WILEY

RESEARCH ARTICLE

A spatial-dependent model for climate emulation

Jingyu Bao[1] | **David J. McInerney[2]** | **Michael L. Stein[3]**

[1]Department of Mathematical Sciences, Tsinghua University, Haidian District Beijing, 100084, China
[2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, 5005, SA, Australia
[3]Department of Statistics, University of Chicago, Chicago, 60637, IL, U.S.A.

Correspondence
Jingyu Bao, Department of Mathematical Sciences, Tsinghua University, Haidian District, Beijing, 100084, China.
Email: jingyubao@galton.uchicago.edu

For studying impacts and policy issues related to climate change, it is often critical to be able to forecast the future climate for a range of forcing scenarios. Complex climate models can be used to study climate change, but they are expensive to run, and thus, can only be used to investigate a limited number of scenarios. For some climate summaries, it is possible to develop a statistical emulator of the climate model that accurately and quickly reproduces the climate model output. Training such an emulator based on a small number of model runs can be challenging, especially when emulating at a fine spatial resolution. This work considers developing such an emulator for a specific climate model, CCSM3, as a function of the past trajectory of atmospheric CO_2 concentrations. We propose a new approach to fitting an emulator for annual temperature at the pixel level of the climate model by combining a spatially varying coefficient model and an infinite distributed lag model. The approach can capture the annual mean temperature at grid-cell level of climate model output in transient climates based on model runs from just a single CO_2 trajectory. We apply the approach to annual temperature emulation over North America and Africa, and show that the resulting emulator predicts annual temperature quite well and that the emulator can be fit in a computationally efficient manner. We show that the emulator outperforms procedures that do not take account of the spatial structure.

KEYWORDS
climate emulation, GCM, space-time model

1 | INTRODUCTION

There is a wide consensus among the scientific community that climate is changing and will almost certainly produce detrimental impacts for human society (IPCC AR5; Stocker et al., 2014). General circulation models (GCMs) are numerical models of the Earth's climate system that can produce climate predictions based on the radiative effects of CO_2 and other anthropogenic forcing agents. These models are important tools in the study of climate variability and climate change. GCM output can be used to evaluate the impacts of climate change and support possible policies to reduce them. However, the computational demands of GCMs limit their use in climate damage estimation, cost–benefit evaluation, and policy decisions, which require repeated iterations of climate projections in response to different forcing trajectories. Statistical models allow us to fit climate model output under some scenarios and emulate behavior of the climate model under a new forcing scenario in much less time (Caldeira &

Myhrvold, 2012; Castruccio et al., 2014). Our work focuses on emulating annual mean temperature at Earth's surface for different trajectories of atmospheric CO_2 concentrations. The goal is to build simple statistical models to emulate climate projections at each grid cell (without aggregation) for a wide range of CO_2 trajectories based on a small set of precomputed climate model runs.

Statistical analysis of climate model output has focused on annual or seasonal summaries aggregated over global, hemispheric, or continental spatial scales and possibly analyzed as time series or with spatial statistical models (e.g., Tebaldi, Smith, Nychkam, & Mearns, 2005; Furrer, Sain, Nychka, & Meehl, 2007; Berliner & Kim, 2008; Smith, Tebaldi, Nychka, & Mearns, 2009; Tebaldi & Sanso, 2009; Buser, Künsch, & Weber, 2010; Kaufman & Sain, 2010; Sain, Nychka, & Mearns, 2011, etc.). Greasby and Sain (2011) proposed a Bayesian hierarchical spatial model along with an intrinsic Markov random field for climate emulation. The model includes spatial effects for some regression parameters but

does not include an effect of CO_2. Castruccio and Stein (2013) introduced a spectral approach to modeling the space-time variations in annual temperatures at the pixel level, but this approach still requires a separate emulator of mean temperatures for a changing climate. Indeed, they essentially use the mean emulator in Castruccio et al. (2014), which shows how a distributed lag model (Judge, Griffiths, Hill, & Lee, 1980) can emulate mean temperature trajectories at the regional level for arbitrary CO_2 scenarios based on a small set of precomputed runs from a GCM. Two key insights in Castruccio et al. (2014) allowed accurate emulation of annual temperatures based on as few as one or two model runs. First, annual mean temperatures at any one time are a function of the past trajectory of forcings and not the specific year under consideration, so that one does not need a separate emulator for different years and thus even a single model run of many years provides substantial information about this function. Second, our understanding of the relationship between CO_2 and temperature implies that mean temperatures in a given year should be nearly linear in the past trajectory of the logarithm of atmospheric CO_2 concentrations, leading to the infinite distributed lag model (Judge et al., 1980). Furthermore, there are both short- and long-term effects of CO_2 on temperature that suggest a specific form for the distributed lag coefficients including only four unknown parameters. Although this approach worked well at the regional level, when applied at the pixel level, the estimated parameters of the distributed lag model suffer from substantial statistical variation. The main contribution of this work is to improve the estimation of the coefficients of the distributed lag model at the pixel level by including spatial dependence in them and including spatial–temporal structure in the error term.

The flexibility and interpretability of varying coefficient models make them important tools for exploring dynamic patterns in many scientific areas, such as economics, politics, epidemiology, medical science, and ecology (Hastie & Tibshirani, 1993; Fan & Zhang, 2008). Spatially varying coefficient models (Gelfand, Kim, Sirmans, & Banerjee, 2003; Gamerman, Moreira, & Rue, 2003; Assunċao, 2003) are natural extensions of varying coefficient models that allow one to exploit the expected spatial structure in relationships between spatial variables. The infinite distributed lag model is widely used in statistics and econometrics when the effect of a change in an independent variable is assumed perpetual, and diminishing over time (Judge et al., 1980; Koyck, 1954). In order to emulate annual mean temperature at the pixel level, we introduce a spatial-dependent model (SDM), a statistical model combining a spatially varying coefficient model and an infinite distributed lag model. This approach distinguishes between short-term and long-term effects of CO_2 at each grid point, where nearby grid points share similar effects based on the CO_2 trajectory, as well as characterizing spatial–temporal error.

The remainder of this paper discusses the structure and performance of the emulator. Section 2 describes the precom-

puted climate model output data. Section 3 provides some statistical background on distributed lag models and spatially varying coefficient models. Section 4 describes our model for climate model output, and Section 5 presents the likelihood approach to estimating parameters and the numerical values of the estimates. Section 6 describes the performance of our emulator and compares the results with other emulators, including models without any spatial dependence. Section 7 discusses possible extensions and improvements to the emulator described in this work. The Appendix explains why estimates of the regression coefficients in the distributed lag model should be very nearly unbiased, which is an assumption underlying one of our approaches to evaluating the effectiveness of the spatially varying coefficient models.

2 | DATASETS FOR EMULATION

To investigate climate model emulation, we use a set of climate model runs based on five different forcing scenarios (Castruccio et al., 2014). The GCM runs differ in their trajectories of future global mean atmospheric CO_2 concentration and have different initial conditions but are performed with the same model and same representation of model physics (Castruccio et al., 2014). Simulations were performed with the Community Climate System Model, version 3 (CCSM3, Yeager, Shields, Large, & Hack, 2006; Collins et al., 2006), at a relatively modest T31 atmospheric resolution ($\approx 3.75°$ × $3.75°$) and nominally $3°$ ocean resolution, a configuration that made it possible to run multiple realizations of a wide range of multicentury scenarios. In order to focus on the effect of CO_2, all other greenhouse gases and aerosol concentrations were held fixed at their preindustrial values in all runs. The five different scenarios of GCM runs used in this work include three scenarios with a gradual rise followed by stabilization of CO_2 and two others with abrupt changes. We denote the five scenarios as "*hi*" (high), "*med*" (moderate), "*low*", "*drop*," and "*jump*" (Figure 1). The scenarios were chosen to explore climate emulation across a wide range of types of CO_2 evolution. In particular, although physically unrealistic, the *drop*
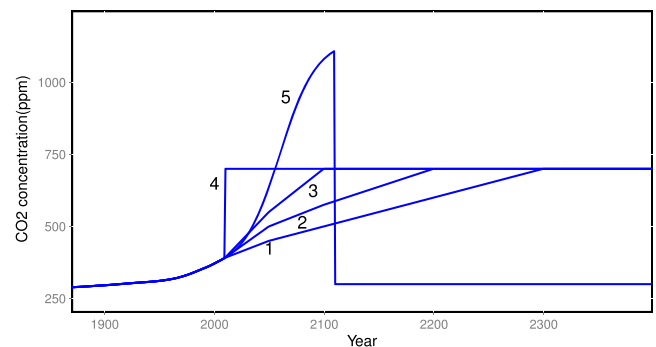

