

Is Institutional Democracy a Good Proxy for Model Independence?

MARTIN LEDUC

Ouranos, and Centre pour l'Étude et la Simulation du Climat à l'Échelle Régionale, Université du Québec à Montréal, Montreal, Quebec, Canada

RENÉ LAPRISE

Centre pour l'Étude et la Simulation du Climat à l'Échelle Régionale, Université du Québec à Montréal, Montreal, Quebec, Canada

RAMÓN DE ELÍA

Ouranos, and Centre pour l'Étude et la Simulation du Climat à l'Échelle Régionale, Université du Québec à Montréal, Montreal, Quebec, Canada


LEO ŠEPAROVIĆ^a

Centre pour l'Étude et la Simulation du Climat à l'Échelle Régionale, Université du Québec à Montréal, Montreal, Quebec, Canada

(Manuscript received 27 October 2015, in final form 27 June 2016)

ABSTRACT

Climate models developed within a given research group or institution are prone to share structural similarities, which may induce resembling features in their simulations of the earth's climate. This assertion, known as the "same-center hypothesis," is investigated here using a subsample of CMIP3 climate projections constructed by retaining only the models originating from institutions that provided more than one model (or model version). The contributions of individual modeling centers to this ensemble are first presented in terms of climate change projections. A metric for climate change disagreement is then defined to analyze the impact of typical structural differences (such as resolution, parameterizations, or even entire atmosphere and ocean components) on regional climate projections. This metric is compared to a present climate performance metric (correlation of error patterns) within a cross-model comparison framework in terms of their abilities to identify the same-center models. Overall, structural differences between the pairs of same-center models have a stronger impact on climate change projections than on how models reproduce the observed climate. The same-center criterion is used to detect agreements that might be attributable to model similarities and thus that should not be interpreted as implying greater confidence in a given result. It is proposed that such noninformative agreements should be discarded from the ensemble, unless evidence shows that these models can be assumed to be independent. Since this burden of proof is not generally met by the centers participating in a multimodel ensemble, the authors propose an ensemble-weighting scheme based on the assumption of institutional democracy to prevent overconfidence in climate change projections.

 Denotes Open Access content.

^a Current affiliation: Atmospheric Numerical Prediction Research Section, Meteorological Research Division, Environment and Climate Change Canada, Dorval, Quebec, Canada.

Corresponding author address: Martin Leduc, Ouranos, 550 Rue Sherbrooke West, West Tower, 19th Floor, Montreal QC H3A 1B9, Canada.
E-mail: leduc.martin@ouranos.ca

1. Introduction

In recent decades, internationally coordinated efforts have been conducted to provide credible ranges of climate change projections to the scientific community. The World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project (CMIP) multimodel datasets CMIP3 (Meehl et al. 2007), CMIP5 (Taylor et al. 2012), and the upcoming CMIP6 (O'Neill et al. 2016) consist of relatively large ensembles of simulations aiming at sampling the main components of

uncertainty that affect climate change projections (Hawkins and Sutton 2011, 2009). The three main sources of uncertainty emerge from the various possible outcomes of greenhouse gas and aerosol (GHGA) emission pathways (Meinshausen et al. 2011; IPCC 2000), the diversity of climate modeling approaches (Haughton et al. 2014), and the internal variability of the climate system (Hawkins and Sutton 2011; Deser et al. 2012; Lorenz 1963). In practice, partitioning overall uncertainty into its main components is a complex task due to the irregular structure of such large ensemble frameworks (Hawkins 2011; Déqué et al. 2007, 2012). But more importantly, assessing model uncertainty is a challenging exercise partly due to the opportunity-based sampling that generated these ensembles (Tebaldi and Knutti 2007; Annan and Hargreaves 2010), the lack of independence between climate models (Knutti et al. 2010; Tebaldi and Knutti 2007), and also the model selection, for instance by modeling centers participating in a CMIP experiment (Haughton et al. 2014).

Agreements between climate change projections from several models are often interpreted as predictors of confidence (e.g., IPCC 2013, 2007; Seager et al. 2007), but such an inference is difficult to defend without any robust measure of model independence (Pirtle et al. 2010). A natural approach to assess the extent of model independence consists of identifying discrepancies from the “truth plus error” paradigm, where the mean of a sample of independent estimates should have an error that converges to zero as the ensemble size becomes very large. The cancellation of errors through simple multimodel averaging has been shown to be less efficient than expected because of correlations between model errors (Jun et al. 2008; Reifen and Toumi 2009; Pennell and Reichler 2011; Knutti et al. 2010; Haughton et al. 2015). Moreover, this approach is problematic because the ensemble mean does not have the same characteristics as the observed climate, particularly with respect to the magnitude of natural climate variability, which is strongly attenuated by the averaging process (Bishop and Abramowitz 2013). Nevertheless, the truth-plus-error paradigm remains the most widely used technique for processing multimodel ensembles (IPCC 2007, 2013). Alternatively, the “indistinguishable” (Annan and Hargreaves 2010, 2011; Sanderson and Knutti 2012) and “replicate Earths” (Bishop and Abramowitz 2013) paradigms now appear to be more suitable for interpreting multimodel ensembles, with the true climate assumed to be among other members, and the ensemble mean as the best estimate for any member.

It is generally accepted that, since modelers share knowledge about climate models and the real physical

system, models are never completely independent and thus common biases have to be expected from their simulations (e.g., Knutti et al. 2010). Investigating correlations between model errors follows the formal definition of statistical independence, which has been extensively used by the community in recent years (e.g., Jun et al. 2008; Reifen and Toumi 2009; Pennell and Reichler 2011; Knutti et al. 2010; Bishop and Abramowitz 2013). Evans et al. (2013) compared the independence ranking developed by Bishop and Abramowitz (2013) with both a climatological and an impact-performance criterion, and found that the independence measure was more efficient for minimizing the size of an ensemble of climate models. This reduced ensemble also preserved important characteristics of the original ensemble (mean and spread) in the context of finding an optimal set of forcing scenarios to drive regional climate model ensembles. More recently, it has been proposed to use the metric of correlation of errors, along with a measure of ensemble dispersion, to better address the issue of climate model independence (Haughton et al. 2014, 2015). Sanderson et al. (2015a,b) has also developed a weighting scheme accounting for both model performance and interdependence.

Despite these important efforts, it remains unclear how our confidence in climate projections can be altered by the knowledge that, in general, climate models do not provide independent representations of the observed climate system. In addition, regional consequences of the lack of model independence have received surprisingly little attention. Steinschneider et al. (2015) have recently made such an attempt by accounting for the effect of model similarities—by building on previous works from Bishop and Abramowitz (2013), Abramowitz and Bishop (2015), and Haughton et al. (2015)—to develop probabilistic climate projections for a select number of spatially limited regions across the United States.

Probably one of the most important findings regarding the independence of climate models is that the correlation of their output is highly related to their genealogy (Masson and Knutti 2011; Knutti et al. 2013). The genealogy of climate models provides insights into their development history—particularly important since different models often share some history (see Edwards 2011). Hence, tracking model history can reveal important information about model similarities, which can manifest in the dynamical core, the physical parameterizations, or the numerical methods that are chosen, and be implemented by the development teams. While no scientific consensus exists on how to quantify model structural differences, models have to

TABLE 1. Research institutes–groups that provided several models or versions to the CMIP3 multimodel archive.

	Name of the institute–group	Country	Label
1	Canadian Centre for Climate Modelling and Analysis	Canada	CGCM
2	Center for Climate System Research (University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC
3	Commonwealth Scientific and Industrial Research Organisation (Atmospheric Research)	Australia	CSIRO
4	U.S. Department of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	United States	GFDL
5	NASA Goddard Institute for Space Studies	United States	GISS
6	National Center for Atmospheric Research	United States	NCAR
7	Met Office Hadley Centre for Climate Prediction and Research	United Kingdom	UKMO

be compared qualitatively. It is known that modelers sometimes share parts of model code, as well as larger model components (e.g., ocean), or even entire models. This sharing process naturally occurs when several models (or versions of the same model) are developed under the same roof, but it can also occur across institutions. For example, in CMIP5, a rather limited diversity of ocean models corresponds to a large number of climate models, notably the Modular Ocean Model (MOM) and the Parallel Ocean Program (POP) (Flato et al. 2013). Another example is the Community Earth System Model (CESM; Hurrell et al. 2013) and the Norwegian Earth System Model (NorESM; Iversen et al. 2013), both of which are derivatives of the Community Climate System Model, version 4 (CCSM4; Gent et al. 2011). Similarly, the Australian Community Climate and Earth-System Simulator (ACCESS) model has the same atmospheric component as the HadGEM2 model, which leads to strong similarities in simulated features (Haughton et al. 2015). Considering the several dozens of climate models in use today, it is an extremely demanding task to catalog all model differences and to assess their independence in this way, which probably explains the very limited amount of literature on this topic.

