Reproducing Internal Variability with Few Ensemble Runs

STEFANO CASTRUCCIO AND ZIQING HU

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, Indiana

BENJAMIN SANDERSON

CERFACS, Toulouse, France

ALICIA KARSPECK

Jupiter Intelligence, Boulder, Colorado

DORIT HAMMERLING

Colorado School of Mines, Department of Applied Mathematics and Statistics, Golden, Colorado

(Manuscript received 15 April 2019, in final form 29 August 2019)

ABSTRACT

While large climate model ensembles are invaluable tools for physically consistent climate prediction, they also present a large burden in terms of computational resources and storage requirements. A complementary approach to large initial-condition ensembles is to train a stochastic generator on fewer runs. While simulations from a statistical model cannot capture the complexity of climate model runs, they can address some specific scientific questions of interest, such as sampling the variability of regional trends. We demonstrate this potential by comparing simulations from a large ensemble and a stochastic generator trained with only four runs, and show that the variability of regional temperature trends is almost indistinguishable. Training stochastic generators on fewer runs might prove especially useful in the context of large climate model intercomparison projects where creating large ensembles for each model is not possible.

1. Introduction

Projections of future climate are inherently uncertain due to sensitivity to and uncertainty about the choice of model, scenario, and parameter values, as well as initial conditions. To assess the uncertainty of climate model projections, the standard approach is to generate ensembles of simulations, a process that rapidly becomes computationally burdensome. In particular, generating a sufficient number of realizations to resolve the range of naturally occurring multidecadal change, and to differentiate it from radiative forced change, requires a considerable computational effort in addition to a heavy storage burden.

Efforts of uncertainty quantification from scenario and physics parameters have been ongoing for decades, with large ensemble projects specifically designed for this purpose, such as the Coupled Model Intercomparison Project phase 5 (CMIP5; Taylor et al. 2012) on a global scale, or the Coordinated Regional Downscaling Experiment (CORDEX; Gutowski et al. 2016) and the North American Regional Climate Change Assessment Program (NARCCAP; Mearns et al. 2009) on a regional scale. In comparison, studies aimed at isolating and quantifying the internal variability have been more recent, and this work is focused on this particular source of uncertainty. The influence of internal processes of the climate systems on the climate response uncertainty has been widely acknowledged (Yoshimori et al. 2005; Nikiema and Laprise 2011; Torn 2016), and many studies have been performed for publicly available ensembles. Hu et al. (2012) used the ensemble from the Coupled Model Intercomparison Project phase 3 to assess the

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-19-0280.s1.

Corresponding author: Stefano Castruccio, scastruc@nd.edu

DOI: 10.1175/JCLI-D-19-0280.1

magnitude of internal variability for temperature, geopotential height, and precipitation, while Lutsko and Takahashi (2018) analyzed the relationship between internal variability and climate sensitivity with energy fluxes and the global-mean surface temperature. Few studies, however, have been specifically designed to quantify this factor. Deser et al. (2012) isolated internal variability with a 40-member ensemble of the Community Climate System Model, version 3 (CCSM3; Gent 2006), by sampling conditions from a control run. Bengtsson and Hodges (2019) used a 100-member ensemble of the Max Planck Institute for Meteorology (MPIMET) for the 1850–2005 period and compared simulated temperature with HadCRUT4 surface temperature data. Kay et al. (2015) proposed a Large Ensemble (LE) of 30 members generated using a single configuration of the Community Earth System Model (CESM; Hurrell et al. 2013). Initial conditions in the atmospheric temperatures were minutely perturbed, creating a plume of diverging simulations until the end of the present century.

The LE required over 10 million CPU hours and over 300 TB in storage, a considerable amount of resource invested for the sole purpose of isolating internal variability. In practice, climate projections are concerned with assessing not just internal variability, but also uncertainty from other model inputs such as future scenarios. Assessing internal variability with the level of detail of the LE while also assessing other sources would require prohibitively large ensembles, given the present computational resources. Therefore, methodologies for reducing both the computational and storage strain, while still providing a sensitivity analysis spanning a sufficiently large area of the model input space, are necessary.

In recent decades, statistical models have been widely used in geoscience as means to quantify uncertainty from scenarios and parameters efficiently. Indeed, a properly validated statistical model can provide a stochastic approximation of climate model output, and is often referred to as metamodel or emulator (Sacks et al. 1989; Kennedy and O'Hagan 2001). Statistical models cannot capture the physical mechanisms and (in most cases) the intervariable relationships of climate models. Given a few ensemble runs, the emulator can, however, be trained to learn the main modes of spatiotemporal variability for some variables of interest, and hence be used as an efficient surrogate of climate model output to assess model sensitivity. Emulators have been extensively used for impact assessment (Harris et al. 2006; Murphy et al. 2007), and the methodology traditionally assumes that the uncertainty depends on the input space via a Gaussian process (Gaussian process emulation; Kennedy and O'Hagan 2001; Rougier et al. 2009) to perform interpolation at unobserved combinations of the input space, known as *kriging* in the statistical literature (Stein 1999). With the noticeable exceptions of a few recent works (e.g., Holden and Edwards 2010; Castruccio et al. 2014), emulation strategies have been focused on sensitivity analysis of physics (Oakley and O'Hagan 2002) or for calibration (Chang et al. 2014; Bounceur et al. 2015; Chang et al. 2016). The value of emulators is widely acknowledged in the climate model community, and their development is explicitly mentioned as a computationally affordable alternative to sensitivity analysis in the reference work for the upcoming CMIP6 ensemble (O'Neill et al. 2016).

While emulators have been used for decades to quantify scenario and parametric uncertainty, they cannot be directly applied to assess internal variability. Indeed, while the scientific objective is still uncertainty quantification as a function of model input, there is a substantial difference: an emulator uses the distance in the input space as the key factor to predict the computer output at an unobserved parameter combination. Initial conditions are, however, by construction considered equivalent, and hence there is no meaningful notion of difference in distance in the input space.

Recently, more appropriate statistical models, referred as *stochastic generators* (SGs; Jeong et al. 2019), propose a stochastic approximation of the original climate model simulation, in order to instantaneously generate surrogate simulations and assess internal variability in a computationally affordable fashion (i.e., orders of magnitude faster). SGs are reminiscent both in terminology and scope of stochastic weather generators (Wilks and Wilby 1999; Ailliot et al. 2015), but are also substantially different in that they focus on larger time scales and simulated data; see Jeong et al. (2019) for a complete discussion on their relative merits. SGs can be also regarded as degenerate versions of emulators with no input space, and have provided useful approximations to the spatiotemporal evolution of annual (Castruccio and Stein 2013; Castruccio et al. 2014; Castruccio 2016; Castruccio and Guinness 2017; Castruccio and Genton 2018) and daily (Poppick et al. 2016) temperature while also imposing a causally consistent dependence on the past CO_2 trajectory. Extensions to three-dimensional temperature (Castruccio and Genton 2016), as well as wind at annual (Jeong et al. 2018), monthly (Jeong et al. 2019), and even daily scales (Tagle et al. 2019), have also been recently proposed.

