) monthly anomalies. In (b), (c), and (d) all grid points are pooled. Based on TAS for 1980–2005.

months—the rank histogram is basically flat indicating that here the distributions follow the indistinguishable interpretation (Fig. 4d). All distributions are relatively symmetric indicating only a small bias.

We want to make sure that the differences noted above are not just a consequence of comparing different number of samples. To this end we define a measure of truth centeredness as the root-mean-square difference between the ranks and $K/2$. For an ensemble where all members are identical to observation this measure would be zero and the larger it is the further the ensemble is from being truth centered. For a perfectly uniform rank histogram this measure would be $K/\sqrt{12}$.¹ For the global annual mean we find this measure to be 4.5. The distributions over all grid points and temporal values of this measure are shown in Fig. 5 for the monthly climatology, monthly means, and monthly anomalies. The measure for the global mean is indicated by the vertical bar. For the monthly climatology about 96% of the grid points are less truth centered than the global mean. This fraction is even larger for monthly means and monthly anomalies. The distributions also become more narrowly peaked around $K/\sqrt{12} \sim 13.0$.

We now continue by fitting the simple statistical model to the rank histograms. Let P_m and P_o be the cumulative distribution functions of models and observations and p_m and p_o the corresponding probability densities. We then have for the distribution of the ranks, $k = 0, 1, \dots, K$, of the observations among the models:

$$r(k) = \int_{-\infty}^{\infty} \binom{K}{k} P_m^k(x) [1 - P_m(x)]^{K-k} p_o(x) dx. \quad (5)$$

This result is general and does not require any assumptions. For the simple statistical model we have $p_o = \mathcal{N}(0, \sigma_{\text{obs}}^2)$ and $p_m = \mathcal{N}(B, \sigma_{\text{mod}}^2)$. Ranks from Eq. (5) are shown in Fig. 3 (cyan curves) where the parameters of the simple statistical model are known. If the parameters are not known they can be estimated from a given rank histogram.

The resulting fits to the histograms are very good as shown by the red curves in Fig. 4. The fitted parameters are shown in Table 2. In addition to the fields discussed above and shown in Fig. 4 we have also in the table included annual climatology, means, and anomalies as well as monthly global means. In general we only find a small bias. It is clear that the indistinguishable interpretation becomes an increasingly better assumption when decreasing the temporal and spatial scales: $\sigma_{\text{obs}}/\sigma_{\text{mod}}$ increases from 0.21 for annual global means to 1.03 for monthly anomalies. The fact that we in general find $\sigma_{\text{mod}} > \sigma_{\text{obs}}$ is an indication of the overdispersion of the model ensemble. Note, that for the anomalies the bias disappears by definition.

We can now calculate the three measures for the error correlation and the ensemble mean both directly from the climate model output and using the analytical expressions, Eqs. (1), (3), and (4) with the fitted parameters. The results are shown in Table 2. We see that the measures calculated from the analytical expressions provide excellent estimates for all fields.

Figure 6 shows the error correlations of all model pairs [$K(K-1)/2 = 990$ for $K = 45$] for the monthly climatology, monthly means, and monthly anomalies.

¹ The mean-square distance is $1/K \int_0^K (r - K/2)^2 dr = K^2/12$.

FIG. 5. The distribution over the grid points of the measure of truth centeredness. (top) Monthly climatology. (middle) Monthly means. (bottom) Monthly anomalies. The vertical line is the value for the annual global means. Based on TAS for 1980–2005.

The mean of the distributions—shown as the black vertical lines and also given in Table 2—are very close to the value expected from the simple statistical model (red vertical line). In all three cases the correlations

are closer to 0.5 (indistinguishable interpretation) than to 0 (truth centered interpretation). When we go from the monthly climatology over the monthly means to the monthly anomalies the mean of the correlations converges toward 0.5 and the distributions become increasingly narrowly centered around this value.

Figure 7 shows the decay of the ensemble mean error [normalized as in Eq. (2)] as function of the ensemble size for the same fields as for the error correlations. The figure is constructed by randomly drawing a subensemble of k members ($k = 1, \dots, K$) from the ensemble of $K = 45$ models. The root-mean-square error of the subensemble mean is calculated. For each k this is done 100 times and the average of the errors plotted (black curve). These curves strongly resemble the analytical curves calculated from Eq. (2) (red curves). The saturation level for large k increases when going from the monthly climatology over the monthly means to the monthly anomalies. For the monthly anomalies the saturation level is close to 1.

We have repeated the analysis with the 500-hPa geopotential height. The rank histograms are shown in Fig. 8 and the fitted parameters and the measures are given in Table 3. The results are rather similar to those of the TAS except that the model ensemble now has a negative bias. In particular for the monthly climatology the simple statistical model does not provide a good fit to the rank histogram. However, the simple statistical model still gives values of the three measures very close to those calculated directly from the model output. The size of the bias relative to σ_{mod} decreases with decreasing temporal and spatial scales and vanishes for monthly anomalies.