To assess the implications of using weakly independent models in regional climate change projections, this paper focuses on the way climate models are built. To this end, we use center origin to form sets of models with varying degrees of similarity. The effect of these model dependencies is investigated in terms of both the similarity of climate change projections and the correlation of their error patterns. A joint use of these two metrics will be proposed to identify situations where the lack of model independence is likely to distort the message conveyed by multimodel ensembles. Here the use of the CMIP3 ensemble better isolates the impacts of same-center dependencies—CMIP5 has a more complex structure that is also affected by many intercenter dependencies (e.g., Haughton et al. 2014).

The manuscript is organized as follows: Section 2a describes nine groups of models according to their structural

similarities and differences. A metric for quantifying model regional agreements in climate change projections is defined in section 2b. In section 3a, the climate change projections, intermodel differences, and the internal variability are compared among the modeling centers for summer surface air temperature. In section 3b, the metric of climate change agreement is compared with the more common pairwise correlation of error patterns. Section 3c describes an application of our approach to assess the effective number of models in the ensemble based on the principle of institutional democracy. Finally, in section 3d, our proposed principle of institutional democracy is implemented as a weighting scheme for ensemble averaging.

2. Methods

a. Groups of same-center models

Since the name of the modeling center is an efficient proxy for model similarities, here we focus on research institutes that contributed simulations from more than one model or model version to the CMIP3 archive. We consider only simulations forced with the A1B emission scenario, resulting in an ensemble of 35 simulations from 15 atmosphere–ocean general circulation models (AOGCMs) and seven modeling groups (Table 1) hosted by four countries.

Table 2 shows the 15 models organized into nine pairs. Each modeling center is represented by one pair, with the exception of GISS, which provided three models to the ensemble (hence three pairs). For each pair, this table provides the details about modeling differences according to the models' main components (A, O, I, L, and C for atmosphere, ocean, sea ice, land surface, and coupling, respectively). The levels of these modeling differences are categorized as either minor (m) or major (M). While this categorization involves some subjectivity, the choices were made as follows. An m was given when model components were known to be developed on the same basis—such as two versions of the same

TABLE 2. Modeling differences between pairs (labeled from 1 to 9 in the first column) of models developed by the same centers in the CMIP3 ensemble. The second and third columns give the name of the institution and the models' identifiers, while the type of modeling differences is detailed in the fourth column. In columns 5–9, model differences are compared according to their main components; refer to the text. More details are provided in the model documentation from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website (<http://www-pcmdi.llnl.gov>).

Pair	Center	Models	Difference	A	O	I	L	C
1	CGCM	T63 and T47	Resolution for A and O.	m	m	—	—	—
2	MIROC	T106 and T42	Resolution for A and O.	m	m	—	—	—
3	GFDL	CM2.1 and CM2.0	Numerical scheme: advection, gravity waves, and damping at the top boundary for A.	m	—	—	—	—
4	CSIRO	3.5 and 3.0	Eddy parameterization (transport coefficient) and mixed-layer treatment (turbulent kinetic energy) for O, numerical scheme for I, wind stress for C, and treatment of surface runoff and river routing scheme for L.	—	m	m	m	m
5	GISS	EH and ER	O component.	—	M	—	—	—
6	GISS	AOM and ER	A, I, C, and L components, and O version.	M	m	M	M	M
7	GISS	AOM and EH	A, O, I, C, and L components.	M	M	M	M	M
8	NCAR	CCSM3 and PCM	Resolution and version for A, O, and I.	m	m	m	M	M
9	UKMO	GEM1 and CM3	Notably resolution, dynamical core, and treatment of aerosols for A.	M	M	M	M	M

code—but that underwent some modifications (e.g., the value of parameters or changes in parameterization packages). Conversely, an M was chosen when the two components appeared to have different code bases (e.g., developed by different teams or with a different name for the component). Understanding the exact nature of the modeling differences in this latter case (M) implies a deeper analysis of all the scientific assumptions being used in both cases, which is beyond the context of the current study.

The pairs 1–4 are formed by models with rather minor structural differences. The CGCM models (pair 1) differ only in atmosphere and ocean resolutions (spectral T63 vs T47), and similarly for MIROC (pair 2) with a larger jump in resolution (spectral T106 vs T42). The GFDL (pair 3) and CSIRO (pair 4) models are different versions of the same models, with minor modifications (which may include model parameters, numerical approximations, or entire parameterization packages) applied to some of their main components, that is, atmosphere for GFDL, and ocean, ice, land, and coupling for CSIRO.

Pairs 5–9 are formed by models that differ in more general structural characteristics. The first GISS (Goddard Institute for Space Studies Model E) pair (pair 5) consists of two models (EH and ER) that have different ocean components (Bleck 2002 vs Russell et al. 1995). In addition, the two ocean models use different spatial resolutions: $2^\circ \times 2^\circ$ and $4^\circ \times 5^\circ$ for EH and ER, respectively. For pairs 6–9, the models differ substantially according to most of their main components (atmosphere, ocean, sea ice, land, and coupling). An apparent similarity, however, exists between the models AOM and ER (pair 6) in which successive versions of the same ocean model are used

(Russell et al. 1995 vs Russell et al. 2000) but with different resolutions ($4^\circ \times 3^\circ$ for AOM). Among the three models from GISS, AOM appears to differ most from the two others (EH and ER), as it uses a different atmosphere component (pairs 6 and 7). Finally, both pairs 8 and 9 include models from different generations with a common history of development within the same institution: the National Center for Atmospheric Research (NCAR) and the Met Office Hadley Centre, respectively. NCAR CCSM3 and PCM are essentially based on the same atmospheric model and also share the same ocean (Parallel Ocean Program) and sea ice (Community Sea Ice Model), although those components correspond to different versions (Washington et al. 2000; Collins et al. 2006b). For the Met Office (UKMO), HadGEM1 was mostly built upon HadCM3, with the aim of being adapted for higher resolutions, as well as an increased complexity in terms of Earth system modeling (Johns et al. 2006).

b. Metric of climate change disagreement

To evaluate the extent to which climate models agree or disagree, here we use a Welch's unequal variances *t* test of the difference between two sample means, that is, each model being represented by the climate change signal averaged over its available members. The magnitude of the models' simulated internal variability is used as the level of noise against which these differences are tested. Given some significance level, the result of the test reads as follows: The rejection of the null hypothesis of equal means is interpreted as a disagreement, whereas an agreement corresponds to a lack of evidence for rejecting the hypothesis.

In the context of simulations run under transient forcing conditions, the internal variability can be assessed

using the spread between simulations (members) from a given model that differ only by slight perturbations in their initial conditions (Deser et al. 2014). However, models participating in large multimodel ensembles are generally represented by very few members. For the CMIP3 dataset, for example, the number of members per model ranges from 1 to 7, thus leaving very few degrees of freedom for addressing statistical significance. We circumvent this issue by assessing the internal variability as the temporal variability around a fitted fourth-degree polynomial trend, similar to the approach adopted in Hawkins and Sutton (2011, 2009). Such a correspondence between temporal and intermember variability assumes ergodicity (Reif 1965) of the climate system, and it neglects the effects of the change of natural variability with GHGA emissions. These are not very strong assumptions for a variable such as temperature (Kay et al. 2015; Holmes et al. 2016).

Let now $P_{m,n}(x, y)$ and $F_{m,n}$ be two mean climate states of a given variable (e.g., surface air temperature), as simulated by the n th member from model m , and where x and y represent the horizontal coordinates. In what follows, P and F are defined as the recent past (1980–2000) and future (2080–2100) periods, respectively. The climate change signal can be defined as $\bar{\Delta}_m(x, y) = \bar{F}_m(x, y) - \bar{P}_m(x, y)$, where $\bar{(\cdot)}_m$ represents the average over all available members from the m th model. Assuming a pair of climate models, $m = [1, 2]$, the t statistic of the difference of the means between the two models is

$$t = \frac{\bar{\Delta}_1 - \bar{\Delta}_2}{\sqrt{\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2}}, \quad (1)$$

where N_m is the sample size (number of members) used in the calculation of $\bar{\Delta}_m$ and $\hat{\sigma}_m^2$ quantifies the internal variability affecting the climate change signal of the m th model. Here, $\hat{\sigma}_m^2$ corresponds to the sum of the variances of the P and F climatic states (i.e., climates of the present and future 20-yr averaging windows, respectively). It is worth noting that the definition of the t statistic in Eq. (1) does not assume equal variances, since the internal variability may be quite different among climate models.