While SGs have been mostly developed in the statistical community as a methodological development in the context of space–time models, their scope can be

enlarged to climate model users whose primary goal is the quantification of internal variability. The goal of this work is to fill this gap by introducing the SGs to the geophysical community, by intentionally focusing on a simpler setting than the aforementioned statistical literature. Indeed, this work aims at simulating the monthly regional temperature from 1920 to 2100 by adopting a simple yet surprisingly powerful time series model as a proof of concept. The SG is trained on only four runs from the LE and generates a statistical ensemble whose regional temperature trends (as well as their associated variability) are comparable to those derived from all 30 ensemble members of the LE. This work focuses on the storage and computational savings implied by using a smaller ensemble than the LE, and then using the SG to assess the internal variability. Our proposed model solely aims at approximating the output from a climate model, and hence we do not perform any validation with ground-based data or reanalysis.

Our work proceeds as follows. Section 2 provides a description of the dataset used, the CESM Large Ensemble. In section 3 we describe the method used for the analysis, as well as the diagnostics for the model and discussion on its limitation. Section 4 discusses the main results by applying our model and comparing the simulations from the ones from the CESM LE. In section 5, we conclude with a discussion.

2. The Large Ensemble

The LE (Kay et al. 2015) is a set of 30 global climate simulations conducted using a nominal 1° coupled ocean, atmosphere, land, and sea ice configuration of the CESM. A single simulation is conducted over the years 1850–1920, after which a small random perturbation in atmospheric temperature is introduced to produce 30 simulations that diverge in their evolution. The world is divided into 47 regions (see Fig. S1 in the online supplemental material) to provide areas with an approximately homogeneous temperature response, following a modification of the regions presented in Ruosteenoja et al. (2007) and used in Castruccio et al. (2014).

Each ensemble member can be considered an independent random sample of the internally generated atmospheric variability in the model, conditional on the climatology, and Fig. S2 in the supplemental material provides evidence for this assumption. The LE is not designed to sample the internal variability of the ocean, given its longer time scales necessary to forget the initial state. Each member is simulated adopting historical emissions from 1920 to 2005 and using the representative concentration pathway 8.5 scenario (RCP8.5; van Vuuren et al. 2011) during 2005–2100. All ensemble members are thus subject to an identical radiative forcing scenario and the ensemble mean can be considered an estimate of the forced response of the model. The model was run assuming a change in all greenhouse gases, and we consider the equivalent CO_2 as the sole contributor for the RCP8.5 radiative forcing, thus excluding the contribution of aerosols and other short-lived compounds. This metric accounts for the combined contribution of all greenhouse gases, weighting them by their global warming potential, as specified by the Intergovernmental Panel on Climate Change (IPCC) (IPCC 2013). In this work, only the LE monthly mean temperatures are considered.

To quantify the internal variability of the different regions, Fig. 1 compares the first 5 years of all runs. The results are shown in the form of a functional boxplot (Sun and Genton 2011), the equivalent of a boxplot for the entire sequence of monthly temperatures. The black curves represent the median temperature profile across all the ensemble members for the different regions, while the blue envelope is indicative of the internal variability.

Three regions—the equatorial Pacific east and west (EPE and EPW) and warm pool equatorial (WPE) regions—display a clearly more variable interannual behavior across runs. The magnitude of the internal variability can be quantified by calculating the ratio between the functional interquartile range and the interannual temperature range, as indicated by the top of every panel in the figure. A ratio greater than one indicates that the internal variability is larger than the interannual range, and for EPE, EPW, and WPE the internal variability is more than twice as large as the range.

3. Methods

For each of the 47 regions, the regional monthly temperature $T(t)$ from 1920 to 2100 is modeled using a time series model. We assume that the evolution of $T(t)$ is influenced by present and past values of the forcing $f(t)$, $f(t-1)$,..., with increased influence from values closer to the present t:

$$T(t) = \mu[f(t), f(t-1), \dots] + \varepsilon(t). \tag{1}$$

This infinite distributed lag model (Judge et al. 1980, ch. 10) assumes that μ is the mean temperature and $\varepsilon(t) = \phi_1\varepsilon(t-1) + \phi_2\varepsilon(t-2) + \eta(t)$. The terms ϕ_1 and ϕ_2 are autocorrelation parameters and $\eta(t)$ is Gaussian white noise with a different variance for every month, estimated from data between 1920 and 2000. The forcing $f(t)$ is specified as $f(t) = \log\{[CO_{2e}](t)/[CO_{2e}]^{(B)}\}$, where $[CO_{2e}](t)$ is the CO_2 equivalent at time t and $[CO_{2e}]^{(B)}$ is the baseline CO_2 equivalent (i.e., the first time point of