In general we see for all three measures that the indistinguishable interpretation becomes more accurate when the temporal and spatial scales decrease. For the monthly anomalies the results are in full agreement with

TABLE 2. A summary of the fitted simple statistical model. Based on TAS for 1980–2005. The first column describes the considered field. Columns two and three show the fitted parameters for the statistical model. The three last columns show the values of the three measures: the error correlation, the relative error of the ensemble mean, and the normalized ensemble mean error in the limit of large K . Here the first number is the empirical value, and the second is the value calculated from the statistical model. The two first rows show theoretical values for the truth centered and the indistinguishable interpretation.

	$\sigma_{\text{obs}}/\sigma_{\text{mod}}$	B/σ_{mod}	Error correlation	Normalized ensemble mean error	Relative error of ensemble mean
Truth centered	0	0	0	0	−1.00
Indistinguishable	1	0	0.5	1	−0.29
Annual global means	0.21	0.00	—	—	—
Annual climatology	0.64	0.06	0.33, 0.29	0.69, 0.66	−0.44, −0.46
Annual means	0.68	0.07	0.36, 0.32	0.73, 0.70	−0.41, −0.43
Annual anomalies	1.05	0.00	0.53, 0.53	1.08, 1.06	−0.27, −0.27
Monthly global means	0.63	0.02	—	—	—
Monthly climatology	0.73	0.10	0.38, 0.35	0.77, 0.75	−0.38, −0.41
Monthly means	0.82	0.08	0.43, 0.40	0.86, 0.84	−0.35, −0.36
Monthly anomalies	1.03	0.00	0.51, 0.52	1.04, 1.04	−0.28, −0.28

FIG. 6. Distribution of the error correlations for all pairs of the 45 CMIP5 models, calculated over all grid points. (top) Monthly climatology. (middle) Monthly means. (bottom) Monthly anomalies. The black vertical line indicates the mean, and the red vertical line indicates the value from Eq. (3) calculated with parameters from the fitted simple statistical model. Based on TAS for 1980–2005.

the indistinguishable interpretation. This transition toward the indistinguishable interpretation is illustrated in Fig. 1 where the position of the different fields is plotted in the space spanned by $\sigma_{\text{obs}}/\sigma_{\text{mod}}$ and B/σ_{mod} . Fields of smaller scales are indicated by smaller symbols and the convergence toward the indistinguishable interpretation (red filled circle) is clearly seen.

4. A conceptual mechanism for the indistinguishable interpretation

As was the case in the previous sections the N -dimensional space can be both spatial, temporal, or a combination. Let us for simplicity consider time series—this could be of monthly or annual means—so that for, e.g., observations we have $\mathbf{o} = (o_1, o_2, \dots, o_N)$,

FIG. 7. The normalized root-mean-square error of the ensemble mean for the CMIP5 models as function of ensemble size. Calculated over all grid points. For each ensemble size k we randomly draw k models (out of the full ensemble of size $K = 45$) with replacement and calculate the root-mean-square error of the ensemble mean of this subensemble. This is done 100 times for each k . Black dots are the 100 individual errors while full black curves are mean of the errors over the 100 draws. (top) Monthly climatology. (middle) Monthly means. (bottom) Monthly anomalies. Red curves are values from Eq. (2) calculated with parameters from the fitted simple statistical model. Based on TAS for 1980–2005.

where the index n refers to time. Assume now that the observations o_n are drawn from a distribution $\mathcal{N}(c_o, \sigma_o^2)$. We implement a separation of scales by letting the large-scale variability be given by c_o and smaller-scale variability modeled by σ_o . In the simplest case c_o is just the temporal mean although more generally it could include short range dependence and depend slowly on the index n . For the k th model we likewise have—before the model is calibrated—that x_n^k is drawn from $\mathcal{N}(c_k, \sigma_k^2)$, where c_k only depends slowly on time. If we interpret the index n as a spatial pointer instead of time, then c_o would be a

FIG. 8. As in Fig. 4, but for 500-hPa geopotential height.

large-scale spatial mean—such as a global mean perhaps including large-scale spatial variations—and σ_o would be the smaller-scale spatial variability.

Assume first that a simple calibration is performed on each temporal value n —as we could assume to be the case if the time series represented, e.g., global means of annual or even longer temporal averages. Then we would for the calibrated values of the k th model y_n^k have $y_n^k = o_n + \xi_n^k$, $n = 1, \dots, N$, where ξ_n^k is random noise. This situation corresponds to small values of σ_o and σ_k and is equivalent to calibrating c_k against c_o . This corresponds to the truth centered interpretation, and if the residual noise is independent of the model then the ensemble mean $\sum_k y_n^k / K$ would converge toward o_n for large K with a root-mean-square error disappearing as the inverse square root of K .