The variance of the climate change signal can be calculated as

$$\hat{\sigma}_m^2 = \frac{2}{N_m} \sum_{n=1}^{N_m} \frac{1}{T-K-1} \sum_{\tau=1}^T [X_{m,n}(x, y, \tau) - \tilde{X}_{m,n}(x, y, \tau)]^2, \quad (2)$$

where $X_{m,n}(x, y, \tau)$ is a time series of mutually exclusive 20-yr averaging windows, τ is the time index, and $\tilde{X}_{m,n}$ is

the fourth-degree polynomial trend associated with climate change (fitted on the time series of 20-yr averaging windows). Calculating the residual mean-square error of the time series ($X_{m,n}$) around the trend ($\tilde{X}_{m,n}$) allows one to roughly estimate the magnitude of internal variability. To do so, the sum of the squared errors is normalized with its number of degrees of freedom $T - K - 1$, where K is the degree of the fitted polynomial. Assuming the independence of the future and past climatic states relative to the trend (as a result of internal variability) and that the internal variability does not change much in time, the variance of the difference between two climatic states is equal to twice the estimated variability, explaining the multiplying factor of 2 in Eq. (2). The internal variability is then averaged over the N_m members.

Since $\hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$, the t statistic in Eq. (1) can be assumed to follow the Student's distribution by using the Welch's approximation to the Behrens–Fisher problem, resulting in a number of degrees of freedom being estimated directly from the data as (Scheffé 1970)

$$df = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{\frac{(\hat{\sigma}_1^2/N_1)^2}{N_1(T-K-1)} + \frac{(\hat{\sigma}_2^2/N_2)^2}{N_2(T-K-1)}}. \quad (3)$$

This approach allows an important increase in the number of degrees of freedom compared to a calculation of the intermember spread over a specific period. It is particularly convenient when very few members are available, resulting in a test with a higher power to reject the null hypothesis. Moreover, this method allows us to include the GFDL, CSIRO, and UKMO groups, whose models ran only a single member.

3. Results

a. Per-institute contributions to a multimodel ensemble

The pairs of same-center models are now investigated according to their projected changes in summer surface air temperature. To enable direct comparison, all simulations were bilinearly interpolated over a common global grid with $4^\circ \times 5^\circ$ of resolution. For each model pair given in Table 2, Fig. 1 gives the mean climate change signal calculated as a 20-yr averaging window for the 2080–2100 period (A1B scenario) relative to the 1980–2000 reference state (20CM3 experiment).

Figure 1 shows the individual contributions of each modeling center to the entire ensemble. It appears that the pairs mostly divide into two categories of climate-sensitivity magnitude. The CGCM, CSIRO, GISS, and NCAR centers are characterized by relatively low climate

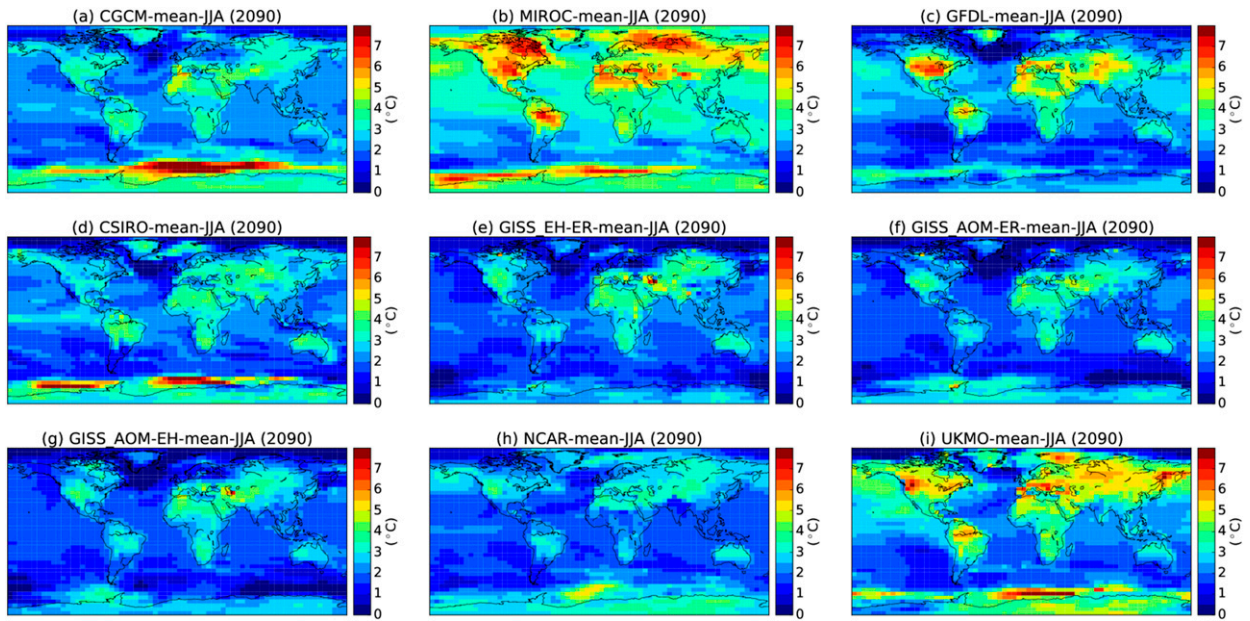


FIG. 1. (a)–(i) Same-center averages of the summer (JJA) surface air temperature ($^{\circ}\text{C}$) change (A1B scenario, 2080–2100 relative to 1980–2000) for the nine pairs of models given in Table 2.

sensitivities over land regions, with values often below 4°C . By comparison, the MIROC, GFDL, and UKMO groups show much higher values, exceeding 7°C in many land areas. The MIROC center contributes to the ensemble as a warm outlier over both land and ocean.

Figure 2 shows the intermodel difference for each pair of climate change signals [i.e., the numerator of Eq. (1)].

As described in section 2b, the statistical significance of these differences is calculated using the sum of the models' standard errors due to internal variability [denominator of Eq. (1)]. As a complement, the average of the natural climate variability, as defined in Eq. (2) for one model, is given for each center in Fig. 3. Colored areas in Fig. 2 represent regions where the null hypothesis

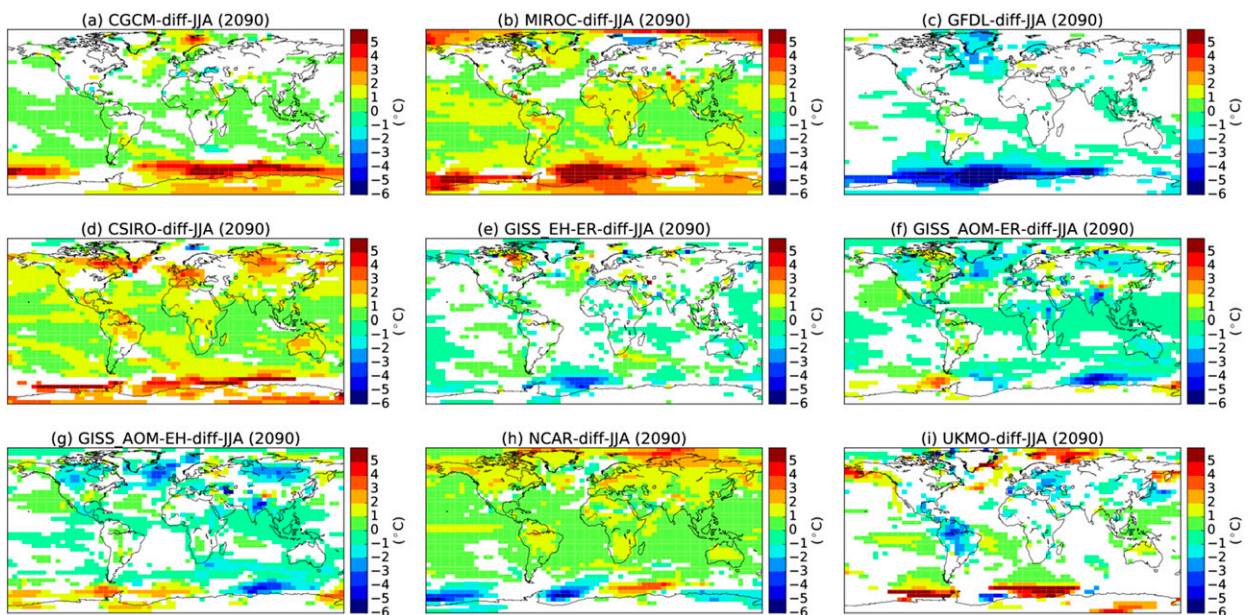


FIG. 2. As in Fig. 1, but for the intermodel differences of the summer surface air temperature change. Differences that are not statistically significant at the 5% level are uncolored.

