Quantifying Sources of Uncertainty in Projections of Future Climate*

PAUL J. NORTHROP AND RICHARD E. CHANDLER

University College London, London, United Kingdom

(Manuscript received 20 March 2014, in final form 31 August 2014)

ABSTRACT

A simple statistical model is used to partition uncertainty from different sources, in projections of future climate from multimodel ensembles. Three major sources of uncertainty are considered: the choice of climate model, the choice of emissions scenario, and the internal variability of the modeled climate system. The relative contributions of these sources are quantified for mid- and late-twenty-first-century climate projections, using data from 23 coupled atmosphere–ocean general circulation models obtained from phase 3 of the Coupled Model Intercomparison Project (CMIP3). Similar investigations have been carried out recently by other authors but within a statistical framework for which the unbalanced nature of the data and the small number (three) of scenarios involved are potentially problematic. Here, a Bayesian analysis is used to overcome these difficulties. Global and regional analyses of surface air temperature and precipitation are performed. It is found that the relative contributions to uncertainty depend on the climate variable considered, as well as the region and time horizon. As expected, the uncertainty due to the choice of emissions scenario becomes more important toward the end of the twenty-first century. However, for midcentury temperature, model internal variability makes a large contribution in high-latitude regions. For midcentury precipitation, model internal variability is even more important and this persists in some regions into the late century. Implications for the design of climate model experiments are discussed.

1. Introduction

Currently, most projections of future global and regional climate are derived from the outputs of coupled atmosphere–ocean general circulation models (GCMs). Projections have historically been conditioned upon ‘‘scenarios’’ of greenhouse gas emissions, each associated with a particular ‘‘storyline’’ of economic and societal development worldwide throughout the twenty-first century (Nakićenović and Swart 2000). Although the most recent set of climate model runs have replaced these storylines with a set of representative concentration pathways (RCPs) that are not associated explicitly with socioeconomic storylines (van Vuuren et al. 2011), from a user perspective their role is similar. In particular, the choice of RCP, like the choice of storyline, will usually be a source of uncertainty in climate projections in the sense that different RCPs or storylines will lead to different projections. From here on, we refer to this uncertainty as ‘‘scenario uncertainty.’’ Other sources of uncertainty include the choice of GCM (different GCMs yield different projections for the same emissions scenario) and the choice of initial conditions for the GCM runs (different initial conditions yield different results). The extent to which results are dependent upon initial conditions can be regarded as a measure of internal variability within the modeled climate system.

The need to characterize uncertainty in projections of future climate is widely accepted, and this requires the use of multiple models, scenarios, and runs to explore the future climate response. However, it is expensive and time consuming to produce projections using a GCM, and it is therefore useful to identify which are the dominant sources of uncertainty in order to understand where to

Denotes Open Access content.

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-14-00265.s1.

Corresponding author address: Paul Northrop, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom.
E-mail: p.northrop@ucl.ac.uk

DOI: 10.1175/JCLI-D-14-00265.1

focus resources. If, for example, internal variability is relatively unimportant as a source of uncertainty, then it may be better to use resources to consider alternative scenarios and GCMs, rather than to produce many runs from the same GCM–scenario combination.

The problem of partitioning uncertainty in climate projections has been considered by several authors, notably Hawkins and Sutton (2009), who characterized projection uncertainty using heuristic measures of variability in ensembles of projections. Yip et al. (2011) took a more formal approach, carrying out an analysis of variance (ANOVA) to partition variability into contributions from different sources. ANOVA is a standard statistical technique for this task and, for balanced data, the decomposition of variability is unique and uncontroversial. In the current context a simple way to create balance is to stipulate that there are equal numbers of runs at each GCM–scenario combination. However, when data are unbalanced it is not clear how best to implement traditional ANOVA since the usual decomposition of variability is not unique (Searle et al. 2006, section 2.3b) so that it can be difficult to identify which sources are dominant.

In this study, we use data from the phase 3 of the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (CMIP3) multimodel dataset (Meehl et al. 2007), downloaded via the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). Yip et al. (2011) used a subset of these data to partition uncertainty in projections of global temperature, based on a classical ANOVA. However, the full dataset is highly unbalanced. Yip et al. (2011) dealt with this by considering only the seven GCMs that have more than one run for each scenario and by choosing exactly two runs for each GCM–scenario combination: a lot of relevant information was therefore discarded in their analysis.

We take a different approach to the problems caused by lack of balance in the data. We use a random-effects ANOVA (Searle et al. 2006; Gelman 2005). The effects (deviations from an overall mean) of individual GCM (G), scenario (S), and GCM–scenario (GS) combinations are treated as being randomly sampled from probability distributions representing respective superpopulations of effects, an idea mentioned in section 2c of Yip et al. (2011). One possible interpretation of such a superpopulation is that it represents the collection of all potential GCMs that could be constructed using combinations of model components and modeling decisions that are consistent with our current understanding of the climate system [see Stephenson et al. (2012) for more discussion of this point]. The superpopulation standard deviation (SD) σ_G of this distribution summarizes variability of effects in the superpopulation of GCMs. We can also estimate the finite-population standard deviation s_G: that is, an SD summarizing variability across the particular GCMs in the ensemble at hand. Section 3b gives more details. Both types of SD are useful: the finite-population SDs summarize variability in the effects of the particular GCMs that have been included in the ensemble under consideration and are thus analogous to the estimates of uncertainty from a classical ANOVA approach (Yip et al. 2011; Hingray et al. 2007; Raisanen 2001). The superpopulation SDs represent variability among the wider population of (actual and notional) GCMs from which the ensemble at hand is considered to be drawn.

In a similar spirit, the random effect associated with a scenario represents variability in a notional population from which the three scenarios represented in the ensemble are considered to have been drawn: the use of a random effect acknowledges that alternative scenarios are possible. With only three scenarios available in the ensemble at hand, however, it is hard to estimate superpopulation quantities with any great precision; thus, we expect, for example, that estimation of the superpopulation SD σ_S here will be subject to much greater uncertainty than that of the finite-population SD s_S.

In section 2, we describe the data and the climate change indices we derive from them. In section 3, we outline some existing approaches to partitioning uncertainty in climate projections and we describe a random-effects ANOVA model. In section 4, we fit this model to indices of mid- and late-twenty-first-century global temperature change. In section 5, we repeat this analysis at a regional scale for each of 22 regions, considering changes both in surface temperature and in precipitation. In section 6, we discuss some implications of our findings for the design of climate model experiments. Computer code to implement this methodology is available online (at http://www.homepages.ucl.ac.uk/~ucakpjn/).

2. CMIP3 data

Since 1992, the CMIP has coordinated several sets of climate model runs from modeling centers around the globe. Although the most recent set is phase 5 of the Coupled Model Intercomparison Project (CMIP5), released in 2013, in the work reported here we analyze the data from CMIP3 to provide absolute comparability with the work of Yip et al. (2011). The generic issues are exactly the same for any ensemble of GCM runs.

The CMIP3 multimodel dataset provides twenty-first-century climate projections from 24 GCMs under three future emissions scenarios developed by the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emissions Scenarios (SRES) (Nakićenović and Swart 2000). These scenarios are generally referred to as

TABLE 1. Numbers of runs for each combination of 24 GCMs and three socioeconomic scenarios (A1B, A2, and B1) for the global temperature experiments in the CMIP3 archive.

GCM No.	GCM acronym	GCM expanded name	A1B	A2	B1
1	BCCR-BCM2.0	Bjerknes Centre for Climate Research Bergen Climate Model, version 2.0	1	1	1
2	CGCM3.1	Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled Global Climate Model, version 3.1	5	5	5
3	CGCM3.1(T63)	Canadian Centre for Climate Modelling and Analysis (T63 spectral resolution)	1	0	1
4	CNRM-CM3	Centre National de Recherches Météorologiques Coupled Global Climate Model, version 3	1	1	1
5	CSIRO Mk3.0	Commonwealth Scientific and Industrial Research Organisation Mark 3.0	1	1	1
6	CSIRO Mk3.5	Commonwealth Scientific and Industrial Research Organisation Mark 3.5	1	1	1
7	GFDL CM2.0	Geophysical Fluid Dynamics Laboratory Climate Model, version 2.0	1	1	1
8	GFDL CM2.1	Geophysical Fluid Dynamics Laboratory Climate Model, version 2.1	1	1	1
9	GISS-AOM	Goddard Institute for Space Studies, Atmosphere–Ocean Model	2	0	2
10	GISS-E2-H	Goddard Institute for Space Studies Model E2, coupled with Hybrid Coordinate Ocean Model	3	0	0
11	GISS-ER	Goddard Institute for Space Studies Model E, coupled with the Russell ocean model	5	1	1
12	FGOALS-g1.0	Flexible Global Ocean–Atmosphere–Land System Model, gridpoint version 1.0	3	0	3
13	ECHAM4	—	1	1	0
14	INM-CM3.0	Institute of Numerical Mathematics Coupled Model, version 3.0	1	1	1
15	INM-CM4.0	Institute of Numerical Mathematics Coupled Model, version 4.0	1	1	1
16	MIROC3.2(hires)	Model for Interdisciplinary Research on Climate, version 3.2 (high resolution)	1	0	1
17	MIROC3.2(medres)	Model for Interdisciplinary Research on Climate, version 3.2 (medium resolution)	3	3	3
18	ECHO-G	ECHAM4 and the global Hamburg Ocean Primitive Equation	3	3	3
19	ECHAM5	—	4	3	3
20	MRI-CGCM2.3.2a	Meteorological Research Institute Coupled Atmosphere–Ocean General Circulation Model, version 2.3.2a	5	5	5
21	CCSM3.0	Community Climate System Model, version 3.0	7	5	8
22	HadCM3	Hadley Centre Coupled Model, version 3	1	1	1
23	HadGEM1	Hadley Centre Global Environment Model, version 1	1	1	0
24	PCM1	Parallel Climate Model, version 1	4	4	4
Total			57	40	48

A1B, A2, and B1 and may be interpreted as relating to low (B1), moderate (A1B), and high (A2) emissions of greenhouse gases over the twenty-first century. Table 1 gives the number of runs available (for surface air temperature) for each GCM–scenario combination. A total of 145 runs are available.