Consider now the situation where the time series represent, e.g., monthly gridpoint values and σ_o and σ_k are large. In this case we would not expect the individual values x_n^k to be calibrated against o_n . Rather, we would expect the mean and variances to be calibrated. We then get $c_k = c_o + \xi_c^k$ and $\sigma_k = \sigma_o + \xi_\sigma^k$ corresponding to the truth centered interpretation for the mean and the variance. However, the individual temporal values y_n^k are now after the calibration drawn from a distribution with random values of the mean and variance. Such a distribution is called a compound distribution and generally has an inflated variance compared to the distribution

with fixed values of the parameters. If we assume that the noise ξ_c^k is distributed as $\mathcal{N}(0, \sigma_c^2)$ and that ξ_σ^k likewise is distributed as $\mathcal{N}(0, \sigma_\sigma^2)$ we can obtain analytical results. In this case we get that y_n^k is distributed as $\mathcal{N}(c_o, \Sigma^2)$, where $\Sigma^2 = \sigma_o^2 + \sigma_c^2 + \sigma_\sigma^2$.

Therefore, the individual calibrated values y_n^k will be drawn from a distribution with the same mean as the observations o_n but with an inflated variance. The situation is schematically illustrated in Fig. 9. With perfect tuning we have $\sigma_c = \sigma_\sigma = 0$ and models and observations will be drawn from the same distribution corresponding to the perfect indistinguishable interpretation. This is similar to a rescaling of the model distribution not unlike the transformation suggested by Bishop and Abramowitz (2013). If the calibrations of the mean and variances are biased this will turn up as a bias in the distribution of y_n^k . Note that the blessings of dimensionality are not used in this conceptual mechanism.

There is a potential alternative explanation of the transition from the truth centered to indistinguishable interpretation based on internal variability. In the simple statistical model of section 2, σ_{mod}^2 is a sum of a contribution from the structural difference between models σ_{struct}^2 and a contribution from the internal variability of each model σ_{var}^2 . One could imagine that the internal variability would dominate with decreasing averaging and would result in a situation corresponding to the indistinguishable interpretation. To estimate the internal variability σ_{var}^2 , we have analyzed the Max Planck Institute Grand Ensemble (MPI-GE) (Maher et al. 2019). This is a 100-member initial condition ensemble performed with a single climate model. We find that σ_{var}^2 is always considerable less than σ_{mod}^2 (from the multimodel ensemble) except for the anomalies. For

² This corresponds to a Bayesian model where the likelihood is $N(c_k, \sigma_k^2)$ and the priors of c_k and σ_k^2 are $N(c_0, \sigma_c^2)$ and $N(\sigma_o^2, \sigma_\sigma^2)$. The distribution of y_n^k is then found by marginalizing the product of the likelihood and the priors over σ_k^2 and c_k .

TABLE 3. As in Table 2, but for 500-hPa geopotential heights.

	$\sigma_{\text{obs}}/\sigma_{\text{mod}}$	B/σ_{mod}	Error correlation	Normalized ensemble mean error	Relative error of ensemble mean
Truth centered	0	0	0	0	−1.00
Indistinguishable	1	0	0.5	1	−0.29
Annual global means	0.24	−0.65	—	—	—
Annual climatology	0.54	−0.56	0.36, 0.38	0.70, 0.79	−0.39, −0.39
Annual means	0.61	−0.48	0.41, 0.38	0.78, 0.79	−0.36, −0.38
Annual anomalies	1.05	0.01	0.51, 0.52	1.04, 1.06	−0.28, −0.27
Monthly global means	0.49	−0.49	—	—	—
Monthly climatology	0.54	−0.46	0.35, 0.34	0.72, 0.73	−0.40, −0.42
Monthly means	0.74	−0.33	0.44, 0.39	0.89, 0.82	−0.33, −0.37
Monthly anomalies	0.99	0.00	0.48, 0.49	0.98, 1.00	−0.30, −0.30

example, for TAS we find $\sigma_{\text{var}}^2/\sigma_{\text{mod}}^2$ is 0.02 and 0.01 for monthly and annual climatologies, respectively. For monthly and annual means the numbers are 0.25 and 0.09. Only for anomalies are these comparable as should be expected as in this case the structural difference has been removed. Also Haughton et al. (2014) found a much smaller spread in initial condition ensembles compared to the CMIP5 multimodel ensemble considering near-surface temperature.

Yokohata et al. (2012) described how tuning may lead to a truth centered interpretation through a series of updates toward observations. Sanderson and Knutti (2012) argued that an ensemble can drift away from the near truth centered situation due to a common model bias in periods not constrained by observations. In contrast, the simple argument above indicates a

mechanism for how a calibration process that leads to truth centered distributions for the large-scale variability c_k can lead to the indistinguishable interpretation for the individual calibrated values y_n^k .

We note that the calibration of climate models is much more complicated than suggested by the simple scheme above. We therefore do not expect the explanation to be perfect or the indistinguishable interpretation to be fully satisfied.