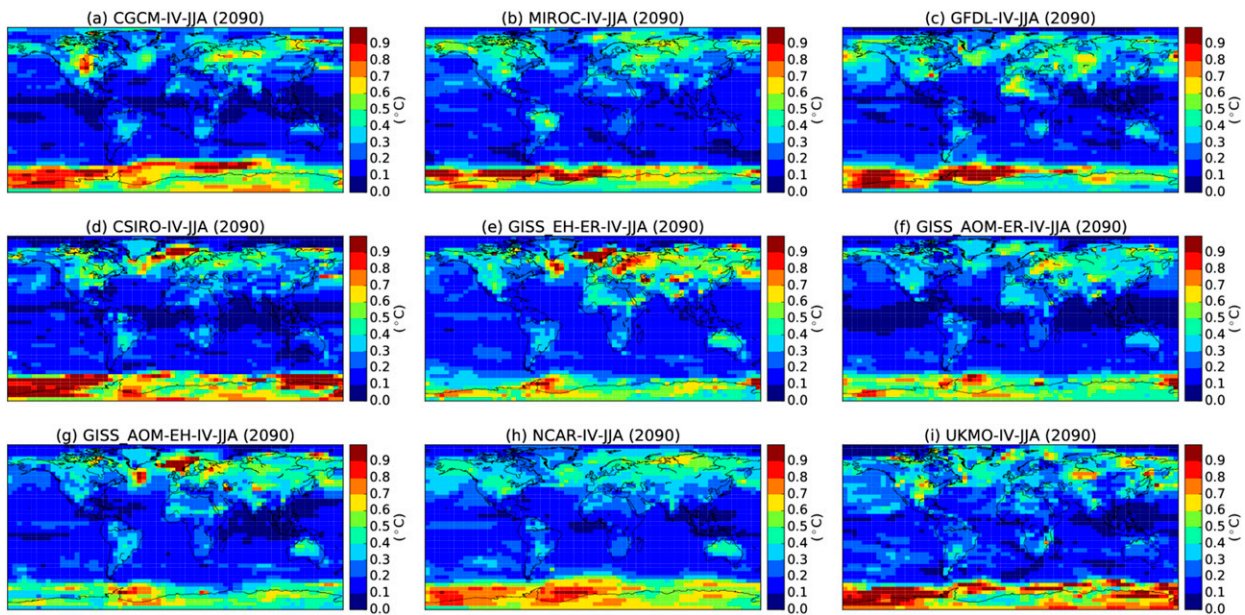


FIG. 3. As in Fig. 1, but for multimodel averaged internal variability in the summer surface air temperature change.

of equal means is rejected at the 5% significance level. The order of calculation of intermodel differences corresponds to that presented in Table 2, where the most recent–sophisticated model appears first. Hence, positive differences in Fig. 2 generally relate to the way climate projections evolve in time through the development of succeeding model versions.

The CGCM models (Fig. 2a), which differ by their resolution, show very low rejection rates over North America, Europe, and Asia, thus implying a statistically insignificant impact on the temperature changes. For the MIROC pair (Fig. 2b), where the change in resolution is more important, the higher-resolution model clearly has a higher climate sensitivity, as seen by the positive significant differences over most of the global domain except in the northern midlatitudes. The intermodel difference is also generally positive for the CGCM pair, but to a smaller extent and magnitude. Since the climate change signals for the CGCM and MIROC pairs are not much different from a statistical point of view over land regions in the northern midlatitudes, considering both models in each pair does not add much supplementary information to the ensemble as compared to the use of a single model version per center. However, since these data are interpolated over the same grid, we note that a sizable part of the potential added value (Di Luca et al. 2013) by the higher-resolution models is not considered here.

Versions 2.0 and 2.1 of the GFDL CM (Fig. 2c) have structural differences that can be understood as minor modifications to the code of their atmospheric model. Expectedly, these show practically no significant differences in

their climate change signal, with the exception of the Southern Ocean. Similar to the GFDL models, Mark 3.0 and 3.5 of the CSIRO model (Fig. 2d) exhibit minor differences in all components other than atmosphere, which remains unchanged. The noticeable difference in climate sensitivities between the CSIRO models appears to be mostly due to changes in the ocean eddy parameterization and mixed-layer treatment, while the other modeling differences (see Table 2 for more details) are expected to play minor roles in this pair.

The five remaining pairs consist of models with more important differences in structural characteristics. The three GISS models are interesting to compare, since one or more of their main components differ. For the EH and ER models (Fig. 2e) with different ocean components, significant positive differences exceeding 3°C are found over Hudson Bay in Canada, and around 1.6°C in the North Atlantic, with a relatively low rejection level elsewhere. GISS-AOM and GISS-ER (Fig. 2f), which differ in all of their components except the ocean, which underwent only a change in version, exhibit smaller differences over Hudson Bay ($<1^{\circ}\text{C}$), while negative differences extend over land (e.g., western Canada, north Asia, India, central Africa, and Australia). For the third pair of GISS models, AOM versus EH (Fig. 2g), all of the models' main components have been changed significantly. It is interesting to note that the large difference for Hudson Bay and the North Atlantic are similar in magnitude (with reverse sign) compared to the GISS EH–ER pair, which represents the same change in the ocean component. Changes in other components appear to have

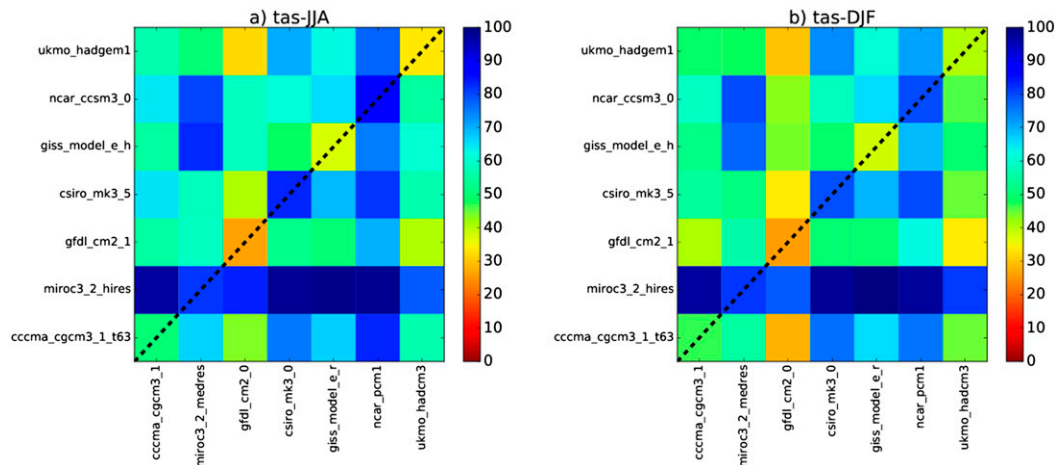


FIG. 4. Pairwise calculation of the climate change model disagreement metric (given as the rejection rate) for temperature (2080–2100 relative to 1980–2000) in (a) summer and (b) winter. Models are organized as one model per center on each axis.

little influence on the maximum difference over Hudson Bay (see the EH–ER pair), but the change in GISS atmospheric component (AOM–ER and AOM–EH pairs) is responsible for most differences over land.

The NCAR models (CCSM3 and PCM1) are distinct models developed within the same institution. The differences between these models are statistically significant over most of the domain. CCSM3 model warms systematically faster than PCM, by about 2°C in the midlatitude land regions and by more than 3°C in the Arctic. Finally, the HadCM3 and HadGEM1 models from the Hadley Centre were compared. It is interesting to note that these two models, which might be considered independent models developed within the same institute, lead to climate change signals that are statistically indistinguishable over a large fraction of land regions (except South America and eastern Europe). These models are known to have very similar global-climate sensitivities (Johns et al. 2006); the relatively low rejection rate of the null hypothesis is comparable to that of GISS-EH and GISS-ER, which share the same atmospheric component (Fig. 2e).

While this should be interpreted carefully, our results suggest that the Hadley Centre models cannot be taken as independent models. Their common history of development may involve shared parameterizations, one example being the radiative transfer parameterization package (Edwards and Slingo 1996), a component that plays an important role in model climate sensitivity to GHGA atmospheric concentration (Collins et al. 2006a). There is also a potential for less obvious dependencies at the institutional level, such as in the choice of the observational dataset used for model validation and tuning. While these higher-level dependencies could

also apply to other same-center models, it should also be taken into account that same-center models may lead to similar responses for the right reasons, that is, a result that is independent of the details embedded within each model and that converges toward the future to be eventually observed (Levins 1966). Although this last conjecture is tempting, a thorough study of model independence between these models is needed before it can be asserted.

b. Discriminating between the same-center models

As shown in the previous section, structural similarities between climate models developed within the same institution can provide some insights into the spatial structure of model agreements in climate change projections. Let us now consider the climate-change “disagreement rate,” which consists of the fraction of the global domain (in surface area) where two climate change projections differ significantly relative to the magnitude of the natural climate variability (i.e., the fraction of colored areas in Fig. 2). The power of this metric to discriminate the same-center model pairs from all other pairs is now assessed within a cross-model comparison framework for temperature change in both summer and winter.

The disagreement rate for projected changes in summer surface air temperature by the end of the twenty-first century is given in Fig. 4a, for all possible combinations between the first and second models of each pair given in Table 1 (GISS-AOM is excluded to simplify the analysis). Pairs of same-center models are represented along the diagonal of the matrix. At first sight, the climate change disagreement metric has a rather low ability to identify the same-center models, according to the current ensemble.

