Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs*

STEFANO CASTRUCCIO+

Department of Statistics, University of Chicago, Chicago, Illinois

DAVID J. MCINERNEY#

Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois

MICHAEL L. STEIN AND FEIFEI LIU CROUCH

Department of Statistics, University of Chicago, Chicago, Illinois

ROBERT L. JACOB

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois

ELISABETH J. MOYER

Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois

(Manuscript received 10 February 2013, in final form 11 July 2013)

ABSTRACT

The authors describe a new approach for emulating the output of a fully coupled climate model under arbitrary forcing scenarios that is based on a small set of precomputed runs from the model. Temperature and precipitation are expressed as simple functions of the past trajectory of atmospheric CO_2 concentrations, and a statistical model is fit using a limited set of training runs. The approach is demonstrated to be a useful and computationally efficient alternative to pattern scaling and captures the nonlinear evolution of spatial patterns of climate anomalies inherent in transient climates. The approach does as well as pattern scaling in all circumstances and substantially better in many; it is not computationally demanding; and, once the statistical model is fit, it produces emulated climate output effectively instantaneously. It may therefore find wide application in climate impacts assessments and other policy analyses requiring rapid climate projections.

1. Introduction

The wide consensus among the scientific community that climate is changing and will almost certainly produce detrimental impacts for humanity [from Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4; Meehl et al. 2007)] means that attention is increasingly turning to evaluating the magnitude of those impacts and possible policies to reduce them. Atmosphere–ocean general circulation models (AOGCMs) are state-of-the-art tools for producing climate predictions based on our best understanding of the radiative effects of CO_2 and other anthropogenic forcing agents and the complex dynamical feedbacks of the earth's climate system. However, the computational demands of AOGCMs preclude or limit

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-13-00099.s1.

+ Current affiliation: CEMSE division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Current affiliation: Department of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia.

Corresponding author address: Elisabeth Moyer, Department of the Geophysical Sciences, University of Chicago, 5734 S. Ellis Ave., Chicago, IL 60637.
E-mail: moyer@uchicago.edu

DOI: 10.1175/JCLI-D-13-00099.1

their use in the context of integrated assessment models (IAMs) used to estimate climate damages and the cost–benefit trade-offs of potential mitigation actions. Analyses that involve optimal policy determination or uncertainty quantification require repeated iterations of climate projections in response to forcing trajectories over the decadal or centennial time scale, which is computationally prohibitive with AOGCMs. For IAMs whose only climate input is global mean temperature (GMT), climate projections can be provided instead by simple energy-balance models tuned to the climate sensitivity of AOGCMs. Climate changes and impacts will not be uniform across the earth, however, and more advanced IAMs may require regional climate predictions. There is increasing need for techniques that can capture the regional information provided by AOGCMs and produce tools useful for the impacts assessment community.

The most common approach for producing such regional projections has been to use ''pattern scaling'' to downscale the projections of simple global energy-balance models. Pattern scaling relies on the assumption that regional climate responses are a linear function of global climate response, so that regional climate evolution can be captured by scaling a single pattern to the global mean temperature. The technique was introduced by Santer et al. (1990) as a means of comparing spatial patterns of climate response from different GCMs and has been widely used in subsequent years (e.g., Hulme and Raper 1995; Hulme and Brown 1998; Cabre et al. 2010; Dessai et al. 2005; Fowler et al. 2007; Harris et al. 2006; Murphy et al. 2007). Different possible techniques for obtaining patterns are reviewed in Mitchell (2003).

The linearity assumption has been shown to be reasonable for centennial-scale projections (e.g., Mitchell et al. 1999; Giorgi 2008), but on some time scales the technique will be inappropriate, since different parts of the earth warm at different rates. Furthermore, if the regional pattern of climate response were a function of the magnitude of warming, a single pattern would also not accurately capture the climate response to arbitrary CO_2 scenarios even in equilibrium cases. Using the Hadley Centre Coupled Model, version 2 (HadCM2), Mitchell (2003) showed that both the rate and the magnitude of forcing changes influence patterns of regional climate and suggested approaches to pattern construction to minimize errors.

We propose to overcome some of the limitations of pattern scaling through an alternative emulation approach based on a collection of precomputed climate model runs that allows us to capture rate dependencies in regional climate evolution. This collection of runs, or training set, is used to obtain estimates of the parameters in simple statistical models that describe temperature and precipitation as a function of past trajectories of radiative forcing due to CO_2. The resulting tool allows us to reproduce (emulate) the output of an AOGCM under a large range of forcing scenarios. Once the emulator is constructed, emulation of a climate scenario is effectively instantaneous, as it would be under pattern scaling. In contrast, climate projection from a state-of-the-art model can still take days to weeks even on the most powerful platforms. Since our training set is used only to estimate statistical parameters, the emulator is determined by a set of regional parameter values and requires negligible data storage. The simplicity and robustness of statistical emulation based on a modest training set makes it a promising tool for impacts assessment. Similar ideas have been previously proposed by Mitchell (2003), though execution was precluded because of lack of suitable collection of model runs, and recently explored by Holden and Edwards (2010) (see section 5 for comparison of approaches).

In the remainder of this paper, section 2 describes the collection of climate runs on which our emulator is based; section 3 introduces the statistical models for annual temperature and precipitation at a regional level and shows an example of emulation; and section 4 develops emulation diagnostics and uses them both to assess the influence of training set size on emulation quality and to compare our emulation to pattern scaling. Finally, section 5 discusses our approach in comparison to other techniques for computer model emulation. We describe the particular requirements and characteristics of climate emulation over forcing scenarios, for which both the inputs and outputs are time series, and provide suggestions to guide future emulation approaches.

2. Precomputed climate runs

To explore the problem of emulating climate under arbitrary forcing scenarios, we built a collection of climate model runs to be used for training and prediction. These runs are driven by different trajectories of future CO_2 concentration and have different initial conditions but all are performed with the same model and same representation of model physics. Simulations were performed with the Community Climate System Model, version 3 (CCSM3; Yeager et al. 2006; Collins et al. 2006), at a relatively modest T31 atmospheric resolution ($\approx 3.75° \times 3.75°$) and nominally 3° ocean resolution, a configuration that allows us to run multiple realizations of a wide range of multicentury scenarios. Since we are interested in capturing the effects of changing CO_2 on climate, in all runs all other greenhouse gases and aerosols are held fixed at their preindustrial values.

The AOGCM runs used in the work described here consist of five scenarios: three with gradual rise and

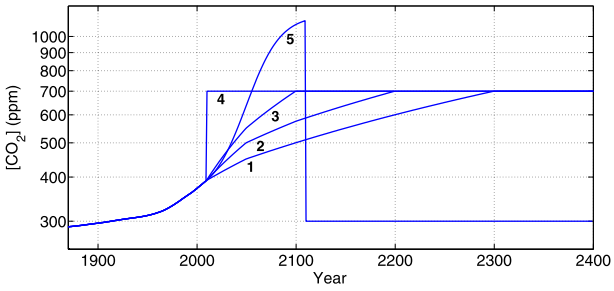

FIG. 1. The CO_2 scenarios used for building the collection of runs. We refer to these throughout the paper as the 1) slow, 2) moderate, 3) fast, 4) jump, and 5) drop scenarios. All scenarios start at year 1870. Some scenarios extend beyond the range shown here: slow, moderate, and fast end at year 2449, whereas jump ends at 2199 and drop ends at 2399.

then stabilization of CO_2 and two with abrupt changes (Fig. 1). All scenarios follow estimated historical CO_2 concentrations from 1870 to 2010 and then branch off into different future trajectories of evolving CO_2 over the subsequent 189–439 years (end years range from 2199 to 2449). We denote the five scenarios as "fast," "moderate," "slow," "jump," and "drop." To enhance our ability to distinguish changes in mean climate from internal variability, we simulated five realizations of each scenario with different initial conditions: specifically, we used restart files from years 410, 420, 430, 440, and 450 of the National Center for Atmospheric Research (NCAR) preindustrial control run b30.048 (Collins et al. 2006). In total, our collection of runs consists of more than 10 000 model years, though individual emulators used in this paper are trained using subsets of the runs.

Multiple realizations of each scenario are useful both in producing emulators and in evaluating emulator performance. We treat the five realizations of each scenario as statistically independent because they were generated with decadally spaced restart files. The chaotic nature of the climate system means that changes in any initial conditions other than those of the deepest ocean are expected to produce essentially independent results after approximately a decade (e.g., Branstator and Teng 2010; Collins 2002; Collins and Allen 2002), so we believe that this assumption of independence is reasonable. For similar reasons, runs under different scenarios but the same restart year should be very nearly independent within a few years after the scenarios diverge but, since all scenarios are identical before 2010, the results for runs with the same restart year are also identical until 2010. We avoid this problem by using runs with different restart years in our training sets.