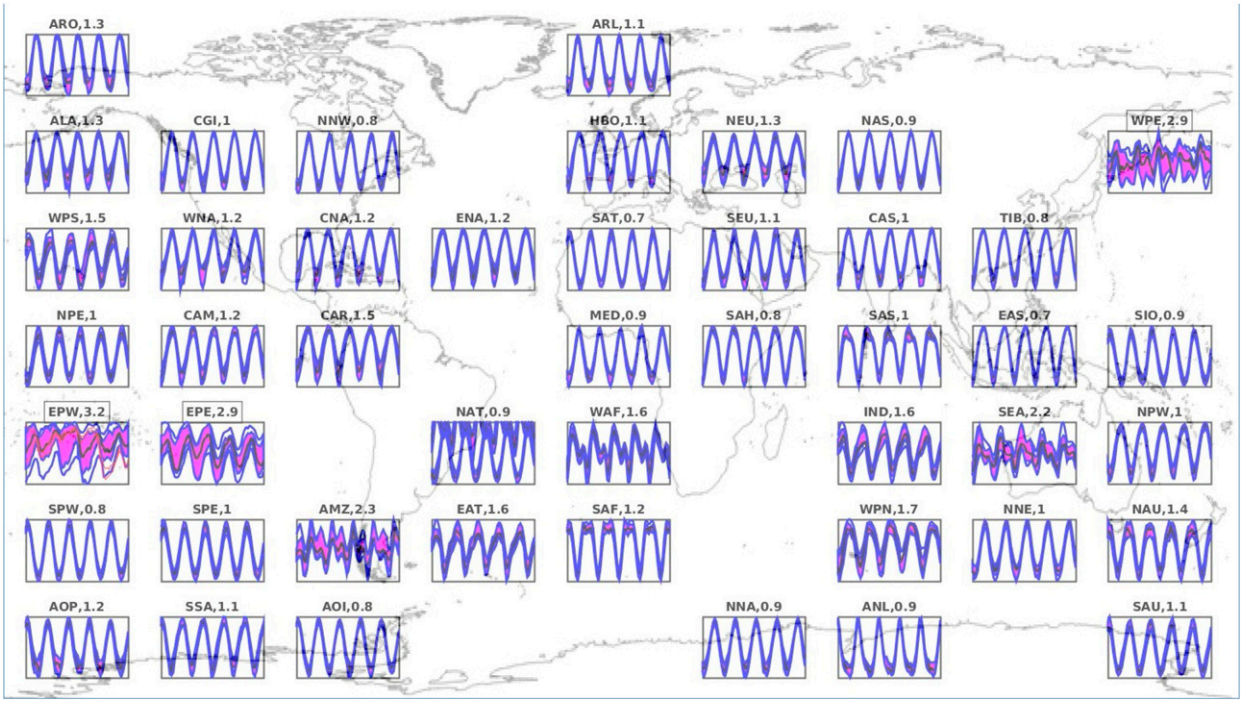

FIG. 1. Functional boxplots (Sun and Genton 2011) of the first 5 years in the large ensemble for all 47 regions. In the title, the ratio between the interquartile range and the range is reported, and the boxed titles represent the three regions with the highest ratio. Regions are arranged to approximate their geographic distribution, and the coastline is overlaid.

the previous time series). For the sake of simplicity, the variance of the response is assumed to be constant across years. This implies a lack of representation of nonseasonal variability to forcing, and thus the SG cannot capture changes in variance throughout the simulation years for processes such as El Niño–Southern Oscillation, which has been shown to change significantly throughout the twenty-first century for the LE (Maher et al. 2018).

The mean $\mu(t)$ comprises an interannual temperature cycle, described through a Fourier series, whose intensity is allowed to change in time as the forcing increases, and a term modeling dependence on the past forcing:

$$\begin{aligned}\mu[f(t), f(t-1), \dots] &= \beta_0 + \beta_1 C(t) \\ &+ \sum_{k=1}^{K}\left\{\gamma_k \cos\left(\frac{2\pi tk}{12}\right) + \zeta_k \sin\left(\frac{2\pi tk}{12}\right)\right\} \\ &+ \sum_{k=1}^{K}\left\{\gamma_k' C(t)\cos\left(\frac{2\pi tk}{12}\right) + \zeta_k' C(t)\sin\left(\frac{2\pi tk}{12}\right)\right\}, \quad (2)\end{aligned}$$

where

$$C(t) = \sum_{m=0}^{\infty}(1-\rho)\rho^m f(t-m).$$

The intercept β_0 is the average annual temperature, the slope β_1 is the coefficient for the temperature response to the past values of the forcing, $\sum_{k=1}^{K}\{\gamma_k \cos(2\pi tk/12) + \zeta_k \cos(2\pi tk/12)\}$ represents the preindustrial annual temperature cycle described through the K Fourier coefficients γ_k and ζ_k, and $\sum_{k=1}^{K}\{\gamma_k' C(t)\cos(2\pi tk/12) + \zeta_k' C(t)\cos(2\pi tk/12)\}$ describes the response of the annual cycle to the past values of the forcing. Finally, $\rho \in [0,1]$ represents how the past contribution of the forcing is weighted (i.e., how much influence the current forcing has with respect to its past values). A large value of ρ represents a strong contribution of the past forcing, while small values account for temperatures dependent only on present or very recent forcing. When $\rho = 0$, $T(t)$ depends just on $f(t)$.

The model is fully specified by the set of parameters $(\beta_0, \beta_1, \gamma_k, \gamma_k', \zeta_k, \zeta_k', \rho, \phi_1, \phi_2, \sigma^2)$, which are estimated for each region independently. The inference procedure consists of two stages. In the first stage the month-specific variance of $\varepsilon(t)$ as well as the autoregressive coefficients are estimated from the four training runs. In the second stage, if ρ was fixed in (2), all other model parameters would be linear and inference could be achieved with a regression model. Since ρ is unknown, an estimate can be obtained conditional on a linear model. This feature allows a fast optimization that can be performed within a few seconds on a modest laptop. The technical details are shown in the

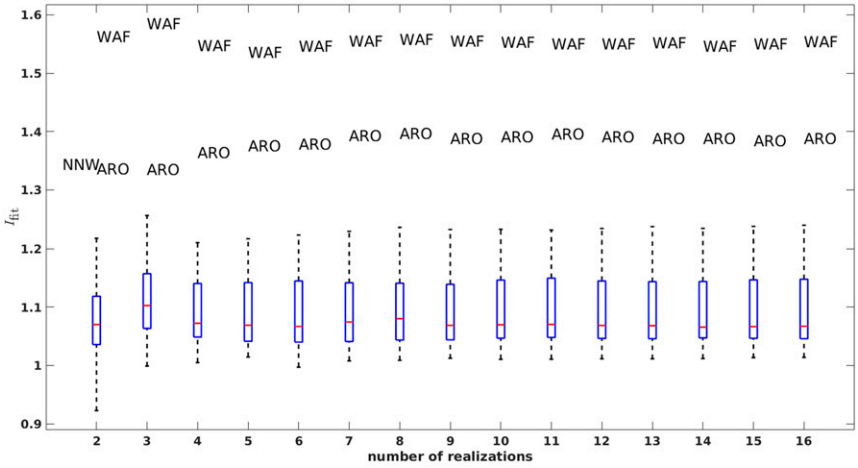

FIG. 2. Lack of fit index, (3), against number of realizations in the training set for all regions. The outliers are indicated with the region code (see Fig. S1 for a complete list).

supplemental material. The statistical inference and simulation is computationally inexpensive. The model (1) can be fit, and realizations can be created, for all regions in less than a minute on a modest laptop.

For each region, we create 30 realizations of monthly temperature from the fitted statistical model. We use these to estimate the boreal winter (DJF) trends discussed in section 4, and we consider the same two 34-yr reference periods (1979–2012 and 2013–46) as in the reference paper for the LE (Kay et al. 2015).

a. Choosing the training set size

The choice of the training size is a balance between stability in the inference of the SG, and benefits for a reduction in computational costs from performing many climate simulations. Indeed, a too-small training set would result in noisy estimates and suboptimal performances, while a large training set would diminish the computational savings implied by using a reduced subset of the LE to train the SG.