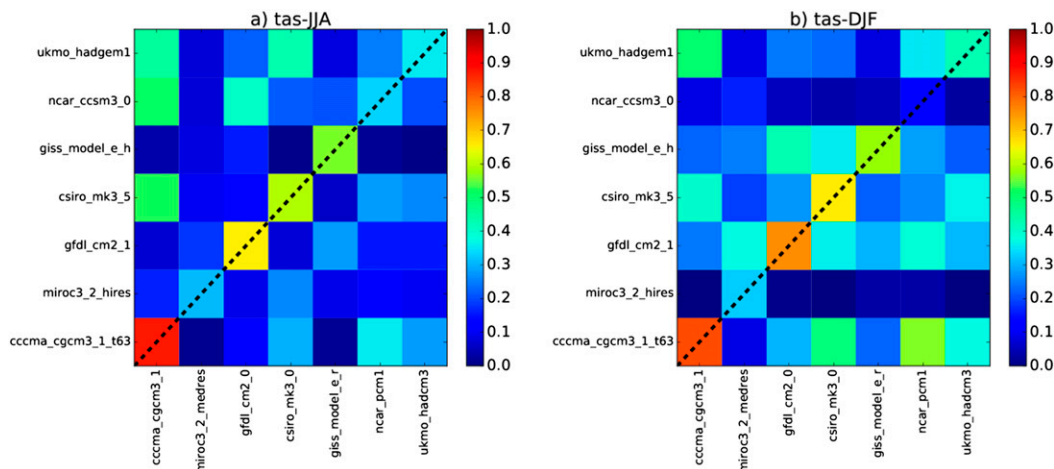


FIG. 5. As in Fig. 4, but for pairwise correlation of model errors relative to the HadCRUT4 dataset for temperature (mean state between 1961 and 1990). One member per model is used.

Nevertheless, three of the four lowest disagreement rates (which represent pairs of models that strongly agree in their projections) appear for same-center models (the GFDL, GISS EH-ER, and UKMO pairs). On the other hand, the MIROC, CSIRO, and NCAR pairs represent strong model disagreements, which are more representative of those found for pairs of models developed by different centers. Overall, similar conclusions are obtained for winter temperature change (Fig. 4b).

It is interesting to compare the previous metric of climate change disagreement with a performance-derived metric, such as pairwise correlation—the more standard approach to addressing issues of model independence (Jun et al. 2008; Reifen and Toumi 2009; Pennell and Reichler 2011; Knutti et al. 2010; Haughton et al. 2015). Here, the pairwise correlation between model errors is calculated using the HadCRUT4 global-observed mean state of 1961–90 surface air temperature (Morice et al. 2012); the results are shown in Figs. 5a and 5b for summer and winter seasons, respectively. While this result could be metric dependent, it can be seen in Fig. 5 that the power to discriminate between the same-center model pairs from other pairs is quite a bit higher in the case of the performance-derived metric as compared to the case of the climate change disagreement metric previously shown in Fig. 4. This further suggests that typical model similarities and model origin act more strongly to constrain agreements between model errors than between regional climate sensitivities to GHGA forcing. Interestingly, the same-center agreements that were previously found using the climate change metric are still detected for the correlation of errors. GFDL and GISS are two striking cases of same-center models that agree in both of their representation of the observed

mean climate and in terms of climate sensitivity. These provide strong examples of dependent models that give similar results, which will be referred to as the “non-informative agreement” in the next sections.

c. Discarding noninformative agreements

When two models provide virtually equal climate change signals, one may be inclined to consider this as noninformative agreement, particularly if the resemblance can be attributed unequivocally to a dependency relationship between the models. Attributing model agreement to a lack of independence is, however, a complex problem. Models that share several components and that are developed within the same institution feed suspicions that this agreement might result solely from common design. In the following, we propose a conservative approach to filter out noninformative agreements from an ensemble, because these may potentially affect ensemble statistics along with our confidence in them.

One popular approach for interpreting multimodel ensembles is known as “one model, one vote” (Knutti 2010), which assumes each model is an equivalent representation of the climate system. Contrasting with this model-democratic point of view, here we propose an “institute democratic” approach to consider multimodel ensembles (i.e., one center, one vote). As a basic rule to be applied on a per-gridpoint basis, two same-center models are considered as a single one when their signals are statistically indistinguishable (otherwise, they are counted as two individual models). It is important to note that since all models are expected to show no signal early in simulations, applying this rule could reduce the information available in the ensemble by discarding informative agreements (i.e., agreements

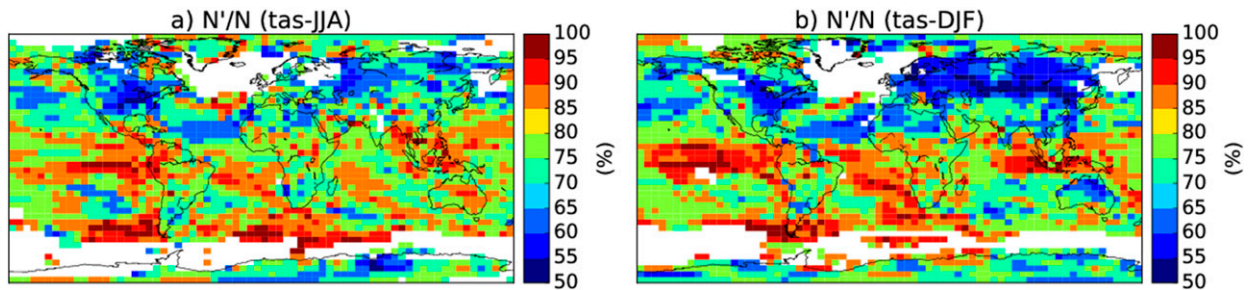


FIG. 6. Ratio (%) of the effective number of models N' to the actual ensemble size $N = 14$. Climate change (2080–2100 relative to 1980–2000) is given for surface air temperature in (a) summer and (b) winter. Uncolored areas are where the signal from at least one of the models is not statistically significant at the 5% level.

between independent models showing no signal): “no signal” in several models should, in fact, increase our confidence that climate change is not occurring—relevant information that should not be filtered out from the ensemble. For this reason, the following analysis focuses only on the cases where the signal is significant relative to the noise in every model considered.

In what follows, we will present an example of discarding noninformative agreements from the ensemble by the strict application of the institute-democratic rule. We will assume that such agreements occur when two same-center models lead to equivalent projections. In reality, however, the same-center models can agree and still be independent; for the sake of simplicity; however, we neglect this possibility and adopt a rather conservative approach. We strongly believe that the burden of proof of the same-center model independence should be at least partly held by the participating centers.

We define the number of independent models N' as the number of models that remain in the ensemble after discarding one model per pair when an agreement is found. It is interesting to compare N' to the full size of the ensemble ($N = 14$, where GISS-AOM was excluded to obtain an even ensemble size). By discarding noninformative agreements related to model similarities, N' can be thought of as the effective ensemble size, which is often referred to be smaller than the actual number of participating models in the ensemble (e.g., Pennell and Reichler 2011; Annan and Hargreaves 2011).

The ratio N'/N for climate projections of temperature in the summer and winter seasons is shown in Fig. 6. The lower limit of this ratio is 50% because N' cannot be smaller than the number of institutions (seven in this ensemble). At the upper limit, the ratio reaches 100% when all intracenter model differences are statistically significant. For temperature change in summer (Fig. 6a), the ratio often has values smaller than 65% over North America and Europe, while such low values extend over most of Asia in winter (Fig. 6b).

Hence, the institute-democratic approach leads to an effective ensemble size that is in general a little more than half of the actual ensemble size over midlatitude land regions. There are also some regions, particularly in the tropics, where the effective ensemble size exceeds 90%. These regions are characterized by a large signal relative to the multidecadal climate variability.

d. Model weighting under institutional democracy

Since the burden of proof that same-center models can be assumed as independent models is generally not met by modeling centers participating in internationally coordinated experiments, we now translate the concept of institutional democracy into a weighting scheme for calculating the ensemble statistics. Based on the same criteria as in section 3c for calculating the number of independent models in the ensemble N' , half and unit weights are assigned to each model depending on whether a noninformative agreement was found with the other model in the same-center pair. Using this scheme for calculating the ensemble statistics of the climate change signal in temperature (Figs. 7a and 7c for summer, and Figs. 7b and 7d for winter), it appears that the model-democratic (arithmetic) and institute-democratic (weighted) ensemble-mean climate change patterns are very similar (the difference between the two types of mean is shown in Figs. 7e and 7f), which is also true for the intermodel spread (not shown). This similarity may appear as highly specific to the current ensemble, which is constructed based on a set of same-center model pairs. However, the use of both a more sophisticated independence metric and a typical ensemble of opportunity appear to lead to similar conclusions (Evans et al. 2013).