The numbers of runs for each scenario reflect choices made by individual modeling groups. Some GCMs have multiple runs per scenario; some have none. As noted above, this lack of balance complicates analysis of the data: the presence of zero cells in Table 1, corresponding to GCMs that provided no runs for a particular scenario, is particularly problematic in this respect. To deal with this in their analysis, Yip et al. (2011) included only seven GCMs (GCMs 2, 17–21, and 24 in Table 1) and chose exactly two runs for each GCM–scenario combination. In section 3b, we consider how to account for the lack of balance and the sparseness of the data in order to utilize all the CMIP3 data.

We consider two climate variables, (surface air) temperature (in °C) and precipitation (converted to mm day^{-1}), because they are the most frequently studied (Giorgi and Francisco 2000a; Giorgi and Mearns 2002; Tebaldi and Knutti 2007). We define indices of change for each variable. For temperature, the indices are the changes in mean temperature in the periods 2020–49 (midcentury) and 2069–98 (late century), each relative to the mean temperature in 1970–99. We use 2098 rather than 2099 because some GCMs did not provide runs for the whole of 2099. The precipitation indices are defined similarly except that we use the percentage—rather than absolute—change from the baseline 1970–99 period. In almost all cases, to ensure that each of our change indices can be regarded as if it was derived from a single long run,

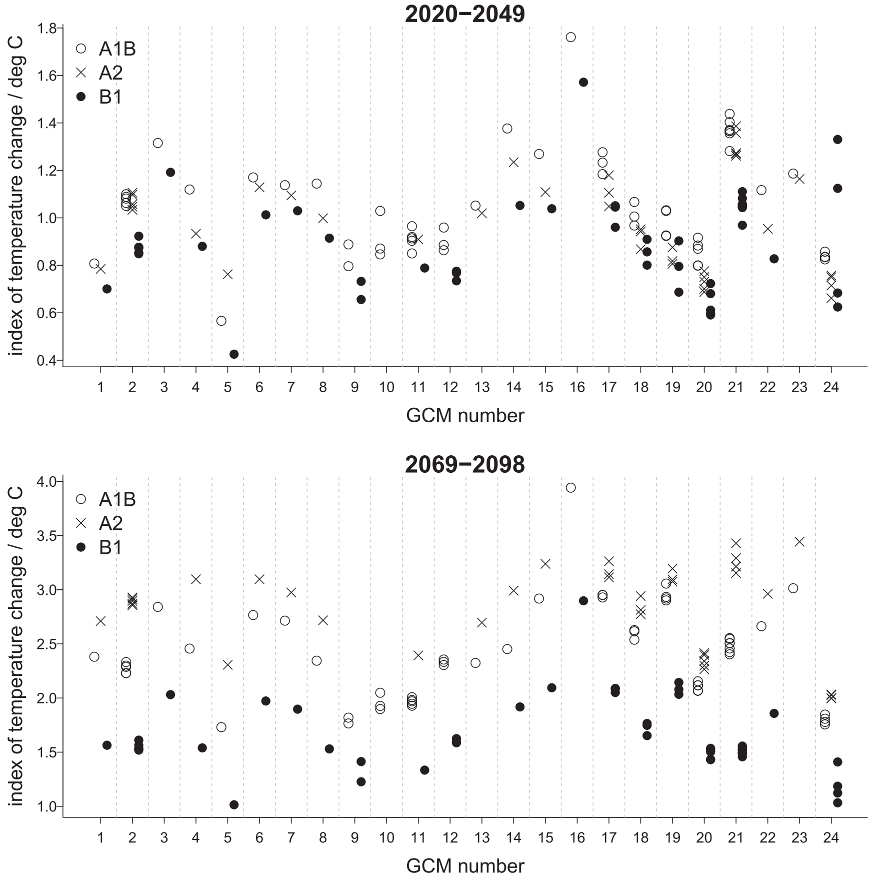

FIG. 1. Global temperature change indices for each available GCM–scenario combination: (top) 2020–49 and (bottom) 2069–98.

we take the 1970–99 data from the twentieth-century climate simulation (20C3M) model runs that were used to initialize the corresponding twenty-first-century runs (see http://www-pcmdi.llnl.gov/ipcc/time_correspondence_summary.htm). There are two exceptions. The first is GCM 15, for which there was an error in the forcings for the 20C3M run used to initialize the scenario runs (here, we have used the corrected 20C3M run, noting that, according to the URL given above, the climate of the year 2000 is very similar in the two runs). The other exception is GCM 24, for which the twenty-first-century runs were not initiated from a twentieth-century run. For this reason (i.e., to avoid using these data inappropriately), we will not include any data from GCM 24 in our analyses. We do, however, include values from GCM 24 in Fig. 1, because it has some relevance to comparisons with the results of Yip et al. (2011) made at the end of section 3a.

In section 4, we consider indices relating to changes in global mean temperature. In section 5, we carry out regional-scale analyses separately for each of the 22 regions considered by Giorgi and Mearns (2002). Giorgi and Francisco (2000b) provide the definitions of 20 of these regions in terms of latitude and longitude. The definitions of the two remaining regions, northern Australia and southern Australia, are given by the IPCC Data Distribution Centre (http://www.ipcc-data.org/sres/scatter_plots/scatterplots_region.html).

The raw data are monthly averages generated on a coarse GCM-specific spatial grid. Following Giorgi and Mearns (2002), for each GCM the data for a given month are spatially interpolated onto a common 0.5° grid using the bicubic spline interpolation function interp() in the R library akima (Gebhardt et al. 2013), before being averaged over each region of interest. Then the monthly averages are converted into averages over the time periods of interest, weighted by the cosine of the latitude of the grid point location, from which the respective indices of change are derived.

Figure 1 summarizes the results of this procedure when applied to surface air temperature over the entire globe. For a given GCM the values under scenario B1 are generally lower than under scenarios A1B and A2. The exception is GCM 24, for which there are two unusually large values for 2020–49: as discussed above, the reason for this is that the twenty-first-century runs for this GCM were not initialized using the corresponding twentieth-century runs and are therefore not directly comparable. Overall, the modeled temperatures for scenarios A1B and A2 are similar for 2020–49, but A2 tends to produce larger values than A1B for 2069–98.

3. Statistical models for partitioning variability

In the following, we let $Y_{ijk}; i=1,\ldots,n_G; j=1,\ldots,n_s$; and $k=1,\ldots,K_{ij}$ be an index of change for GCM i, scenario j, and run k. For the CMIP3 data, $n_G=23$, $n_S=3$, and K_{ij} varies between 0 and 8 (see Table 1).

a. A fixed-effects ANOVA model

Hawkins and Sutton (2009) proposed some heuristic ways of partitioning variability. Yip et al. (2011) put that work on a more formal footing using a statistical model: namely, a two-way fixed-effects ANOVA,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad i=1,\ldots,n_G,$$
$$j=1,\ldots,n_S, k=1,\ldots,K_{ij}, \qquad (1)$$

where μ is the overall mean change in the index, over all GCMs and scenarios; α_i is an adjustment for GCM i; β_j is an adjustment for scenario j; γ_{ij} is a scenario-specific additional adjustment for GCM i; and the error terms ϵ_{ijk} are independent identically distributed random variables with mean 0 and variance σ^2, representing residual variability between runs: this can be considered as variability that is internal to the modeled system.

The GCM–scenario interaction effects $\{\gamma_{ij}\}$ are referred to as interaction effects and measure how variability over GCMs changes with scenario or, equivalently, how the variation between scenarios differs between GCMs: as, for example, with the mid-twenty-first-century projections of global temperature in Fig. 1, where one GCM seems to rank scenario B1 differently from most of the others, as discussed above.

Consider a balanced design with $K_{ij}=K>1$ for all i,j (i.e., K runs for each GCM–scenario combination) and constraints $\sum_i\alpha_i=\sum_j\beta_j=0$, $\sum_j\gamma_{ij}=0$ for $i=1,\ldots,n_G$ and $\sum_i\gamma_{ij}=0$ for $j=1,\ldots,n_S$ to avoid parameter redundancy. If $K=1$ it is not possible to estimate both the interaction effects and the error variance σ^2 without making extra assumptions about the form of the interaction effects.

We define the overall mean $\overline{Y}_{\cdots}=(1/n_G n_S K)\sum_i\sum_j\sum_k y_{ijk}$. Using a similar "bar–dot" notation, where dots are used to indicate suffices over which averaging has taken place, GCM-specific means are $\overline{Y}_{i\cdot\cdot}$, $i=1,\ldots,n_G$; scenario-specific means are $\overline{Y}_{\cdot j\cdot}$, $j=1,\ldots,n_S$; and means for specific GCM–scenario combinations are $\overline{Y}_{ij\cdot}$, $i=1,\ldots,n_G$, $j=1,\ldots,n_S$. The least squares estimates of the quantities in (1) are $\hat{\mu}=\overline{Y}_{\cdots}$, $\hat{\alpha}_i=\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots}$, $\hat{\beta}_j=\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots}$, $\hat{\gamma}_{ij}=\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot j\cdot}+\overline{Y}_{\cdots}$, and $\hat{\epsilon}_{ijk}=Y_{ijk}-\overline{Y}_{ij\cdot}$ (Yip et al. 2011).