The choice of scenarios for the precomputed runs was not based on any formal design criteria and is not meant to be optimal in any sense. We deliberately chose some scenarios that were somewhat realistic and others with large changes in CO_2 in order to be able to distinguish short- and long-term effects, but in general we sought simply to reproduce the kind of runs that would typically be available in preexisting archives of climate model output. Impacts assessments often require emulation of multiple AOGCMs, but it would be prohibitively difficult for an individual research group to run multiple climate models to generate optimal libraries for emulation. It is therefore useful to develop emulation techniques that are not critically sensitive to the characteristics of their training sets and that can make use of existing community multimodel resources such as the archive from phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012).

3. Statistical models for temperature and precipitation

In this work, we emulate annual mean temperature and precipitation in climate projections with simple statistical models that involve a mean function that varies in time plus a stochastic term. For the mean function, we chose simple functional forms relating temperature T and precipitation P to past trajectories of CO_2 that capture physically justified relationships. We train emulators based on various subsets of our precomputed climate model runs, fitting the parameters of the statistical models using standard statistical methods (see supplementary material for more details). The resulting emulators can then predict annual temperature and precipitation for arbitrary climate forcing scenarios. In the emulations shown here, we fit the statistical models not at native climate model spatial resolution (48×96 grid points for T31 resolution) but aggregated at subcontinental scale in 47 regions. The regions are modifications of those defined by Ruosteenoja et al. (2003), subdivided over the oceans to ensure that we separately emulate regions of qualitatively different precipitation response (e.g., see Fig. 4 or Fig. S1 in the supplemental material for regional codes). Without regional aggregation, obtaining a stable fit of the statistical models parameters for T and P would require a significantly larger training set. Emulation can be extended to the grid scale through regional pattern scaling (see section 4).

a. Temperature

A long body of research suggests that within the range of CO_2 concentrations likely to be produced by anthropogenic activity, equilibrium global mean temperature change is proportional to $\log[CO_{2r}]$, where $[CO_{2r}]$ is the ratio between current and preindustrial CO_2 concentrations (Manabe and Wetherald 1967; Forster et al. 2007). For policy analysis purposes, however, emulating

equilibrium climate is less relevant than understanding the spatiotemporal climate changes that populations will face over the next century. We seek here to emulate the transient climate response when climate is a function not only of the present value of $[CO_{2r}]$ but also of its past history. As mentioned before, even if pattern scaling were sufficient to reproduce equilibrium climate (i.e., if the equilibrium spatial distribution of temperature were linear with $\log[CO_{2r}]$), it would not be sufficient in transient climates. Because different regions of the earth warm at different rates, the spatial distribution of temperature anomalies in a given year during warming will not be a multiple of the equilibrium pattern.

For emulation of temperature, we propose a representation that captures this dependence on past trajectories of CO_2 via an infinite distributed lag model (Judge 1980, chapter 10) in which current temperature is dependent on a weighted sum of past $\log[CO_{2r}](t)$,

$$T(t) = \beta_0 + \beta_1 \frac{1}{2}\{\log[CO_{2r}](t) + \log[CO_{2r}](t-1)\} + \beta_2 \sum_{i=2}^{+\infty} w_i \log[CO_{2r}](t-i) + \varepsilon(t), \quad (1)$$

where $T(t)$ is the temperature at year t. Because temperature may show some autocorrelation, we assume the stochastic term $\varepsilon(t)$ is an autoregressive model of order 1: $\varepsilon(t) = \phi\varepsilon(t-1) + \nu(t)$, where ν is a Gaussian white noise with unknown variance σ^2. This model is able to capture the modest dependence in temperatures across years.

The β coefficients in Eq. (1) are physically interpretable: β_0 is preindustrial temperature, β_1 is the near-term response to changes in CO_2, and β_2 is the slower response dependent on CO_2 levels in prior years. This form gives us the flexibility to represent a temperature response characterized by multiple adjustment time scales and is especially important when emulating scenarios with abrupt CO_2 changes. Using the average $\log[CO_{2r}]$ over years t and $t-1$ for the short-term effect is somewhat arbitrary, but we have experimented with other forms for this term and not found anything clearly superior. Because we expect the influence of past radiative forcing to decrease as we go back in time, the weights w_i in the long-term component should be chosen to decrease with the trajectory year i. We choose here a simple exponential decay of the weighting of past years: $w_i = \rho^{-2}(1-\rho)\rho^i$ with $0 < \rho < 1$ so $\sum_{i=2}^{\infty} w_i = 1$. [Note that we could also have taken the infinite sum in Eq. (1) to start at 0 rather than 2. The resulting fitted models would be negligibly different.] The model parameters are then the three β_j's, ρ, ϕ, and σ^2. The first four parameters capture the mean evolution of the climate system averaged over initial conditions, a deterministic function of CO_2 trajectory, and the final two parameters describe the stochastic variability in the climate state about this mean, which differs between realizations (initial conditions). We discuss emulation of the stochastic behavior of both temperature and precipitation in section 3c.

It is important to point out several assumptions implicit in the choice of our functional form for temperature. First, the model assumes that, on average, equilibrium spatial temperature patterns are linear with $\log[CO_{2r}]$ since, when sufficient time has passed after stabilization of CO_2 concentration, emulated mean temperature approaches

$$\beta_0 + (\beta_1 + \beta_2)\log[CO_{2r}]_{\text{stab}},$$

where the subscript "stab" indicates the CO_{2r} level after stabilization. This assumption would likely break down in cases of extreme CO_2 changes. Second, our functional form is appropriate only for centennial-scale or shorter emulation scenarios. Although in principle our approach allows us to emulate climate in any year for arbitrary CO_2 scenarios, Eq. (1) should not be used for emulating considerably beyond the several-century time span of the training runs. This constraint arises not only because statistical models cannot be expected to capture processes not represented in the training set but also because the simple exponential weights used here do not capture well the combined behavior of the decadal/centennial-scale warming of the upper ocean and the long-tail warming of the deep ocean over thousands of years (see Fig. S4 in the supplemental information).

To construct an emulator, we derive parameter estimates from one or more training runs. (By "run" we mean a climate projection driven by a given scenario and begun from given initial conditions.) Throughout this manuscript, we focus on an emulator generated with a training set consisting of two runs: one realization each of the fast and jump scenarios with different restart years. The resulting emulator appears to track accurately the overall trend of out-of-training set climate scenarios. Figures 2a,b show emulations of the mean temperature trajectory for the slow and drop scenarios, superimposed with all five realizations of actual CCSM3 output for these scenarios. Emulation of the drop scenario does show slight misfit immediately following the sudden drop in CO_2. This misfit can be reduced by using a more complex functional form, but introducing additional terms can lead to instability of the fit and we consider the emulation of this physically extreme scenario to be reasonably good under the circumstances. (See section 4 for a more extensive evaluation of emulation fidelity, and see Table S1 in the supplemental material for parameter estimates and their standard errors for all regions.)

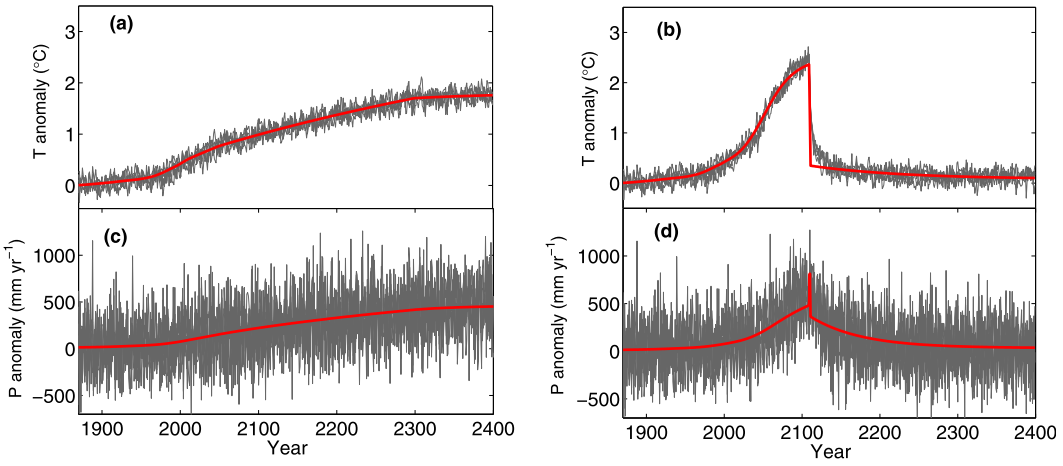

FIG. 2. Examples of (a),(b) temperature emulation for the North Pacific west (NPW) region, chosen as representative of a region with significant change, and (c),(d) precipitation emulation for the equatorial Pacific west (EPW) region, chosen to highlight interesting transient precipitation behavior. Panels (a) and (c) show the emulated slow scenario, and (b) and (d) show the drop scenario. The emulator was trained by one realization each of the fast and jump scenarios. The solid red line represents the emulated mean function and the gray lines show the five CCSM3 realizations for the scenarios. Emulation captures expected transient precipitation behavior in which precipitation anomaly is a function of the rate of change in radiative forcing. Note that the trend in temperature is larger relative to stochastic variability than it is for precipitation. We define diagnostics of emulation goodness-of-fit I_1 and trend-vs-variability I_2 in section 4a. Values of (I_1, I_2) for the emulations shown here in (a)–(d) are (1.01, 11.23), (1.94, 35.82), (1.02, 1.18), and (1.09, 1.41), respectively; I_2 is much larger for temperature, as expected.

