Stochastic Environmental Research and Risk Assessment manuscript No.
(will be inserted by the editor)

A Bayesian hierarchical approach for spatial analysis of climate model bias in multi-model ensembles

Maeregu Woldeyes Arisido · Carlo Gaetan · Davide Zanchettin · Angelo Rubino

Received: date / Accepted: date

Abstract Coupled atmosphere-ocean general circulation models are key tools to investigate climate dynamics and the climatic response to external forcings, to predict climate evolution and to generate future climate projections. Current general circulation models are, however, undisputedly affected by substantial systematic errors in their outputs compared to observations. The assessment of these so-called biases, both individually and collectively, is crucial for the models' evaluation prior to their predictive use. We present a Bayesian hierarchical model for a unified assessment of spatially referenced climate model biases in a multi-model framework. A key feature of our approach is that the model quantifies an overall common bias that is obtained by synthesizing bias across the different climate models in the ensemble, further determining the contribution of each model to the overall bias. Moreover, we determine model-specific individual bias components by characterizing them as non-stationary spatial fields. The approach is illustrated based on the case of near-surface air temperature bias in the tropical Atlantic and bordering regions from a multi-model ensemble of historical simulations from the fifth phase of the Coupled Model Intercomparison Project. The results demonstrate the improved quantification of the bias and interpretative advantages allowed by the posterior distributions derived from the proposed Bayesian hierarchical framework, whose generality favors its broader application within climate model assessment.

M.W. Arisido (✉)
E-mail: maeregu.arisido@unive.it
Tel.: +393665067068

M. W. Arisido · C. Gaetan · D. Zanchettin · A. Rubino
Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Via Torino 155, 30172 Venice, Italy.

Keywords Bayesian hierarchical method · Climate biases · Climate model uncertainty · Gaussian kernels · Posterior distribution · spatial model.

1 Introduction

Coupled atmosphere-ocean general circulation models (GCMs) use mathematical approximations of the laws of fluid dynamics, thermodynamics and chemistry to simulate the mass and energy transfers and the radiative exchanges within and across the global climate system (Flato et al. 2013). Climate simulations performed with such models provide quantitative estimates of geophysical quantities such as temperature and precipitation, which are used for both investigation of climate dynamics and to produce historical and paleoclimate simulations as well as projections of future climate, where climate changes by virtue of natural as well as anthropogenic forcings can be assessed (e.g., Tebaldi et al. 2005; Flato et al. 2013).

Despite the continuing improvement of climate models, simulations performed with the current generation of GCMs involve substantial uncertainties. The use of so-called multi-model ensembles is a common practice in contemporary climate science, as they allow to overcome the peculiarities of individual simulations, like those linked to the chosen initial conditions and applied external forcing, and the deficiencies of individual models, by combining the information into a multi-model consensus (Lambert and Boer 2001; Neuman 2003; Tebaldi et al. 2005; Sain and Furrer 2010; Kang et al. 2012). The Coupled Model Intercomparison Project phase 5 (CMIP5, Taylor et al. 2012) provides the largest collection of multi-model experiments with state-of-the-art GCMs. It demonstrated that current climate simulations are affected by large systematic errors of the mean state and variability, or biases, i.e., discrepancies between observed and simulated characteristics over extensive regions (Wang et al. 2014). These biases are largely attributed to the limited understanding of many of the interactions and feedbacks in the climate system (Jun et al. 2008), inadequate representation of well known processes in climate models and, to some extent, the unpredictability of the climate system itself (Keller 2009; Leith and Chandler 2010). One of the most severe biases shared by different models is the warm sea-surface temperature bias in the southeastern tropical Atlantic (Flato et al. 2013). Multiple causes have been identified at its origin, in different models, including local factors, such as the along-shore windstress and surface heat fluxes (e.g., Wahl et al. 2015; Milinski et al. 2016), and larger-scale or even remote phenomena, such as the propagation into the southeastern tropical Atlantic of downwelling anomalies generated at the equator (e.g. Toniazzo and Woolnough 2014).

In this paper we focus on assessing climate model biases in multi-model ensembles. It is debated how the information brought by the different models in a multi-model ensemble should be optimally combined to generate consensus: current climate models have been developed by sharing model components (Jun et al. 2008; Flato et al. 2013), and so they are not always independent from each other (e.g., Knutti 2010). The consequent weighting models based on arguments such as model independence can substantially affect the estimation of multi-model consensus and associated uncertainty (Knutti 2010; Flato et al. 2013).

We present a spatial analysis based on the Bayesian hierarchical model that provides a unified assessment of the biases within a multi-model context. Specifically, the proposed probabilistic approach allows to estimate the overall bias component, i.e., the component of the bias which is the same for all models, and the individual model biases, i.e., the components of the bias that are specific of each model, further characterizing each model's contribution to the overall bias and related uncertainty. We describe the different bias components as non-stationary spatial fields.