The quantities used by Yip et al. (2011) to quantify uncertainty attributable to model, scenario, model–scenario interaction, and internal (between run) variation are $M=\sum_i\hat{\alpha}_i^2/n_G$, $S=\sum_j\hat{\beta}_j^2/n_S$, $I=\sum_i\sum_j\hat{\gamma}_{ij}^2/n_G n_S$, and $V=\sum_i\sum_j\sum_k\hat{\epsilon}_{ijk}/n_G n_S K$, respectively. These quantities are proportional to the usual mean squares in an ANOVA framework. However, standard formulas for the expected values of these mean squares show that interpretation of these quantities is not as straightforward as it first appears. For example, to compare the relative contributions of scenario choice and internal variability to the total uncertainty under model (1), a natural measure is $[n_S^{-1}\sum_j\beta_j^2]/\sigma^2$, and it is tempting to estimate this as S/V. However, the results of Searle et al. (2006, section 4.3) show that the expected value of S is $(n_S-1)(n_G n_S K)^{-1}\sigma^2+n_S^{-1}\sum_{i=1}^{n_G}\beta_j^2$, and the expected value of V is $(K-1)K^{-1}\sigma^2$. It is clear that the ratio S/V will tend to overestimate the quantity of interest, therefore, because of the presence of $(n_S-1)(n_G n_S K)^{-1}\sigma^2$ in the expected value of S and the factor $(K-1)K^{-1}$ in the expected value of V. Bias in the estimation of $n_S^{-1}\sum_j\beta_j^2$ by S may be small if σ^2 is small. However, V may very substantially underestimate σ^2 if K is small (e.g., by a factor of 2 if $K=2$), an issue noted by Déqué et al. (2007). Similar issues arise in the comparison of other measures of variability from model (1).

When the design is unbalanced interpretation of M, S, I, and V is even less clear. This motivates use of a related framework under which these problems can be addressed without discarding data. We achieve this using a random-effects ANOVA model (see section 3b), defining explicitly the quantities to be inferred from data.

In Fig. 2, we compare the results from the methodology of Yip et al. (2011, their Figs. 3a and 4b) with the corresponding plots obtained using our approach described below. In reproducing the Yip et al. (2011) plots we have used a slightly different baseline period (1970–99 rather than 1971–2000), because for some of the models the SRES experiment data start in 2000; further, since they included runs from GCM 24, we have done the same here, selecting the two runs that did not result in obviously anomalous behavior. Although there are some differences in the relative contributions from internal variation, in this case σ^2 is small enough that its presence in the expected value of S has little impact. The most interesting difference is that in Yip et al. (2011) (and in Hawkins and Sutton 2009) scenario uncertainty becomes more important than GCM uncertainty at approximately 2050, whereas under our approach this does

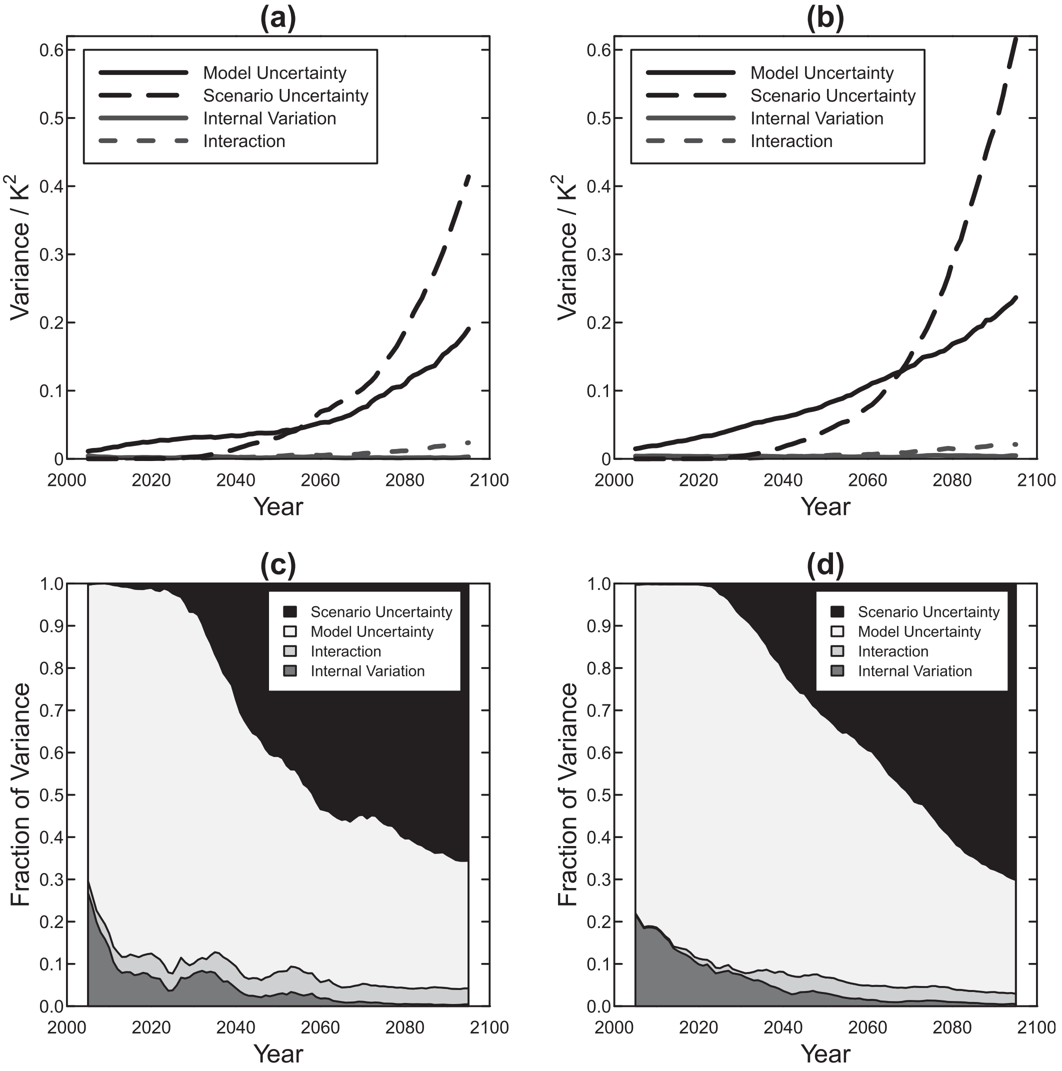

FIG. 2. Comparison of (left) Yip et al. (2011) with (right) the proposed approach. Sources of uncertainty in global, decadal CMIP3 projections of global temperature change, relative to the baseline period 1970–99. (a),(b) Estimates of variances M, S, I, and V (Yip et al. 2011) and finite-population variances s_G^2, s_S^2, s_{GS}^2, and s_R^2. (c),(d) Estimates as a fraction of the sum of the variances.

not occur until approximately 2070. These differences are the result of using different statistical approaches and different data and will depend on the behavior of the datasets involved. However, the plots in the top row of Fig. 2 suggest that, in this particular case, the overall effect of including data from more models, as well as more runs from existing models, is to increase estimates of model uncertainty and (toward the end of the twenty-first century when the differences between scenarios become apparent) scenario uncertainty.

b. A random-effects ANOVA model

In the fixed-effects ANOVA model (1), standard analysis methods focus on the specific effects $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\gamma_{ij}\}$ for the ensemble under consideration. By contrast, in a random-effects version of the same model (see, e.g., Searle et al. 2006) individual effects are not of direct interest but rather are considered to be sampled from some larger population, and it is the variability within this larger population that is the focus of the analysis.

Equation (1) still applies: that is,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, n_G,$$
$$j = 1, \ldots, n_S, k = 1, \ldots, K_{ij}, \qquad (2)$$

but now α_i is considered to be a normally distributed random variable with mean zero and variance σ_G^2 [we write $\alpha_i \sim N(0, \sigma_G^2)$, with similar notation for other random variables]; $\beta_j \sim N(0, \sigma_S^2)$; $\gamma_{ij} \sim N(0, \sigma_{GS}^2)$; $\epsilon_{ijk} \sim N(0, \sigma_R^2)$; and all random variables are assumed to be independent.

The terms in (2) have the same interpretation as those in (1) but now, instead of focusing on the individual effects (the α, β, and γ) we are interested in (the relative magnitudes of) the superpopulation SDs σ_G, σ_S, σ_{GS}, and σ_R as representing a partitioning of uncertainty that acknowledges the potential for additional GCMs and emissions scenarios that are not represented in the data available. We also examine the relative magnitudes of the finite-population SDs s_G, s_S, s_{GS}, and s_R, where, for example, s_G is defined via $s_G^2 = (1/22)\sum_{i=1}^{23}(\alpha_i - \overline{\alpha})^2$ and $\overline{\alpha} = (1/23)\sum_{i=1}^{23}\alpha_i$, as these summarize variabilities within the ensemble at hand. The presence of the error terms ϵ_{ijk} in (2) mean that parameters in the finite-population SDs cannot be observed exactly; they can only be estimated and the estimates have some uncertainty associated with them. However, α_i, for example, may be estimated precisely if there are many runs under GCM i. For a discussion of superpopulation and finite-population effects, see Gelman and Hill (2003, section 21.2).

c. *Statistical inference for random-effects ANOVA*

In situations where there are reasonably large numbers of groups (corresponding to GCMs and scenarios here) but where the data are unbalanced, random-effects models such as (2) are often fitted using restricted maximum likelihood (REML) estimation (Patterson and Thompson 1971; Harville 1977), with standard errors and confidence intervals computed using simulation (e.g., a 95% confidence interval for a particular parameter such as σ_G is obtained from the 2.5% and 97.5% sample percentiles of fits from simulated datasets). We illustrate this procedure below, using the R library lme4 (Bates et al. 2014) to perform REML estimation.