FIGURE 1 Scenarios of global mean atmospheric CO_2 concentration used for training and testing. The five different scenarios are labeled as (1) *low*, (2) *med*, (3) *hi*, (4) *jump*, and (5) *drop*. All scenarios start at year 1870. *Hi*, *med*, *low*, and *drop* end at 2399 and *jump* ends at year 2199. The atmospheric CO_2 concentrations for the five scenarios are the same for 1870–2009

and *jump* scenarios provide important information for distinguishing between short- and long-term effects of changing atmospheric CO_2 concentration. In all scenarios, the same historical values are used for atmospheric CO_2 concentrations in the years of 1870–2009, and subsequently diverge into five different trajectories for 190–390years (end years range from 2199 to 2399 and the total number of model years range from 330 to 530). We have five realizations under each scenario with different initial conditions, which were obtained using restart files from years 410, 420, 430, 440, and 450 of the National Center for Atmospheric Research preindustrial control run b30.048 (Collins et al., 2006; Castruccio et al., 2014). We treat each model run, including five realizations of each scenario, as statistically independent, which is reasonable given the extreme sensitivity to initial conditions in GCM runs (Castruccio et al., 2014).

This paper focuses on emulating annual mean temperature at grid cell level by understanding the relationship between local annual average temperature and the past trajectory of atmospheric CO_2 concentrations. We train emulators based on various subsets of GCM runs and estimate the parameters of the SDM using the methods described in the following sections. The resulting emulators are then used to predict annual mean temperature at particular locations under arbitrary CO_2 scenarios. We mainly use the five realizations of a single scenario as the training set and then test the emulator on all five scenarios, but we also consider the performance of the emulator when fewer than five realizations are used. Compared with oceans, climates over continents have different responses to CO_2 change and are of greater interest for impact assessment, so we consider the climate model output for the US and Canada (hereafter North America, NA) and Africa (AF). These two different continents have very different latitude ranges and allow us to show the broad validity of our approach. There are 198 pixels in NA and 184 pixels in AF (out of 48×96 CCSM3 grid cells for T31 resolution, see Figure 2). Our collection of runs then consists
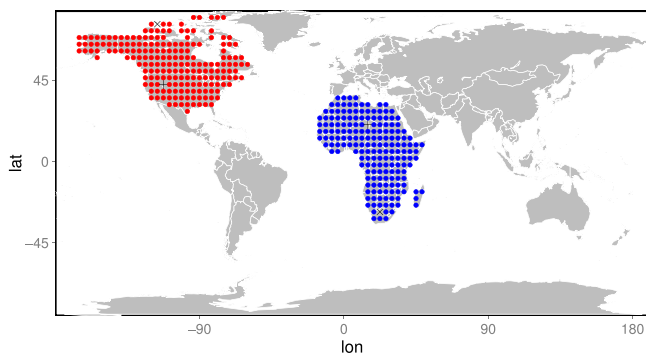

FIGURE 2 The locations used for emulation: 198 pixels in North America (red) and 184 in Africa (blue). Each location shown represents a grid cell at T31 resolution ($\approx 3.75° \times 3.75°$). Emulation is performed on annual average temperature data in these grid cells. The symbols denote specific pixels referred to in other figures ("×": Figure 6; "+" : Figure 8)

of about 100,000 spatial–temporal annual temperatures in each scenario run and each continent.

3 | STATISTICAL ASSUMPTIONS

Koyck (1954) and Judge et al. (1980) introduced infinite distributed lag models, which are of the form

$$Y(t) = \sum_{j=0}^{\infty} w_j X(t-j) + \epsilon(t), \tag{1}$$

where t denotes time, $Y(t)$ is a variable that depends on lagged values of the independent variable $X(t)$, the w_j's are lag weights, and the $\epsilon(t)$ is random error. Thus, the current value of $Y(t)$ depends on all the past values of X. Changes in atmospheric CO_2 concentrations not only affect temperature by the immediate change in the radiative properties of the atmosphere but also have long-term effects due to the slow warming of the oceans. We use the infinite distributed lag model to account for both these short- and long-term effects of changing atmospheric CO_2 concentrations on annual average temperature. As in Castruccio et al. (2014), we assume the w_j's in (1) for $j > 1$ decay exponentially to model the long-term CO_2 effect. Thus, our model has one nonlinear parameter in it to account for the rate of this exponential decay.

A simple way to fit this model at the pixel level would be to do the estimation separately at every pixel of interest, ignoring any spatial information. Another extreme would be to assume the parameters of the model do not vary across pixels within some region. A good compromise between these two extremes is to use a version of the spatially varying coefficients model described in Gelfand et al. (2003) for spatiotemporal data, which has the form

$$Y(s, t) = \sum_{i=1}^{p} X_i(s, t)\beta_i(s) + \epsilon(s, t), \tag{2}$$

where s indicates spatial location, t indicates time, $Y(s, t)$ is the outcome of interest (here, temperature), $X_1(s, t), \cdots, X_p(s, t)$ are covariates (here, functions of the past CO_2 trajectory and the exponential decay parameter), $\beta_1(s), \cdots, \beta_p(s)$ are unknown random coefficients from some multivariate spatial process model, and $\epsilon(s, t)$ are random errors from some spatiotemporal process model. Note specifically that we do not let $\beta_i(s)$ depend on t, which is a consequence of the distributed lag model.

We will assume that $\beta_1(s), \cdots, \beta_p(s)$ and $\epsilon(s, t)$ are independent Gaussian processes. Assuming that $\epsilon(s, t)$ is Gaussian is perhaps reasonable because we are considering annual averages of temperature (Castruccio & Stein, 2013), but assuming the $\beta_i(s)$'s are Gaussian and independent is for convenience.

We briefly review some material for Gaussian processes. Gaussian processes are stationary if the mean is a constant and covariance only depends on the separation at

two points. Suppose $Z(\boldsymbol{x})$ is a stationary Gaussian process on $\mathbb{R}^2$, has mean $\mu = E[Z(\boldsymbol{x})]$, and covariance function $K(\boldsymbol{h}) = E[(Z(\boldsymbol{x} + \boldsymbol{h}) - \mu)(Z(\boldsymbol{x}) - \mu)]$. A class of covariance functions widely used in geoscience and environmental studies is the Matérn covariance family (Fuentes, 2001; Nychka, Wikle, & Royle, 2002), which are specified by parameters (σ^2, α, ν) (Stein, 1999, see supplementary material for more mathematical detail). Here, $\sigma^2 > 0$ is a scale parameter, and $\alpha > 0$ is a spatial scale parameter, whose inverse, $1/\alpha$, is sometimes referred to as a correlation length (Gneiting et al., 2012). The parameter ν measures the smoothness of Z: $Z(\boldsymbol{x})$ is m times mean square differentiable if and only if $\nu > m$. If the smoothness parameter ν equals an integer plus $\frac{1}{2}$, the Matérn function reduces to the product of an exponential function and a polynomial (Stein, 1999). For example, if the smoothness parameter $\nu = 1/2$, the corresponding Matérn function is

$$K(\boldsymbol{x}) = \sigma^2 \alpha^{-1} e^{-\alpha|\boldsymbol{x}|}, \tag{3}$$

which is a exponential covariance function and if $\nu = 3/2$, the corresponding Matérn function is

$$K(\boldsymbol{x}) = \sigma^2 \alpha^{-3} e^{-\alpha|\boldsymbol{x}|}(1 + \alpha|\boldsymbol{x}|). \tag{4}$$

Model (4) is smoother than Model (3). We only use these two Matérn functions with $\nu = 1/2$ and $3/2$ throughout this paper.