Our approach represents therefore a step forward compared to previous assessments of climate model biases based on Bayesian hierarchical modeling, which dealt with spatially aggregated geophysical data (e.g., Christensen et al. 2008; Buser et al. 2009) or grid-points individually (e.g., Boberg and Christensen 2012).

We illustrate the method by using observational reference data and an ensemble of six historical full-forcing climate simulations contributing to CMIP5. We focus on an application involving spatially referenced near-surface air temperature averaged over the years 1950-2005, and covering the tropical Atlantic Ocean and bordering regions.

In the following section, we describe the data and present our definition of climate model bias. Section 3 discusses the Bayesian hierarchical method tailored for a unified assessment of climate model biases in a multi-model framework, while section 4 illustrates the results. We provide a concluding discussion in section 5.

2 Data and climate model biases

The dataset comprises observational reference and climate model outputs. Although the latter are obtained from deterministic numerical models, it is a common practice to consider the model output as 'data', which may not represent the traditional statistical definition of data.

2.1 Observations and GCM output

We use monthly-mean data obtained from the NCEP reanalysis (Kalnay et al. 1996; Kistler et al. 2001) as our observational reference data. Reanalysis data are the output of a state-of-the-art analysis/forecast system with data assimilation using past data from 1948 to the present. The data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. Reanalysis data are therefore not direct observations, yet they facilitate the purposes of this study by providing gridded records of absolute temperatures. This is an advantage compared to other observational products that provide anomalies as main gridded output, such as the temperature series produced by the Climatic Research Unit of the University of East Anglia (Brohan et al. 2006). Our climate model outputs are based on monthly-mean data from an ensemble of six historical full-forcing climate simulations contributing to CMIP5. An overview of the models' characteristics is provided in Table 1, see Zanchettin et al. (2015) for more details on the models and the simulations.

The analysis is for the period 1950-2005 CE for which we derive climatologies of annual-mean values starting from the monthly-mean time series of both observations and simulations over the tropical Atlantic region. Geographically the tropical Atlantic is defined here as the region covering the latitude range 35°S to 15°N and the longitude range 40°W to 20°E.

2.2 Climate model biases

Climate model bias is determined by comparing output data against observations. We let $Y(s)$ represent the temperature observations and $X_j(s)$ denote the temperature simulated by the climate model j at the spatial location $s \in D$ for the domain $D \subset \mathbb{R}^2$. One crucial aspect of the complexity inherent in the assessment of climate model biases is the spatial misalignment between observations and model output. The fact that model output and observations are provided on different grids may hinder statistical analysis of the bias at the grid-point level. To tackle this issue, we interpolated the output

Table 1 The six general circulation models (GCMs) utilized for this study. BCC stands for BCC-CSM1-1 and GISS: GISS-E2-R, IPSL: IPSL-CM5A-LR, MPI: MPI-ESM-P and MIROC: MIROC-ESM. The ensemble has also been used by Zanchettin et al. (2015).

GCMs	Atmospheric resolution	Research center
CCSM4	1.25°N ×0.94°E	National Center for Atmospheric Research (USA)
BCC	2.81°N ×2.75°E	Beijing Climate Center (China)
IPSL	3.75°N ×1.90°E	Institute Pierre Simon Laplace (France)
MPI	1.88°N ×1.90°E	Max Planck Institute for Meteorology (Germany)
GISS	2.50°N ×2.00°E	NASA/Goddard Institute for Space Studies (USA)
MIROC	2.81°N ×2.75°E	Center for Climate System Research (Japan)

data on the regular observational grid to ensure that $Y(s)$ and $X_j(s)$ are aligned on the same grid (see, e.g., Jun et al. 2008; Banerjee et al. 2014). Empirical climate model biases are then calculated as

$$B_j(s) = Y(s) - X_j(s), \quad j = 1, \ldots, 6 \tag{1}$$

where $B_j(s)$ denotes the bias of climate model j relative to the observation at spatial location s. For n sites in D, we observe the biases, namely $\{B_j(s_i), \ldots, B_j(s_n)\}$. Figure 1 summarizes the bias fields of near-surface air temperature in the tropical Atlantic region from the six climate models. Clearly, the different GCMs produce similar spatial features of the bias. For instance, all models produce a warm bias over the Angola-Benguela front region. We also note distinct features for each model bias. For instance, the above mentioned warm bias in the Angola-Benguela front region has different severity in the different models, with peak values ranging from 3 kelvin in CCSM4 to 5 kelvin on MIROC, and can extend either to the north, like in CCSM4, GISS and IPSL, or to the south, like in MIROC and BCC. Also, the south Atlantic mid-latitudes can feature either an extensive negative bias, like in CCSM4, BCC and MIROC, or an extensive positive bias, like in GISS and IPSL. The remainder of this paper devotes to quantifying the shared bias and the individual components and associated uncertainties across the different climate models.