However, with only three scenarios there is little information in the data about σ_S. In such situations Gilmour and Goos (2009) argue against the use of REML because σ_S can be underestimated, estimates of zero may be produced, and estimates of uncertainty tend to underestimate the true uncertainty. In such situations, a Bayesian analysis may be preferable.

In Bayesian inference (Bernardo and Smith 2003) the parameter vector θ of a model, here $\theta = (\mu, \sigma_G, \sigma_S, \sigma_{GS}, \sigma_R)$, is treated as a random variable. A prior distribution $\pi(\theta)$, representing uncertainty about θ in the absence of the data y, is specified. Let $L(\theta; y)$ denote the likelihood function: that is, probability density of y as a function of θ. Then inference is based on the posterior distribution which is proportional to $L(\theta; y)\pi(\theta)$. In all but the simplest problems, an explicit expression for the posterior distribution is not available. However, samples from it may be obtained using Markov chain Monte Carlo (MCMC) techniques (Gilks et al. 1996; Gelman and Rubin 1992); by drawing sufficiently large samples, we may characterize any aspect of the posterior distribution (e.g., the mean, median, and percentiles) to any desired degree of accuracy. In particular, a 95% credible interval for any quantity of interest is determined by the 2.5% and 97.5% percentiles of its posterior distribution. In the work reported below, MCMC sampling is carried out using WinBUGS (Lunn et al. 2000) via the R library arm (Gelman et al. 2010).

Operationally, perhaps the most obvious difference between Bayesian and other methods of statistical inference is the incorporation of the prior distribution $\pi(\theta)$. Ideally, this represents the analyst's uncertainty about the model parameters θ in the absence of any data; often a noninformative prior is used in an attempt to ensure that the results are not influenced by what are seen as the analyst's subjective judgments. In situations such as that considered here, however, the data themselves may provide relatively little information about some model parameters such as σ_S. In such cases, it may be worth specifying a weakly informative prior (Gelman 2006) that encapsulates some basic constraints on the parameters but that otherwise allows the likelihood component to dominate the posterior distribution. If the prior is approximately flat over the range of θ values that are consistent with the data, then the results from a Bayesian analysis will be dominated by the contribution of the likelihood to the posterior in this range so that the influence of the prior can be considered unimportant: in such situations, we expect good agreement between REML estimates and the mode of the posterior distribution [i.e., the value θ for which the posterior density $\pi(\theta \mid y)$ is maximized]. In a Bayesian setting however, estimation is usually based on the mean of the posterior distribution rather than its mode [the reasons for this are set out in Gilmour and Goos (2009)] and, if the posterior is highly skewed, its mean and mode can be very different. This in itself can account for differences between Bayesian and frequentist analyses. In the current context, the posterior distribution of σ_S is highly positively skewed (see Figs. 3 and 4), so the posterior mean is much greater than the posterior mode and a Bayesian analysis is probably preferable, as explained in Gilmour and Goos (2009).

Gelman (2006) considers what kind of prior distribution should be placed on a variance component σ when the available data provide only limited information about it: in the present context, this is the situation for σ_S because data are available for just three emissions scenarios. He shows that a weakly informative prior is

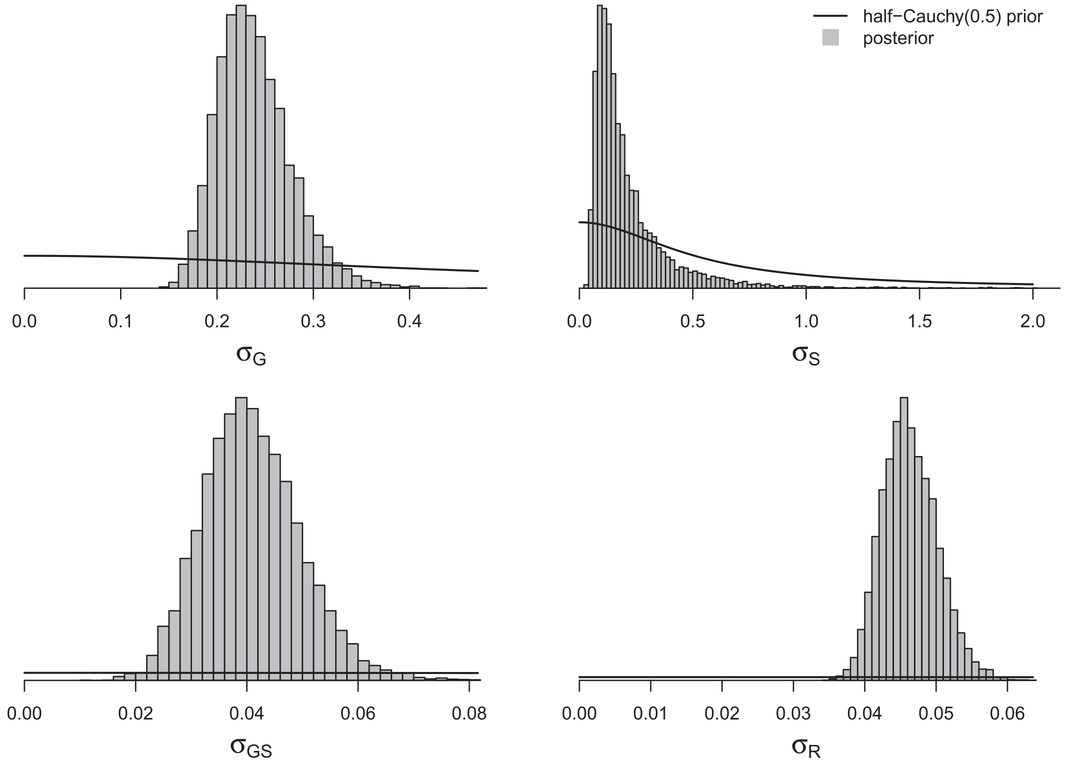

FIG. 3. Global temperature change 2020–49. Posterior distributions of σ_G, σ_S, σ_{GS}, and σ_R (histogram) are based on half-Cauchy (0.5) prior distributions (solid lines).

necessary, to downweight the posterior probability of physically implausible values. Gelman (2006) and N. G. Polson and J. G. Scott (2012) demonstrate that a half-Cauchy (A) prior, with probability density function (pdf),

$$p(\sigma) = \frac{2}{\pi A}\left(1 + \frac{\sigma^2}{A^2}\right)^{-1}, \quad \sigma > 0, \qquad (3)$$

is more appropriate than the commonly used uniform, log-uniform, or inverse-gamma priors. For a suitable value of A the half-Cauchy prior encourages the posterior distribution for σ to place high probability on a realistic range. However, the prior has a ‘‘heavy tail’’ (i.e., the pdf decays slowly as σ increases), and this prevents the prior from having an undue influence if the data suggest a larger than anticipated value of σ. This behavior is demonstrated in the plots relating to σ_S in Figs. 3 and 4.

4. Global temperature change

In this section, we present the results of REML and Bayesian analyses of the global temperature data using model (2). For the latter, we use independent half-Cauchy priors for the variance components, choosing the scale parameter A to provide weak prior information. For the 2020–49 indices we use $A = 0.5$°C and for 2069–98 we use $A = 1$°C, based on the following reasoning: Suppose that two of the sources (GCM, scenario, and simulation run) of uncertainty are kept constant while the other is varied. It would be very unlikely that the resulting projections would have a range of more than, say, 10°C in their temperature projections by the midcentury and more than, say, 20°C by the end of the century. Under model (2), as a rule of thumb the range of the random effects from each source of uncertainty can be considered to correspond very roughly to four standard deviations (or 4σ); thus, for each source of uncertainty, we judge that the corresponding random-effects standard deviation σ does not exceed 2.5°C for the midcentury projections and 5°C for end-of-century projections. The chosen values of A place small ($\approx$0.13) prior probability on these eventualities. We use the same prior distribution for all the σ parameters. We use a noninformative $N(0, 10^6)$ prior for the mean parameter μ.

a. Results: Sources of uncertainty

Figures 3 and 4 show the prior distributions and posterior distributions of the superpopulation SDs. For σ_G, σ_{GS}, and σ_R the information provided by the data via the likelihood has dominated, as was intended. This can be