5. Conclusions

We have presented a quantitative study of the structure of multimodel ensembles and their relation to observations. Results were interpreted using a simple

FIG. 9. Schematic representation of the possible effect of calibration on large scales. The cyan curve is the observations, black curves are different uncalibrated models, and orange curves are the calibrated models. In the right panel the horizontal axis represents time or space; in the left panel the distributions are shown. The uncalibrated models have considerably different means and variances, but the calibration makes the ensemble means and model variances more closely centered around the observed values with small additive errors in both. The individual values (spatial or temporal) of the multimodel ensemble will after the calibration (red) be distributed almost as the observation (cyan) but with inflated variance.

statistical model describing the observations and models as drawn from two different distributions differing in variance and including a bias. The indistinguishable interpretation and the truth centered interpretation which are often used as null hypotheses when studying the independence of climate models—appear as two different limits of this model.

We considered three measures that are all based on the model errors and have been addressed in the previous literature. These measures are the limit of the ensemble mean error for large ensemble size, the correlation between individual model errors, and the relative error of the ensemble mean relative to the typical error of the individual models. In high dimensions—i.e., when the error is calculated as the root-mean-square or correlation over an extended spatial region or a long period—analytical expressions for the three measures can be found under the simple statistical model. These expressions only depend on the size of the bias and the ratio of the variances.

Using these three measures and the rank histograms we studied TAS and 500-hPa geopotential height in a 45-member multimodel CMIP5 ensemble of historical experiments using the NCEP analysis as observations. We studied different averages of the fields stretching from annual global means to monthly gridpoint anomalies. We found that the less averaging we applied the better the indistinguishable interpretation was fulfilled. In fact, the monthly gridpoint anomalies followed the indistinguishable interpretation almost perfectly. This conclusion follows both from the behavior of the three measures and the rank histograms.

The change toward a more perfect fulfillment of the indistinguishable interpretation when the spatial and temporal scales become smaller prompted us to formulate a simple conceptual mechanism explaining the occurrence of indistinguishable ensembles. While the origin of a truth centered ensemble is easy to imagine, as it is basically a measurement noise model, the origin of an indistinguishable ensemble is somewhat harder to comprehend. Our mechanism assumes that the calibration of climate models is focused on large-scale quantities such as means and variances. We emphasize that this is only meant as a conceptual mechanism. The calibration of climate models is much more complicated than the calibration of simple statistical models and other explanations might be possible.

The structure of the multimodel ensemble does have important consequences. In the indistinguishable interpretation (and even more when $\sigma_{\text{obs}} > \sigma_{\text{mod}}$) the ensemble mean error saturates quickly with increasing K , suggesting that a smaller number of models may be sufficient. However, also note that in high dimensions

under the indistinguishable interpretation the ensemble mean is radically different from the individual models (Christiansen 2018a). When studying the independence of models, the choice of null hypotheses is often between the indistinguishable interpretation and the truth centered interpretation. The error correlations are often used as statistics in such studies. However, under the truth centered interpretation the error correlations of independent models are distributed around zero while they are distributed around 0.5 under the indistinguishable interpretation. Therefore, the error correlations should be used with care under the indistinguishable interpretation as the positive mean value is intrinsic and not a reflection of a common bias. In this situation correlations between models are a more suitable statistic.

We also note that the analytical results from the simple statistical model is remarkable successful. This is surprising considering that the model basically only includes two parameters and considering that all components (e.g., different grid cells) of all the individual models are drawn from the same distribution with a single variance (likewise all components of the observations are drawn from the same distribution). The analytical results rely on the simplifications of the high-dimensional limit and at least a part of the success hinges on the power of the blessing of dimensionality and the central limit theorem.

In our analysis we used 45 different climate models but did not consider the fact that they might not all be independent. For example, the different CESM1 models are close in the genealogy tree of models and might not be considered independent. The expression Eq. (3) for the correlations does not depend on K . For the relative error of the ensemble mean the analytical expression Eq. (4) is valid in the limit of large K . However, as shown in Christiansen (2018a) this convergence is fast and 10–20 models should be enough. The mean-square error between observations and ensemble mean, Eq. (2), depends on the ensemble size through the factor $1/K$. As seen in Fig. 7 we get a very good agreement between the simple statistical model and the values calculated directly from the model ensemble using the nominal value of the size of the subensemble k . At least for small k this should be expected from the bootstrap method used to produce Fig. 7. If the number of models k is less than the independent number of models in the full ensemble, then the number of independent models in the subensemble will be close to k . For values of k near K the decay has saturated and the effect of the exact value of K is small [Eq. (2)]. Thus, there is not necessarily any conflict between the results of Fig. 7 and the result that the number of independent models is often found to be less than half or the nominal ensemble size (Pennell and

FIG. B1. Distributions of $\sum_i^N x_i^2/N$ for data drawn from an exponential covariance matrix $\sigma_i \sigma_j \theta^{|i-j|}$ with $\theta = 0.6$ (4000 realizations). The different panels have $N = 50, 200, 400$, and 800 . Anisotropy is included by setting $\sigma_i^2 = 1 + 5i/N$. The effective degrees of freedom estimated with the ξ^2 method (N^*) are 23, 94, 188, and 376. The vertical line indicates the average variance, $\sigma_a^2 = 3.5$.