3 Bayesian hierarchical approach for climate model biases

Our aim is to obtain a statistical representation of climate model biases in a multi-model ensemble that separates an overall common bias from the individual components. We present a Bayesian hierarchical model formulated based on three levels: *data*, *process*, and *parameters* (Berliner 2003). The data model captures the information given in the form of empirically measured biases, conditional on a hidden spatial bias process. The process level models the spatial structure and links the hidden spatial process to a set of parameters. In the parameter model, prior distributions are specified for the parameters. The three levels are specified in terms of probability distributions in a hierarchical structure

$$\begin{cases} [\text{data}|\text{process}]: & \text{Data model} \\ [\text{process}|\text{parameters}]: & \text{Process model} \\ [\text{parameters}]: & \text{Parameter model,} \end{cases}$$

where $[A|B]$ denotes a conditional probability distribution of A given B and $[A]$ denotes the probability density of A.

3.1 Data model

We assume that the empirical bias $B_j(s)$ can be decomposed into two components: a spatial component $M_j(s)$ and a noise component $\varepsilon_j(s)$, namely

$$B_j(s) = M_j(s) + \varepsilon_j(s), \quad j = 1 \ldots, 6 \tag{2}$$

Here $\{\varepsilon_j(s)\}$ is a Gaussian white noise with zero mean and variance $\sigma^2_{\varepsilon,j}$, independent from $\{\varepsilon_k(s)\}$, for $k \neq j$. Additionally, the noise component $\{\varepsilon_j(s)\}$ is assumed independent from $\{M_j(s)\}$. Thus, conditionally on the spatial process $\{M_j(s)\}$, the observed bias $B_j(s)$ has a Gaussian distribution with mean $M_j(s)$, and variance $\sigma^2_{\varepsilon,j}$ that represents the data model level.

3.2 Process model

GCM ensemble members feature biases which may originate from different factors including parameterizations, discretization to solve the numerical equations, resolution level and imposed boundary conditions. The spatial process $\{M(s)\}$, with $M(s) = (M_1(s), \ldots, M_6(s))'$ is multivariate, and can be modeled in different ways (Gelfand et al. 2010). Here we assume that the climate bias can be additively decomposed into two components