The choice of the training set can be informed by a simple index that leverages on the independence of the LE members conditional on their climatology, and quantifies the fit between the LE members and the SG (Castruccio and Stein 2013; Castruccio et al. 2014):

$$I_{\text{fit}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \hat{T}(t)]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \overline{T}(t)]^2}, \tag{3}$$

where $\overline{T}(t) = 1/R\sum_{r=1}^{R} T_r(t)$ is the average across realizations, and $\hat{T}(t)$ is the fitted model according to the SG. [The adjustment factor $R/(R-1)$ accounts for the dependence between $T_r(t)$ and $\overline{T}(t)$.] When I_{fit} is close to 1, this indicates that the statistical model is able to perform as well as the LE mean. The extent of the departure of I_{fit} from 1 measures the inability of the SG to capture the variability against the LE mean. The validity and interpretation of I_{fit} hinges on the Gaussian assumption of the data, which at monthly time scale has been verified to be appropriate; see the diagnostics section (section 3c). For higher temporal resolutions or other variables with non-Gaussian behavior, a different assessment metric must be chosen.

Figure 2 shows the variability of I_{fit} across all 47 regions in the form of a boxplot and as a function of the training set size, with outliers indicated by region code (see Fig. S1). Note that I_{fit} is very close to 1 for many regions, an indication that the SG proposed is able to capture the variability around the mean climate as well as the LE mean (the outliers will be discussed in section 3c). The same figure also shows how for a training set size of four or more LE members, very little to no additional improvement is achieved by increasing the sample size, as apparent from the very stable boxplots of I_{fit}. Based on these results, for each of the 47 regions we use only four randomly selected simulations (out of the possible 30 runs in the LE) to obtain the estimates for the statistical model.

The model can be further improved to achieve values of I_{fit} even closer to one, but only at the cost of an increase in SG complexity, which would have impacted parameter interpretability. In this work, we opted for a small loss in fit in an effort to provide a simple and intuitive model.

While I_{fit} assesses the goodness of fit, it does not measure the strength of the trend, so a second index

aimed at quantifying it must be introduced. As in Castruccio et al. (2014) we propose

$$I_{\text{trend}} = \frac{\frac{T}{T-1}\sum_{r=1}^{R}\sum_{t=1}^{T}[T_r(t)-\overline{T}_r]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{T}[T_r(t)-\overline{T}(t)]^2}, \tag{4}$$

where $\overline{T}_r = (1/T)\sum_{t=1}^{T} T_r(t)$ is the average across time for each realization. This index indeed measures the relative magnitude of the signal in time against internal variability. For monthly temperature, the intra-annual variability is clearly the dominating feature of the time series and is much larger than internal variability for all but a few regions, and Fig. S3 shows how regions with the strongest interannual variability tend to be the ones with a larger index. Overall, all indices are considerably larger than one, hence underlying the existence of a temporal trend.

b. *Diagnostics of the model*

1) GAUSSIANITY

The SG as defined in (1) and (2) hinges on several assumptions that need to be verified. Normality of $\varepsilon(t)$ in (1) can be assessed by performing inference according to the previous section, and obtaining the residuals. Figure S4 shows the histogram of the residuals, with a normal probability density function overlaid for the (a) southern Europe and (b) Atlantic Ocean regions. The two curves have a close resemblance, so our assumption of Gaussianity is justified.

2) TEMPORAL STRUCTURE

The proposed time series model assumed an autoregressive structure with two lags. Figure S5 shows the estimated autocorrelation coefficients for an AR(4), where the red regions represent significant p values at 5%. While the first coefficient is always significant (Fig. S5a), the second is only in some areas (Fig. S5b), and as the lag increases the number of significant areas decreases to only a few regions (Figs. S5c and S5d). A choice of AR(3) would also have been possible, at least for some areas, but a simpler model was preferred.

c. *Relative merits of the stochastic generator*

1) DEPENDENCE ON THE BACK TRAJECTORY

The model (1) assumes a dependence from present and past concentration of CO_2 equivalent. This assumption allows a more sensible extrapolation in time than simpler parametric trend models. The inadequacy of simpler trend models could be assessed by comparing the extrapolation on a longer time scale. The LE does not have extended RCP, so the Community Climate System Model, version 4 (CCSM4; Gent et al. 2011), first CMIP5 realization was used, from 2005 to 2300. Our SG and a model with a linear and quadratic increase was fit from 2005 and 2100 (plus the same interannual cycle as the SG), and then extrapolations were performed from 2101 to 2300. Figure S6 shows the results for two regions, the Arctic Ocean and southern Europe. For both regions, it is apparent how a model with no forcing dependence misses completely the trajectory for years beyond 2100, in the case of the Arctic Ocean by an order of magnitude. Our model, while not perfect, predicts a more sensible future temperature closer to the simulations.

Allowing for dependence on the forcing back trajectory also allows the SG to be readily generalizable to potential extensions of the LE (or other ensembles) where multiple forcings are used.

2) CONDITIONAL SIMULATIONS

The SG simulations over the next section are provided conditionally on the estimated parameters being the true one. Given the length of each time series, the estimates have very high precision so this practice is justified. Indeed, Fig. S7 shows a map of the median signal-to-noise ratio for all linear coefficients across regions (i.e., a ratio between the estimated linear coefficient its standard deviation). The estimates are overall stronger than the noise, and hence the uncertainty propagation is not expected to play a major role, as will be shown in the next section. Alternatively, one could embed the parametric uncertainty into the simulations, a Bayesian model could be defined, and simulations could be performed conditional on draws from the parameters' posterior.

3) MODEL INADEQUACY AGAINST EXTERNAL FORCING

While the SG is able to capture the climatology for temperature patterns forced directly by greenhouse gas concentrations, it is not able by construction to capture change in patterns due to either indirect effects such as melting of ocean ice or short-term extreme meteorological events whose effect is sizable on the monthly temperature. In Fig. S8, we show a comparison similar to that of Fig. 3 for the two regions with the highest I_{fit} according to Fig. 2, the West Africa (WAF) and Arctic Ocean (ARO) regions. For ARO, the model sensibly (but systematically) underestimates the minimum temperature, as the scatterplot in the top-left corner shows. This lack of fit is dictated by the lack of information about polar ice coverage, which is foreseen to be zero