Despite the previous similarity, the level of confidence associated with these statistics (ensemble mean and spread) should, however, highly depend on the chosen approach. Recalling the truth-plus-error paradigm, an ensemble of N models is generally interpreted as a sample of N independent and identically distributed

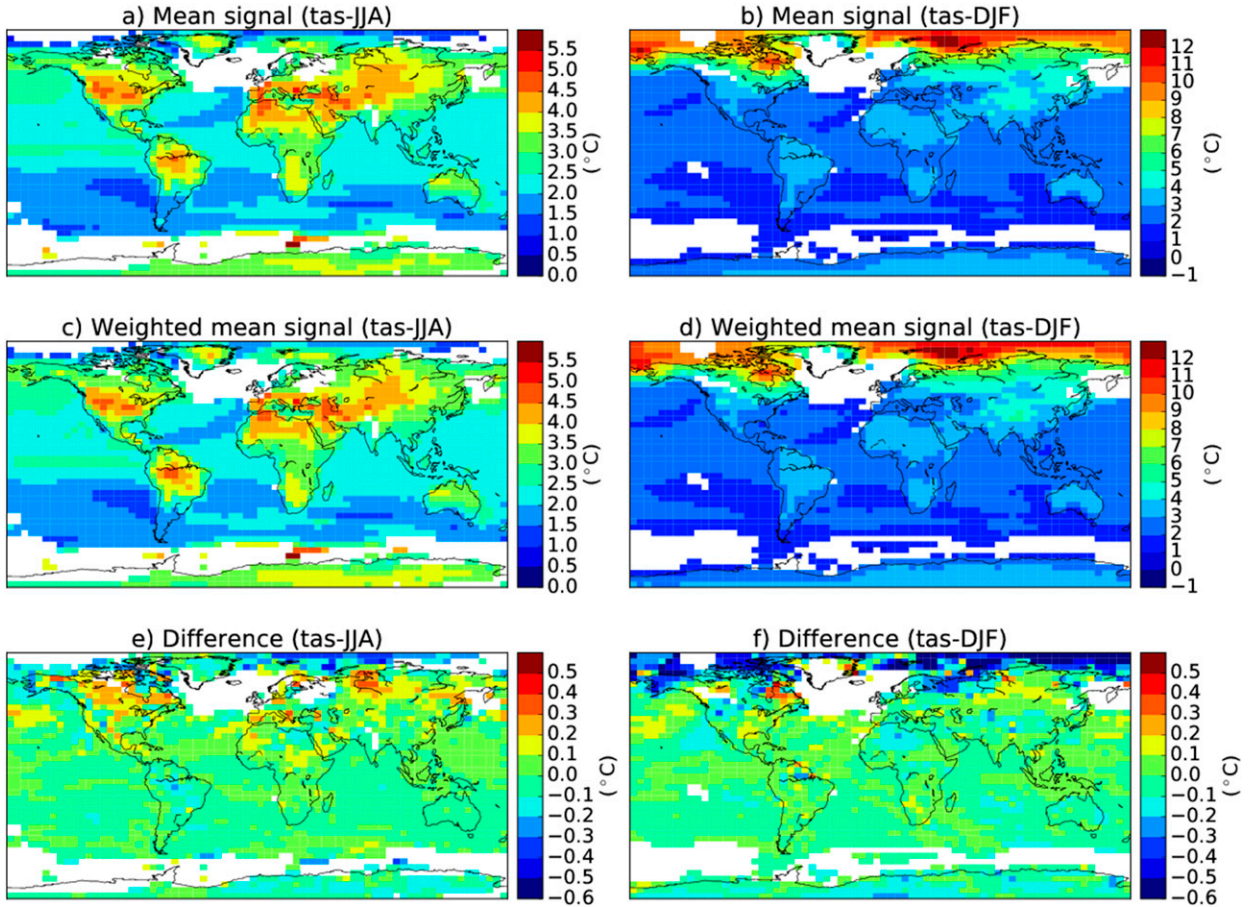


FIG. 7. (a) Model-democratic (arithmetic) and (c) institute-democratic (weighted) ensemble-mean climate change patterns for summer surface air temperature (2080–2100 relative to 1980–2000) with (e) their difference (arithmetic minus weighted). (b),(d),(f) As in (a),(c),(e), but for winter. Uncolored areas are where the signal from at least one of the models is not statistically significant at the 5% level.

climate change estimates drawn from a larger population (e.g., Stephenson et al. 2012; von Storch and Zwiers 2013). Under this paradigm, the model-democratic sample mean $\hat{\Delta}$ (one model, one vote) has a standard error (i.e., the standard deviation of the sample mean) of

$$\text{Std}(\hat{\Delta}) = \sqrt{\frac{\hat{\sigma}^2}{N}}, \quad (4)$$

where $\hat{\sigma}^2$ is the intermodel sample variance of the climate change signal and N is the ensemble size. Similarly, if we now interpret Eq. (4) from the institute-democratic perspective, the variance of the weighted sample mean $\hat{\Delta}'$ is

$$\text{Std}(\hat{\Delta}') = \sqrt{\frac{\hat{\sigma}'^2}{N'}}, \quad (5)$$

where $\hat{\sigma}'^2$ is the weighted intermodel sample variance of the climate change signal and N' is the effective ensemble size. Equations (4) and (5) reflect different avenues under

the truth-plus-error paradigm to assess the statistical uncertainty in the sampling of a multimodel mean. Since (although partly due to the way the current ensemble is built) this weighting technique leads to $\hat{\sigma}'^2 \approx \hat{\sigma}^2$, the ratio N'/N is approximately equal to the ratio of the squared standard errors [$\text{Var}(\hat{\Delta})/\text{Var}(\hat{\Delta}')$]. For example, an N'/N ratio of 60% involves inflating the model-democratic standard error (standard deviation) by about 30% to reach the institute-democratic value.

A convenient way to analyze the impact of this weighting scheme is by comparing the ensemble-mean signal with the statistical uncertainty (or standard error) associated with it. The signal-to-uncertainty ratio [$\hat{\Delta}/\text{Std}(\hat{\Delta})$ or $\hat{\Delta}'/\text{Std}(\hat{\Delta}')$] of climate change projections in summer temperature is shown in Figs. 8a and 8c for the model-democratic and institute-democratic approaches, respectively. The signal becomes weaker relative to uncertainty under the institute-democratic approach. The relative error of the statistical uncertainty, [$\text{Std}(\hat{\Delta}') - \text{Std}(\hat{\Delta})/\text{Std}(\hat{\Delta})$], can be interpreted as the