$$M_j(s) = \mu(s) + \eta_j(s), \quad j = 1, \ldots, 6, \tag{3}$$

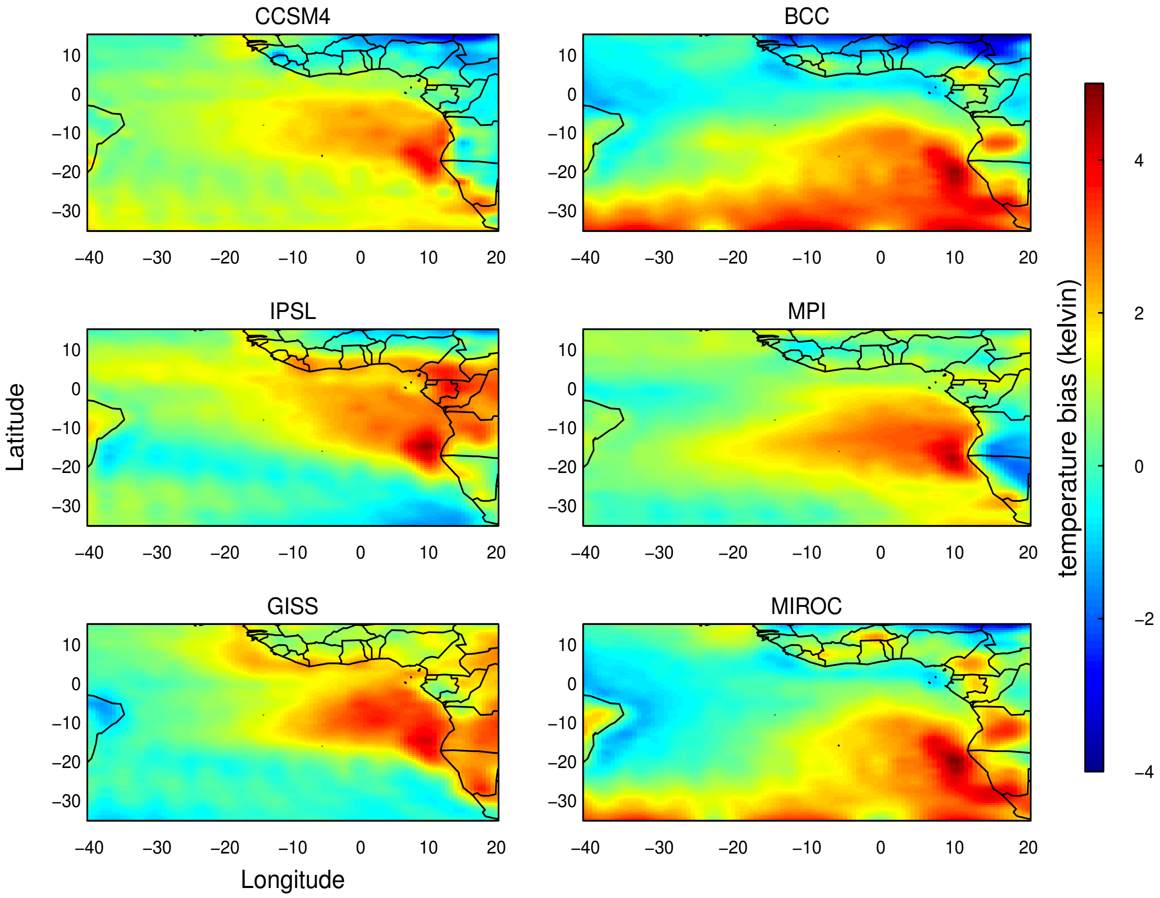

Fig. 1 Empirical bias for simulated near-surface air temperature for each of the six GCMs in the ensemble relative to the observed temperature over the tropical Atlantic region.

where $\mu(s)$ is the overall common bias capturing shared large-scale features for all climate models, while $\eta_j(s)$ describes the jth model-specific features. Based on this interpretation, specification (3) can be viewed as a version of a random effect model (see Furrer et al. 2007; Kaufman and Sain 2010; Kang et al. 2012, for examples of applications in climatology). To model the two spatial components $\mu(s)$ and $\eta_j(s)$, we adopt an approach based on kernel basis functions (see, e.g., Higdon 1998) and we suppose that

$$\mu(s) = \mathbf{w}(s)'\alpha_k \qquad \eta_j(s) = \mathbf{w}^*(s)'\nu_j, \tag{4}$$

where $\mathbf{w}(s) = \{w_1(s), \ldots, w_p(s)\}'$, $\mathbf{w}^*(s) = \{w_1^*(s), \ldots, w_{p^*}^*(s)\}'$ are vectors of Gaussian kernels and $\alpha = (\alpha_1, \ldots, \alpha_p)'$ and $\nu_j = \{\nu_{j,1}, \ldots, \nu_{j,p^*}\}'$ are vectors of parameters. The shape and number of kernels associated to $\mathbf{w}(s)$ and $\mathbf{w}^*(s)$ are different. Since the individual components $\{\eta_j(s) : j = 1, \ldots, 6\}$ aim to capture local-scale features, a larger number p^* of kernels with a narrower spatial bandwidth are expected to be required with respect to that necessary to describe the overall common bias $\mu(s)$, i.e., $p < p^*$. However, the number of kernels p and p^* will be much less than the number of data points n. The choice of the kernels and their shapes is further discussed in section 3.4. The parameters α and $\{\nu_j,\ j = 1, \ldots, 6\}$ are considered as random. More precisely α is multivariate Gaussian $\alpha \sim \text{Gau}(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is the $p \times p$ covariance matrix, and $\{\nu_j,\ j = 1, \ldots, 6\}$ are mutually independent zero mean Gaussian processes, $\nu_j \sim \text{Gau}(\mathbf{0}, \tau_j^2 I_{p^*})$, where $\tau_j^2 I_{p^*}$ is the covariance matrix and I_{p^*} is the $p^* \times p^*$ identity matrix. With this setup $\eta_j(s)$ is a Gaussian random variable with zero mean and variance $\text{var}(\eta_j(s)) = \tau_j^2 \mathbf{w}^*(s)'\mathbf{w}^*(s)$. Thus the parameters τ_j^2 measure how each climate model bias varies about the overall common bias. More specifically, different values of τ_j^2 across the various models indicate different levels of departure from the common bias. Alternatively, similar values of τ_j^2 for different models indicate that they vary similarly about the overall common bias, suggesting that the contribution of each climate model in estimating the overall common bias is similar. Under these hypotheses we have constructed a non-stationary spatial process for $M_j(s)$ with covariance function

$$\begin{aligned}\mathrm{cov}(M_j(s),M_j(s')) = \sum_{m=1}^{p}\sum_{k=1}^{p} G_{mk} w_m(s) w_k(s') + \\ \tau_j^2 \sum_{m=1}^{p^*}\sum_{k=1}^{p^*} w_m^*(s) w_k^*(s')\end{aligned} \tag{5}$$

where $G_{mk} = \mathrm{cov}(\alpha_m, \alpha_k)$ is the m,k entry of the covariance matrix $\mathbf{G}$, and cross-covariance function

$$\mathrm{cov}(M_j(s),M_l(s')) = \sum_{m=1}^{p}\sum_{k=1}^{p} G_{mk} w_m(s) w_k(s'), \quad j \neq l. \tag{6}$$