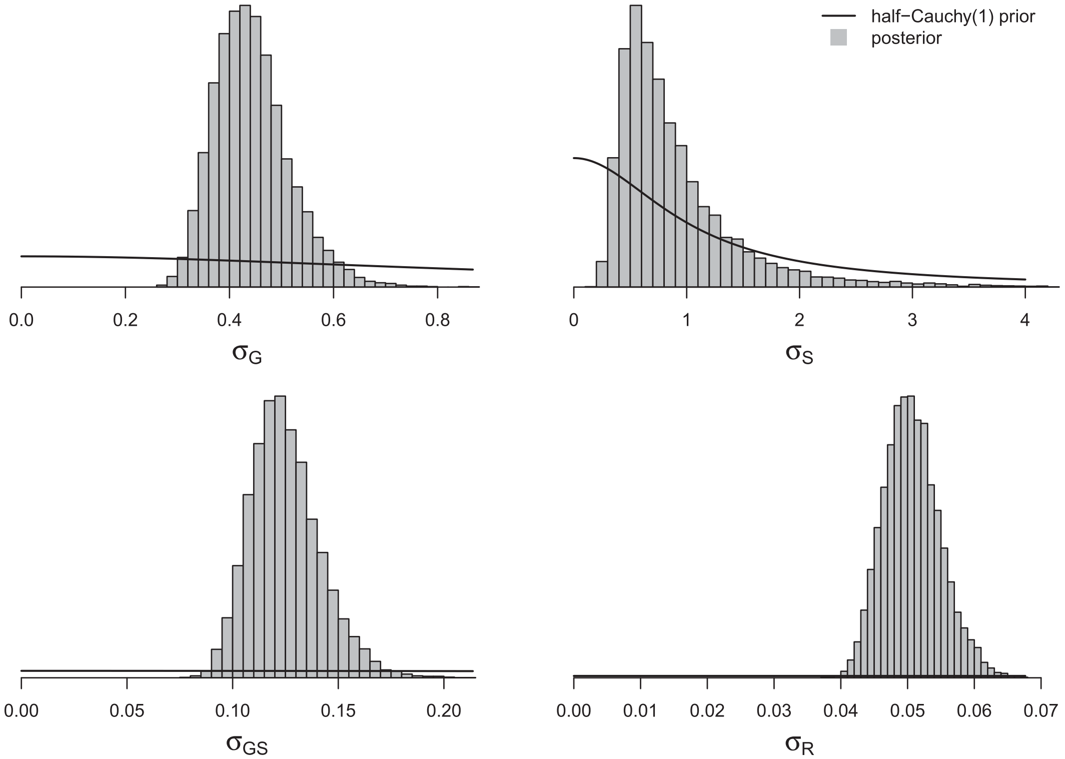

FIG. 4. Global temperature change 2069–98. Posterior distributions of σ_G, σ_S, σ_{GS}, and σ_R (histogram) are based on half-Cauchy (1) prior distributions (solid lines).

inferred from the fact that the priors are virtually flat over intervals for which the posteriors are nonnegligible. This is not the case for σ_S: the half-Cauchy prior provides weak information to prevent very unrealistic values of σ_S from appearing in the posterior simulations.

Table 2 summarizes inferences from the REML and Bayesian analyses. As anticipated from the discussion in section 3c, the REML point estimates of σ_S are much smaller than the Bayesian point estimates (and similar to the posterior modes in Figs. 3 and 4) and the REML-based confidence intervals are much narrower than their Bayesian equivalents. As noted by Gilmour and Goos (2009), these features of the REML inferences are not desirable: they reflect a lack of information about σ_S rather than strong evidence that its value is small. A further consequence is that the Bayesian interval estimates for μ are wider than those from the REML analysis. For the other variance components there are

TABLE 2. REML and Bayesian inferences of the global temperature data based on model (2). REML: estimate, standard error, and 95% confidence interval. Bayes: posterior mean (median), posterior standard deviation, and 95% credible interval.

		2020–49			2069–98		
	Analysis	Estimate/posterior mean (median)	Std error/ SD	95% confidence/ credible interval	Estimate/posterior mean (median)	Std error/ SD	95% confidence/ credible interval
μ	REML	1.085	0.072	(0.941, 1.220)	2.473	0.353	(1.775, 3.167)
	Bayes	1.091 (1.090)	0.176	(0.754, 1.422)	2.474 (2.479)	0.581	(1.339, 3.553)
σ_G	REML	0.231	0.036	(0.156, 0.301)	0.433	0.069	(0.297, 0.565)
	Bayes	0.239 (0.234)	0.038	(0.177, 0.327)	0.445 (0.437)	0.074	(0.329, 0.614)
σ_S	REML	0.097	0.046	(0.009, 0.190)	0.601	0.279	(0.093, 1.167)
	Bayes	0.217 (0.156)	0.204	(0.058, 0.723)	0.804 (0.660)	0.564	(0.311, 2.160)
σ_{GS}	REML	0.039	0.009	(0.018, 0.055)	0.120	0.016	(0.089, 0.150)
	Bayes	0.041 (0.040)	0.009	(0.025, 0.060)	0.123 (0.121)	0.017	(0.095, 0.161)
σ_R	REML	0.045	0.004	(0.038, 0.053)	0.050	0.004	(0.041, 0.058)
	Bayes	0.046 (0.046)	0.004	(0.040, 0.055)	0.051 (0.050)	0.004	(0.043, 0.059)

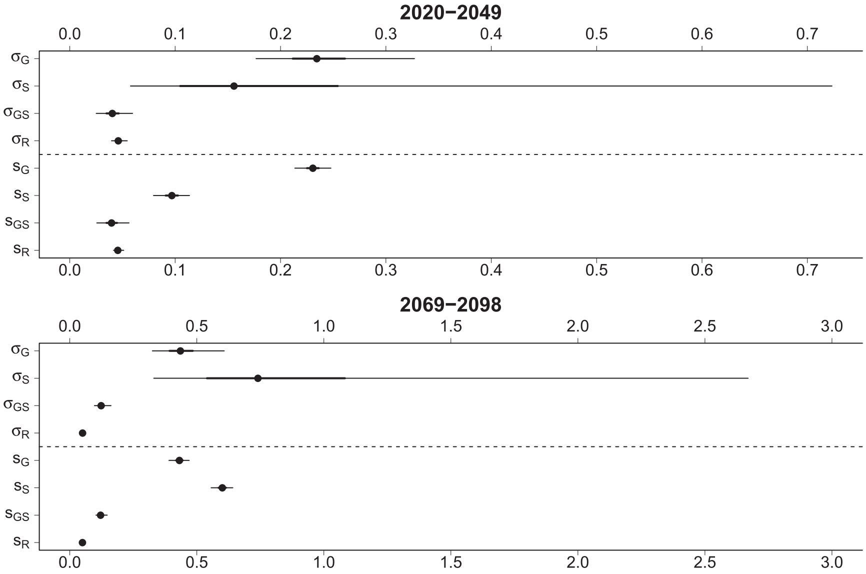

FIG. 5. Global temperature change. Summaries of the posterior distributions for the superpopulation standard deviations (σ_G, σ_S, σ_{GS}, and σ_R) and finite-population standard deviations (s_G, s_S, s_{GS}, and s_R) are shown. Medians (dots), 50% intervals (thick lines), and 95% intervals (thin lines) are plotted. The 50% intervals extend from the 25th to 75th percentiles of the respective posterior distributions; 95% intervals extend from the 2.5th to 97.5th percentiles. Periods are (top) 2020–49 and (bottom) 2069–98.

only small differences between the REML and Bayesian inferences.

Figure 5 summarizes the posterior distributions of the superpopulation and finite-population standard deviations. The posteriors of the superpopulation SDs are positively skewed: relatively small values of these quantities are inconsistent with the data, but relatively large values cannot be ruled out. As expected (section 3c), the finite-population SDs have smaller posterior medians and narrower interval estimates than the superpopulation SDs. This is especially pronounced for scenario (σ_S and s_S).

In any Bayesian analysis, it is helpful to explore the sensitivity of results to a plausible range of prior distributions. Here, we have repeated the analysis above for different values of A in the prior (3). We find that only inferences about σ_S are sensitive to this choice. For example, for 2020–49, the estimated median and 95% credible interval for σ_S are 0.136°C and (0.054, 0.542)°C for $A = 0.25$ and 0.166°C and (0.058, 0.875)°C for $A = 1$. For σ_G these values are 0.234°C and (0.176, 0.330)°C for $A = 0.25$ and 0.243°C and (0.182, 0.342)°C for $A = 1$. The estimated posterior distributions of the finite-population SDs are virtually constant over a wide range of values of A.

The superpopulation standard deviation estimates in Table 2 and Fig. 5 reveal the dominant sources of uncertainty in projections of global mean temperature. For the 2020–49 time horizon, σ_G has the largest posterior mean value, closely followed by σ_S. This suggests that over this time period the dominant source of uncertainty is the choice of GCM, followed by the choice of emissions scenario. Later in the century (2069–98), however, variability over scenarios is greater than variability over GCMs. The same general findings apply if we compare finite-population SDs, although the importance of scenario is reduced because the posterior medians of s_S are smaller than the respective posterior medians of σ_S. In both time periods, internal variability (represented by the residual standard deviation σ_R) is estimated to be much smaller than variability over GCMs and over scenario. Internal variability is estimated to be smaller in 2069–98 than in 2020–49. It could be that global warming reduces variability in global temperature. By 2069–98, variability attributable to GCM–scenario interaction has overtaken internal variability.

b. *Results: Individual GCMs and scenarios*

The top plots in Figs. 6 and 7 summarize the posterior distributions of the effects of individual GCMs and scenarios on our climate indices: that is, $(\mu + \alpha_1, \ldots, \mu + \alpha_{23})$ and $(\mu + \beta_1, \mu + \beta_2, \mu + \beta_3)$, respectively. We can see that GCM 16 [MIROC3.2(hires)] gives atypically high projections of global temperature change and that projected temperature changes are greatest under scenario A1B in the midcentury and under scenario A2 in the late century. The plots on the bottom left in these figures show how GCM-specific effects [the terms $\{\mu + \gamma_{ij}\}$ in model (2)] vary with scenario. For clarity here, we