4 | STATISTICAL MODELING FOR CLIMATE MODEL OUTPUT

Let $T(\boldsymbol{s}, t)$ be the annual average temperature at location $\boldsymbol{s}$ and year t, $\boldsymbol{s} \in D$, where D is some region of interest (here NA or AF). Historical research suggests that equilibrium global mean temperature change is proportional to $\log[CO_{2r}](t)$ (Manabe & Wetherald, 1967; Forster et al., 2007), where $\log[CO_{2r}](t)$ is the logarithm of the ratio between atmospheric CO_2 concentration in year t and its preindustrial value. Exploratory analyses show that the annual average temperature at nearby grid points share similar short-term and long-term effects from the change of atmospheric CO_2 concentration. To emulate the annual mean temperature pixel-wise, we propose a model that captures this dependence of nearby grid points and past trajectory of CO_2 via combining a spatially varying coefficient model (2) and infinite distributed lag model (1):

$$T(\boldsymbol{s}, t) = \beta_{0,\boldsymbol{s}} + \beta_1(\boldsymbol{s})S(t) + \beta_2(\boldsymbol{s})L(t) + \epsilon(\boldsymbol{s}, t), \tag{5}$$

where

$$S(t) = \frac{1}{2}\{\log[CO_{2r}](t) + \log[CO_{2r}](t-1)\}$$

is the short-term CO_2 covariate at year t and

$$L(t) = \sum_{i=2}^{\infty} w_i \log[CO_{2r}](t-i)$$

is the long-term CO_2 covariate at year t, with $w_i = \rho^{i-2}(1-\rho)$ decaying exponentially (and normalized to sum to 1) (see further discussion in supplementary material). The parameter ρ captures the decreasing impact of CO_2 from further in the past on current temperatures (we set the atmospheric CO_2 concentration level to equal the preindustrial level for years before the start of the model run). Here, $\beta_1(\boldsymbol{s})$ and $\beta_2(\boldsymbol{s})$, which are modeled as two independent spatially stationary Gaussian processes, give the values for the short-term and long-term effect of CO_2. These forms for the short-term and long-term effects are the same as in Castruccio et al. (2014). We assume $\beta_1(\boldsymbol{s})$ and $\beta_2(\boldsymbol{s})$ have mean μ_1, μ_2, and covariance function K_1, K_2 correspondingly, where K_1 and K_2 are Matérn covariance function with parameters $(\sigma_1^2, \alpha_1, \nu_1)$ and $(\sigma_2^2, \alpha_2, \nu_2)$ and $\nu_1 = \nu_2 = 1.5$. This choice corresponds to a differentiable process and reflects our expectation that these parameters should vary smoothly in space. For the intercepts, we find that $\beta_{0,\boldsymbol{s}}$ can be estimated well without taking account of spatial structure, so treat them as fixed effects. We model the spatial–temporal error term $\epsilon(\boldsymbol{s}, t)$ by

$$\epsilon(\boldsymbol{s}, t) = \phi \epsilon(\boldsymbol{s}, t-1) + \nu(\boldsymbol{s}, t), \tag{6}$$

where $\nu(\boldsymbol{s}, t)$ is white in time and colored in space, so that $\nu(\boldsymbol{s}, t)$ is independent from year to year but correlated across locations. For any given year t, $\nu(\boldsymbol{s}, t)$ is assumed to be a stationary Gaussian process with mean 0 and covariance K_ν, where K_ν is Matérn covariance function with parameters $(\sigma_\nu^2, \alpha_\nu, \nu_\nu)$ and $\nu_\nu = 0.5$ (see supplementary materials for further discussion on modeling the error term). We considered more complex models for $\epsilon(\boldsymbol{s}, t)$ (for example, allowing $\epsilon(\boldsymbol{s}, t)$ to depend on $\epsilon(\boldsymbol{s}', t-1)$ for $\boldsymbol{s}'$ a neighbor to $\boldsymbol{s}$), but (6) performs almost the best in the sense of mean squared error of the emulator, so we present results for this simple model.

SDM (5) distinguishes short-term and long-term effects of changes in CO_2. To demonstrate how the model changes with the decay rate ρ, we plot the short-term covariate $S(t)$ and long-term covariate $L(t)$ as functions of t for different ρ in SDM (5) under the *low* and *drop* scenarios (Figure 3). The high multicollinearity between short-term and long-term covariates when the decay rate ρ is less than 0.95, especially for the *low* scenario, makes the model difficult to fit. With this simple model, we need to find a procedure that can estimate ρ accurately and stably at all grid points. We tried different models for the decay rate ρ including modeling ρ as a function of latitude. The results show no benefit compared with fixing ρ throughout a continent. Thus, we assume ρ is the same throughout each continent. We set $\rho = e^{\rho_0}/(1 + e^{\rho_0})$ and fit ρ_0 instead of ρ to avoid problems with estimating the standard error of a parameter near its boundary value.

In order to explore the performance of adding spatial structure in the short-term and long-term effect and fixing ρ for each continent, we compare models with and without spatial
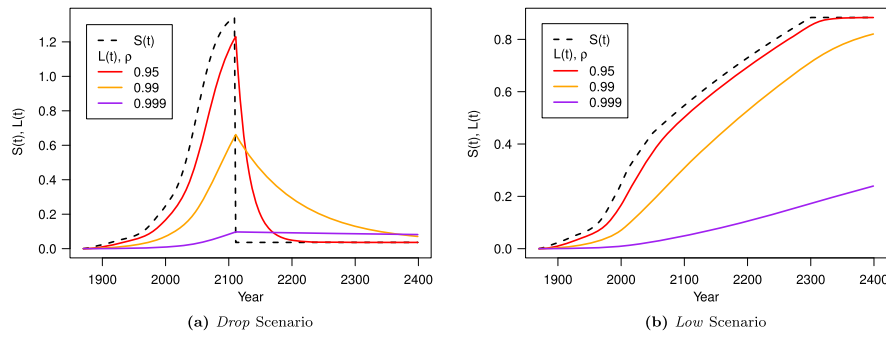
FIGURE 3 (a), (b) show the covariates short-term $S(t)$ and long-term $L(t)$ in SDM (5) for different values of the decay rate ρ under the *drop* and *low* scenarios. The black dashed line indicates the short-term $S(t)$. Red, orange, purple solid lines represent long-term $L(t)$ for different values of ρ. When ρ is smaller than 0.95, the collinearity between the short-term and long-term covariates is very strong, especially for the *low* scenario

TABLE 1 Different models for comparison

Model	Description
Model V	No spatial structure in coefficients with varying rate parameter ρ.
Model F	No spatial structure in coefficients with fixed rate parameter ρ.
Model SV	Spatial structure in coefficients with varying rate parameter ρ.
Model SF	Spatial structure in coefficients with fixed rate parameter ρ.

structure in the coefficients and fixed or varying ρ in Section 6. We denote these models as in Table 1 (the mathematical definitions of these models are in the supplementary material). Note that Model SF is just another name for SDM (5).

5 | PARAMETER ESTIMATION

To assess the emulator, we need to estimate the parameters in SDM (5) first. Because each model run (except for the shorter *jump* runs) has about 100,000 observations, none of which are independent, brute force calculations of the likelihood are not feasible. The nonlinear decay rate parameter ρ further complicates the estimation. Therefore, we use a two-stage maximum likelihood method and the mixed model equation (Robinson, 1991) to estimate the parameters and coefficients (see supplementary materials for computational details). The computational effort to fit one of these emulators (one scenario, five realizations) takes about 25min on one Linux PC (16-core CPU: Intel(R) Xeon(R) E5-2670 2.60GHz) using the R language. Once the emulator parameters are estimated, computing the emulated mean temperatures for a new scenario is effectively instantaneous.

5.1 | Estimated coefficients based on five realizations of single scenario

To emphasize the point that we do not need many scenarios to train the emulator, we only present training sets with a single scenario. We report results when all five realizations of a scenario are included in the training set in this subsection in order to get accurate parameter estimates, which makes it

easier to see patterns in these estimates across scenarios (see Table 2). In the next subsection, we consider emulation using fewer than five realizations.

If our emulator provided a completely accurate statistical description of annual average temperatures, we should not see any consistent differences in parameter estimates obtained using different scenarios. Indeed, the estimated parameters in all scenarios are fairly consistent (see Table 2). However, there are some notable patterns in the parameter estimates that are consistent across both NA and AF. For example, the estimated rate parameter $\hat{\rho}$ is smaller in the *drop* scenario than other scenarios in both continents, which means the emulated annual average temperature is less effected by the distant past when fitted using this scenario. Furthermore, for both continents, $\hat{\sigma}_1^2$ is smallest for the *med* and *low* scenarios, and $\hat{\sigma}_2^2$ is largest for these scenarios. That is, the scenarios with the most gradual changes in atmospheric CO_2 concentrations have the smallest spatial variability in the short-term covariate and the largest spatial variability in the long-term covariate.