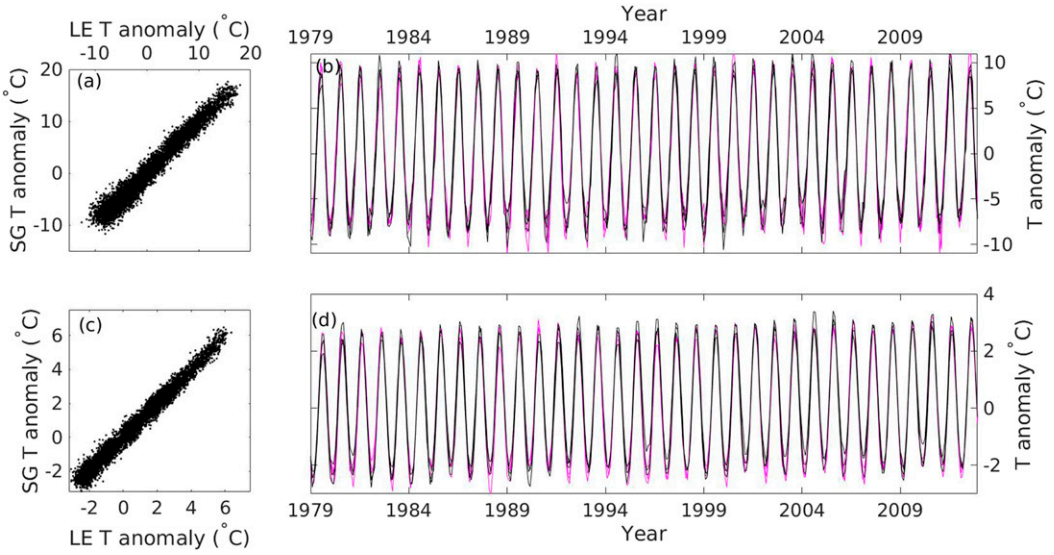

FIG. 3. A comparison of three randomly selected runs from the LE (magenta) against three randomly selected superimposed runs of the SG (black), for the (a),(b) southern Europe and (c),(d) North Atlantic regions, showing (a),(c) the scatterplot of LE against SG data and (b),(d) the time series for a 34-yr reference period (1979–2012) used in the trend analysis.

well before 2100 under RCP8.5. A joint model for ice coverage and temperature is expected to improve the marginal temperature simulations. For WAF, the SG is able to reproduce the main temporal features of the LE, but the bimodal interannual trend of the monthly temperature has a strong variability due to the transition between dry and wet seasons, and is not captured effectively by the SG. A relative lack of fit does not, however, imply that the SG should not be used for these regions. (Indeed, Fig. 6 shows that the long-term trend is captured in these regions as well.)

4. A stochastic generator for the Large Ensemble

While our main goal is to compare the variability in long-term temperature trends, we start by investigating the time-evolving regional monthly distributions. We choose two of the 47 regions, namely the southern Europe (SEU) and North Atlantic (NAT) regions (see Fig. S1 for the area locations and Table S1 for their bounds), to illustrate the monthly distributions results, but similar considerations apply for the other regions. Figure 3 shows the time series of three randomly selected realizations from the LE, which were not used in the training set of the statistical model, superimposed on three realizations from the statistical model (1) for the SEU and NAT regions. For both regions, the three superimposed realizations are visually indistinguishable, and the SG is able to capture the climatological trend and the seasonal cycle correctly, as is apparent from the entire time series (Figs. 3a,c) and from the detailed time series for the first 34-yr reference period (Figs. 3b,d).

The time series plots indicate visually indistinguishable results, but they are not ideally suited to analyze how well the internal variability for monthly temperature is reproduced by (1). The boxplots shown in Fig. 4 provide more insights on this feature by comparing the spread by month. Three runs from the LE (not in the training set) are compared against three runs from the SG for each month between 1920 and 2000. The statistical model is able to generate a variability very similar to the internal variability of monthly temperatures from the climate model runs. In the previous section, we stated that uncertainty quantification is not a strong factor here, and in Fig. S9 we show the same results as in Fig. 4, with an additional boxplot as generated by a SG, which accounts for uncertainty propagation from the linear parameters, and the results confirm how uncertainty propagation is not strong in this case due to the length of the time series. In Fig. S10 we also report a boxplot of the interannual range (i.e., maximum annual temperature minus minimum annual temperature) for the same regions and time window. The SG shows a similar range of variability to the LE also according to the interannual metric, both across land and ocean.

Figure 5 shows the variability in multidecadal trends from the 30 LE simulations compared to an equal number of runs generated from the SG in (1) for boreal winter (DJF) temperature trends over the two 34-yr reference periods. Similar to the findings in Kay et al.