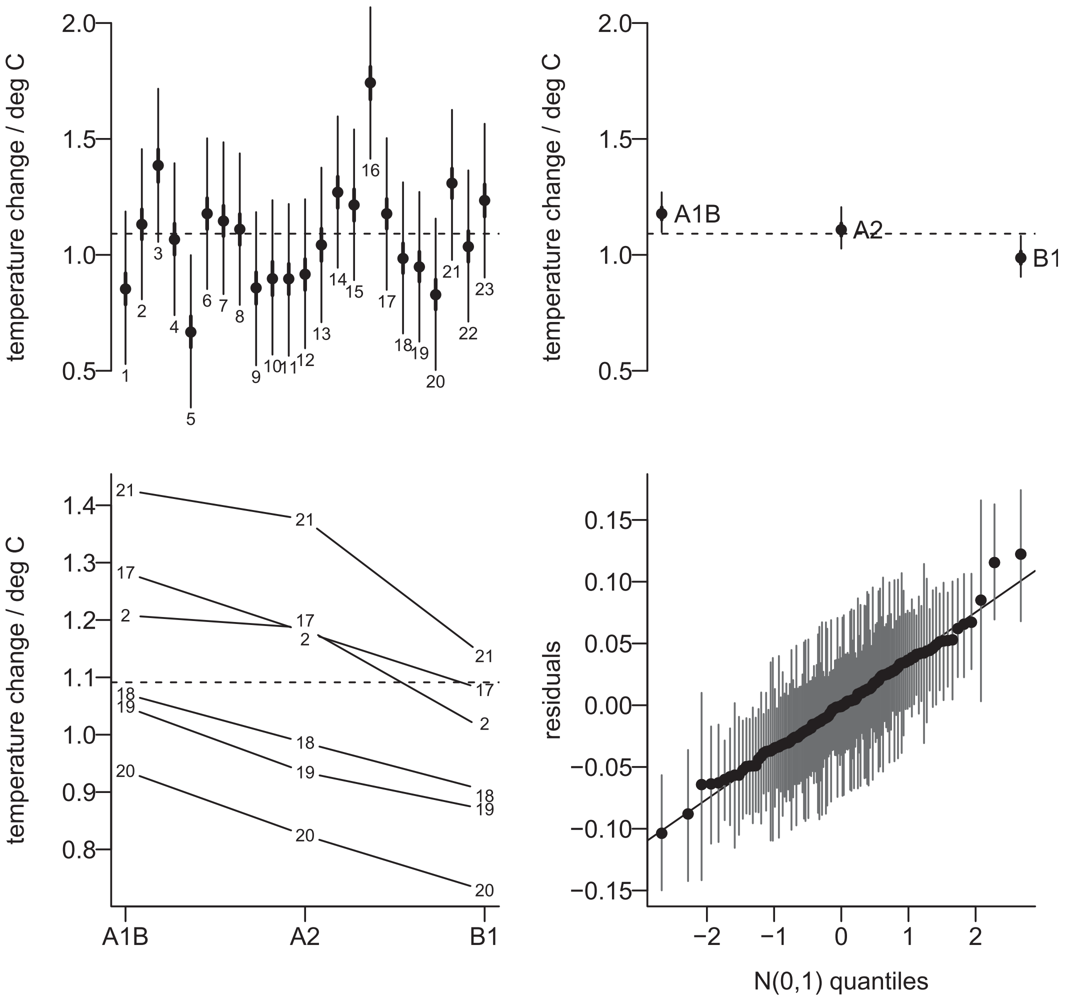

FIG. 6. Summaries of posterior distributions for global temperature change for 2020–49. (top) Medians (dots), 50% intervals (thick lines), and 95% intervals (thin lines) for each (left) GCM and (right) scenario. (bottom left) GCM–scenario interaction plot of posterior medians by GCM and scenario, for selected GCMs. The labels give the GCM number. The dashed horizontal lines are drawn at the posterior mean of the overall mean change μ. (bottom right) Normal QQ plot of the posterior median of data-level errors. The vertical gray lines are 95% intervals for the data-level errors.

have plotted only the six GCMs considered by Yip et al. (2011) that remain after the exclusion of GCM 24.

The bottom right of Figs. 6 and 7 are normal quantile–quantile (QQ) plots of the posterior medians of the data-level errors $\epsilon_{ijk} = y_{ijk} - (\mu + \alpha_i + \beta_j + \gamma_{ij})$; $i = 1, \ldots, 23$; $j = 1, 2, 3$; and $k = 1, \ldots, K_{ij}$: their purpose is to check the assumption in model (2) that these errors are normally distributed, since in this case the points on the QQ plots should lie roughly on a straight line. For 2020–49 (Fig. 6) the points lie remarkably close to the line. For 2069–98 (Fig. 7) the curvature might suggest a slight (positive) skew in the error distribution: this is driven by single runs with values that are greater than those of their counterparts (see, e.g., scenario A2 for GCM 21 and scenario A1B for GCM 19 in Fig. 1).

c. *Model checking*

We have also carried out posterior predictive checks (see, e.g., Gelman et al. 2003, chapter 6) to assess whether the model is consistent with the data. We compare the real data to 10 000 datasets simulated from the posterior predictive distribution under the model. The real data should not behave very differently to the simulated datasets. To examine this, we choose test summaries to reflect important aspects of the data. The test summaries we use are based on derived datasets containing (i) all responses; (ii) the 23 mean responses for each GCM; (iii) the 3 mean responses for each scenario; (iv) the 64 mean responses for each GCM–scenario combination; and (v) for GCM–scenario

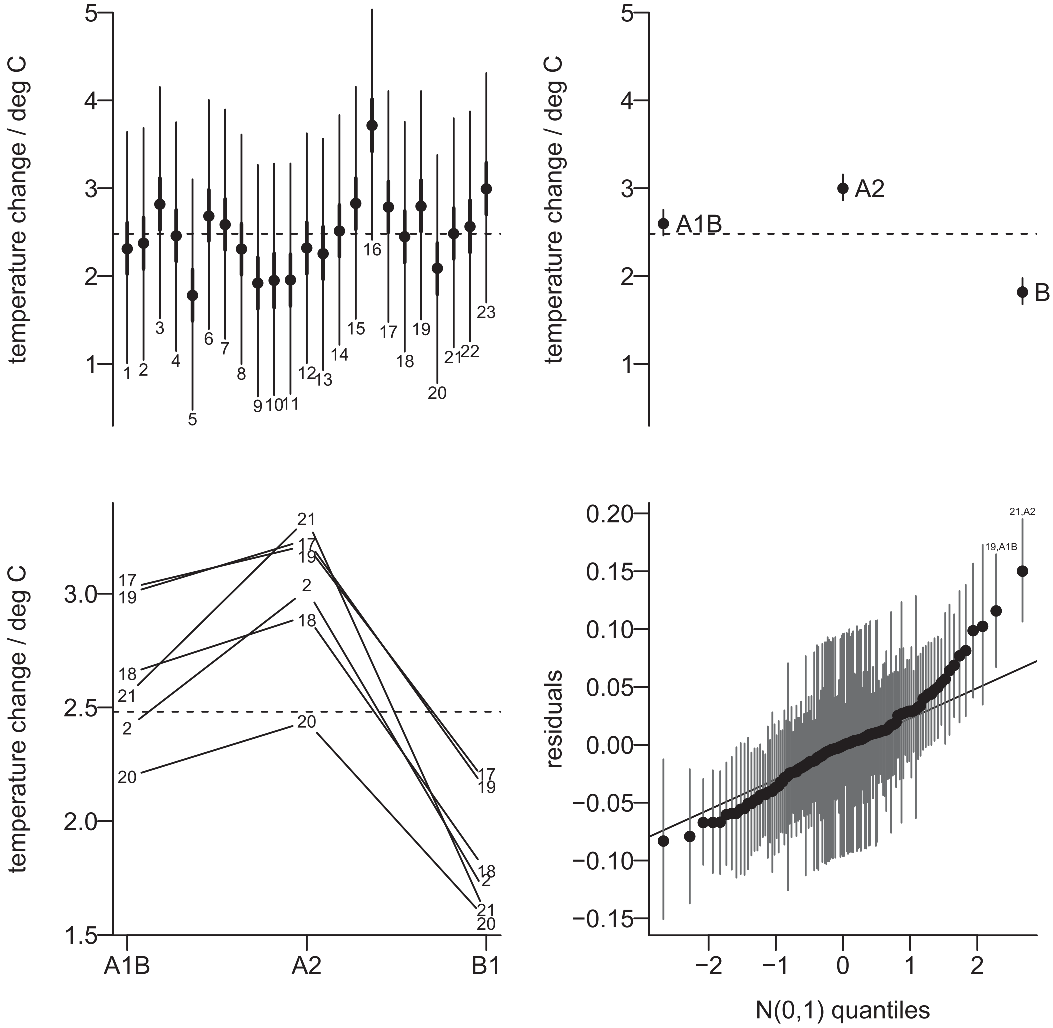

FIG. 7. Summaries of posterior distributions for global temperature change 2069–98. (top) Medians (dots), 50% intervals (thick lines), and 95% intervals (thin lines) for each (left) GCM and (right) scenario. (bottom left) GCM–scenario interaction plot of posterior medians by GCM and scenario, for selected GCMs. The labels give the GCM number. The dashed horizontal lines are drawn at the posterior mean of the overall mean change μ. (bottom right) Normal QQ plot of the posterior median of data-level errors. The vertical gray lines are 95% intervals for the data-level errors.

combinations with more than 1 run, residuals, defined as the differences between the responses and the corresponding mean value in (iv). The datasets (i)–(v) are chosen as summaries of total variability, variability across GCM, scenarios, GCM–scenario combinations and runs respectively. For each simulated derived dataset we calculate eight statistics: minimum, the quartiles, maximum, interquartile range, mean, and standard deviation. We also calculate these statistics for the derived datasets based on the real data. For each statistic, we calculate the proportion of the simulated values that are greater than the corresponding statistic from the real data to give a posterior predictive p value. Formal treatment of these p values is complicated by the fact that if the model is true the p value is more likely to be near 0.5 than near 0 or 1 (Meng 1994), but values near 0 or 1 may highlight a potential discrepancy between model and data. We find (details are provided as supplementary material) that these checks indicate good agreement between model and data, lending support to the conclusions from the modeling exercise.