There are some systematic differences in parameter values between NA and AF. The autoregressive parameter $\hat{\phi}$, while small over both continents, is distinctly larger over NA for all five scenarios. All the estimated spatial scale parameters $\hat{\alpha}_v$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, whose reciprocals are correlation lengths (Gneiting, Kleiber, & Schlather, 2012), are larger in AF than NA. This implies the short- and long-term effects of CO_2 and the error term have stronger correlation in AF. AF may have stronger correlation in spatial parameters because the temperature in AF has less spatial variation than the temperature in NA. NA spans from lower-mid latitudes to the lower Arctic and has huge temperature gradients. AF is in the tropics and subtropics and has a much smaller temperature gradient. The estimated short-term, long-term, and total mean CO_2 effects

TABLE 2 Estimates and standard errors of parameters in the spatial-dependent model (5) for five scenarios and two continents

Parameters	Estimation (standard error, confidence interval)				
	Hi	*Med*	*Low*	*Drop*	*Jump*
North America					
ϕ	0.0778 (0.0014)	0.0774 (0.0014)	0.0815 (0.0014)	0.0770 (0.0014)	0.0730 (0.0018)
σ_v^2	0.1666 (0.0003)	0.1677 (0.0003)	0.1686 (0.0003)	0.1708 (0.0003)	0.1684 (0.0004)
α_v (6.21×10^{-3}/km)	0.1186 (0.0024)	0.1191 (0.0024)	0.1152 (0.0023)	0.1148 (0.0023)	0.1158 (0.0030)
σ_1^2	1.3269 (0.1068)	1.0002 (0.0908)	0.8205 (0.0945)	1.2960 (0.0868)	1.3149 (0.0894)
α_1 (6.21×10^{-3}/km)	0.7461 (0.1094)	0.4907 (0.0964)	0.5013 (0.0960)	0.7609 (0.1031)	0.6773 (0.0992)
σ_2^2	0.8241 (0.0991)	0.9907 (0.1105)	1.0576 (0.1230)	0.5493 (0.0459)	0.4494 (0.0496)
α_2 (6.21×10^{-3}/km)	0.9069 (0.1137)	0.6130 (0.1113)	0.8912 (0.1229)	0.8352 (0.0934)	0.8628 (0.1107)
ρ_0	4.2444 (0.0956)	4.1671 (0.0879)	4.3367 (0.1147)	3.7331 (0.0502)	4.2149 (0.1644)
ρ ($e^{\rho_0}/(1 + e^{\rho_0})$)	0.9859	0.9847	0.9871	0.9766	0.9854
CI for ρ	(0.9830, 0.9882)	(0.9819, 0.9871)	(0.9839, 0.9897)	(0.9743, 0.9788)	(0.9800, 0.9894)
Africa					
ϕ	0.0521 (0.0014)	0.0545 (0.0014)	0.0570 (0.0014)	0.0558 (0.0014)	0.0472 (0.0018)
σ_v^2	0.1565 (0.0003)	0.1553 (0.0003)	0.1555 (0.0003)	0.1509 (0.0003)	0.1549 (0.0004)
α_v (6.21×10^{-3}/km)	0.7454 (0.0069)	0.7339 (0.0068)	0.7350 (0.0068)	0.7161 (0.0068)	0.7303 (0.0087)
σ_1^2	1.0589 (0.1028)	0.8568 (0.0807)	0.4569 (0.1218)	1.4269 (0.1043)	1.1852 (0.1075)
α_1 (6.21×10^{-3}/km)	1.9043 (0.1580)	1.6540 (0.1441)	2.4656 (0.6685)	2.1270 (0.1445)	2.0523 (0.1603)
σ_2^2	1.3059 (0.1259)	1.7270 (0.1615)	1.8219 (0.1635)	0.6578 (0.0715)	0.9682 (0.1079)
α_2 (6.21×10^{-3}/km)	2.7700 (0.2154)	2.9399 (0.2183)	2.1426 (0.1931)	2.7263 (0.2178)	2.5657 (0.2192)
ρ_0	4.0039 (0.0570)	4.3931 (0.0677)	4.0289 (0.0761)	2.7660 (0.0561)	3.3110 (0.0860)
ρ ($e^{\rho_0}/(1 + e^{\rho_0})$)	0.9821	0.9878	0.9825	0.9408	0.9648
CI for ρ	(0.9800, 0.9839)	(0.9861,0.9893)	(0.9798, 0.9849)	(0.9344, 0.9466)	(0.9586, 0.9701)

Note. Confidence intervals for ρ are based on transforming asymptotic confidence interval for ρ_0. Estimates for each scenario are based on five realizations (see the supplementary material for computational details).
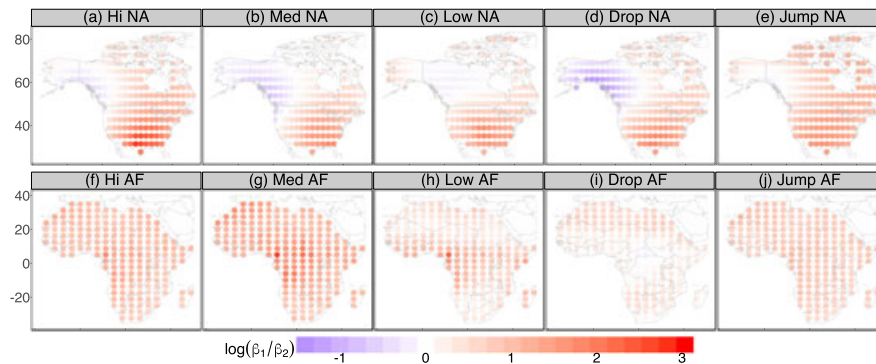

FIGURE 4 The log ratios of estimated short-term and long-term coefficients $\log(\hat{\beta}_1/\hat{\beta}_2)$ of a spatial-dependent model (5) in North America (NA) and Africa (AF) for different scenarios. Estimates are based on five realizations of a single scenario. Red indicates the estimated short-term effect is stronger than the long-term effect and blue means weaker. Upper row gives log ratios of coefficients in NA and bottom row gives log ratios of coefficients in AF

also have some significant differences between the two continents and five scenarios (see supplementary materials for detailed discussion).

The log ratios of estimated coefficients in NA and AF, $\log(\hat{\beta}_1/\hat{\beta}_2)$ (Figure 4) show how the short-term and long-term effects of atmospheric CO_2 concentration vary smoothly with spatial locations. The log ratios of effects in northwest NA are smaller than other regions in NA (Figure 4(a)–(e)), which means the CO_2 in the two most recent years has less influence in northwest NA than other locations compared with CO_2

in earlier years. However, the short-term effects are stronger than long-term effects in southern NA. In AF, the log ratios of estimated coefficients are all positive and fairly evenly distributed in all five scenarios (Figure 4(f)–(j)). The log ratios are smaller in the *low* and *drop* scenarios than other scenarios in AF, but are still almost all positive. Meanwhile, the estimated coefficients (see supplementary material) show clearly different patterns in NA and AF. Thus, it would be a mistake to pool model output across the two continents to estimate the parameters of the emulator.

5.2 | Estimated coefficients based on fewer than five realization of one scenario

This subsection demonstrates the advantages of using spatial models for estimating the regression coefficients $\beta_1(s)$ and $\beta_2(s)$. We focus on what happens when one has only a single model run in order to make the advantages of using spatial information clearest and because we then have a model-free way of assessing this advantage. Figure 5 shows boxplots of empirical standard deviations for NA across pixels of the estimated coefficients in Model V and F, which do not account for spatial structure, and Model SV and SF, which do. The main comparisons of interest are between V and SV and between F and SF, because it is not clear how to compare estimates of especially $\beta_2(s)$ when one model has spatially varying ρ and the other a single ρ for the whole continent. Whether one considers models with spatially varying or a single ρ, the spatial models generally show much smaller variability, especially for the more physically realistic scenarios *low*, *med*, and *hi*. (The same plot for AF is shown in supplementary material.)