3.3 Parameter model

In the parameter level, we specify prior probability distributions for the model parameters $\{(\sigma^2_{\varepsilon,1}, \tau^2_1), \ldots, (\sigma^2_{\varepsilon,6}, \tau^2_6), \mathbf{G}\}$. Prior distributions for these parameters are generally taken to be non-informative. For $\sigma^2_{\varepsilon,j}$, we assign a proper uniform prior on the standard deviation scale $\sigma_{\varepsilon,j} \sim \mathrm{Unif}(a,b)$ for each j independently. The values of the hyperparameters a and b are chosen so as to obtain an approximately non-informative prior. For $\{\tau_j^2\}$, we use the Half-Cauchy (HC) prior with scale parameter θ. We avoid using the usually implemented inverse-gamma priors, since these priors do not yield a proper posterior if the priors are taken to be non-informative. This was confirmed by our preliminary assessment (not shown) and supported by Gelman (2006) and Polson and Scott (2012). We specify the HC prior as $\tau_j \sim \mathrm{HC}(\theta)$ for each j independently. Large but finite values of θ represent approximately non-informative prior distributions. See the appendix for further details on prior and hyperparameter choices. We also need to specify the prior distribution for the covariance matrix $\mathbf{G}$. The inverse Wishart (IW) prior has been proposed for covariance matrices like $\mathbf{G}$, with scale parameter the identity matrix $\mathbf{I}_p$ and $p+1$ degrees of freedom. Although computationally convenient, the IW family is found to be quite constraining as p is the only 'tuning parameter' available to express uncertainty in the elements of $\mathbf{G}$ (Gelman and Hill 2006; Leith and Chandler 2010). We use the modified version of the IW (see, e.g., Gelman and Hill 2006; O'Malley et al. 2008) which is based on the decomposition $\mathbf{G} = \Gamma \mathbf{Q} \Gamma$, where Γ is a diagonal matrix with the scaling elements $\{\omega_k^2\}$ being given non-informative uniform priors over a wide range, and $\mathbf{Q} \sim \mathrm{IW}(p+1, \mathbf{I}_p)$. We then determine $\mathbf{G}$ by computing its diagonal and off-diagonal elements, $G_{kk} = \omega_k^2 Q_k$ and $G_{kl} = \omega_k^2 \omega_l^2 Q_{kl}$ for $k,l = 1,\ldots,p$.

3.4 The choice of the kernels

Several types of kernel functions have been used in the literature, including Gaussian kernels (Stroud et al. 2001) and bisquare functions (Kang et al. 2012). In this paper we have considered a Gaussian kernel specified as

$$w_k(s) \propto \exp\{-(s-c_k)'\Sigma^{-1}(s-c_k)/2\}, \tag{7}$$

where c_k denotes the center of the kernel and Σ determines the shape. The number of kernels, p or p^*, their locations and shapes must be chosen. These choices are often based on the presence of prior information such as smoothness and spatial dependence related to the spatial process (Stroud et al. 2001). If we choose spherically shaped kernels, i.e., $\Sigma = \kappa I_2$ on R^2 and $\kappa > 0$, and the centers belong to a regular grid over an unbounded domain, (5) approximates a covariance function of a stationary isotropic process when the number of kernels is very large. Alternatively, a geometrically anisotropic process may be obtained if we choose non-spherical Gaussian kernels. One way to investigate whether the spatial biases are direction-dependent or not is to perform variogram analyses of the biases for different directions (Cressie 1993). A variogram provides a descriptive statistic of the spatial continuity of a data set. Empirical variograms are calculated by averaging the semi-variances over all pairs of available observations, with a specified separation distance and direction. Figure 2 illustrates the empirical variograms of the six GCM biases for the directions: $0°, 45°, 90°, 135°$ (i.e. North, Northeast, East and Southeast direction, respectively). Observing the plots of the variograms within each panel does not reveal strong anisotropy in the four directions at small distances since the patterns are not largely different. This suggests that we can safely choose a spherical kernel.

Figure 3 shows the two different sets of centers which are used for our main analysis. Panel (a) shows $p = 36$ equally-spaced Gaussian kernels with scale $\Sigma = 0.6\mathbf{I}_2$ on R^2, which are used to model $\mu(s)$. Panel (b) shows $p^* = 45$ unequally-spaced Gaussian weighting kernels with the smaller scale $\Sigma = 0.4\mathbf{I}_2$, which are used to model $\eta_j(s)$. In section 4 we present a sensitivity analysis for the kernel choice, and discuss the advantages and drawbacks of different choices.