b. *Precipitation*

Precipitation in transient climates has been frequently described as a combination of a fast response that is a function of the changed forcing agent and a slow linear response to evolving temperature. The fast response is negative in the case of CO_2, so that, in scenarios with rising CO_2, precipitation at a given temperature is lower than it would be at equilibrium for that temperature (Andrews and Forster 2010). The transient precipitation response was first discussed in detail by Allen and Ingram (2002), and the fast–slow framework became commonly accepted in later works (e.g., Bala et al. 2010; Cao et al. 2011). These findings motivate the following regression model for precipitation [though see McInerney and Moyer (2012) for further discussion of underlying physics],

$$P(t) = \gamma_0 + \gamma_1 \hat{T}(t) + \gamma_2 \log[\mathrm{CO}_{2r}](t) + \eta(t), \qquad (2)$$

where $\gamma_1 \hat{T}(t)$ and $\gamma_2 \log[\mathrm{CO}_{2r}](t)$ are the slow and fast terms, respectively, and $\hat{T}(t)$ is the mean emulated temperature from Eq. (1). We use $\hat{T}(t)$ rather than $T(t)$, the actual temperature in year t, because the physical processes underlying the model are likely distinct from those driving stochastic interannual variability. Since we found no clear evidence for dependence in the stochastic terms for precipitation in this model, the stochastic term $\eta(t)$ is simply assumed to be Gaussian white noise with unknown variance τ^2. Once $\hat{T}(t)$ is obtained from fitting Eq. (1), the parameters in Eq. (2) are estimated using linear regression. Joint emulation of temperature and precipitation including their stochastic components would require modeling the corresponding stochastic terms $\varepsilon(t)$ for temperature and $\eta(t)$ for precipitation jointly, which we do not attempt here.

The resulting emulated mean precipitation again matches well the overall trend in the CCSM3 output, although variability in precipitation is much larger than in temperature and trend prediction is therefore less informative (Figs. 2c,d). We chose to show the equatorial west Pacific in Fig. 2 because this region demonstrates one feature of our emulation that stands out in scenarios of abrupt CO_2 change: a sharp spike in precipitation coincident with a drop in CO_2 (Fig. 2d), such that precipitation momentarily increases even while temperature is decreasing. This effect has a well-founded physical interpretation and has been shown clearly above variability in AOGCM output in more extreme scenarios in several recent works (e.g., Wu et al. 2010; McInerney and Moyer 2012). Linear pattern scaling with global mean temperature change cannot capture this effect.

c. *Stochastic temperature and precipitation components*

While the mean emulations shown in Fig. 2 capture the dependence of temperature and precipitation on

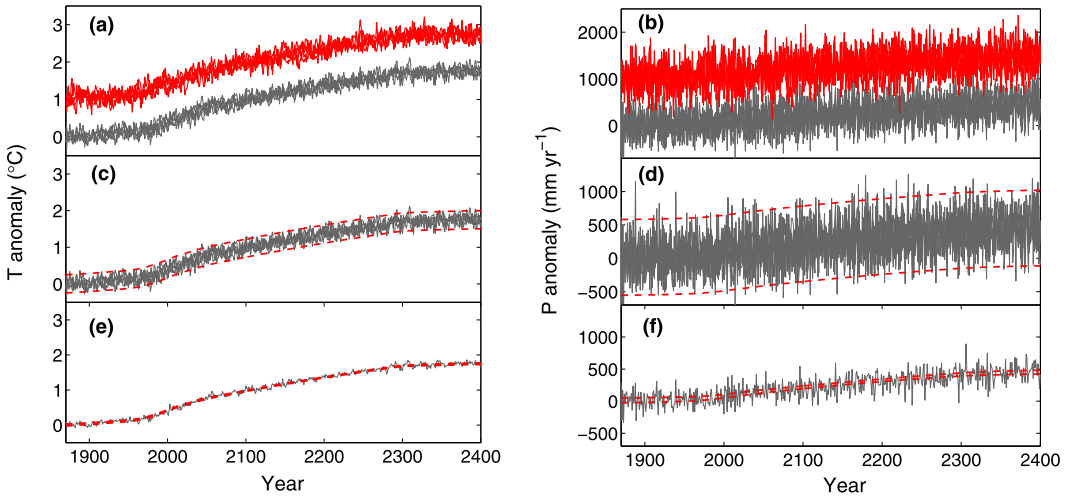

FIG. 3. Examples of uncertainty quantification (a),(c),(e) for temperature in the North Pacific west (NPW) region and (b),(d),(f) for precipitation emulation for the equatorial Pacific west (EPW) region. All panels show the emulated slow scenario. The emulator was trained by one realization each of the fast and jump scenarios. In (a),(b), an example of emulated realizations is shown. The gray lines represent the five CCSM3 realizations and the red lines represent the five emulated realizations (with an offset of 1°C for temperature and 1000 mm yr^{-1} for precipitation). The actual runs and those simulated via the emulator appear to be qualitatively similar. In (c),(d), the five superimposed CCSM3 realizations are shown in gray, and the dashed red lines denote the 95% prediction bands from the emulator. Empirical coverage is 0.9531 for (c) and 0.9545 for (d), very close to the nominal 95% level. In (e),(f), the mean across the five CCSM3 realizations of the slow scenario is shown in gray, and the dashed red lines represent the pointwise 95% confidence bands based on the emulator. The bands are very narrow, especially for temperature, highlighting the ability of the emulator to capture the mean trend with very high precision.

CO_2 trajectories, impacts assessments may require emulation that fully reproduces an actual climate simulation, including short-term variability. Many applications would therefore require addition of stochastic components to the mean emulator. A simple initial approach is to simulate this variability from our stochastic models and estimated parameters. This method implicitly assumes that the statistical characteristics of the error terms are invariant over time for any scenario and are the same for all scenarios. That assumption is unlikely to be exactly true but appears to provide a satisfactory approximation for most regions in the scenarios tested here. That is, the simple stochastic model appears to capture the variability in the actual realizations of the CCSM3 temperature and precipitation (Figs. 3a,b, which show emulated full simulations including stochastic components for the cases of Figs. 2a,c, along with corresponding actual CCSM3 realizations). More quantitatively, CCSM3 output can be compared with the 95% prediction bands based on the emulators (Figs. 3c,d). For the cases shown, the empirical coverage of the prediction intervals are 0.9531 and 0.9545 for temperature and precipitation, respectively, very close to the nominal coverage of 0.95. Figure S3 in the supplemental material shows empirical coverage for temperature for all regions and both the slow and drop scenarios; the results are close to 95% in all regions other than the Southern Ocean. The fact that our model does not provide an accurate substitute for CCSM3 output in the Southern Ocean is not unexpected because upwelling from the deep ocean complicates temperature evolution there. Misfit for the Southern Ocean is evident in multiple diagnostics of emulation performance (see section 4).

4. Diagnostics, training set size, and comparison with pattern scaling

a. *Evaluating the fit*

The appropriate evaluation of emulator performance depends on the purpose for which the emulator is used. For impacts assessments that have previously relied on global pattern scaling, one possible performance criterion is exceeding the emulation fidelity provided by pattern scaling. Other criteria could be that emulation error is small relative to differences in climate projections between AOGCMs or small relative to initial conditions uncertainty in the emulated AOGCM. We discuss here various approaches to evaluating emulator performance. Evaluations are aided by having multiple realizations for each prediction scenario, allowing us to distinguish the mean climate trajectories from the stochastic component without assuming our mean model is

correct. The test of empirical coverage of 95% prediction intervals discussed in section 3 is one type of emulator evaluation, but not the most relevant for the main focus of this work, emulation of change in mean climate. We therefore seek additional diagnostics.

Even if our emulation model Eq. (1) were strictly correct for all scenarios, the mean emulator generated from it would retain some uncertainty because of the limited size of the training set used to estimate the model parameters. Confidence bands for the estimated regression function provide a natural way to quantify this uncertainty. Figures 3e,f shows the pointwise 95% confidence bands along with the average of the five available CCSM3 realizations. (See supplementary materials for details.) The widths of these bands are small relative to internal variability and agree well by eye with the average of the five CCSM3 realizations.