These comparisons, while stark, are possibly misleading because they take no account of any biases in the estimates based on spatial methods. We can take advantage of our multiple realizations of each scenario to obtain effectively unbiased estimates of differences in mean squared errors for estimates of regression coefficients. For $k = 1, 2$ and scenario i, define $\beta_k^{(i)}(s)$ to be the limiting value of $\beta_k(s)$ we would obtain using a pixel-wise fit as the number of realizations tends to ∞. If the distributed lag model were correct, $\beta_k^{(i)}(s)$ would be the same for every scenario i, but even when the model is not correct, we can still think of $\beta_k^{(i)}(s)$, $k = 1, 2$ as the "best-fitting" parameter values for that scenario. Let $\hat{\beta}_{j,m,k}^{(i)}(s)$ denote the estimate

of $\beta_k^{(i)}(s)$ using scenario i, run j (one realization), and model m at location s. Let $\hat{\beta}_{-j,m,k}^{(i)}(s)$ denote the estimate using scenario i, all four runs except j, and model m at location s. For each scenario i and region S (here NA or AF), define, for $k = 1, 2$,

$$D_k^{(i)}(m, m') = \frac{1}{5N_S} \sum_{s \in S} \sum_{j=1}^{5} \left(\hat{\beta}_{j,m,k}^{(k)}(s) - \hat{\beta}_{-j,m',k}^{(i)}(s) \right)^2, \quad (7)$$

as a criterion of the difference between using one run of model m and four runs of m', where N_S is the number of pixels in region S. If $\hat{\beta}_{-1,m',k}^{(i)}(s)$ is unbiased for $\beta_k^{(i)}$, then under our assumption that the different realizations under a single scenario are iid realizations of a stochastic process,

$$
\begin{aligned}
E\left\{ D_k^{(i)}(m, m') \right\} &= \frac{1}{N_S} \sum_{s \in S} E\left\{ \left(\hat{\beta}_{1,m,k}^{(i)}(s) - \beta_k^{(i)}(s) \right) \right. \\
&\quad \left. - \left(\hat{\beta}_{-1,m',k}^{(i)}(s) - \beta_k^{(i)}(s) \right) \right\}^2 \\
&= \frac{1}{N_S} \sum_{s \in S} E\left(\hat{\beta}_{1,m,k}^{(i)}(s) - \beta_k^{(i)}(s) \right)^2 \\
&\quad + \frac{1}{N_S} \sum_{s \in S} E\left(\hat{\beta}_{-1,m',k}^{(i)}(s) - \beta_k^{(i)}(s) \right)^2,
\end{aligned}
$$

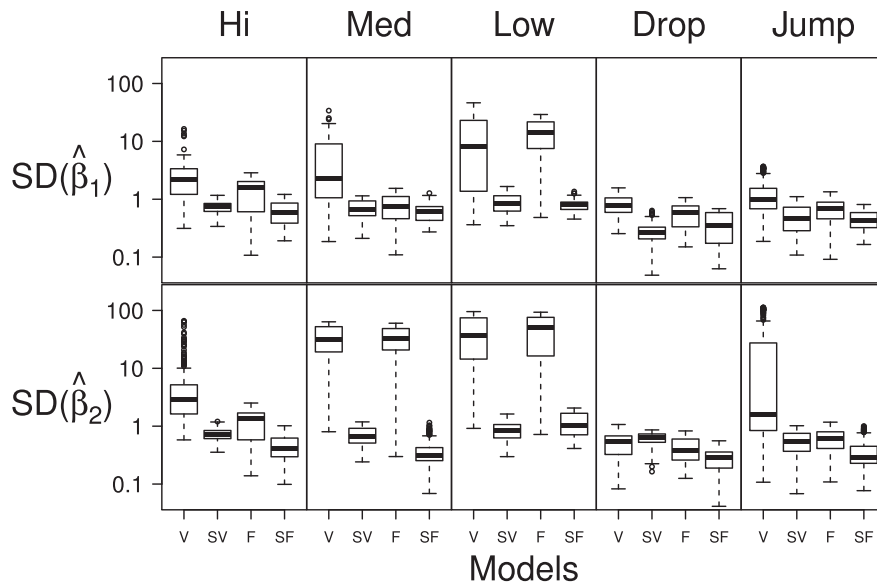because the cross-term cancels because of the independence across runs and the unbiasedness of $\hat{\beta}_{-1,m',k}^{(i)}(s)$. It follows that

$$
\begin{aligned}
E\left\{ D_k^{(i)}(m, m') - D_k^{(i)}(m', m') \right\} \\
= \frac{1}{N_S} \sum_{s \in S} E\left(\hat{\beta}_{1,m,k}^{(i)}(s) - \beta_k^{(i)}(s) \right)^2 \\
- \frac{1}{N_S} \sum_{s \in S} E\left(\hat{\beta}_{1,m',k}^{(i)}(s) - \beta_k^{(i)}(s) \right)^2.
\end{aligned}
\quad (8)
$$

FIGURE 5 Boxplots of pixel-wise sample standard deviations of estimates of β_1 and β_2 in North America. The estimates are based on using a single realization of the indicated scenario, and the standard deviations are sample standard deviations over the five realizations of that scenario. Results shown for four statistical models and all five scenarios. Both β_1 and β_2 are more variable in models without spatial structure (V and F) than models taking account of spatial structure (SV and SF), sometimes dramatically so

TABLE 3 For North America, values of $D_k^{(i)}(m, m')$, defined in (7)

		$D_1^{(i)}(m, m')$					$D_2^{(i)}(m, m')$				
m (1 runs)	*m′* (4 runs)	*Hi*	*Med*	*Low*	*Drop*	*Jump*	*Hi*	*Med*	*Low*	*Drop*	*Jump*
Model V	Model V	13.76	79.06	446.31	1.08	2.95	257.80	2964.05	4684.23	0.49	1816.05
Model SV	Model V	3.02	5.67	91.35	0.83	0.70	27.09	1033.78	2742.87	0.63	30.40
Model F	Model F	2.31	1.03	301.85	0.52	0.52	1.73	1391.36	5684.90	0.31	0.48
Model SF	Model F	0.34	1.03	1.44	0.58	1.01	0.27	0.63	2111.16	0.16	0.46

Note. As argued in the Appendix, $D_k^{(i)}(m, m') - D_k^{(i)}(m', m')$ should be a nearly unbiased estimator of the average (across pixels) mean square error of $\beta_k^{(i)}(s)$. We see that for the three scenarios without sudden changes in CO_2 levels, the spatial methods SV and SF are far superior to their nonspatial counterparts, V and F.

Thus, $D_k^{(i)}(m, m') - D_k^{(i)}(m', m')$ is an unbiased estimate of the difference in mean square errors for the regression coefficients between methods m and m'. In the Appendix, we argue that if model m' is a model with no spatial component, then $\hat{\beta}_{-1, m', k}^{(i)}(s)$ should be very nearly unbiased for $\beta_k^{(i)}$ under the iid assumption for runs under a single scenario.

Table 3 shows that, if model m is SV or SF and model m' is model V or F, respectively, the values of $D_k^{(i)}(m, m') - D_k^{(i)}(m', m')$ are always non-negative, and sometimes hugely so, for the three scenarios *hi*, *med*, and *low* without sudden jumps. For the *drop* and *jump* scenarios, the results are mixed. Scenarios with sudden jumps in CO_2 are physically unrealistic and are rarely available for most GCMs, so the main message here is that if there is only a single model run available, there can be a large advantage in using the spatial methods, even if one assumes ρ is constant across a continent.

6 | EMULATION PERFORMANCE

Our main objective in this work is to emulate annual mean temperature accurately for a wide range of CO_2 trajectories, which requires distinguishing between short- and long-term effects. Emulating the *drop* scenario provides the most severe test for making this distinction, because the CO_2 concentration of the *drop* scenario has the largest range, from 289 to about 1100 ppm. Training the emulator based on the *low* scenario makes the test particularly difficult, because atmospheric CO_2 concentrations change the most slowly in this scenario. The large variances of estimated coefficients in Figure 5 and Table 3 also give a glimpse of the difficulty of generating an emulator based on the *low* scenario. This section mostly shows the results for emulators of the *drop* and the *jump* scenario trained with five realizations of the *low* or the *drop* scenario, although we also considered other combinations of training and test scenarios.