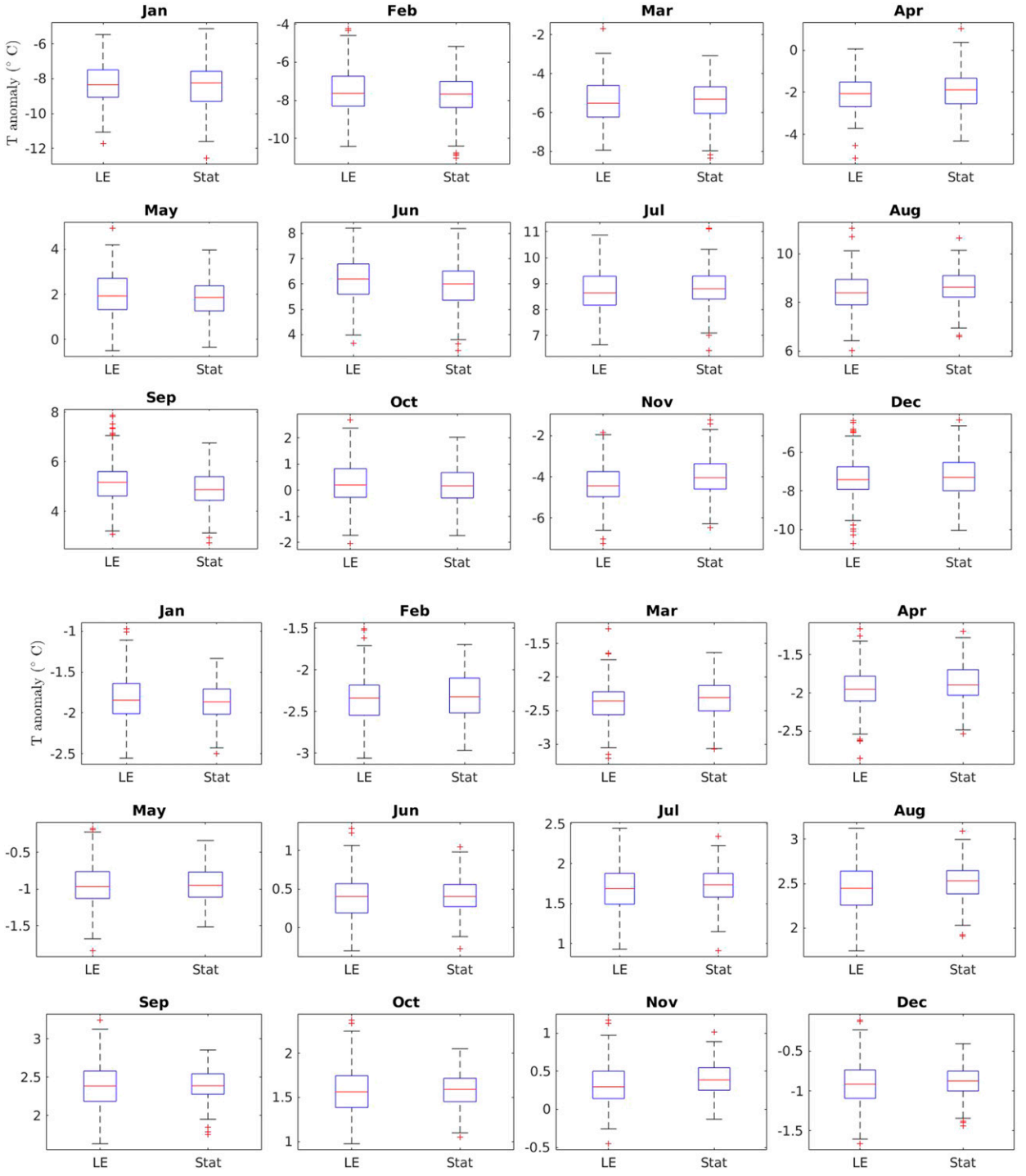

FIG. 4. Boxplot comparison of the distribution of temperature for the (top) southern Europe (SEU) and (bottom) North Atlantic (NAT) regions. Three runs for the large ensemble are considered from 1920 to 2000 (left boxplot) and compared against three runs generated from the statistical model in (1) (right boxplot) for each month.

(2015), there is a large spread in temperature trends due to internal variability. There is some bias in the southern Europe region in Figs. 5a and 5b, which could be partly attributable by the lack of characterization of aerosol forcing in the input. A small underestimation of the uncertainty of the North Atlantic in Fig. 5c is attributable to the lack of a time-varying, nonseasonal variability in the SG, which cannot capture the changes in

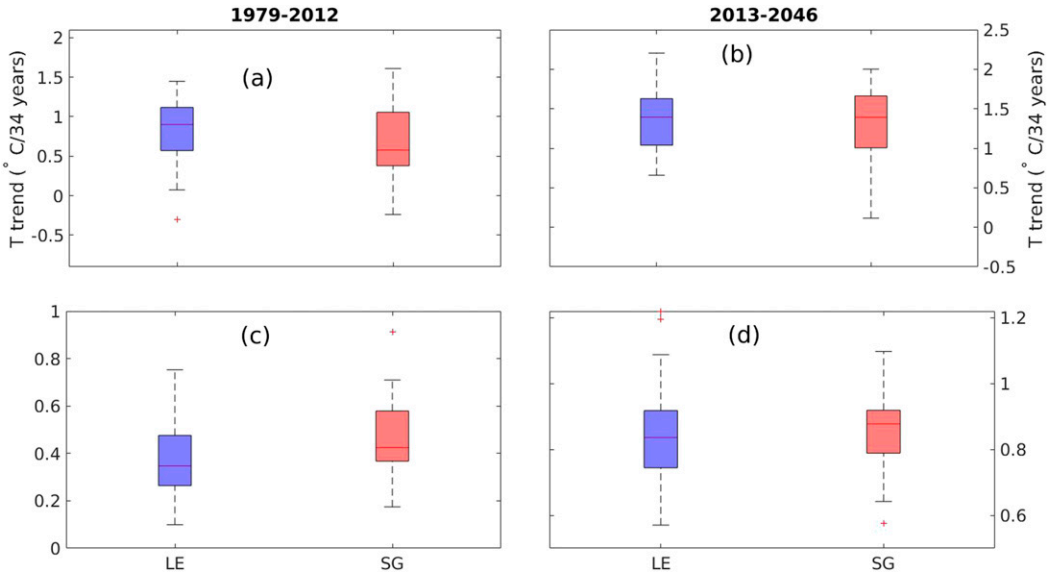

FIG. 5. Boreal winter (DJF) surface air temperature trends [°C (34 yr)$^{-1}$] for the 30 LE members (blue) and 30 simulations generated from the SG (red). Results are for the (a),(b) southern Europe (SEU) and (c),(d) North Atlantic (NAT) regions for the two 34-yr reference periods. The means and standard deviations from these 30 runs for the future 34-yr reference period in (b) and (d) are shown in Fig. 6.

variability of El Niño–Southern Oscillation observed in the LE (Maher et al. 2018), as well as its teleconnections. Figure S11 also shows results similar to Fig. 5 with an additional boxplot from a SG that accounts for uncertainty propagation; as for the previous analysis, it is apparent how this factor does not play a major role.

Figure 6 summarizes this finding for all regions by showing the mean trends for the future reference period, for the LE and the SG, along with the associated standard deviation. Figures S12–S15 provide the individual trend results for the first seven simulations for both ensembles and reference time periods. For the past reference period, the regions that show significant variability differences between the LE and the SG ensemble are the two equatorial Pacific regions, EPW and EPE. For the future reference period the only region exhibiting different variability is northern Europe (NEU). In all three cases the inferred variability is larger in the SG than in the LE. This is due to the statistical model not being able to perfectly capture interannual physical processes, and hence inflating the error term. For all other regions, for the past and future reference periods, the variability in the trends is indistinguishable between the LE and the SG.

5. Discussion and conclusions

We propose a time series model trained on only four climate model runs from a large initial-condition ensemble as a tool for reproducing the internal variability of regional temperature trends. We found that realizations from the stochastic generator are visually very similar to the original simulations, and reproduce the uncertainty of monthly regional temperature and derived multidecadal trends. The proposed modeling approach is comparatively simple and computationally inexpensive, and requires minimal storage. The functional form of the SG was chosen to be as simple as possible to allow for straightforward interpretation of the statistical parameters in (1) and (2). As a result, some interannual features such as the change in variability from El Niño–Southern Oscillation are not captured, but a modification of the error structure would have allowed us to incorporate this feature into the statistical model. The SG simulations are effective in capturing the uncertainty around the climatological mean of monthly regional temperatures, but other important properties such as changes in temperature extremes or threshold exceedances would require a more sophisticated, non-Gaussian statistical model, and likely a larger training set.

While the proposed model does not account for interregional dependence, work has been done to extend such methodology to account for spatial dependence, in order to generate spatially coherent patterns. Such a SG would require a considerably more sophisticated model to account for spatial dependence (Castruccio and Stein 2013; Castruccio and Guinness 2017), as well as a substantial increase in the computational burden for grid resolution, given the large dimensionality of the output.