Reichler 2011; Leduc et al. 2016; Sanderson et al. 2015). However, at least some methods to estimate the number of independent models is very sensitive to the effective degrees of freedom in the spatial field itself. It can also be shown that the spread of the bootstrapped values around the mean (Fig. 7) is determined by the effective degrees of freedom in the monthly climatology and goes to zero as this number increases. We will discuss these issues in a later paper.

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 Agreement 776613). 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 of this paper) 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. The author thanks the editor and the anonymous reviewers whose comments/suggestions helped improve and clarify this manuscript.

APPENDIX A

Derivations of Eqs. (1), (3), and (4)

In this appendix we present derivations of the analytical expressions, Eqs. (1), (3), and (4), discussed in section 2c. Recall from the discussion of high dimensional spaces in section 2b that we can set the dot product of independent vectors to zero and substitute the length of a random vector with its expectation value. See also Christiansen (2019) and references therein.

We first consider the mean-square error. We have

$$\begin{aligned}\|\mathbf{o} - \bar{\mathbf{x}}\|^2/N &= \|\mathbf{o} + \mathbf{b} + \bar{\mathbf{x}} - \mathbf{b}\|^2/N \\ &= \|\mathbf{o}\|^2/N + \|\mathbf{b}\|^2/N + \|\bar{\mathbf{x}} - \mathbf{b}\|^2/N = \quad (\text{A1})\end{aligned}$$

$$\|\mathbf{o}\|^2/N + \|\mathbf{b}\|^2/N + 1/K^2 \sum_k \|\mathbf{x}^k - \mathbf{b}\|^2/N. \quad (\text{A2})$$

In the last two steps we have used the orthogonality. Using now the property of the constant lengths, $\|\mathbf{o}\|^2/N = \sigma_{\text{obs}}^2$ and $\|\mathbf{x}^k - \mathbf{b}\|^2/N = \sigma_{\text{mod}}^2$ we get Eq. (1).

To derive Eq. (3) we let $\langle \rangle$ denote averaging over N . The error correlation is

$$\text{corr}(\mathbf{e}^k, \mathbf{e}^m) = \frac{(\mathbf{e}^k - \langle \mathbf{e}^k \rangle) \cdot (\mathbf{e}^m - \langle \mathbf{e}^m \rangle)}{\|\mathbf{e}^k - \langle \mathbf{e}^k \rangle\| \|\mathbf{e}^m - \langle \mathbf{e}^m \rangle\|}. \quad (\text{A3})$$

With $\mathbf{z}^k = \mathbf{x}^k - \mathbf{b}$ we get $\mathbf{e}^k - \langle \mathbf{e}^k \rangle = \mathbf{z}^k + \mathbf{b} - \mathbf{o}$ as $\langle \mathbf{b} \rangle$ is zero for large N and $\langle \mathbf{o} \rangle$ is zero by definition (no lack of generality). Using that $\mathbf{z}^k, \mathbf{z}^m, \mathbf{o}$, and $\mathbf{b}$ are orthogonal and that $\|\mathbf{o}\|^2 = \sigma_{\text{obs}}^2 N$, we have $(\mathbf{z}^k + \mathbf{b} - \mathbf{o}) \cdot (\mathbf{z}^m + \mathbf{b} - \mathbf{o}) = B^2 N + \sigma_{\text{obs}}^2 N$. Likewise we obtain $\|\mathbf{z}^k + \mathbf{b} - \mathbf{o}\|^2 = \sigma_{\text{mod}}^2 N + \sigma_{\text{mod}}^2 N + B^2 N$ by using the orthogonality and that $\|\mathbf{x}^k\|^2 = \sigma_{\text{mod}}^2 N$. Therefore,

$$\begin{aligned}\text{corr}(\mathbf{e}^k, \mathbf{e}^m) &= \frac{(\mathbf{z}^k + \mathbf{b} - \mathbf{o}) \cdot (\mathbf{z}^m + \mathbf{b} - \mathbf{o})}{\|\mathbf{z}^k + \mathbf{b} - \mathbf{o}\| \|\mathbf{z}^m + \mathbf{b} - \mathbf{o}\|} \\ &= \frac{B^2 + \sigma_{\text{obs}}^2}{\sigma_{\text{mod}}^2 + \sigma_{\text{mod}}^2 + B^2}. \quad (\text{A4})\end{aligned}$$

For the relative error, Eq. (4), we proceed as follows. From Eq. (1) we have $\|\mathbf{o} - \bar{\mathbf{x}}\|^2/N = B^2 + \sigma_{\text{obs}}^2$ for large K . We also see, most easily seen by setting $K = 1$ in Eq. (1), that $\|\mathbf{o} - \mathbf{x}^k\|^2/N = B^2 + \sigma_{\text{obs}}^2 + \sigma_{\text{mod}}^2$. This expression is independent of k and is therefore also the median error. Finally,

$$\frac{\|\mathbf{o} - \bar{\mathbf{x}}\| - \|\mathbf{o} - \mathbf{x}^k\|}{\|\mathbf{o} - \mathbf{x}^k\|} = \frac{\sqrt{B^2 + \sigma_{\text{obs}}^2} - \sqrt{B^2 + \sigma_{\text{obs}}^2 + \sigma_{\text{mod}}^2}}{\sqrt{B^2 + \sigma_{\text{obs}}^2 + \sigma_{\text{mod}}^2}}, \quad (\text{A5})$$

from which Eq. (4) follows by reducing by σ_{mod}^2 .