3.5 MCMC simulations

Parameter estimation and inference is based on a Bayesian context by sampling from the posterior probability distribution, which is generalized as

$$\begin{aligned}&[\text{process, parameters}|\text{data}] \propto [\text{data}|\text{process}] \times \\ &\qquad [\text{process}|\text{parameters}][\text{parameters}].\end{aligned}$$

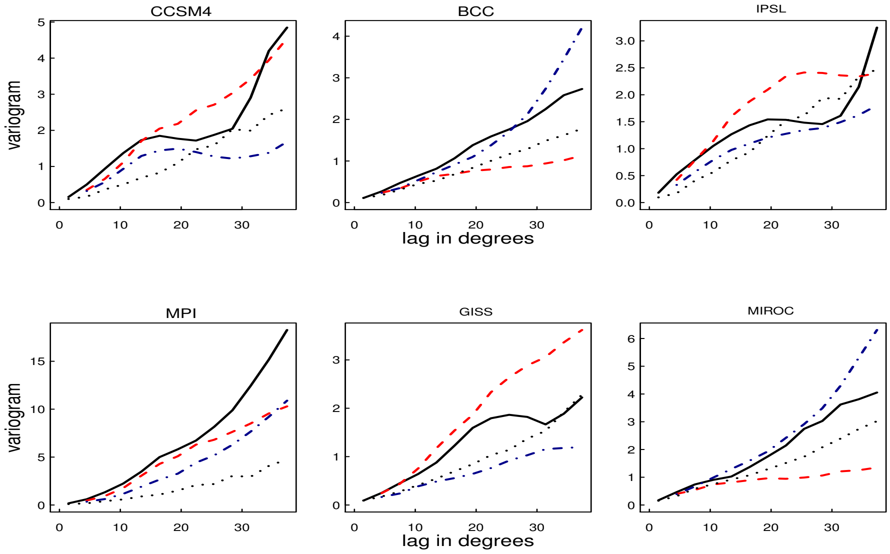

Fig. 2 Empirical variograms of the six GCM biases of the tropical Atlantic region for four different directions (black solid 0°, red dashed 45°, gray dotted 90°, blue dash-dotted 135°). The variograms were analyzed using the robust estimator as given by Cressie (1993).

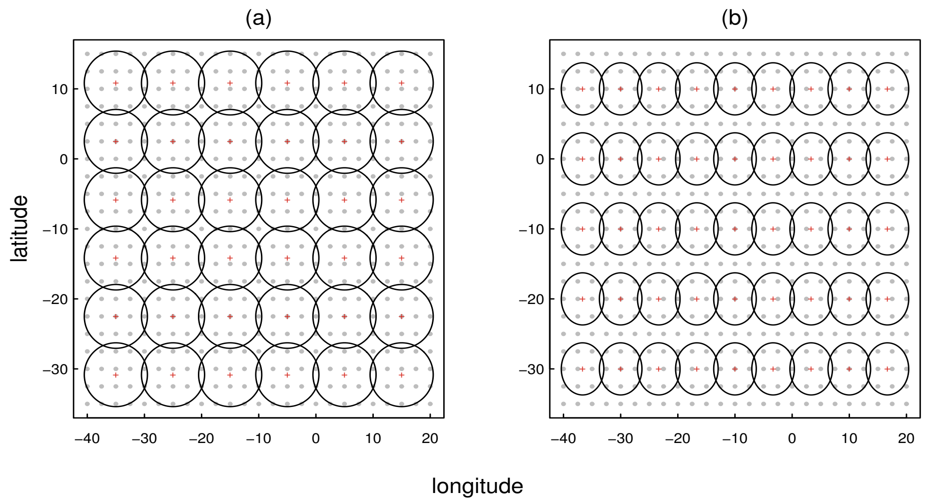

Fig. 3 (a) The 36 equally-spaced Gaussian kernels that are used to model $\mu(s)$; (b) The 45 unequally-spaced Gaussian kernels that are used to model $\eta_j(s)$. Gray color bullets indicate data locations and red color crosses indicate centers of the kernels.

The posterior distributions corresponding to $\mu(s)$ and $\eta_j(s)$ cannot be obtained in closed form, so we use the Markov Chain Monte Carlo (MCMC) method (Gilks et al. 1996) with Gibbs sampler that adopt full conditional distributions. For the MCMC simulations we used three chains, each with overdispersed starting values. We performed 50000 simulations discarding the first 20000 as burn-in. The remaining samples were thinned at every tenth step to reduce autocorrelations of successive samples, from which the remaining 3000 draws were used for posterior analyses. We performed the computations by using the OpenBUGS (version 3.2.3) statistical software package. The computation time depends mainly on the size of the kernel vectors. For example, if we use 36 Gaussian kernels to describe both $\mu(s)$ and $\eta_j(s)$, the computations take about 10 hours on a 64-bit OS X 10.10.5 Intel Core i5 1.6 GHz. Posterior convergence was assessed by inspecting the simulation history of a sample of parameters using graphical tools and the Gelman-Rubin formal convergence diagnostic (Cowles and Carlin 1996). We then summarized the MCMC draws in terms of mean, median and standard deviation to make posterior inference about the unknowns.