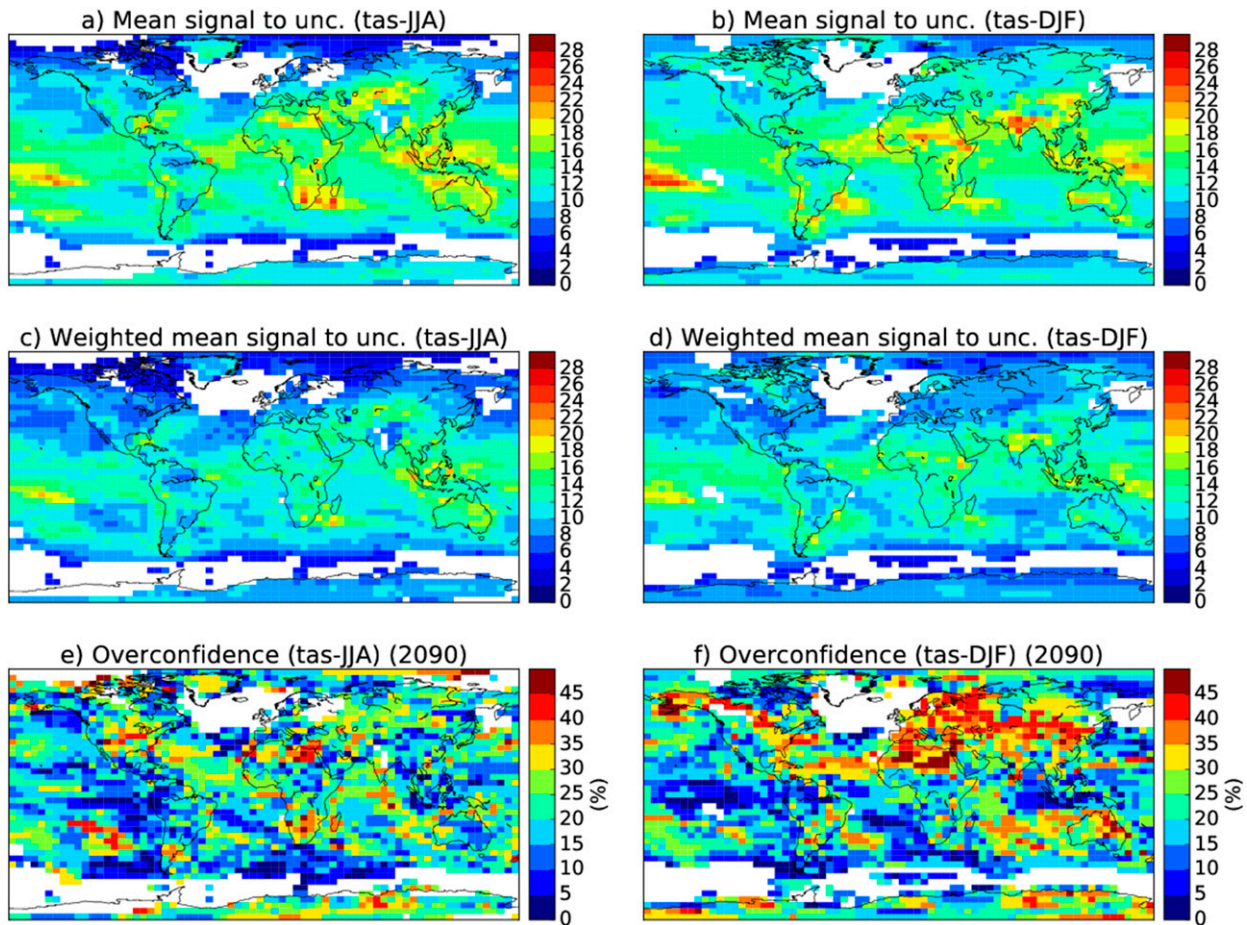


FIG. 8. Ensemble-mean signal (2080–2100 relative to 1980–2000) to uncertainty ratio for summer surface air temperature calculated using (a) the model-democratic [$\hat{\Delta}/\text{Std}(\hat{\Delta})$] and (c) the institute-democratic [$\hat{\Delta}'/\text{Std}(\hat{\Delta}')$] approaches. (e) The overconfidence given as the relative error of the statistical uncertainty $\{[\text{Std}(\hat{\Delta}') - \text{Std}(\hat{\Delta})]/\text{Std}(\hat{\Delta})\}$. (b), (d), (f) As in (a), (c), (e), but for winter. Uncolored areas are where the signal from at least one of the models is not statistically significant at the 5% level.

overconfidence associated with the arithmetic ensemble mean that might be prevented by considering institutional democracy. As shown in Fig. 8e, overconfidence is larger than 25% over several regions (e.g., North America, Asia, South Africa), while South America and Australia exhibit lower overconfidence ($<15\%$). Overconfidence exceeds 40% in winter (Fig. 8f) over several regions (parts of Canada, North Africa, Europe, and eastern Asia). These results provide some insight into the global spatial structure of ensemble overconfidence, while Steinschneider et al. (2015) have recently shown a similar impact of the effective ensemble size on the width of probability density functions for climate projections over some regions distributed across the United States.

4. Discussion and conclusions

In this study, we have investigated the impact of model similarities on regional climate change projections by

focusing on institutes that contributed more than one model (or version) to the CMIP3 multimodel ensemble. Typical differences among models are changes in resolution; numerical scheme; ocean eddy parameterization; or even entire components, such as atmosphere or ocean.

In the first part of the analysis, the climate change signals of surface air temperature at the end of the twenty-first century are averaged for each modeling center. The differences in climate change signals were then calculated for these pairs of same-center models and tested against the noise of internal variability. Strong model agreements (i.e., indistinguishable climate change signals) were found over most of the land regions for the CGCM, GFDL, GISS EH-ER, and UKMO pairs of models. On the other hand, model disagreements showed regional characteristics more or less easily attributable to the model differences. An interesting example is that of GISS EH-ER pair, which showed significantly different responses over

Hudson Bay and the North Atlantic. While these models share the same atmosphere, sea ice, land, and coupling components, they do differ in their ocean model. These results suggest that the lack of model independence can have an important impact on projections according to a given simulated variable and geographical location, while it could not be so for other cases. Another interesting result is that of the Hadley Centre models, which despite their important differences—they belong to different generations—nevertheless show striking agreements in their projections. This is consistent with [Rauser et al. \(2015\)](#), who showed that successive generations of climate models can overlap significantly in terms of performance. While the nature and meaning of similarities between Hadley Centre models are beyond the scope of this study, several questions are nonetheless prompted: from the role played by the shared radiative transfer parameterization package ([Edwards and Slingo 1996](#)) to the impact of developing models under the same “culture,” including here validation and tuning methods.

In the second part of the analysis, the robustness of the same-center hypothesis (i.e., whether same-center models should give similar results) was investigated by comparing the same-center and different-center model pairs using 1) the proposed metric of climate change disagreement and 2) a metric based on the correlation of model-error patterns ([Pennell and Reichler 2011](#)). The latter was shown to be a successful method to quantify model independence and has been used on several occasions ([Jun et al. 2008](#); [Reifen and Toumi 2009](#); [Pennell and Reichler 2011](#); [Knutti et al. 2010](#); [Haughton et al. 2015](#)). These metrics were compared in a cross-comparison context adapted for validating the same-center hypothesis (i.e., by using one model per family on each axis). The same-center models were more clearly discriminated in the model-evaluation metric than under the climate change disagreement metric. This result could be interpreted in different ways: it may suggest that our criterion for future climate needs further refinement to become a better discriminator, or it may suggest that the issue of model independence is of lesser importance in climate change projections than it is for simulating the observed mean state of climate. In essence, this would mean that model differences tend to be expressed more readily in climate-sensitivity differences than in present climate averages.

A finding related to the previous discussion is that strong agreements in climate change signals from same-center models tend to correspond to a strong correlation of errors in evaluation mode, whereas the opposite is not true: strong error correlations do not imply an indistinguishable climate change signal. Strong agreements in both climate change and evaluation mode are likely to be noninformative, given the underlying model

dependencies; caution should be exercised during the analysis of such an ensemble of climate change projections.

Comparing groups of same-center models with other arbitrary groups showed that model agreements in either climate change or evaluation mode may occur in groups of models created by different centers, as well. This is not surprising, because typical model similarities found using the same-center proxy method may also appear among the models of different institutions. This limitation of the same-center criterion can be twofold: models from different groups may share parts of their code, and they may share similar scientific premises based on different codes. The former is sometimes documented, but the information tends to be scattered across the climate modeling literature. The latter is much more difficult to establish and would entail a meticulous comparison of the codes describing each component, as well as their development histories.

Given the considerable scientific challenge in determining how model similarities play a causality role for model agreements in their response to GHGA forcing, we have taken for the sake of simplicity the default position of identifying as “noninformative agreements” the case when models that share multiple components or were developed within the same institution provide statistically indistinguishable projections. Hence, we have used the occurrence of such noninformative agreements in the ensemble to downweigh the same-center models on a per-cell basis. This procedure bears some resemblance to the calculation of an effective ensemble size ([Pennell and Reichler 2011](#); [Annan and Hargreaves 2011](#)) based on the pairwise correlations of model errors. These estimates suggest an effective ensemble size that could be as low as 25% of the actual number of models populating the CMIP3 archive. By applying the institute-democratic criterion to temperature change, we have found the current CMIP3 subset to be smaller than 65% of the actual ensemble size over several land regions by the end of the century.

Assuming the same-center criterion as an alternative definition of the effective ensemble size, it was then implemented within an ensemble-weighting scheme. Using such a technique, noninformative agreements can be filtered out from the ensemble to favor diversity of projections over the ensemble size. It was then shown how using the same-center criterion increased confidence intervals in the ensemble statistics. This result should be interpreted as an apparent loss of confidence, since the model-democratic case tends to provide overconfident results. Despite this change in confidence, the application of both types of averaging lead to ensemble statistics (mean and variance) that are very similar. This was partly attributed to the construction of

the current ensemble, but it is worth noting that an independence ranking is an efficient way to reduce the size of an ensemble while preserving its initial characteristics (Evans et al. 2013).

Recalling that the issue of model independence goes beyond the same-center context in the CMIP3 ensemble (Knutti et al. 2010), this is increasingly true for its successors (CMIP5 and CMIP6). These ensembles are less institution centric since cross-institutional dependencies become more common, as shown by the few examples cited in the introduction (e.g., CESM and NorESM being both forked from CCSM4). One strategy that could be undertaken by the climate modeling community to circumvent noninformative model agreements is to develop criteria of the kind suggested here to identify models of very similar nature. For example, the current framework—that applies the rule of institutional democracy to pairs of models—could be generalized for arbitrarily sized groups. Institutions were interpreted here as modeling centers, but one may think of more general definitions, such as groups formed from models with a recent or significant overlap in their development histories.