APPENDIX B

Demonstration of Concentration of Measure for Dependent Data

Here we demonstrate the concentration of measure and the relation to the effective degrees of freedom. In Christiansen (2018a) we considered univariate Gaussian and uniform distributions of different dimensions N but

FIG. B2. The variance of $\sum_1^N x_i^2/N$ as function of the independent degrees of freedom N^* estimated with the ξ^2 method. Data drawn (4000 realizations) from an exponential covariance matrix with $N = 500$ and $\theta = 0.1, 0.5, 0.8, 0.95$, and 0.995 . Filled circles are Gaussian marginals with anisotropy given by $\sigma_i^2 = 1 + 5i/N$. Filled squares are uniform marginals with $\sigma_i^2 = 1 + 5i/N$. Filled triangles are Gaussian marginals with stronger anisotropy: $\sigma_i^2 = 10.2(i/N)^2 + 0.1$. In all cases the average variance, $\sigma_a^2 = \sum_i \sigma_i^2/N$, is 3.5. The straight line is $y = 2\sigma_a^4/N^*$.

only for isotropic and independent data. To illustrate the convergence for dependent data with different effective degrees of freedom, we generate multivariate Gaussian data with the exponential covariance function $\sigma_i \sigma_j \theta^{|i-j|}$, $i, j = 1, \dots, N$, and zero mean. Here, the diagonal elements σ_i^2 give the variance of the different components. The amount of correlation between the components is given by θ : $\theta = 0$ corresponds to uncorrelated components while $\theta = 1$ corresponds to perfectly correlated components. To include anisotropy in the variances we let $\sigma_i^2 = 1 + 5i/N$, so that the variances vary with a factor of 5.

For a given θ and N we draw Gaussian distributed $\mathbf{x}^k$, $k = 1, \dots, K$, with the specified covariance matrix and calculate $\|\mathbf{x}^k\|^2/N$. In Fig. B1 the distributions of this quantity (over $K = 4000$ realizations) are shown for $\theta = 0.6$ and $N = 50, 200, 400$, and 800 . As expected, the distributions are centered around $\sigma_a^2 = \sum_1^N \sigma_i^2/N = 3.5$ (the average variance). Likewise, the distributions of the angles ϕ —from $\mathbf{x}^i \cdot \mathbf{x}^j = \|\mathbf{x}^i\| \|\mathbf{x}^j\| \cos(\phi)$ —are centered around $\pi/2$ (not shown). More importantly, the widths of the distributions decrease with increasing N .

We have confirmed that this behavior is not restricted to the specific form of the covariance matrix by using, for example, a Cauchy covariance function $(1 + |i-j|^2/\theta^2)^{-1}$ (not shown). We have also confirmed that it is not restricted to Gaussian distributed data. This is done by rank transforming the Gaussian data followed by an appropriate scaling. We then get a distribution that has marginal uniform distributions and the same rank-correlation structure as the original Gaussian data.

There are several definitions and algorithms for the effective degrees of freedom (Bretherton et al. 1999; Wang and Shen 1999). Some of these were compared by Wang and Shen (1999) when applied to the surface temperature. The methods all have different strengths and weaknesses—e.g., some do not take anisotropy into account—and generally agree only within a factor of 2. The ξ^2 method may be particularly relevant here as we consider sum of squares. For identically and independently distributed data we have the well-known result $\sum_1^N x_i^2/N \sim \mathcal{N}(\sigma^2, 2\sigma^4/N)$ for large N (Ferguson 1996, chapter 7). In fact, this is a simple example of the concentration of measure as for large N the variance of $\sum_1^N x_i^2/N$ goes to zero as $1/N$ and $\sum_1^N x_i^2/N$ itself becomes σ^2 . The ξ^2 method is based on the assumption that for dependent data this decay should hold when substituting N with N^* (Lorenz 1969; Toth 1995). The effective degrees of freedom becomes $N^* = N^2/\sum_1^N \lambda_i^2$ where λ_i , $i = 1, \dots, N$, are the eigenvalues of the correlation matrix (Fraedrich et al. 1995).

For the example in Fig. B1, where we know the analytical covariance matrix, we get $N^* = 23, 94, 188$, and 376. Figure B2 more systematically shows the relation between N^* and the variance of the distributions of $\sum_1^N x_i^2/N$. The $1/N^*$ behavior is clearly seen with deviations only for the smallest N^* . The proportionality constant varies with the details of marginal distributions and the amount of anisotropy. In these examples the ξ^2 method was applied to analytical covariance matrices. Unfortunately, the ξ^2 method tends to strongly underestimate the degrees of freedom when applied to empirical covariance matrices (Wang and Shen 1999).