5. Regional analyses of temperature and precipitation

For many purposes, uncertainties in projections of global temperature change are less relevant than those in projections of regional changes; regional precipitation

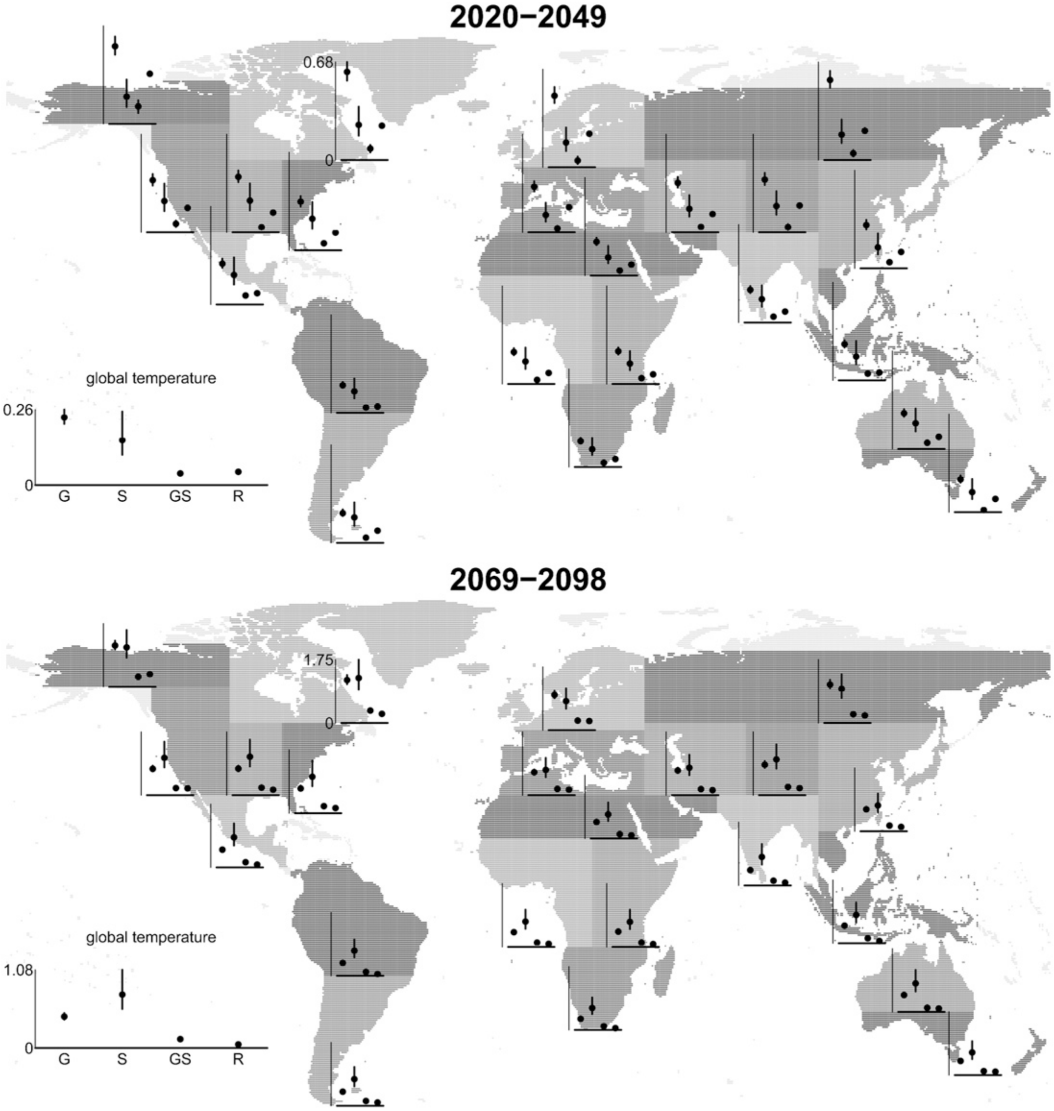

FIG. 8. Regional analyses of change in mean surface temperature from 1980–99 to (top) 2020–49 and (bottom) 2069–89. Posterior quartiles: median (dots) and central 50% credible intervals of the superpopulation standard deviations. The global analysis is summarized in the bottom left. From left to right, the ordering is GCMs, scenarios, GCM–scenario interaction, and runs. The vertical scales are different for the global and regional analyses.

changes are also likely to be critically important. In this section, therefore, we repeat the analysis of the previous section for both temperature and precipitation changes, within each of the 22 regions outlined in section 2.

a. *Regional temperature*

We repeat within each region the Bayesian analysis of section 4 using the same half-Cauchy priors: $A = 0.5$ for 2020–49 and $A = 1$ for 2069–98. Figure 8 summarizes (using the posterior median and 50% central credible intervals) the estimated posterior distributions of the superpopulation standard deviations σ_G, σ_S, σ_{GS}, and σ_R, globally and in each region, for 2020–49 and 2069–98. We use 50% intervals to prevent the large uncertainty in σ_S from dominating the plots. If we compare the posterior medians of the superpopulation SDs we find that, for 2020–49, variability over GCMs is greater than variability over scenarios and runs for each region and variability over runs is greater than variability over scenarios in some regions, predominantly in the north. For 2069–98 we find that the scenario is a greater source of variability than earlier in the century, and the scenario contributes at least as much variability as GCM in most regions and much more in many regions. The corresponding figures for the finite-population standard deviations s_G, s_S, s_{GS}, and s_R (not shown but available

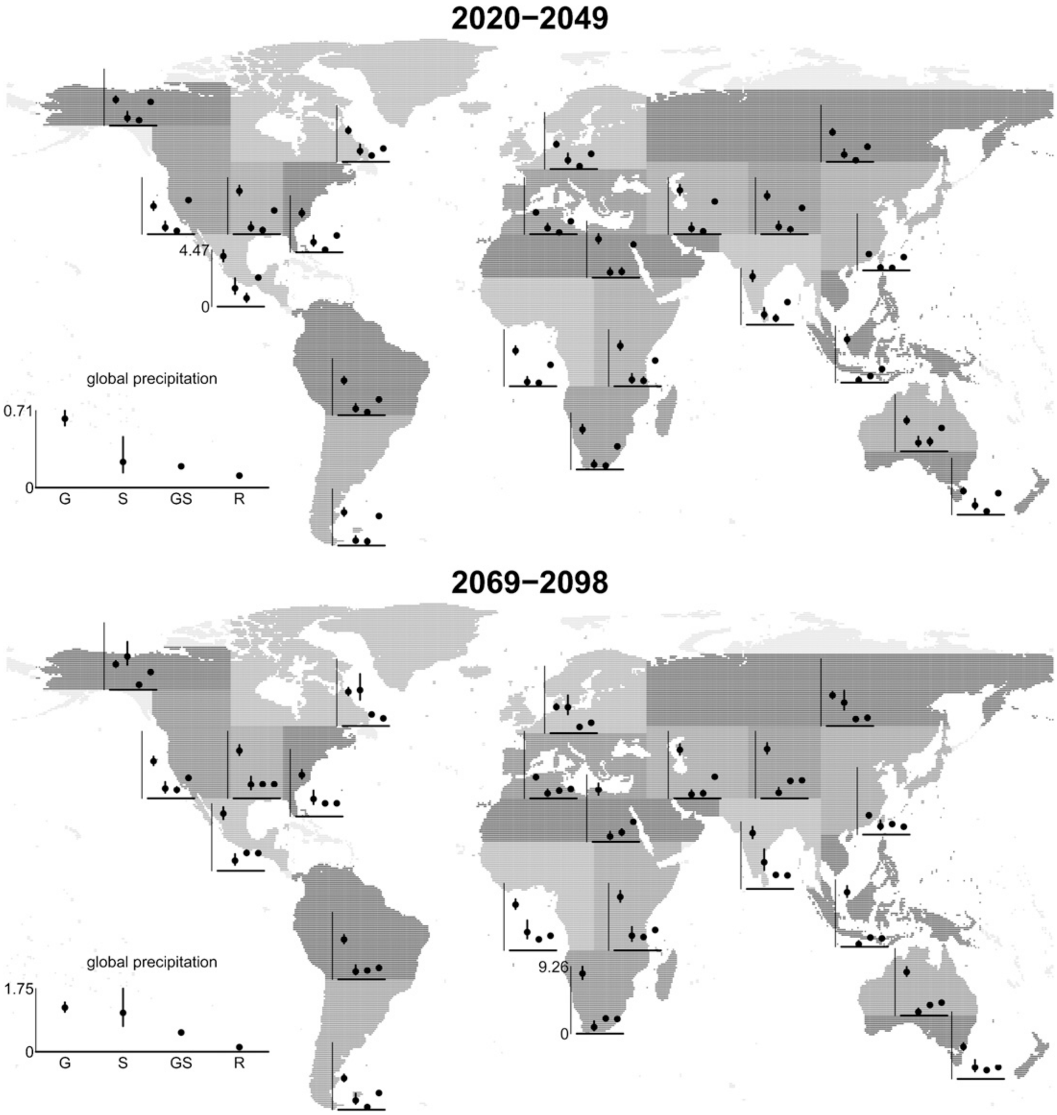

FIG. 9. Regional analyses of percentage change in mean precipitation from 1980–99 to (top) 2020–49 and (bottom) 2069–89. Posterior quartiles: median (dots) and central 50% credible intervals of the superpopulation standard deviations. The global analysis is summarized in the bottom left. From left to right, the ordering is GCMs, scenarios, GCM–scenario interaction, and runs. The vertical scales are different for the global and regional analyses.

as supplementary material) provide the same general findings.

b. *Regional precipitation*

We repeat the Bayesian analyses for the precipitation indices (percentage changes from the 1980–99 mean), using a half-Cauchy scale parameter of $A = 2.5$ for 2020–49 and $A = 5.0$ for 2069–98. These values are chosen using the same argument used for temperature in section 4: here we consider as very unlikely a range of 50% points by midcentury and 100% points by the end of the century.