4 Results

Figure 4 summarizes the posterior results with respect to the overall climate model bias $\mu(s)$. Panel (a) presents the posterior mean of the overall common bias $\mu(s)$ using the 36 Gaussian kernels that are shown in Figure 3(a). The posterior standard deviations of $\mu(s)$ is shown in Figure 4(b). To better understand $\mu(s)$, Figure 4(c) shows the empirical bias which is estimated as a simple average of the biases from the six GCMs with the underlying assumption that all GCMs have equal weight in synthesizing the overall common bias. The posterior mean of the overall bias and its associated empirical estimate agree well on the general features of the bias over the whole tropical Atlantic region. Common features include the warm error over the southeastern tropical Atlantic, which reaches peak values exceeding 4 kelvin over the Angola-Benguela front region and extends westward as far as 10°W. Both estimates capture a cold error of similar severity over the western tropical Atlantic ocean, along the South American coast. Shared features over landmassess include the cold error over the subsaharan region and warm errors over major near-coastal mountainous African regions, such as the Cameroon line and the Namib desert. Compared to the empirical estimate, the posterior mean of the overall bias intensifies the cold errors over the western Kalahari and over the Congo river, while reducing the cold error over the subsaharan region. The posterior standard deviations of the overall common bias (Figure 4(b)) suggest that its estimate is largely uncertain in the southeastern tropical Atlantic, over the Angola-Benguela front region, where the largest bias is observed. Uncertainty in the common bias estimate is large also along the African coast, possibly reflective of the diversity in the representation across models of coastal topography and/or freshwater discharge processes. The bias estimate is, conversely, more certain in regions affected by cold errors, such as the subsaharan region and the western tropical Atlantic Ocean.

Overall, the posterior mean estimate of the overall bias has a smoother spatial pattern than the corresponding empirical estimate, which changes more rapidly, in the longitude-latitude space. The similarity of the bias patterns in Figure 3(a) and (c) suggests that the proposed method highlights the same common features of the bias that are reflected in the empirical bias estimate. Nonetheless, the Bayesian approach allows to gain deeper insights about how much each climate model varies around the overall common bias. As pointed out in section 3.2, the variance parameters $\{\tau_j^2 : j = 1,\ldots,6\}$ are useful to assess how each climate model bias varies about the overall common bias. Figure 4(d) depicts the posterior medians of τ_j along with the 25th and 75th percentiles, which show a marked difference across the individual GCMs about the overall common bias. CCSM4 varies the least, whereas IPSL and GISS vary the most about the overall common bias. Thus, in terms of weighting the contributions of each GCMs in synthesizing the overall common bias, CCSM4 is ranked first, whereas IPSL and GISS have smaller weights. One benefit of the Bayesian hierarchical method is that it allows to determine the heterogeneity across the climate models, highlighting the limitations of the equal weight assumption often adopted in the traditional empirical estimate.

We now provide posterior assessments of the individual bias components $\{\eta_j(s) : j = 1,\ldots,6\}$. These individual components measure the departure of each climate model bias from the overall common bias $\mu(s)$. As compared to $\mu(s)$, $\eta_j(s)$ describe model-specific local features. Thus, we use a relatively large number of kernels $p^* = 45$, which are shown in Figure 3(b). Figure 5 shows the posterior means of $\{\eta_j(s) : j = 1,\ldots,6\}$. The values of $\eta_j(s)$ for CCSM4 are overall the smallest among all models in the ensemble, suggesting that the most prominent features of CCSM4 go to the overall common bias. This is consistent with our previous result that CCSM4 varies the least about the overall common bias, see Figure 4(d). Similarly, as expected from Figure 4(d), IPSL shows large departures from the overall common bias, followed by GISS. All models show warm errors over the Angola-Benguela front region in their individual bias components. This counterintuitive result is explained by the different location of the peak warm error across the different models, i.e., all models feature a warm bias in the Angola-Benguela front region captured by $\mu(s)$, but each with model-distinctive intensity and spatial structure, which