6.1 | Performance of SDM

Figure 6 shows the emulated annual mean temperatures for the *drop* scenario at two locations (see "×" in Figure 2 for the locations) based on emulators trained on either *drop* or *low* scenario model runs using SDM (5). The emulated temperature of the *drop* scenario based on the *drop* scenario
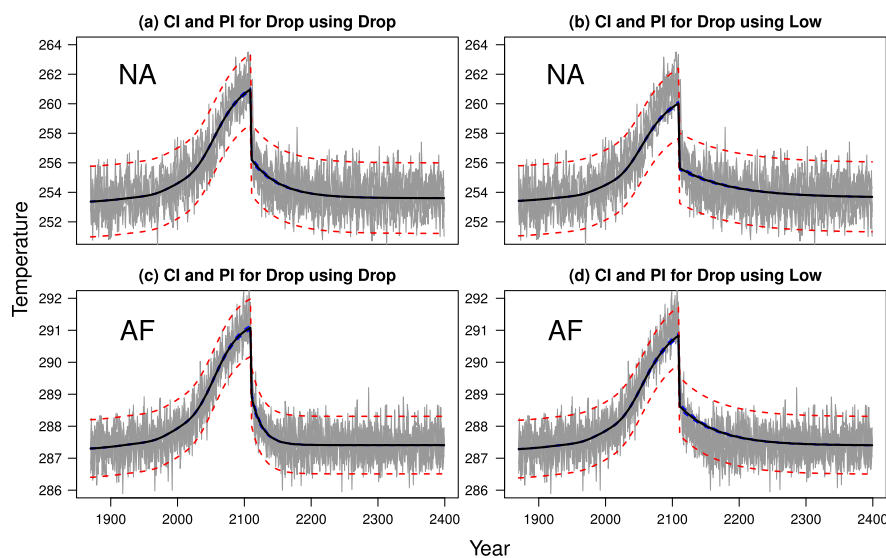

FIGURE 6 Emulated temperatures of the *drop* scenario using the spatial-dependent model (5) based on five realizations of the *drop* or *low* scenarios in North America (NA) and Africa (AF). Gray lines present the five CCSM3 realizations for the chosen scenario, black lines are emulated annual mean temperature, blue dashed lines are (pointwise) 95% confidence intervals for the emulator (difficult to see because intervals are so narrow), and red dashed lines denote 95% prediction bands for the emulator (see supplementary material for computational details). Plots (a) and (b) are for a pixel in NA, plots (c) and (d) are for a pixel in AF (the pixels marked by "×" in Figure 2). Plots (a) and (c) show results for an emulator trained on five realizations of the *drop* scenario and (b) and (d) show similar results for an emulator trained on the *low* scenario
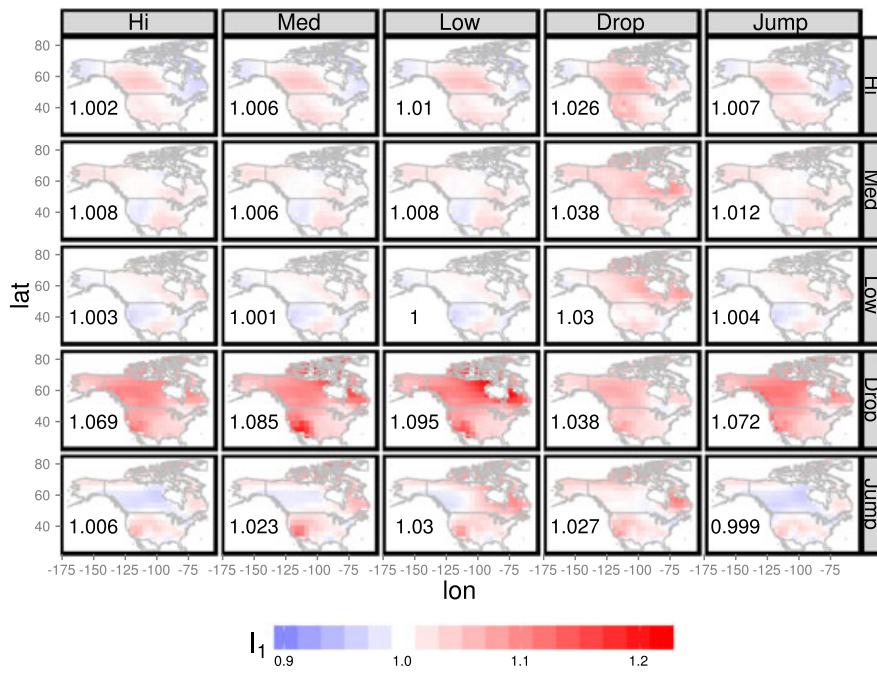
FIGURE 7 Values of the "emulation optimality" index I_1 under the spatial-dependent model (5) in North America. Each plot uses the five realizations of the column-labeled scenario to emulate the row-labeled scenario. Red points indicate that I_1 values are larger than 1 and deeper red means worse emulation performance. Values of I_1 near 1 indicate an excellent fit and I_1 values less than 1 are because of stochastic variation. All I_1's are less than 1.22, which means the emulator performs well in every scenario picked. Numbers in lower left corner of each plot give the mean of I_1 in each case. All indices have been computed between the year 2010 and the farthest time point available from the model run
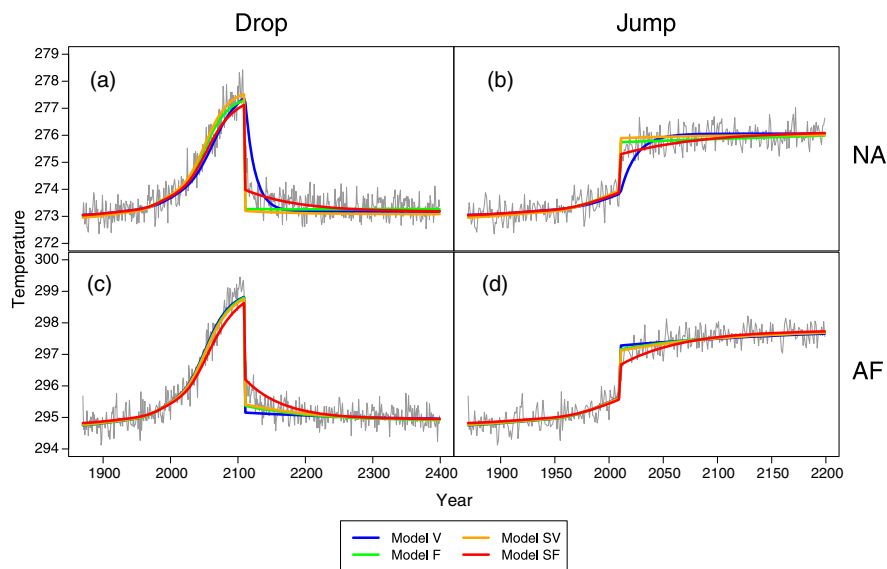

FIGURE 8 Emulator at individual pixels for the *drop* and the *jump* scenarios generated by four models based on the *low* scenario in North America (NA) ((a),(b)) and Africa (AF) ((c),(d)) (see locations marked "+" in Figure 2). The gray lines are average temperatures of five realizations of the *drop* and *jump* scenarios. Emulator for *drop* scenario generated by Model SF (the red solid line in (a),(c)) provides a smooth curve that captures the trend of sudden drop of annual average temperature well and a slight misfit before the sudden drop. Other models (the blue, orange, and green solid lines in (a),(c)) do not capture the trend after the sudden drop. All models in (b),(d) show a good emulation before the sudden jump but only Model SF (red solid line) captures the trend of mean temperature following the jump

(Figure 6(a) and 6(c)) shows almost perfect fit. In this situation, the training and test scenarios are the same, which implies the SDM (5) fairly accurately models the annual mean temperature. The emulated temperature of the *drop* scenario based on the *low* scenario shows some misfit before and

after the sudden drop of CO_2 concentration in both continents (Figure 6(b) and 6(d)), although the misfit is modest after the drop in NA. A misfit after the sudden drop for an emulator trained on the *low* scenario is not surprising. The misfit before the drop is not unexpected either, because the

atmospheric CO_2 concentration before the drop in the *drop* scenario is much higher than the highest level in the *low* scenario, so it requires an extrapolation from the training set. Overall, the emulator performs quite well even in this extreme *drop* case. In other situations not shown here, the emulator of mean temperature under SDM (5) also performs well.