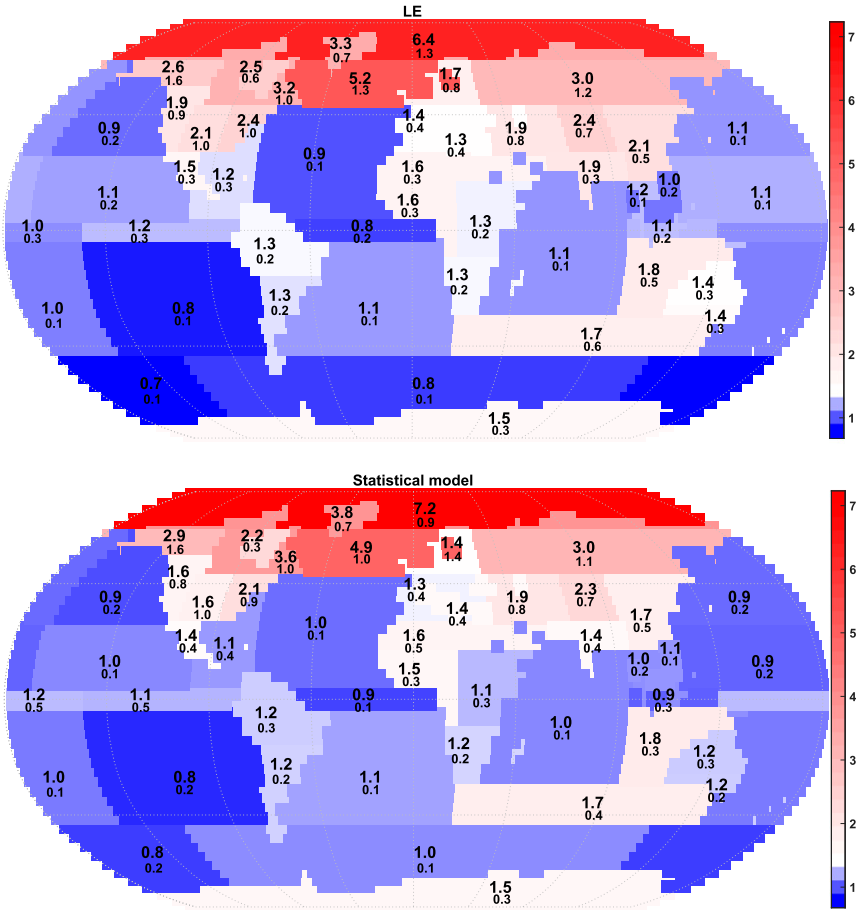

FIG. 6. Mean boreal winter (DJF) surface air temperature trends [°C (34 yr)$^{-1}$] of the (top) 30 large ensemble runs and (bottom) 30 runs generated from the statistical model for the 2013–46 reference period. The numbers below the mean values in each region indicate the standard deviations of the trend estimates. Results for a subsample of seven individual runs from the Large Ensemble and the statistical model for both reference periods are shown in Figs. S12–S15.

Similarly, extensions to three-dimensional profiles of temperatures (Castruccio and Genton 2016) have been developed, but the computational challenges associated with such methods are substantial compared to this work. The proposed framework has also been extended to other variables such and wind at annual scale (Jeong et al. 2018) and to time scales such as monthly (Jeong et al. 2019) and even daily scale (Tagle et al. 2019), which require a more articulated non-Gaussian modeling. Extensions to account for dependence across multiple variables (such as temperature and precipitation) are of high interest but have not been developed yet.

The SG is trained with changing greenhouse gases but no aerosol forcing, hence our choice of CO_2 equivalent concentration. If aerosol is present, a more general approach with radiative forcing as the input could be proposed. We anticipate that similar results should be achievable, with the added benefit of being able to capture short-term cooling events such as volcanic eruptions.

This model relies on Gaussianity and a particular choice of a functional for of the mean, but such assumptions do not limit the extent of applicability of our framework. Indeed, it is possible to develop statistical models tailored to capture internal variability, as long as the different runs convey equal information. If one is interested in another variable, another spatiotemporal scale, or another measure of interest of the climate system such as threshold exceedances or extremes, another distribution can be proposed and the proposed method could be applied along the same lines. The only requirement is that the different ensemble members are independent identically distributed samples from some distribution whose parameters are to be estimated

from a training set. Formally, we assume that each ensemble member Y is

$$Y \overset{\text{iid}}{\sim} \mathcal{F}(\boldsymbol{\theta}),$$

where $\mathcal{F}$ is a generic probability distribution (which can be non-Gaussian and/or nonstationary), and $\boldsymbol{\theta}$ is estimated from the training set.

In summary, we propose a new framework that is appropriate for quantifying internal variability with a statistical model, with a model choice that is fast, computationally affordable, widely generalizable, and may be useful for training statistical models for selected regional variables in large climate model intercomparison projects as a means of inexpensively expanding effective ensemble sizes.

Acknowledgments. CESM is sponsored by the National Science Foundation (NSF) and the U.S. Department of Energy (DOE). Administration of the CESM is maintained by the Climate and Global Dynamics Laboratory (CGD) at the National Center for Atmospheric Research (NCAR). The data are freely available at the Earth System Grid Federation.

REFERENCES

Ailliot, P., D. Allard, V. Monbet, and P. Naveau, 2015: Stochastic weather generators: An overview of weather type models. *J. Soc. Fr. Stat.*, **156**, 101–113.

Bengtsson, L., and K. I. Hodges, 2019: Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability? *Climate Dyn.*, **52**, 3553–3573, https://doi.org/10.1007/s00382-018-4343-8.

Bounceur, N., M. Crucifix, and R. D. Wilkinson, 2015: Global sensitivity analysis of the climate–vegetation system to astronomical forcing: An emulator-based approach. *Earth Syst. Dyn.*, **6**, 205–224, https://doi.org/10.5194/esd-6-205-2015.

Castruccio, S., 2016: Assessing the spatio-temporal structure of annual and seasonal surface temperature for CMIP5 and reanalysis. *Spat. Stat.*, **18**, 179–193, https://doi.org/10.1016/j.spasta.2016.03.004.

——, and M. L. Stein, 2013: Global space–time models for climate ensembles. *Ann. Appl. Stat.*, **7**, 1593–1611, https://doi.org/10.1214/13-AOAS656.

——, and M. G. Genton, 2016: Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. *Technometrics*, **58**, 319–328, https://doi.org/10.1080/00401706.2015.1027068.

——, and J. Guinness, 2017: An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *J. Roy. Stat. Soc.*, **66C**, 329–344, https://doi.org/10.1111/rssc.12167.