Centers that can afford to provide multiple models or model versions to a CMIP experiment should make an effort to show how the supplementary models add new information to the ensemble. From the statistical point of view, it is always better to have more simulations available, but it is also true that many users do not recognize the fact that the ensembles of opportunity, such as CMIP, do not necessarily consist of independent models and that the model-democratic approach is likely to yield overconfidence. “Flagging” models that exhibit noninformative agreements based on the same-center hypothesis proposed here could help users to process the information provided by the ensembles of opportunity. Similarly, parallel development of the same model by a different institution should also be made more explicit, as end users of climate model data are not always aware of model origin beyond the institution’s name.

Acknowledgments. This research was done within the Centre pour l’Étude et la Simulation du Climat à l’Échelle Régionale (ESCLER), funded by the Canadian Foundation for Climate and Atmospheric Sciences (CFCAS Grant NW CRCMD) and the Ouranos Consortium for Regional Climatology and Adaptation to Climate Change. This paper was also funded by the Canadian Network for Research and Innovation in Machining Technology (CRDPJ 386153-09) through the Natural Sciences and Engineering Research Council of Canada. The authors thank Ben Sanderson for the insightful discussions and suggestions, Mr. Mourad Labassi

for maintaining a user-friendly local computing facility, and Blaise Gauvin St-Denis for managing the CMIP3 data at Ouranos. We also acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP’s Working Group on Coupled Modelling (WGCM) for its role in making available the WCRP CMIP3 multimodel dataset. Support for this dataset is provided by the Office of Science, U.S. Department of Energy.

REFERENCES

- Abramowitz, G., and C. H. Bishop, 2015: Climate model dependence and the ensemble dependence transformation of CMIP projections. *J. Climate*, **28**, 2332–2348, doi:10.1175/JCLI-D-14-00364.1.
- Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, doi:10.1029/2009GL041994.
- , and —, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate*, **24**, 4529–4538, doi:10.1175/2011JCLI3873.1.
- Bishop, C. H., and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.*, **41**, 885–900, doi:10.1007/s00382-012-1610-y.
- Bleck, R., 2002: An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. *Ocean Modell.*, **4**, 55–88, doi:10.1016/S1463-5003(01)00012-9.
- Collins, W. D., and Coauthors, 2006a: Radiative forcing by well-mixed greenhouse gases: Estimates from climate models in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). *J. Geophys. Res.*, **111**, D14317, doi:10.1029/2005JD006713.
- , and Coauthors, 2006b: The Community Climate System Model version 3 (CCSM3). *J. Climate*, **19**, 2122–2143, doi:10.1175/JCLI3761.1.
- Déqué, M., and Coauthors, 2007: An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections. *Climatic Change*, **81**, 53–70, doi:10.1007/s10584-006-9228-x.
- , S. Somot, E. Sanchez-Gomez, C. Goodess, D. Jacob, G. Lenderink, and O. Christensen, 2012: The spread amongst ENSEMBLES regional scenarios: Regional climate models, driving general circulation models and interannual variability. *Climate Dyn.*, **38**, 951–964, doi:10.1007/s00382-011-1053-x.
- Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, doi:10.1007/s00382-010-0977-x.
- , A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**, 2271–2296, doi:10.1175/JCLI-D-13-00451.1.
- Di Luca, A., R. Elia, and R. Laprise, 2013: Potential for small scale added value of RCM’s downscaled climate change signal. *Climate Dyn.*, **40**, 601–618, doi:10.1007/s00382-012-1415-z.
- Edwards, J. M., and A. Slingo, 1996: Studies with a flexible new radiation code. I: Choosing a configuration for a large-scale model. *Quart. J. Roy. Meteor. Soc.*, **122**, 689–719, doi:10.1002/qj.49712253107.
- Edwards, P. N., 2011: History of climate modeling. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 128–139, doi:10.1002/wcc.95.

- 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, doi:10.1088/1748-9326/8/4/044050.
- 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.
- Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, doi:10.1175/2011JCLI4083.1.
- Haughton, N., G. Abramowitz, A. Pitman, and S. J. Phipps, 2014: On the generation of climate model ensembles. *Climate Dyn.*, **43**, 2297–2308, doi:10.1007/s00382-014-2054-3.
- , —, —, and —, 2015: Weighting climate model ensembles for mean and variance estimates. *Climate Dyn.*, **45**, 3169–3181, doi:10.1007/s00382-015-2531-3.
- Hawkins, E., 2011: Our evolving climate: Communicating the effects of climate variability. *Weather*, **66**, 175–179, doi:10.1002/wea.761.
- , and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1107, doi:10.1175/2009BAMS2607.1.
- , and —, 2011: The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dyn.*, **37**, 407–418, doi:10.1007/s00382-010-0810-6.
- Holmes, C. R., T. Woollings, E. Hawkins, and H. de Vries, 2016: Robust future changes in temperature variability under greenhouse gas forcing and the relationship with thermal advection. *J. Climate*, **29**, 2221–2236, doi:10.1175/JCLI-D-14-00735.1.
- Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, doi:10.1175/BAMS-D-12-00121.1.
- IPCC, 2000: *IPCC Special Report on Emissions Scenarios*. Cambridge University Press, 570 pp.
- , 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 996 pp.
- , 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp.
- Iversen, T., and Coauthors, 2013: The Norwegian Earth System Model, NorESM1-M—Part 2: Climate response and scenario projections. *Geosci. Model Dev.*, **6**, 389–415, doi:10.5194/gmd-6-389-2013.
- Johns, T. C., and Coauthors, 2006: The new Hadley Centre Climate Model (HadGEM1): Evaluation of coupled simulations. *J. Climate*, **19**, 1327–1353, doi:10.1175/JCLI3712.1.
- Jun, M., R. Knutti, and D. W. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.*, **103**, 934–947, doi:10.1198/016214507000001265.
- Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.
- Knutti, R., 2010: The end of model democracy? *Climatic Change*, **102**, 395–404, doi:10.1007/s10584-010-9800-2.
- , R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, doi: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, doi:10.1002/grl.50256.
- Levins, R., 1966: The strategy of model building in population biology. *Amer. Sci.*, **54**, 421–431.
- Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
- Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, doi:10.1029/2011GL046864.
- Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.*, **88**, 1383–1394, doi:10.1175/BAMS-88-9-1383.
- Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, **109**, 213–241, doi:10.1007/s10584-011-0156-z.
- Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.
- O’Neill, B. C., and Coauthors, 2016: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geosci. Model Dev.*, **9**, 3461–3482, doi:10.5194/gmd-9-3461-2016.
- Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, doi: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, doi:10.1016/j.envsci.2010.04.004.
- Rausser, F., P. Gleckler, and J. Marotzke, 2015: Rethinking the default construction of multimodel climate ensembles. *Bull. Amer. Meteor. Soc.*, **96**, 911–919, doi:10.1175/BAMS-D-13-00181.1.
- Reif, F., 1965: *Fundamentals of Statistical and Thermal Physics*. McGraw-Hill Series in Fundamentals of Physics, McGraw-Hill, 651 pp.
- Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704, doi:10.1029/2009GL038082.
- Russell, G. L., J. R. Miller, and D. Rind, 1995: A coupled atmosphere-ocean model for transient climate change studies. *Atmos.–Ocean*, **33**, 683–730, doi:10.1080/07055900.1995.9649550.
- , —, —, R. A. Ruedy, G. A. Schmidt, and S. Sheth, 2000: Comparison of model and observed regional temperature changes during the past 40 years. *J. Geophys. Res.*, **105**, 14 891–14 898, doi:10.1029/2000JD900156.
- Sanderson, B. M., and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, **39**, L16708, doi:10.1029/2012GL052665.
- , —, and P. Caldwell, 2015a: Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*, **28**, 5150–5170, doi:10.1175/JCLI-D-14-00361.1.
- , —, and —, 2015b: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, doi:10.1175/JCLI-D-14-00362.1.
- Scheffé, H., 1970: Practical solutions of the Behrens-Fisher problem. *J. Amer. Stat. Assoc.*, **65**, 1501–1508, doi:10.1080/01621459.1970.10481179.
- Seager, R., and Coauthors, 2007: Model projections of an imminent transition to a more arid climate in southwestern North America. *Science*, **316**, 1181–1184, doi:10.1126/science.1139601.
- 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, doi:[10.1002/2015GL064529](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, doi:[10.1002/env.2153](https://doi.org/10.1002/env.2153).
- 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, doi:[10.1175/BAMS-D-11-00094.1](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. London*, **A365**, 2053–2075, doi:[10.1098/rsta.2007.2076](https://doi.org/10.1098/rsta.2007.2076).
- von Storch, H., and F. Zwiers, 2013: Testing ensembles of climate change scenarios for statistical significance. *Climatic Change*, **117**, 1–9, doi:[10.1007/s10584-012-0551-0](https://doi.org/10.1007/s10584-012-0551-0).
- Washington, M. W., and Coauthors, 2000: Parallel climate model (PCM) control and transient simulations. *Climate Dyn.*, **16**, 755–774, doi:[10.1007/s003820000079](https://doi.org/10.1007/s003820000079).