There are four fewer runs for precipitation than temperature: under scenario A1B, GCM 11 (GISS-ER) has three fewer runs and GCM 12 (FGOALS-g1.0) has one fewer. Figure 9 summarizes the estimated posterior distributions of the superpopulation standard deviations σ_G, σ_S, σ_{GS}, and σ_R globally and in each region, for 2020–49 and 2069–98.

The findings are quite different to those for temperature. For 2020–49, globally variability over GCMs is greatest, but there is relatively high variability over different runs from the same GCM. Regionally, there is a similar picture, but in many areas variability over runs is similar to variability over GCMs (in western North America the former is greater than the latter) and in many regions variability over scenarios seems relatively unimportant. For 2069–98, globally the scenario is more important than in 2020–49, but in many regions the scenario still seems relatively unimportant. A reviewer has pointed out that scenario uncertainty is very low in regions, such as Southeast Asia and South Africa, where large precipitation changes are projected but there is no consensus among the models on the sign of the change. The result is a multimodel mean that is close to zero under all scenarios. In such regions the uncertainty attributed to model–scenario interaction tends to be greater than scenario uncertainty, suggesting that uncertainty due to model depends on scenario. In contrast, in areas like Alaska and Greenland, all models indicate an increase in precipitation that increases with increasing greenhouse gas emissions (analogous to the situation that applies, in all regions, for temperature), leading to large scenario uncertainty. The corresponding figures for the finite-population standard deviations s_G, s_S, s_{GS}, and s_R (not shown but available as supplementary material) provide the same general finding. These results show that relative contributions to climate uncertainty of GCM, scenario, and internal variability depend on climate variable, region, and time horizon.

6. Discussion

Running climate simulations is a time-consuming exercise so it is important to make the outputs as useful as possible. Statistical models, with parameters that relate to scientific questions of interest, can help to inform the design of future climate experiments. They can answer questions like the following: How can fixed computational resources be allocated in order to estimate parameters with greatest precision? What data would be needed to estimate the parameters with desired precision? One possible objection to the models of the type we consider is that the uncertainties due to, for example, scenario and scenario-specific GCM run are fundamentally different in their nature. However, our analysis does quantify the implications of making choices between different models, scenarios, and simulation runs.

For model (2) in section 3b, choosing a good design is difficult because optimal designs depend on the relative sizes of the superpopulation SDs, which are unknown (Khuri 2000). Thus, some prior information (perhaps based on the results in sections 4 and 5) or a design that adapts to incoming data is necessary. In the current context, for situations where internal variability is relatively unimportant, it is not worth running many simulations per GCM–scenario combination.

The results in this paper can be used to infer where the major sources of variation lie. For example, the analysis of global temperature in section 4 suggests that variability between runs, for a given GCM–scenario combination, is far smaller than between GCMs and scenarios. Therefore, it is more important to devote resources to quantifying variation over GCMs and scenarios than over such runs. For global temperature it is better to use multiple GCMs and scenarios than multiple runs at single GCM–scenario combinations. However, the analyses reported in section 4a show that for some regions variability over runs is greater than variability over scenarios, particularly for 2020–49; multiple runs for each GCM–scenario combination are therefore desirable to quantify uncertainty if interest lies in these regional quantities. In the precipitation analyses of section 5b we find that variability over runs is generally of greater importance than in the temperature analyses. In some instances it is the largest source of variability and in some regions it is a greater source of variability than scenarios even in 2069–98.

Thus, different climate variables can have competing design requirements and compromise may be necessary in designing climate experiments to meet several objectives. These results do not provide any clear guidance for something like the CMIP experiments, which have multiple potential uses. However, they do provide guidance for users who might want to select a small subset of CMIP runs to assess, for example, the potential impacts of climate change. Impacts studies often involve the selection of a relatively small number of GCM runs to drive their models: the methodology introduced here can provide guidance on how to ensure that the dominant sources of uncertainty are represented in such an exercise.

Acknowledgments. We acknowledge the modeling groups, the PCMDI and the WCRP Working Group on Coupled Modelling (WGCM) for their roles in making the WCRP CMIP3 multimodel dataset available. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. We are very grateful to three anonymous referees: their comments and suggestions have improved the original manuscript.

REFERENCES

Bates, D., M. Maechler, B. Bolker, and S. Walker, cited 2014: lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. [Available online at http://CRAN.R-project.org/package=lme4.]

Bernardo, J. M., and A. F. M. Smith, 2003: *Bayesian Theory.* 2nd ed. Wiley, 640 pp.

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.

Gebhardt, A., T. Petzoldt, and M. Maechler, cited 2013: akima: Interpolation of irregularly spaced data. [Available online at http://CRAN.R-project.org/package=akima.]

Gelman, A., 2005: Analysis of variance—Why it is more important than ever. *Ann. Stat.,* **33,** 1–53, doi:10.1214/009053604000001048.

——, 2006: Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.,* **1,** 515–533.

——, and D. B. Rubin, 1992: Inference from iterative simulation using multiple sequences. *Stat. Sci.,* **7,** 457–472, doi:10.1214/ss/1177011136.

——, and J. Hill, 2003: *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, 648 pp.

——, J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003: *Bayesian Data Analysis.* 2nd ed. Chapman and Hall, 696 pp.

——, Y.-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng, cited 2010: arm: Data analysis using regression and multilevel/hierarchical models. [Available online at http://CRAN.R-project.org/package=arm.]

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, Eds., 1996: *Markov Chain Monte Carlo in Practice*. 2nd ed. Chapman and Hall, 486 pp.

Gilmour, S. G., and P. Goos, 2009: Analysis of data from non-orthogonal multistratum designs in industrial experiments. *J. Roy. Stat. Soc.,* **58C,** 467–484, doi:10.1111/j.1467-9876.2009.00662.x.

Giorgi, F., and R. Francisco, 2000a: Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.,* **27,** 1295–1298, doi:10.1029/1999GL011016.

——, and ——, 2000b: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.,* **16,** 169–182, doi:10.1007/PL00013733.

——, and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method. *J. Climate,* **15,** 1141–1158, doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

Harville, D. A., 1977: Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc.,* **72,** 320–340, doi:10.1080/01621459.1977.10480998.

Hawkins, E., 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.

Hingray, B., A. Mezghani, and T. Buishand, 2007: Development of probability distributions for regional climate change from uncertain global mean warming and an uncertain scaling relationship. *Hydrol. Earth Syst. Sci.,* **11,** 1097–1114, doi:10.5194/hess-11-1097-2007.

Khuri, A. I., 2000: Designs for variance components estimation: Past and present. *Int. Stat. Rev.,* **68,** 311–322, doi:10.1111/j.1751-5823.2000.tb00333.x.

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter, 2000: WinBUGS—A Bayesian modelling framework: Concepts, structure and extensibility. *Stat. Comput.,* **10,** 325–337, doi:10.1023/A:1008929526011.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multi-model dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394, doi:10.1175/BAMS-88-9-1383.

Meng, X. L., 1994: Posterior predictive *p*-values. *Ann. Stat.,* **22,** 1142–1160, doi:10.1214/aos/1176325622.

Nakićenović, N., and R. Swart, Eds., 2000: *Special Report on Emissions Scenarios.* Cambridge University Press, 559 pp.

Patterson, H. D., and R. Thompson, 1971: Recovery of interblock information when block sizes are unequal. *Biometrika,* **58,** 545–554, doi:10.1093/biomet/58.3.545.

Polson, N. G., and J. G. Scott, 2012: On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.,* **7,** 887–902.

Raisanen, J., 2001: CO_2-induced climate change in CMIP2 experiments: Quantification of agreement and role of internal variability. *J. Climate,* **14,** 2088–2104, doi:10.1175/1520-0442(2001)014<2088:CICCIC>2.0.CO;2.

Searle, S. R., G. Casella, and C. E. McCulloch, 2006: *Variance Components.* 2nd ed. Wiley-Blackwell, 536 pp.

Stephenson, D., 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.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London,* **365A,** 2053–2075, doi:10.1098/rsta.2007.2076.

van Vuuren, D. P., and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change,* **109,** 5–32, doi:10.1007/s10584-011-0148-z.

Yip, S., C. A. T. Ferro, D. B. Stephenson, and E. Hawkins, 2011: A simple, coherent framework for partitioning uncertainty in climate predictions. *J. Climate,* **24,** 4634–4643, doi:10.1175/2011JCLI4085.1.