REFERENCES

- Abramowitz, G., and Coauthors, 2019: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.*, **10**, 91–105, <https://doi.org/10.5194/esd-10-91-2019>.
- Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, <https://doi.org/10.1029/2009GL041994>.
- , and —, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate*, **24**, 4529–4538, <https://doi.org/10.1175/2011JCLI3873.1>.
- , and —, 2017: On the meaning of independence in climate science. *Earth Syst. Dyn.*, **8**, 211–224, <https://doi.org/10.5194/esd-8-211-2017>.
- Bishop, C., 2007: *Pattern Recognition and Machine Learning*. 2nd ed. Springer-Verlag, 738 pp.
- Bishop, C. H., and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.*, **41**, 885–900, <https://doi.org/10.1007/s00382-012-1610-y>.
- Blum, A., J. Hopcroft, and R. Kannan, 2018: *Foundations of Data Science*. Cornell University, 454 pp., <https://www.cs.cornell.edu/jeh/book.pdf>.
- Boé, J., 2018: Interdependency in multimodel climate projections: Component replication and result similarity. *Geophys. Res. Lett.*, **45**, 2771–2779, <https://doi.org/10.1002/2017GL076829>.
- 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).
- Chazottes, J.-R., 2015: Fluctuations of observables in dynamical systems: From limit theorems to concentration inequalities. *Nonlinear Dynamics New Directions*, Springer, 47–85.
- Cherkassky, V. S., and F. Mulier, 2007: *Learning from Data: Concepts, Theory, and Methods*. 2nd ed. John Wiley and Sons, 560 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>.
- , 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 “Comment on ‘Ensemble averaging and the curse of dimensionality.’” *J. Climate*, **31**, 9017–9019, <https://doi.org/10.1175/JCLI-D-18-0416.1>.
- , 2019: Analysis of ensemble mean forecasts: The blessings of high dimensionality. *Mon. Wea. Rev.*, **147**, 1699–1712, <https://doi.org/10.1175/MWR-D-18-0211.1>.
- Clusel, M., and E. Bertin, 2008: Global fluctuations in physical systems: A subtle interplay between sum and extreme value statistics. *Int. J. Mod. Phys.*, **22B**, 3311–3368, <https://doi.org/10.1142/S021797920804853X>.
- Donoho, D. L., 2000: High-dimensional data analysis: The curses and blessings of dimensionality. *Conf. on Math Challenges of the 21st Century*, Los Angeles, CA, American Mathematical Society.
- 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).
- Evans, J. P., F. Ji, G. Abramowitz, and M. Ekström, 2013: Optimally choosing small ensemble members to produce robust climate simulations. *Environ. Res. Lett.*, **8**, 044050, <https://doi.org/10.1088/1748-9326/8/4/044050>.
- Ferguson, T. S., 1996: *A Course in Large Sample Theory*. Chapman and Hall, 245 pp.
- 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.
- Fraedrich, K., C. Ziehmann, and F. Sielmann, 1995: Estimates of spatial degrees of freedom. *J. Climate*, **8**, 361–369, [https://doi.org/10.1175/1520-0442\(1995\)008<0361:EOSDOF>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0361:EOSDOF>2.0.CO;2).
- Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>.
- Gleckler, P., K. Taylor, and C. Douthiaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, <https://doi.org/10.1029/2007JD008972>.
- Gorban, A. N., I. Y. Tyukin, 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>.
- , V. A. Makarov, and I. Y. Tyukin, 2020: High-dimensional brain in a high-dimensional world: Blessing of dimensionality. *Entropy*, **22**, 82, <https://doi.org/10.3390/e22010082>.