To assess the fit quantitatively, we exploit our multiple realizations of each scenario to compute a pure lack-of-fit index whose validity does not depend on knowing the form of the mean function (Montgomery, 2013; Castruccio et al., 2014). This lack-of-fit index $I_1(s)$ (or just I_1) measures emulation performance at location s relative to the optimal emulation possible given initial condition uncertainty. Let $\hat{T}(s, t)$ denote the emulated mean temperature at location s and year t ($t = 1$ corresponding to the year 2010 where the scenarios diverge). We compare the sum of squared deviations of the actual realizations from the emulated temperature $\hat{T}(s, t)$ and the average across realizations $\bar{T}(s, t)$,

$$I_1(s) = \frac{\sum_{r=1}^{R} \sum_{t=1}^{m} [T_r(s, t) - \hat{T}(s, t)]^2}{\frac{R}{R-1} \sum_{r=1}^{R} \sum_{t=1}^{m} [T_r(s, t) - \bar{T}(s, t)]^2}.$$

The numerator measures the performance of the emulator, and the denominator gives an unbiased estimator of mean squared error of a "perfect" emulator (i.e., if we had an infinite number of realizations under the scenario of interest). Under the assumption that the different realizations of a given scenario are independent and identically distributed (and that realizations of different scenarios are independent), the mean of the numerator of I_1 must be at least as large as the mean of the denominator, except possibly when the training and test scenarios are the same. Thus, values of I_1 near 1 indicate an excellent fit.

The values of I_1 in NA using SDM (5) when the emulator is estimated by each of the five scenarios are shown in Figure 7, in which the column labels indicate the training scenario and the row labels indicate the test scenario. For all pairs of training and test scenarios and all pixels in NA, I_1 is at most 1.22, and if we exclude the *drop* scenario as a test scenario (the fourth row of Figure 7), the largest I_1 value is 1.13. Within each row, as we might expect, the mean I_1 value is smallest when the training scenario and test scenario are the same. The emulator performs quite well throughout NA even when the scenarios differ. SDM (5) also performs well in AF in all cases (see the values of I_1 in AF in supplementary material).
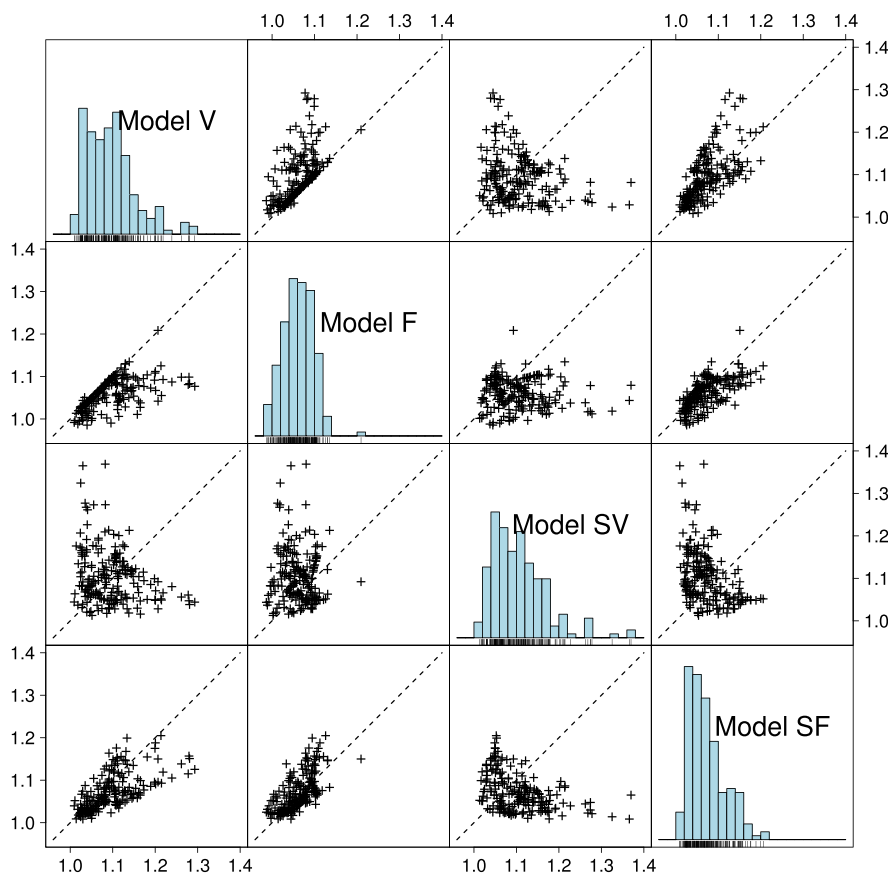

FIGURE 9 I_1 values for the *drop* scenario training on the *low* scenario for different statistical models in NA. The diagonal figures are histograms of I_1 values; Model SF has the smallest values overall. Dashed lines are the 1–1 line. Points under these dashed lines indicate pixels for which the emulator under the row-labeled model performs better than the column-labeled model

6.2 | Performance of different emulators

We now consider results for all four emulator models in Table 1 in Section 4. Figure 8 shows the annual mean temperature trajectory for the *drop* and *jump* scenarios at two locations (see "+" in Figure 2 for locations) for these four emulators, compared with the average temperature of five realizations in each scenario. Model SF (SDM (5)) captures the trend of mean temperature following the sudden change of CO_2 for both the *drop* and *jump* scenarios, but does slightly worse before the sudden change in the *drop* scenario. The other three models (Model V, Model SV, and Model F) cannot track well the sudden change of the annual mean temperature trend. Figure 9 gives scatterplots of I_1 values for the *drop* scenario in NA based on the *low* scenario for all pairs of the four different models (see supplementary material for same figure in AF) and shows the emulator generated by Model SF performs better than the other three emulators as measured by I_1 for almost all grid points.

To explore further the ability of an emulator to capture the trend following a sudden change of CO_2, we consider emulation for the *drop* scenario for illustration and calculate I_1 for the 30 years following the sudden drop (from the year 2111 to 2140, the sudden drop is at year 2110), and denote the index as $I_1^{(Drop30)}$. The index $I_1^{(R)}$ is I_1 calculated over the remaining years (from the year 2010 to 2110 and 2141 to 2399). The choice of 30 years after the drop is somewhat arbitrary, but the results are not qualitatively changed by moderate changes in the length of this period (results not shown here).

Table 4 shows emulator Model SF has the smallest average I_1 value among the four emulators for both NA and AF. Results for $I_1^{(Drop30)}$ and $I_1^{(R)}$ show that essentially all of the advantage of Model SF over the other emulator models occurs in the 30-year post-drop period. Indeed, all four emulators

perform well and with little difference between them when one excludes this period (see supplementary material for plots of I_1^{Drop30}, $I_1^{(R)}$ for NA and AF).

When we use all five realizations of a given scenario in the training sets, the emulators that do not use spatial information arguably work well enough for many purposes. If we use fewer realizations, the differences between the emulators that use spatial information and those that do not become much more dramatic (see the supplementary materials for more details).

7 | DISCUSSION

We have shown the effectiveness of including spatial dependence in some of the regression parameters of an emulator of annual mean temperature at the pixel level of a GCM. We fit parameters separately for the two continents NA and AF, and the clear differences in hyperparameters in Table 2 show that one would not want to assume these parameters were equal over all land regions. We would expect some parameters to differ even more over water, especially ρ, which controls the decay of the effect of past CO_2 values on the present climate. Thus, if one wanted to develop a unified global emulator, one would need to take account of differences between continents and between land and ocean. It is not clear to us that there would be much benefit in developing such an emulator for annual mean temperatures. If, as would often be the case, one wants also to emulate variations about the mean, then a global model would be critical to capture spatial dependencies at all scales, in which case, an approach such as the one in Castruccio et al. (2014) would be relevant.

The very simple form for the distributed lag coefficients given in (5) allows one to estimate the parameters of the emulator well with as few as one GCM run in some circumstances. However, this very simple form cannot be expected to be exactly correct and can thus be a source of the bias when trying, for example, to emulate the *drop* scenario when training on the *low* scenario. Indeed, Table 2 shows some evidence for this bias in the consistent patterns across training scenarios for NA and AF. Thus, when one has a sufficiently large and diverse training set, one may want to use a more complex model than in SDM (5) by, for example, allowing nonlinear terms in $\log[CO_{2r}]$. In this case, one might also want to consider allowing the rate parameter ρ to vary within a continent by, for example, being larger for coastal pixels. It then might also make sense to include additional terms in the model, for example, to include a "medium-term" effect in addition to the short- and long-term effects in the model presented here. We have tried including such an effect in the situations studied here and found that the multicollinearity between the resulting covariates made parameter estimation very difficult.

Another approach to reducing bias due to model misspecification would be to allow the training scenarios used for an emulator to depend on the scenario being emulated.