These confidence bands assume that the underlying statistical model is correct. We consider two additional indices whose validity does not depend on knowing the form of the mean function. The index I_1 measures emulation performance relative to the optimal emulation possible given initial condition uncertainty and I_2 measures the trend in the data relative to initial condition uncertainty (i.e., how much of the variation in a climate time series could be explained by an emulator).

The first index is related to what statisticians call the lack-of-fit statistic (e.g., see Montgomery 2012). Let $T_r(t)$ denote temperature for year $t = 1, \ldots, n$ (here $t = 1$ corresponds to the year 2010, the year the scenarios diverge) and realization $r = 1, \ldots, R$ (here $R = 5$). We compare the sum of squared deviations of the actual realizations from the emulated mean temperatures $\hat{T}(t)$ to the sum of squared deviations of realizations from the average across realizations $\overline{T}(t) = 1/R\sum_{r=1}^{R} T_r(t)$,

$$I_1 = \frac{\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \hat{T}(t)]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \overline{T}(t)]^2} = \frac{N_1}{O_1}. \quad (3)$$

The numerator N_1 measures the actual performance of the emulator. The denominator O_1 makes use of the multiple realizations we have under each scenario to give an unbiased estimate of the sum of squared errors for a hypothetical ''perfect'' emulator that, for each year t, reproduces the average temperature over an infinite number of realizations. The factor of $R/(R-1)$ in O_1 takes account of the fact that we do not know this perfect emulator but use $\overline{T}(t)$ as an estimate of it. A value of 1 for I_1 is therefore the best possible performance from an emulator. (Occasional values less than 1 may however arise because of random variation in N_1 and O_1.)

A value for I_1 close to 1 has different implications depending on the noise in the model output being emulated. In particular, if the noise is large compared to the trend in the data, then I_1 will likely be close to 1 even if the emulation poorly captures the small underlying trend. To quantify the degree of variation in the data attributable to the trend, we construct an index whose denominator is that of I_1 but whose numerator now describes the trend itself,

$$I_2 = \frac{\frac{n}{n-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \overline{T}_r]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \overline{T}(t)]^2}, \quad (4)$$

where $\overline{T}_r$ is the mean across time of each realization, $\overline{T}_r = 1/n\sum_{t=1}^{n} T_r(t)$. Note that this index depends only on the AOGCM data and is completely independent of the emulation. If the mean AOGCM data show no trend, then the numerator and the denominator are unbiased estimates of the same quantity and I_2 should be close to 1. The conditions $I_2 \gg 1$ and $I_1 \approx 1$ would mean that there is a trend to emulate and that the emulator captures it well. If I_1 is comparable to I_2, then the emulator would not be useful for tracking the evolution of the mean. As interannual variability in precipitation is larger relative to trend than it is in temperature (e.g., Fig. 2; see also Deser et al. 2012), I_2 values tend to be much smaller for precipitation than for temperature (cf. Figs. 4 and 9).

These indices suggest that the temperature emulator described previously in section 3 (trained by one realization each of the fast and jump scenarios) produces near-optimal mean emulation of nearly all regions in the physically reasonable slow stabilization scenario and only modestly degraded quality in the extreme drop scenario (Fig. 4 shows I_1 and I_2 values for all regions). For the slow scenario, the emulated mean functions are essentially optimal (I_1 very nearly 1) throughout the Northern Hemisphere and equatorial region and close to optimal ($I_1 \leq 1.13$) everywhere except in part of the Southern Ocean. For the drop scenario, unsurprisingly, the emulator predictions perform substantially worse in all regions, but even here we believe this lack of fit may be small compared to other possible sources of error in forecasting climate, such as differences between AOGCMs or differences between AOGCMs and reality and so would still serve as a useful emulator. The largest discrepancies arise for both scenarios in a single portion of the Southern Ocean. Values of I_1 substantially larger than 1 are not necessarily associated with a poor skill of the emulator relative to other techniques but do indicate

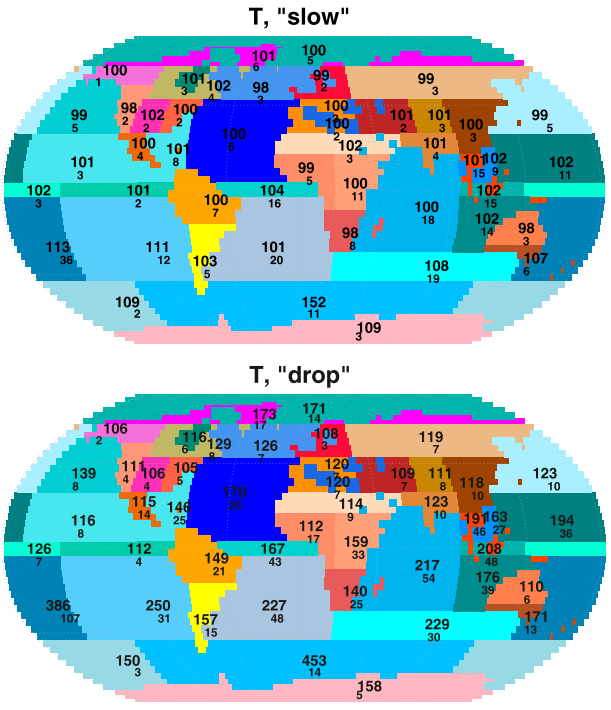

FIG. 4. Emulation indices for all regions for the regional temperature emulation described in the text and shown in Fig. 2. The value in large font is the "emulation optimality" index I_1 (×100) and in small font below is the trend index I_2. Low I_2 means there is little trend relative to noise and the I_1 index is not informative, even if close to 100 (optimal emulation). Shown are the (top) slow and (bottom) drop scenarios. Emulation is worse for the physically extreme drop scenario, as expected, but is generally close to the optimal value of 1 in most inhabited regions. All indices have been computed between the year 2010 and the farthest time point (2449 for slow and 2399 for drop).

that the statistical model for the region could be improved.

In the end, whether an emulator of an AOGCM is adequate will depend on the specific application. Because we make no effort to capture spatial dependence in the stochastic terms between regions, the emulator would be less appropriate for studies that involve large-scale spatial correlations in weather; for example, global droughts or jet stream shifts. [See Castruccio and Stein (2013) for one approach to emulating the stochastic component of annual temperatures in climate model output that captures both spatial and temporal dependence.] We also do not capture any dependence between the stochastic components of temperature and precipitation within a region. However, for an impacts assessment requiring annual temperatures in a given region, any differences between the emulated temperature and the AOGCM temperature showed in, for example, Fig. 3a would most likely be inconsequential.

b. *Training set size: How many scenarios/realizations?*

One of the advantages of our approach is that it permits emulation with a relatively small training set of precomputed runs. To determine the trade-off between size of the training set and goodness of fit, we examined the performance of the emulator with a varying number of scenarios and realizations. Investigating the impact of the number of realizations on emulation quality is the more straightforward test, involving computing I_1 for temperature emulation over a range of number of realizations used. Figure 5b shows results from an experiment in which the moderate scenario was emulated with from 1 to 5 realizations of the fast scenario as the training set. Increasing the number of realizations of each training scenario produces more accurate emulations, but the difference between the use of even 1 and 2 realizations is small and there is diminishing return gained from further increasing the number of realizations in the training set. Increasing the number of realizations further also does not reduce the misfit of the outlier regions with highest I_1 values, which all lie in the Southern Ocean.

Testing the value added by additional scenarios is a less well-defined problem, since different choices of scenarios will affect the emulation differently. Nevertheless, we attempt a test by conducting emulations with increasing numbers of scenarios. Again we emulate temperature in the moderate scenario beginning with a training set consisting of a single realization of slow and successively adding to the training set fast, jump, and drop (Fig. 5a), which is a rough attempt to order the training scenarios from most to least similar to the prediction scenario. The results show that the addition of scenarios first improves and then degrades the emulation. We interpret this result as implying that our simple statistical model cannot perfectly represent all scenarios; that is, the best values of β_0, β_1, β_2, and ρ in Eq. (1) vary somewhat with the scenario. Including scenarios in the training set very different from the one emulated can then result in worse performance. Figure 5 shows that even a single slow or a single fast realization yields a fairly good emulator of the moderate scenario. However, we would be cautious about building emulators when AOGCM output is available for only one scenario since that would leave no opportunity to check for stability of the regression parameters across scenarios.

Our tests suggest that the choice of training set is not especially crucial if prediction and training scenarios are similar, but more care would be needed for emulating extreme scenarios. One approach might be to choose different training sets according to the prediction scenario.