——, and M. G. Genton, 2018: Principles for statistical inference on big spatio-temporal data from climate models. *Stat. Probab. Lett.*, **136**, 92–96, https://doi.org/10.1016/j.spl.2018.02.026.

——, D. J. McInerney, M. L. Stein, F. Liu Crouch, R. Jacob, and E. J. Moyer, 2014: Statistical emulation of climate model projections based on precomputed GCM runs. *J. Climate*, **27**, 1829–1844, https://doi.org/10.1175/JCLI-D-13-00099.1.

Chang, W., M. Haran, R. Olson, and K. Keller, 2014: Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.*, **8**, 649–673, https://doi.org/10.1214/14-AOAS733.

——, ——, P. Applegate, and D. Pollard, 2016: Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Stat. Assoc.*, **111**, 57–72, https://doi.org/10.1080/01621459.2015.1108199.

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, https://doi.org/10.1007/s00382-010-0977-x.

Gent, P. R., 2006: Preface to Special Issue on Community Climate System Model (CCSM). *J. Climate*, **19**, 2121–2630, https://doi.org/10.1175/JCLI9020.1.

——, and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, https://doi.org/10.1175/2011JCLI4083.1.

Gutowski, W. J., Jr., and Coauthors, 2016: WCRP Coordinated Regional Downscaling Experiment (CORDEX): A diagnostic MIP for CMIP6. *Geosci. Model Dev.*, **9**, 4087–4095, https://doi.org/10.5194/gmd-9-4087-2016.

Harris, G. R., D. M. H. Sexton, B. B. B. Booth, M. Collins, J. M. Murphy, and M. J. Webb, 2006: Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Climate Dyn.*, **27**, 357–375, https://doi.org/10.1007/s00382-006-0142-8.

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, https://doi.org/10.1029/2010GL045137.

Hu, Z.-Z., A. Kumar, B. Jha, and B. Huang, 2012: An analysis of forced and internal variability in a warmer climate in CCSM3. *J. Climate*, **25**, 2356–2373, https://doi.org/10.1175/JCLI-D-11-00323.1.

Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. T. F. Stocker et al., Eds., Cambridge University Press, 1535 pp.

Jeong, J., S. Castruccio, P. Crippa, and M. G. Genton, 2018: Reducing storage of global wind ensembles with stochastic generators. *Ann. Appl. Stat.*, **12**, 490–509, https://doi.org/10.1214/17-AOAS1105.

——, Y. Yan, S. Castruccio, and M. G. Genton, 2019: A stochastic generator of global monthly wind energy with Tukey *g*-and-*h* autoregressive processes. *Stat. Sin.*, **29**, 1105–1126, https://doi.org/10.5705/SS.202017.0474.

Judge, G., W. E. Griffiths, R. Carter Hill, H. Lütkepohl, and T.-S. Lee, 1980: *The Theory and Practice of Econometrics*. Wiley, 793 pp.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kennedy, M. C., and A. O'Hagan, 2001: Bayesian calibration of computer models. *J. Roy. Stat. Soc.*, **63B**, 425–464, https://doi.org/10.1111/1467-9868.00294.

Lutsko, N. J., and K. Takahashi, 2018: What can the internal variability of CMIP5 models tell us about their climate sensitivity? *J. Climate*, **31**, 5051–5069, https://doi.org/10.1175/JCLI-D-17-0736.1.

Maher, N., D. Matei, S. Milinski, and J. Marotzke, 2018: ENSO change in climate projections: Forced response or internal variability? *Geophys. Res. Lett.*, **45**, 11 390–11 398, https://doi.org/10.1029/2018GL079764.

Mearns, L. O., W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes, and Y. Qian, 2009: A regional climate change assessment program for North America. *Eos, Trans. Amer. Geophys. Union*, **90**, 311, https://doi.org/10.1029/2009EO360002.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos. Trans. Roy. Soc.*, **365A**, 1993–2028, https://doi.org/10.1098/RSTA.2007.2077.

Nikiema, O., and R. Laprise, 2011: Budget study of the internal variability in ensemble simulations of the Canadian Regional Climate Model at the seasonal scale. *J. Geophys. Res.*, **116**, D16112, https://doi.org/10.1029/2011JD015841.

Oakley, J., and A. O'Hagan, 2002: Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, **89**, 769–784, https://doi.org/10.1093/biomet/89.4.769.

O'Neill, B. C., and Coauthors, 2016: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geosci. Model Dev.*, **9**, 3461–3482, https://doi.org/10.5194/gmd-9-3461-2016.

Poppick, A., D. J. McInerney, E. J. Moyer, and M. L. Stein, 2016: Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances. *Ann. Appl. Stat.*, **10**, 477–505, https://doi.org/10.1214/16-AOAS903.

Rougier, J., D. M. H. Sexton, J. M. 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, https://doi.org/10.1175/2008JCLI2533.1.

Ruosteenoja, K., H. Tuomenvirta, and K. Jylhä, 2007: GCM-based regional temperature and precipitation change estimates for Europe under four SRES scenarios applying a super-ensemble pattern-scaling method. *Climatic Change*, **81** (S1), 193–208, https://doi.org/10.1007/s10584-006-9222-3.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989: Design and analysis of computer experiments. *Stat. Sci.*, **4**, 409–423, https://doi.org/10.1214/ss/1177012413.

Stein, M. L., 1999: *Statistics for Spatial Data: Some Theory for Kriging*. Springer, 247 pp.

Sun, Y., and M. G. Genton, 2011: Functional boxplots. *J. Comput. Graph. Stat.*, **20**, 316–334, https://doi.org/10.1198/jcgs.2011.09224.

Tagle, F., S. Castruccio, P. Crippa, and M. G. Genton, 2019: A non-Gaussian spatio-temporal model for daily wind speeds based on a multivariate skew-*t* distribution. *J. Time Ser. Anal.*, **40**, 312–326, https://doi.org/10.1111/jtsa.12437.

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

Torn, R. D., 2016: Evaluation of atmosphere and ocean initial condition uncertainty and stochastic exchange coefficients on ensemble tropical cyclone intensity forecasts. *Mon. Wea. Rev.*, **144**, 3487–3506, https://doi.org/10.1175/MWR-D-16-0108.1.

van Vuuren, D. P., and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change*, **109**, 5–31, https://doi.org/10.1007/s10584-011-0148-z.

Wilks, D. S., and R. L. Wilby, 1999: The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.*, **23**, 329–357, https://doi.org/10.1177/030913339902300302.

Yoshimori, M., T. F. Stocker, C. C. Raible, and M. Renold, 2005: Externally forced and internal variability in ensemble climate simulations of the Maunder Minimum. *J. Climate*, **18**, 4253–4270, https://doi.org/10.1175/JCLI3537.1.