- Hall, P., J. S. Marron, and A. Neeman, 2005: Geometric representation of high dimension, low sample size data. *J. Roy. Stat. Soc.*, **67B**, 427–444, <https://doi.org/10.1111/j.1467-9868.2005.00510.x>.
- Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2).
- Houghton, N., G. Abramowitz, A. Pitman, and S. J. Phipps, 2014: On the generation of climate model ensembles. *Climate Dyn.*, **43**, 2297–2308, <https://doi.org/10.1007/s00382-014-2054-3>.
- Herger, N., G. Abramowitz, R. Knutti, O. Angélil, K. Lehmann, and B. M. Sanderson, 2018: Selecting a climate model subset to optimise key ensemble properties. *Earth Syst. Dyn.*, **9**, 135–151, <https://doi.org/10.5194/esd-9-135-2018>.
- Hersbach, H., and Coauthors, 2019: Global reanalysis: Goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, No. 159, ECMWF, Reading, United Kingdom, 17–24, <https://www.ecmwf.int/en/eLibrary/19027-global-reanalysis-goodbye-era-interim-hello-era5>.
- Hourdin, F., and Coauthors, 2017: The art and science of climate model tuning. *Bull. Amer. Meteor. Soc.*, **98**, 589–602, <https://doi.org/10.1175/BAMS-D-15-00135.1>.
- Jones, P. D., and B. 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*, Global Environmental Change, Vol. 41, Springer-Verlag, 625–644.
- Jun, M., R. Knutti, and D. W. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence. *J. Amer. Stat. Assoc.*, **103**, 934–947, <https://doi.org/10.1198/016214507000001265>.
- 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](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>.
- , D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, <https://doi.org/10.1002/grl.50256>.
- , J. Sedláček, B. M. Sanderson, R. Lorenz, E. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.*, **44**, 1909–1918, <https://doi.org/10.1002/2016GL072012>.
- Kontorovich, L., and K. Ramanan, 2008: Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.*, **36**, 2126–2158, <https://doi.org/10.1214/07-AOP384>.
- Kumar, S., J. L. Kinter, Z. Pan, and J. Sheffield, 2016: Twentieth century temperature trends in CMIP3, CMIP5, and CESM-LE climate simulations: Spatial-temporal uncertainties, differences, and their potential sources. *J. Geophys. Res. Atmos.*, **121**, 9561–9575, <https://doi.org/10.1002/2015JD024382>.
- Leduc, M., R. Laprise, R. de Elia, and L. Šeparović, 2016: Is institutional democracy a good proxy for model independence? *J. Climate*, **29**, 8301–8316, <https://doi.org/10.1175/JCLI-D-15-0761.1>.
- Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, [https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2).
- Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.*, **11**, 2050–2069, <https://doi.org/10.1029/2019MS001639>.
- Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, <https://doi.org/10.1029/2011GL046864>.
- Mauritsen, T., and Coauthors, 2012: Tuning the climate of a global model. *J. Adv. Model. Earth Syst.*, **4**, M00A01, <https://doi.org/10.1029/2012MS000154>.
- 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>.
- Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**, 71–116, <https://doi.org/10.1088/0034-4885/63/2/201>.
- , R. Buizza, R. Hagedorn, A. Lorenze, M. Leutbecher, and S. Lenny, 2006: Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, No. 106, ECMWF, Reading, United Kingdom, 10–17, <https://www.ecmwf.int/sites/default/files/eLibrary/2006/18024-ensemble-prediction-pedagogical-perspective.pdf>.
- Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, <https://doi.org/10.1175/2010JCLI3814.1>.
- Pirtle, Z., R. Meyer, and A. Hamilton, 2010: What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environ. Sci. Policy*, **13**, 351–361, <https://doi.org/10.1016/j.envsci.2010.04.004>.
- Potempski, S., and S. Galmarini, 2009: *Est modus in rebus*: Analytical properties of multi-model ensembles. *Atmos. Chem. Phys.*, **9**, 9471–9489, <https://doi.org/10.5194/acp-9-9471-2009>.
- 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>.
- , —, and P. Caldwell, 2015: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, <https://doi.org/10.1175/JCLI-D-14-00362.1>.
- , M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geosci. Model Dev.*, **10**, 2379–2395, <https://doi.org/10.5194/gmd-10-2379-2017>.
- Sansom, P. G., D. B. Stephenson, C. A. T. Ferro, G. Zappa, and L. Shaffrey, 2013: Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. *J. Climate*, **26**, 4017–4037, <https://doi.org/10.1175/JCLI-D-12-00462.1>.
- Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part I. Model evaluation in the present climate. *J. Geophys. Res. Atmos.*, **118**, 1716–1733, <https://doi.org/10.1002/JGRD.50203>.
- Steinschneider, S., R. McCrary, L. O. Mearns, and C. Brown, 2015: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. *Geophys. Res. Lett.*, **42**, 5014–5044, <https://doi.org/10.1002/2015GL064529>.
- Stephenson, D. B., M. Collins, J. C. Rougier, and R. E. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, **23**, 364–372, <https://doi.org/10.1002/env.2153>.
- Swart, N. C., J. C. Fyfe, E. Hawkins, J. E. Kay, and A. Jahn, 2015: Influence of internal variability on Arctic sea-ice trends. *Nat. Climate Change*, **5**, 86–89, <https://doi.org/10.1038/nclimate2483>.
- 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>.

- Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF.
- 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>.
- Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **365A**, 2053–2075, <https://doi.org/10.1098/rsta.2007.2076>.
- Toth, Z., 1995: Degrees of freedom in Northern Hemisphere circulation data. *Tellus*, **47A**, 457–472, <https://doi.org/10.3402/tellusa.v47i4.11531>.
- 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).
- Yokohata, T., J. D. Annan, M. Collins, C. S. Jackson, M. Tobis, M. J. Webb, and J. C. Hargreaves, 2012: Reliability of multi-model and structurally different single-model ensembles. *Climate Dyn.*, **39**, 599–616, <https://doi.org/10.1007/s00382-011-1203-1>.