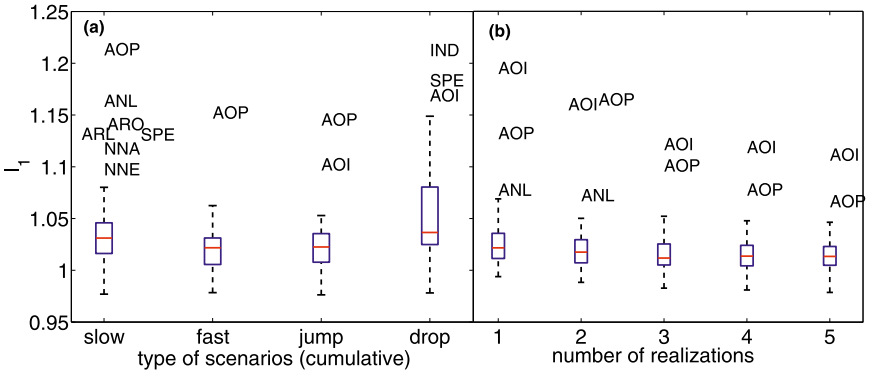

FIG. 5. Boxplots of the fit index I_1 for various (b) numbers of realizations and (a) scenarios in the training set for mean emulation of the moderate stabilization scenario. The training sets for the realization test (b) are made up of 1–5 realizations of the fast scenario; those for the scenario test (a) are made up of a single realization of slow and then adding, successively, one realization of fast, jump, and drop. Adding realizations of a single scenario offers a modest benefit as shown in (b), and adding scenarios too dissimilar from the test case can actually degrade emulator performance as shown in (a). Box-and-whisker plots exclude severe outliers, which are shown with their regional codes. Four of the five outliers lie in the polar regions (see Fig. S1 in the supplemental material for locations).

In this case, one algorithm might be to 1) order the available forcing scenarios in the training set by their similarity to the prediction scenario; 2) fit the emulator using first only the nearest training scenario, then the two nearest, and so on; and 3) choose the emulator with the smallest training set that offers stable parameter estimations as measured by the width of the 95% confidence bands for the mean emulator (e.g., Figs. 3e,f). Further research would be needed to actually apply this approach in the context of integrated assessments over many possible scenarios, both to define the notion of similarity and to automate implementation. In this work we have focused simply on demonstrating that, in some circumstances, emulation requires only a limited training set of a few scenarios and realizations. This finding supports the utility of statistical emulation based on modest training sets for uses such as policy analysis or model intercomparison.

c. *Comparison with pattern scaling*

One of the motivations for our approach to statistical emulation is to offer an improvement on pattern scaling by capturing the dependencies on rate of forcing change that make transient climates different from equilibrium ones. We therefore test the fidelity of our mean emulation against pattern scaling to global mean temperature. To provide a direct comparison, we first evaluate performance of the regional climate projections generated by our statistical mean emulator to regional projections generated by pattern scaling to GMT. Second, we evaluate an extension of our approach that allows us to emulate climate at native model spatial resolution, again comparing to GMT pattern scaling. The latter test may be more relevant for policy analysis purposes, since impacts assessments often require finescale climate projections. We perform grid-scale emulation by a hybrid approach, by first statistically emulating regional temperature and precipitation and then downscaling by pattern scaling to the regional mean temperatures.

For the comparison of regional emulation, we construct patterns of temperature and precipitation for our 47 regions from all realizations in our training set (fast and jump). Pattern scaling assumes that all regional temperature anomalies $T_i(t) - T_{i,\mathrm{PI}}$ are linear with global mean temperature anomaly $T_{\mathrm{GM}}(t) - T_{\mathrm{GM,PI}}$ (subscripts PI and GM denote preindustrial values and global mean, respectively). We derive the pattern by linear regression on all data in the training set assuming

$$T_i(t) - T_{i,\mathrm{PI}} = \alpha_i[T_{\mathrm{GM}}(t) - T_{\mathrm{GM,PI}}] + \varepsilon_i(t) \qquad (5)$$

and estimating α_i by least squares. Patterns for temperature and precipitation are shown in Figs. 6 and 7, with the fitted relationship between the regional climate variable and GMT shown in red. These figures provide a visual check on the linearity assumption behind pattern scaling and on the variability in regional temperature and precipitation.

GMT in a typical pattern-scaling emulation would usually be obtained by running an energy-balance model tuned to match the climate sensitivity of the AOGCM to be emulated. Here we forgo the use of an additional external model and instead simply use the GMT from our statistical emulator. This simplification gives pattern

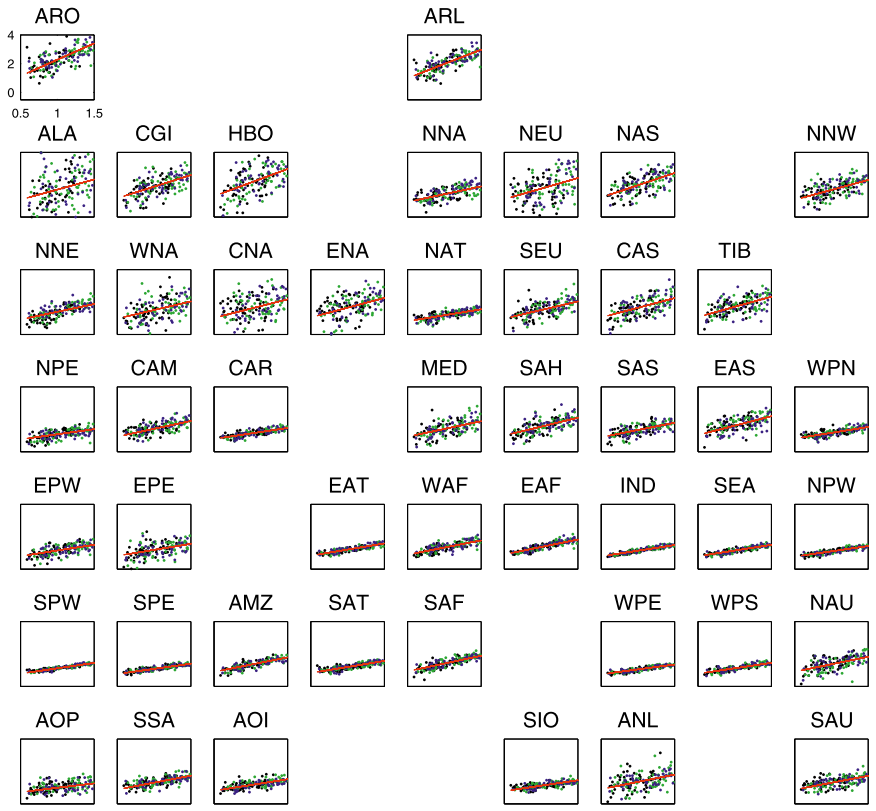

FIG. 6. Construction of regional pattern scaling for temperature: linear regressions of regional temperature anomalies on GMT. Data used are 60 yr from 2010 to 2070 (we picture a subset of the data in this figure for visualization purposes) for all 47 regions in the standard training set consisting of the fast and jump scenarios. The two scenarios are shown in different colors. Regions are arranged to approximate their geographic distribution (north at top) to give an idea of spatial patterns. Panels share a consistent y-axis scale, so that differences in warming rate and variability may be seen by eye.

scaling a slight artificial advantage over a more realistic comparison. Nevertheless, when comparing to emulation of temperature in the same scenarios shown previously (slow and drop), statistical emulation matches or outperforms pattern scaling in most regions (Fig. 8). Comparing Figs. 4 and 8, we see for the slow scenario, which has the smallest transient response and emulation is easiest, the regional differences in performance for our emulator and pattern scaling as measured by I_1 are small; these differences are much larger for the more challenging drop scenario. For precipitation, I_2 values are much smaller than for temperature (see Fig. 9), so the differences in I_1 values for the two emulators are unsurprisingly smaller. Nevertheless, in both prediction scenarios used here, statistical emulation conveys an advantage in most regions outside of the Southern Ocean (which is problematic for both methods).

For a grid-scale comparison, we use a hybrid approach, emulating regional temperature and precipitation and then downscaling by applying pattern scaling at the regional level. This approach consists of four steps:

1) For each region i, use the training set to fit parameters for regional T_i and P_i.
2) With those parameters, statistically emulate regional T_i and P_i for the prediction scenario.
3) For each region i, use the training set to obtain regional patterns of grid-scale T and P.
4) Predict grid-scale T and P by multiplying the regional patterns by emulated regional T_i.

This approach retains the benefits of statistical emulation in capturing nonlinearities in regional climate evolution but allows projections at small spatial scale.

Step 3, estimating for each region i a grid-resolution pattern that scales with respect to regional temperature, is mathematically similar to the global pattern scaler described previously, where we obtained a regional-resolution pattern that scales with respect to global

FIG. 7. As in Fig. 6, but for precipitation. Because precipitation anomalies differ widely between regions, *y*-axis scales are shown in percent separately for each panel.

mean temperature. For T emulation, we use all data in the training set to fit the parameters in

$$T[(L,\ell),t] - T_{\mathrm{PI}}(L,\ell) = \alpha_{(L,\ell)}[T_i(t) - T_{i,\mathrm{PI}}] + \varepsilon_{(L,\ell)}(t), \tag{6}$$

where $T[(L,\ell),t]$ is temperature at a model grid point at latitude L and longitude ℓ and the i subscript again refers to subcontinental regions. The grid-level parameters $\alpha_{(L,\ell)}$ are estimated by least squares. We compare this hybrid pattern-scaling emulator with the simple global pattern scaling described previously: the pattern is at grid level and the scaler is GMT, which we obtain from our statistical emulation. In the case of temperature emulation, the very simple hybrid approach outperforms pattern scaling for most grid points outside of the polar regions, particularly for the continental areas of greatest interest for impacts assessment (Fig. 10).