TABLE 4 Sample means and standard deviations across pixels of I_1, $I_1^{(Drop30)}$ and $I_1^{(R)}$ for the *drop* scenario training on the *low* scenario in NA and AF. These three indices are calculated in different time periods. I_1 is from year 2010 to 2399, $I_1^{(Drop30)}$ is from year 2111 to 2140 (the sudden drop is at year 2110) and $I_1^{(R)}$ is from year 2010 to 2110 and 2141 to 2399. A value close to 1 means perfect fit

		NA		AF	
Index	**Method**	**mean**	**sd**	**mean**	**sd**
I_1	Model V	1.275	0.254	1.290	0.203
	Model F	1.159	0.064	1.316	0.303
	Model SV	1.215	0.123	1.279	0.210
	Model SF	1.095	0.044	1.215	0.104
$I_1^{(Drop30)}$	Model V	3.302	2.601	3.182	1.669
	Model F	2.321	0.512	3.568	2.960
	Model SV	2.530	0.945	3.500	2.244
	Model SF	1.354	0.148	2.147	1.017
$I_1^{(R)}$	Model V	1.094	0.057	1.140	0.151
	Model F	1.061	0.035	1.142	0.152
	Model SV	1.105	0.064	1.105	0.099
	Model SF	1.073	0.043	1.142	0.079

Presumably, one should choose scenarios for the training set that are similar in some relevant sense to the scenario being emulated. However, how to define similarity of scenarios is not so clear because we might want to use only parts of one scenario within the training set. For example, one might use the *jump* scenario only up to the time of the jump to train an emulator for the *hi* scenario. This approach deserves further investigation (see the supplementary materials for more details).

Castruccio et al. (2014) also developed an emulator for annual precipitation levels, and we expect that including spatial effects would be particularly helpful here for pixel-wise emulation due to the substantial amount of interannual variation in precipitation at the pixel level. Both this work and Castruccio et al. (2014) only look at annual averages, but for many impacts, seasonal or even shorter term averages would be of greater interest. In principle, the methodologies developed here could be applied to seasonal averages, but the noise would necessarily be larger and one might want to use a model that explicitly includes a seasonal component that changes slowly over time rather than developing a separate model for each season. In addition, as one moves to shorter time scales, temporal dependencies will be stronger and more attention should be paid to modeling the space-time dependence in the error term.

Although the focus here has been on the emulation of the mean of annual average temperature at the pixel level, our model for the space-time variations about the mean could also be of value for climate model emulation. In particular, by developing a separate model for every continent (or other appropriately defined region) rather than a global model as in Castruccio and Stein (2013), we avoid the challenge of capturing nonstationarities at land/ocean boundaries, perhaps yielding simple and effective regional emulators of the annual average temperature process and not just its mean.

APPENDIX

This appendix considers situations under which our estimates of the regression coefficients $\beta_1^{(i)}(s)$ and $\beta_2^{(i)}(s)$ should be effectively unbiased. This unbiasedness is the key assumption underlying (8), which says we can then use the multiple realizations to unbiasedly estimate differences in mean square errors of these coefficients between spatial and nonspatial methods. It suffices to consider the results for a single pixel, so we drop the dependence of any quantity on location s. Suppose $Y_1, \ldots, Y_r$ represent r iid realizations of the response vector (average annual temperatures at the pixel), which have common mean θ and common covariance matrix V. If we assume $\theta = X\beta$ (note that the covariate matrix X is the same for all realizations), then the generalized least squares estimate of β is $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \bar{Y}$, where $\bar{Y} = \frac{1}{r} \sum_{j=1}^r Y_j$. When $\theta = X\beta_0$, then $E\hat{\beta} = \beta_0$. However, even when the mean model is incorrectly specified, if we define β_0 as the minimizer of $(X\beta - \theta)^T V^{-1}(X\beta - \theta)$, then we still have $E\hat{\beta} = \beta_0$.

Thus, in this idealized setting, the generalized least squares estimate for β is unbiased for β under this definition of the "true" value of β even if the mean model is misspecified.

Following standard practice in the climate modeling literature, we have been assuming that the different realizations of a GCM under the same scenario produce iid realizations of a stochastic process, so the assumption that $Y_1, \ldots, Y_r$ are iid is well grounded. However, there are two complications with the argument of the preceding paragraph. First, the covariance matrix V needs to be estimated because of the presence of the autoregressive coefficient and, second, our mean model includes the nonlinear parameter ρ. As Table 2 shows, the autoregressive coefficient is quite small (at least when estimated by model SDM (5)) and, hence, should have little impact on the estimates of the mean parameters. The need to estimate ρ should have little impact on the expected value of estimates of the linear mean parameters as long as ρ is well estimated. At least when ρ is assumed constant throughout the continent, ρ should be very stably estimated, which, Table 2 shows, is the case under model SDM (5). Thus, we believe that the comparisons between spatial and nonspatial estimation methods in Table 3 are fair.

REFERENCES

Assunċao, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, *14*(5), 453–473.

Berliner, L. M., & Kim, Y. (2008). Bayesian design and analysis for superensemble-based climate Forecasting. *Journal of Climate*, *21*(9), 1891–1910.

Buser, C. M., Künsch, H. R., & Weber, A. (2010). Biases and uncertainty in climate projections. *Scandinavian Journal of Statistics*, *37*(2), 179–199.

Caldeira, K., & Myhrvold, N. P. (2012). Temperature change vs. cumulative radiative forcing as metrics for evaluating climate consequences of energy system choices. *Proceedings of the National Academy of Sciences*, *109*(27), E1813–E1813.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., & Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed GCM runs. *Journal of Climate*, *27*(5), 1829–1844.

Castruccio, S., & Stein, M. L. (2013). Global space–time models for climate ensembles. *The Annals of Applied Statistics*, *7*(3), 1593–1611.

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, GB, Bretherton, C. S., Carton, J. A., & Henderson, T. B. (2006). The community climate system Model Version 3 (CCSM3). *Journal of Climate*, *19*(11), 2122–2143.

Fan, J., & Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, *1*(1), 179–195.

Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D. W., & Myhre, G. (2007). Changes in atmospheric constituents and in radiative forcing. Chapter 2, *Climate Change 2007. The Physical Science Basis* (pp. 129–234). UK, and New York: Cambridge University Press, Cambridge.

Fuentes, M. (2001). A high frequency Kriging approach for non-stationary environmental processes. *Environmetrics*, *12*(5), 469–483.

Furrer, R., Sain, S. R., Nychka, D., & Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics*, *14*(3), 249–266.

Gamerman, D., Moreira, A. R., & Rue, H. (2003). Space-varying regression models: Specifications and simulation. *Computational Statistics & Data Analysis*, *42*(3), 513–533.

Gelfand, A. E., Kim, H. J., Sirmans, C., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, *98*(462), 387–396.

Gneiting, T., Kleiber, W., & Schlather, M. (2012). Matern Cross-Covariance Functions for Multivariate Random Fields. *Journal of the American Statistical Association*, *105*(491), 1167–1177.

Greasby, T. A., & Sain, S. R. (2011). Multivariate spatial analysis of climate change projections. *Journal of agricultural, biological, and environmental statistics*, *16*(4), 571–585.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(4), 757–796.

Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T. C. (1980). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

Kaufman, C. G., & Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, *5*(1), 123–149.

Koyck, L. M. (1954). *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland Publishing Company.

Manabe, S., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *Journal of Atmospheric Sciences*, *24*(3), 241–259.

Montgomery, D. C. (2013). *Design and Analysis of Experiments* (8th ed.). New York: Wiley.

Nychka, D., Wikle, C., & Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, *2*(4), 315–331.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32.

Sain, S. R., Nychka, D., & Mearns, L. (2011). Functional ANOVA and regional climate experiments: A statistical analysis of dynamic downscaling. *Environmetrics*, *22*(6), 700–711.

Smith, R. L., Tebaldi, C., Nychka, D., & Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, *104*(485), 97–116.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Stocker, T., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., & Midgley, P. M. (2014). *Climate Change 2013: The Physical Science Basis*. Cambridge, UK, and New York: Cambridge University Press.

Tebaldi, C., & Sansó, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 83–106.

Tebaldi, C., Smith, R. L., Nychka, D., & Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, *18*(10), 1524–1540.

Yeager, S. G., Shields, C. A., Large, W. G., & Hack, J. J. (2006). The low-resolution CCSM3. *Journal of Climate*, *19*(11), 2545–2566.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.