5. Alternative emulation strategies

In the previous section we compared our climate model emulation approach to pattern scaling, the most commonly used approach for emulation of climate model output in the impacts assessment community (see, e.g., Santer et al. 1990; Hulme and Raper 1995; Hulme and Brown 1998; Cabre et al. 2010; Dessai et al. 2005; Fowler et al. 2007; Harris et al. 2006; Murphy et al. 2007). However, interest is growing in alternative approaches, and it is therefore useful to compare our technique with more complex emulation strategies proposed in the recent literature (Rougier et al. 2009; Holden and Edwards 2010; Wilks 2012; Vecchi et al. 2011; Murphy et al. 2007). These strategies include the empirical orthogonal function (EOF) regression of Holden and Edwards (2010) and Gaussian process (GP) modeling, a standard method for emulating the output of deterministic computer models (Sacks et al. 1989; Santner et al. 2003; Kennedy and O'Hagan 2001; Oakley and O'Hagan 2002; Rougier et al. 2009; O'Hagan 2006). For climate models, Gaussian processes have mainly been used to emulate over physical parameters, although Holden and Edwards (2010) raise the prospect of using Gaussian processes for forcing scenario emulation. Williamson et al. (2012) use Gaussian processes for forcing scenario emulation, but only emulate a single output from the model, not a time series. A number of authors have built emulators over physical parameters in order to

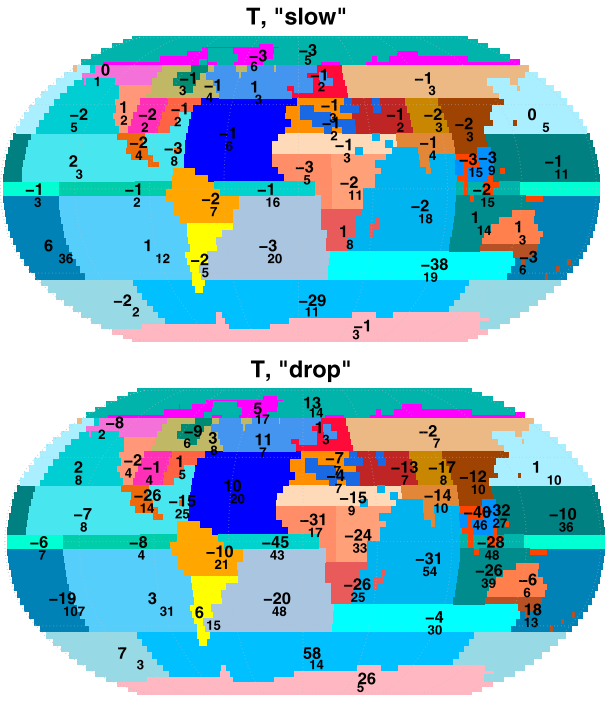

FIG. 8. Comparison between statistical emulation and pattern scaling for regional temperature. Training set, predicted scenarios, and time range for calculating indices are as in Fig. 4. The top number shown in each region is the log ratio of the temperature fit indices I_1 for the statistical model (numerator) and pattern scaling (denominator), multiplied by 100 for clarity. Negative numbers mean that statistical emulation outperforms pattern scaling. The small type gives the trend index I_2, which does not depend on the emulator. (top) For the slow scenario, the median log ratio across all the regions times 100 is −1.35 (with 10% and 90% quantiles of −2.94 and 0.93, respectively), indicating a modest advantage from statistical emulation. (bottom) Statistical emulation provides stronger benefits for the drop scenario: the median log ratio is −7.42 (with 10% and 90% quantiles of −30.08 and 10.68, respectively).

calibrate a climate model (Sanso et al. 2008; Sanso and Forest 2009; Sham Bhat et al. 2012; Drignei et al. 2008).

The GP approach to computer model emulation assumes that the output of interest is a Gaussian process in some set of inputs that vary across model runs. Among others, Challenor et al. (2010) and Rougier (2008) have discussed extensions to the GP approach to multivariate climate output, and several authors have proposed approaches for multivariate, time-dependent output: projection on a lower dimensional space via principal component analysis (Wilkinson 2010; Higdon et al. 2008) or wavelet decomposition (Bayarri et al. 2007), choice of a single representative output (Challenor et al. 2006) or a spatial aggregated average of it (Hankin 2005), kernel mixing and matrix identities (Sham Bhat et al. 2012), and dynamically autoregressive models (Fei and West 2009).

Using Gaussian processes to emulate computer models is attractive in many circumstances because it does not require the prior assumption of any particular parametric form for the relationship between inputs and outputs and provides an internally consistent approach to estimating the uncertainties of the emulator based on the GP model (Sacks et al. 1989; Oakley and O'Hagan 2002). This flexibility comes at some cost, since it is intrinsically difficult to estimate an arbitrary function nonparametrically in high dimensions. Nevertheless, to give a specific example, Challenor et al. (2006) fit a GP emulator to climate model output with 17 input parameters and only 100 model runs. This fitting is aided by the fact that most of the input parameters appear to have little impact on the output of interest. Emulation over physical parameters that are globally constant has been done with very few model runs by exploiting the information available in a spatially resolved climate model that provides many informative outputs about these parameters from each run (Sanso et al. 2008; Sanso and Forest 2009; Sham Bhat et al. 2012). In contrast, for the forcing scenario emulation, we should not assume that any of the statistical parameters in our emulators Eqs. (1) and (2) are constant across all regions, since accounting for regional differences in patterns of climate change is the whole point of our approach. We instead exploit the multiple observations in time rather than in space to build an emulator with few runs.

In our view, emulating a long time series of spatially resolved climate variables over a wide range of forcing scenarios is a highly specialized problem, and general techniques for multivariate computer model emulation are not the most appropriate tools to approach it. Choosing an appropriate emulation strategy requires recognition of three key issues: 1) the desired output variables are a function of the previous history of CO_2 or other forcings and so the emulator inputs should be functions of past trajectories; 2) because climate response is dependent only on these past trajectories, the statistical model that relates model inputs to outputs is the same for any given year [i.e., the β_j's and ρ in Eq. (1) do not depend on t]; and 3) the appropriate means of reducing the dimensionality of the problem is not to limit the inputs, which would reduce the types of forcing trajectories that can be emulated, but instead to reduce the number of parameters that need to be fit by using a structured model of the functional form describing climate response.

Reducing climate emulation to a tractable problem necessarily involves some compromises. The trade-offs of different choices are illustrated by comparing our approach to that of Holden and Edwards (2010), whose goal is the most similar to ours among published works on climate model emulation of which we are aware. Holden and Edwards (2010) share our motivation of using a collection of climate runs and relatively simple

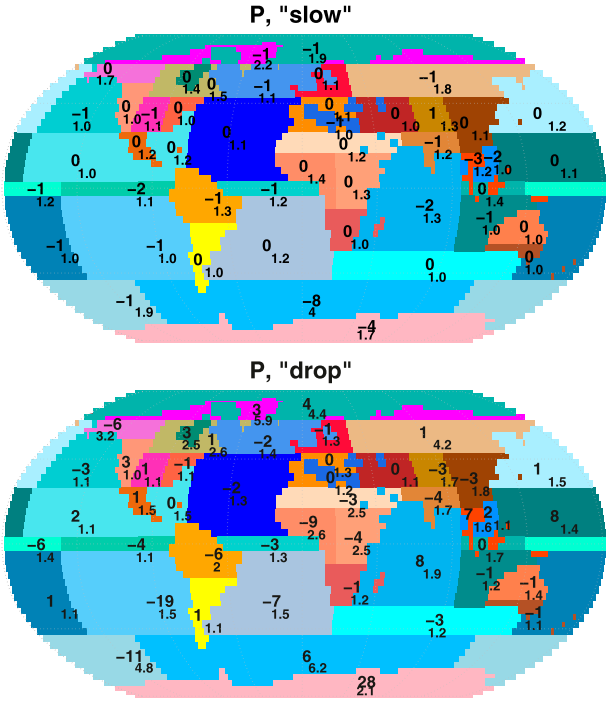

FIG. 9. As in Fig. 8, but for precipitation. The high variability in precipitation leads to smaller I_2 values and reduces the distinction between emulation methods. For the slow scenario, the median log ratio of I_1 across all regions ($\times 100$) is -0.40 (with 10% and 90% quantiles of -1.74 and 0.23, respectively); for the drop scenario it is -0.93 (with 10% and 90% quantiles of -5.70 and 5.91, respectively).

statistical techniques to produce computationally efficient climate predictions for the purposes of integrated assessment modeling, although they include both forcing scenarios and 19 climate model parameters as inputs, whereas we only consider forcing scenarios. Both their approach and our approach limit the number of parameters that need to be estimated in the statistical model, although with some noticeable differences.

Holden and Edwards (2010) emulate decadal average temperature at a single time period (2100) based on annual CO_2 levels between 2005 and 2105. If one were to directly regress each output for this problem (temperature changes for each pixel of the model) on the 100 inputs (CO_2 in each year from 2005 to 2105), the resulting parameter estimates would likely be unstable and yield problematic predictions under some CO_2 trajectories. To obtain outputs with a higher signal to noise ratio, Holden and Edwards (2010) consider just the five principal EOFs rather than results for each individual grid point as the outputs. To reduce the number of regression parameters that need to be estimated for each output, they consider only CO_2 trajectories following a specific functional form (a cubic polynomial), so that the regression is made on the three polynomial parameters (the polynomial is constrained to equal a fixed value in 2005) rather than on each of the 100 yr of the CO_2 time series. The emulation problem thereby simplifies to a regression of five outputs on three parameters of a CO_2 trajectory. This simplicity permits Holden and Edwards (2010) to extend their analysis to include emulation over physical parameters.

These choices make emulation possible, but with several limitations. Reducing spatial dimensionality of grid-level output by using EOFs rather than our use of subcontinental regions is a reasonable choice, though we believe the regional approach makes interpretation of results somewhat easier. However, restricting CO_2 trajectories to some simple functional form described by a small number of parameters (e.g., cubic polynomials) forgoes the flexibility needed for integrated assessment problems in which CO_2 emissions must be allowed to vary with economic activity, whose own growth may be complex. The restriction to cubic polynomials also precludes modeling scenarios with abrupt changes in CO_2 levels.

A more fundamental set of limitations results from formulating the output as a function of the CO_2 concentrations for a fixed set of years (which we call a fixed time-frame trajectory) rather than as a past trajectory of CO_2 concentrations. Specifically, when using fixed time-frame trajectories, the only model output that can be used for emulation are results for those years over which the prediction is sought. In contrast, using past trajectories permits use of any model runs covering any years to build a single emulator that allows predictions for all years. The limitation is less apparent in Holden and Edwards (2010) because they make only a single prediction in time (a change in decadal averages). If, however, their collection of climate model runs were used to predict temperature in an earlier period such as 2021–30, then the fixed time-frame approach would require excluding all available model output after 2030. Furthermore and perhaps more importantly, with a fixed time-frame trajectory, one would have to build and fit a new statistical model for each time point at which one wants to predict, whereas past trajectories can be used to generate a single emulator for predictions at all time points. Because the past trajectory approach uses all information in the training runs to build a single emulator, we can produce a stable emulator with much fewer training data. In some circumstances, we were able to build an effective emulator based on a single run (see Fig. 5) and can predict a whole series of annual average temperatures, whereas Holden and Edwards (2010) use 245 runs and predict only a single temperature (itself a decadal average). As we have noted, Holden and Edwards (2010) also include variation in climate model parameters but, even with a fixed climate model parameterization, they

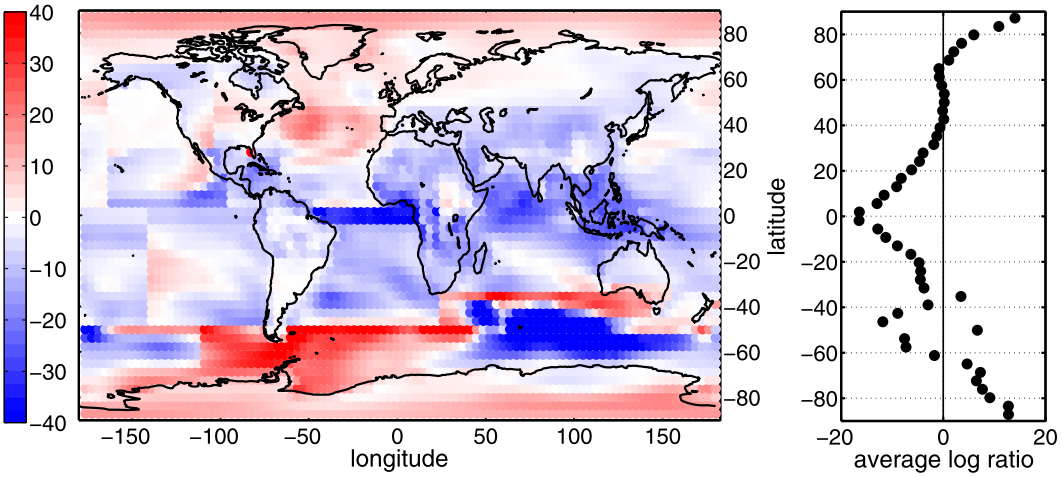

FIG. 10. Emulating temperature at grid resolution and comparison with pattern scaling. (left) The log ratio (×100) of the fit index I_1 for statistical emulation of the drop scenario over pattern scaling. This is the grid-scaled case of the bottom panel in Fig. 8. Negative values (blue) indicate that statistical emulation outperforms pattern scaling. (right) The average log ratio for different latitude bands. Statistical emulation generally outperforms pattern scaling outside the polar regions.

would need at least three runs to estimate the three parameters related to their cubic polynomial representation of the forcing scenario. The requirement for a large training set in turn led Holden and Edwards (2010) to use a climate model of only intermediate complexity, Grid Enabled Integrated Earth System Model, version 2 (GENIE-2; Lenton et al. 2007).

While the functional form we chose in Eq. (1) is somewhat arbitrary, no further increase in complexity seemed warranted. With the runs available to us, explorations with several more complex functional forms did not yield substantially better emulation performance (lower I_1) for centennial-scale predictions. On the other hand, models with fewer parameters than Eq. (1) that we have considered resulted in noticeable degradation of prediction skills for some scenarios. Our finding that temperature emulations in the somewhat realistic slow scenario yield I_1 values very near 1 in nearly all regions (e.g., Fig. 4a) implies that even the simple approach we describe leaves little room to further improve emulation of the mean temperature evolution over time scales typical of impacts assessments.

Although our emulators of mean trajectories worked very well in some circumstances, there is still room for improvement in several categories: for precipitation (where trend is small relative to variability), for scenarios with extremely rapid CO_2 changes, and for longer-time-scale scenarios. In all cases, a larger collection of climate model runs would be necessary to explore these issues. Multiple millennial-scale training runs would allow adding a second lag term in the statistical models to account for the qualitatively different climate response at long time scales. Runs with substantial jumps in consecutive years could address the misfit after rapid CO_2 changes by allowing separate contributions from each of the two most recent years rather than taking their average. Finally, a larger collection of scenarios might make it feasible to allow the regression parameters to vary smoothly in some way with the prediction scenario or, more in keeping with the approach here, the past trajectory. That is, we could construct a model that views these parameters as a function of the past trajectory, possibly as a multivariate GP after some dimension reduction on the past trajectory.

6. Conclusions

Statistical emulation of climate model output from computationally demanding AOGCMs has the potential to make climate projections capturing the full temporal dynamics of transient climates readily available for impacts assessment, policy analysis, and other applications. Developing methods that can function reasonably well with very small training sets is essential, however, to permit emulation to be a widely useful tool. The simple statistical approach we have outlined here permits us to credibly emulate climate model output with a very small training set, even in some cases of severe scenario extrapolations. Small training set size is permitted by two key aspects of our approach: treating emulation inputs (CO_2 concentrations here) as past trajectories rather than fixed time-frame trajectories and using simple, physically based statistical models that capture the relationships between CO_2 and temperature

or precipitation. The consequence is that a small training set produces rich results.

While the collection of runs used here was based on a fairly coarse spatial resolution climate model, the proven efficiency of our emulator should permit its use for emulating more state-of-the-art models based on quite small training sets. This approach performs at least as well as pattern scaling in all circumstances we have examined and substantially better in many. It therefore can be seen as a natural alternative for fast climate impacts assessments, saving orders of magnitude in computational time over running a full AOGCM.

Acknowledgments. The authors thank Ray Pierrehumbert and participants in the University of Chicago's 2008 Workshop on Modeling Uncertainty in Integrated Climate Assessment Models for valuable discussion that helped initiate this project and Matt Huber, Lan Zhao, and Wonjun Lee for computational assistance. We also thank the reviewers for many helpful recommendations leading to improvements in the substance and presentation in this paper. This work is part of the Center for Robust Decision-making on Climate and Energy Policy (RDCEP): funding was provided by grants from the University of Chicago (UC) Energy Initiative; from the University of Chicago and the Department of Energy under section H.44 of DOE Contract DE-AC02-07CH11359 awarded to Fermi Research Alliance, LLC; from STATMOS, an NSF-funded Network (NSF-DMS Awards 1106862, 1106974, and 1107046); and from the NSF Decision Making Under Uncertainty program (NSF Grant SES-0951576). Simulations were performed on "Fusion," a 320-node computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory and on TeraGrid resources operated by Purdue University. Gary Strand (NCAR) provided CCSM3 restart files. Data storage was provided by PADS (NSF Grant OCI-0821678) at the Computation Institute, a joint initiative between the UC and ANL.

REFERENCES

Allen, M., and W. J. Ingram, 2002: Constraints on future changes in climate and the hydrologic cycle. *Nature,* **419,** 224–232.

Andrews, T., and P. M. Forster, 2010: The transient response of global-mean precipitation to increasing carbon dioxide levels. *Environ. Res. Lett.,* **5,** 025212, doi:10.1088/1748-9326/5/2/025212.

Bala, G., K. Caldeira, and R. Nemani, 2010: Fast versus slow response in climate change: Implications for the global hydrological cycle. *Climate Dyn.,* **35,** 423–434.

Bayarri, M., and Coauthors, 2007: Computer model validation with functional output. *Ann. Stat.,* **35,** 1874–1906.

Branstator, G., and H. Teng, 2010: Two limits of initial-value decadal predictability in a CGCM. *J. Climate,* **23,** 6292–6311.

Cabre, M., S. Solman, and M. Nunez, 2010: Creating regional climate change scenarios over southern South America for the 2020's and 2050's using the pattern scaling technique: Validity and limitations. *Climatic Change,* **98** (3–4), 449–469.

Cao, L., G. Bala, and K. Caldeira, 2011: Why is there a short-term increase in global precipitation in response to diminished CO_2 forcing? *Geophys. Res. Lett.,* **38,** L06703, doi:10.1029/2011GL046713.

Castruccio, S., and M. Stein, 2013: Global space-time models for climate ensembles. *Ann. Appl. Stat.,* **7,** 1593–1611.

Challenor, P., R. Hankin, and R. Marsh, 2006: Towards the probability of rapid climate change. *Avoiding Dangerous Climate Change,* H. Schellnhuber et al., Eds., Cambridge University Press, 55–63.

——, D. McNeall, and J. Gattiker, 2010: Assessing the probability of rare climate events. *The Oxford Handbook of Applied Bayesian Analysis,* A. O'Hagan and M. West, Eds., Oxford University Press, 403–430.

Collins, M., 2002: Climate predictability on interannual to decadal time scales: The initial value problem. *Climate Dyn.,* **19,** 671–692.

——, and M. R. Allen, 2002: Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate,* **15,** 3104–3109.

Collins, W. D., and Coauthors, 2006: The Community Climate System Model: CCSM3. *J. Climate,* **19,** 2122–2143.

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.

Dessai, S., X. Lu, and M. Hulme, 2005: Limited sensitivity analysis of regional climate change probabilities for the 21st century. *J. Geophys. Res.,* **110,** D19108, doi:10.1029/2005JD005919.

Drignei, D., C. E. Forest, and D. Nychka, 2008: Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. *Ann. Appl. Stat.,* **2,** 1217–1230.

Fei, L., and M. West, 2009: A dynamic modelling strategy for Bayesian computer emulation. *Bayesian Anal.,* **4,** 393–412.

Forster, P., and Coauthors, 2007: Changes in atmospheric constituents and in radiative forcing. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 129–234.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.,* **27,** 1547–1578.

Giorgi, F., 2008: A simple equation for regional climate change and associated uncertainties. *J. Climate,* **21,** 1589–1604.

Hankin, R., 2005: Introducing BACCO, an R bundle for Bayesian analysis of computer code output. *J. Stat. Software,* **14,** 1–21.

Harris, G., D. Sexton, B. Booth, M. Collins, J. Murphy, and M. Webb, 2006: Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Climate Dyn.,* **27,** 357–375.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley, 2008: Computer model calibration using high dimensional output. *J. Amer. Stat. Assoc.,* **103,** 570–583.

Holden, P. B., and N. R. Edwards, 2010: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling. *Geophys. Res. Lett.,* **37,** L21707, doi:10.1029/2010GL045137.

Hulme, M., and S. Raper, 1995: An integrated framework to address climate change (ESCAPE) and further developments of the global and regional climate modules (MAGICC). *Energy Policy,* **23,** 347–355.

——, and O. Brown, 1998: Portraying climate scenario uncertainties in relation to tolerable regional climate change. *Climate Res.,* **10,** 1–14.

Judge, G., Ed., 1980: *The Theory and Practice of Econometrics.* Wiley, 810 pp.

Kennedy, M. C., and A. O'Hagan, 2001: Bayesian calibration of computer models. *J. Roy. Stat. Soc.,* **63B,** 425–464.

Lenton, T., and Coauthors, 2007: Effects of atmospheric dynamics and ocean resolution on bi-stability of the thermohaline circulation examined using the Grid Enabled Integrated Earth system modelling (GENIE) framework. *Climate Dyn.,* **29,** 591–613.

Manabe, S., and R. T. Wetherald, 1967: Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *J. Atmos. Sci.,* **24,** 241–259.

McInerney, D., and E. Moyer, 2012: Direct and disequilibrium effects on precipitation in transient climates. *Atmos. Chem. Phys. Discuss.,* **12,** 19 649–19 681.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–845.

Mitchell, J. F. B., T. C. Johns, M. Eagles, W. J. Ingram, and R. A. Davis, 1999: Towards the construction of climate change scenarios. *Climatic Change,* **41,** 547–581, doi:10.1023/A:1005466909820.

Mitchell, T. D., 2003: Pattern scaling: An examination of the accuracy of the technique for describing future climates. *Climatic Change,* **60,** 217–242, doi:10.1023/A:1026035305597.

Montgomery, D., 2012: *Design and Analysis of Experiments.* 8th ed. Wiley, 730 pp.

Murphy, J., B. Booth, M. Collins, G. Harris, D. Sexton, and M. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos. Trans. Roy. Soc.,* **365A,** 1993–2028.

Oakley, O., and A. O'Hagan, 2002: Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika,* **89,** 769–784.

O'Hagan, A., 2006: Bayesian analysis of computer code output: A tutorial. *Reliab. Eng. Syst. Saf.,* **91,** 1290–1300.

Rougier, J., 2008: Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Stat.,* **17,** 827–843.

——, D. Sexton, J. Murphy, and D. Stainforth, 2009: Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *J. Climate,* **22,** 3540–3557.

Ruosteenoja, K., T. Carter, K. Jylha, and H. Tuomenvirta, 2003: Future climate in world regions: An intercomparison of model-based projections for the new IPCC emissions scenarios. Finnish Environment Institute Tech. Rep., 83 pp.

Sacks, J., W. Welch, T. Mitchell, and H. Wynn, 1989: Design and analysis of computer experiments. *Stat. Sci.,* **4,** 409–423.

Sanso, B., and C. Forest, 2009: Statistical calibration of climate system properties. *J. Roy. Stat. Soc.,* **58C,** 485–503.

——, ——, and D. Zantedeschi, 2008: Inferring climate system properties using a computer model. *Bayesian Anal.,* **3,** 1–37.

Santer, B. D., T. M. L. Wigley, M. E. Schlesinger, and J. F. B. Mitchell, 1990: Developing climate scenarios from equilibrium GCM results. Max-Planck-Institut-für-Meteorologie Tech. Rep. 47, 29 pp.

Santner, T. J., B. J. Williams, and W. I. Notz, 2003: *The Design and Analysis of Computer Experiments.* Springer-Verlag, 283 pp.

Sham Bhat, K., M. Haran, R. Olson, and K. Keller, 2012: Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics,* **23,** 345–362.

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.

Vecchi, G. A., M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. Held, and R. Gudgel, 2011: Statistical–dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Wea. Rev.,* **139,** 1070–1082.

Wilkinson, R. D., 2010: Bayesian calibration of expensive multivariate computer experiments. *Large-Scale Inverse Problems and Quantification of Uncertainty,* L. Biegler et al., Eds., John Wiley & Sons, 195–215.

Wilks, D. S., 2012: "Superparameterization" and statistical emulation in the Lorenz '96 system. *Quart. J. Roy. Meteor. Soc.,* **138,** 1379–1387.

Williamson, D., M. Goldstein, and A. Blaker, 2012: Fast linked analyses for scenario-based hierarchies. *J. Roy. Stat. Soc.,* **61,** 665–691.

Wu, P., R. Wood, J. Ridley, and J. Lowe, 2010: Temporary acceleration of the hydrological cycle in response to a CO_2 rampdown. *Geophys. Res. Lett.,* **37,** L12705, doi:10.1029/2010GL043730.

Yeager, S., C. Shields, W. Large, and J. Hack, 2006: The low-resolution CCSM3. *J. Climate,* **19,** 2545–2566.