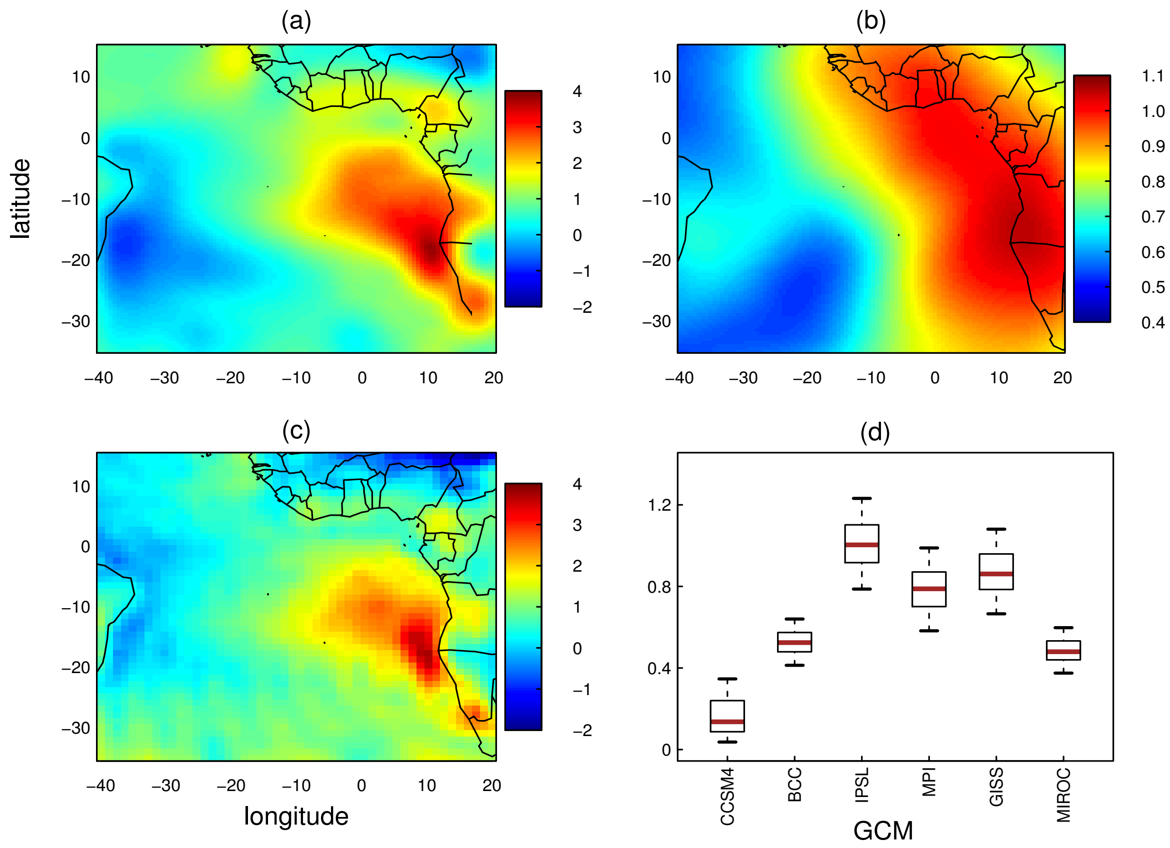

Fig. 4 (a) Posterior mean of the overall common bias $\mu(s)$; (b) associated posterior standard deviation of the overall common bias; (c) empirical estimate of climate model bias, obtained by naively averaging the six climate models assuming the same weight for all of them; (d) Boxplots of the posterior samples of the standard deviation parameters $\{\tau_j : j = 1,\dots,6\}$ where the bold solid horizontal bars denote the medians, the lower and upper bars of the boxes indicate the 25th and 75th percentiles respectively.

are captured by $\eta_j(s)$. The extratropical South Atlantic is the region where the largest variability in $\eta_j(s)$ value is detected, particularly due to the large values of opposite sign in GISS and BCC.

4.1 Model assessment

We now investigate the adequacy of our modeling approach to the choices of Gaussian weighting kernels and hyperparameters. In particular, we recall that our model fitting requires specifying (1) the number of weighting kernels, (2) the scale of the kernels (Σ) and (3) the locations or centers of the kernels. To assess the robustness of the results with respect to these choices, we perform a sensitivity analysis for the overall bias $\mu(s)$ using three different numbers of kernels, that is $p \in \{15, 28, 48\}$, three different choices of scale of the kernels, that is $\Sigma \in \{0.1\mathbf{I}_2, 1.2\mathbf{I}_2, 5\mathbf{I}_2\}$ where $\mathbf{I}_2$ is the identity matrix on R^2, and three different sets of kernel locations. Figure 6 presents the contour plots of the overall common bias $\mu(s)$ associated to the different choices of p and Σ. The three panels in the upper row show the contour plots of $\mu(s)$ fixing $\Sigma = 0.5\mathbf{I}_2$ while varying p. A value of $p = 15$ results in a smooth pattern of $\mu(s)$, which features a peak warm bias of 1.5 kelvin in the Angola-Benguela front region, which is displaced westward compared to the empirical estimate as well as to Bayesian hierarchical estimates obtained with larger p values. The pattern also misses many of the topographic characteristics recognizable from Figure 4. With a larger number of kernels ($p = 48$), the overall common bias appears to be more jagged lacking enough smoothness, while it produces a more detailed spatial pattern. The choice of $p = 28$ (panel b) produces smoothed contour lines and a warmer bias of about 3 kelvin in the Angola-Benguela front region, which is closer to the empirical average as well as to the Bayesian estimate. The three panels in the lower row display the contour plots for the overall common bias estimated by fixing $p = 15$ while varying $\Sigma \in \{0.1\mathbf{I}_2, 1.2\mathbf{I}_2, 5\mathbf{I}_2\}$. We use a low value for $p = 15$ in order to amplify the effect of changes in Σ. The choice of Σ seems to have the opposite impact of the choice of p:

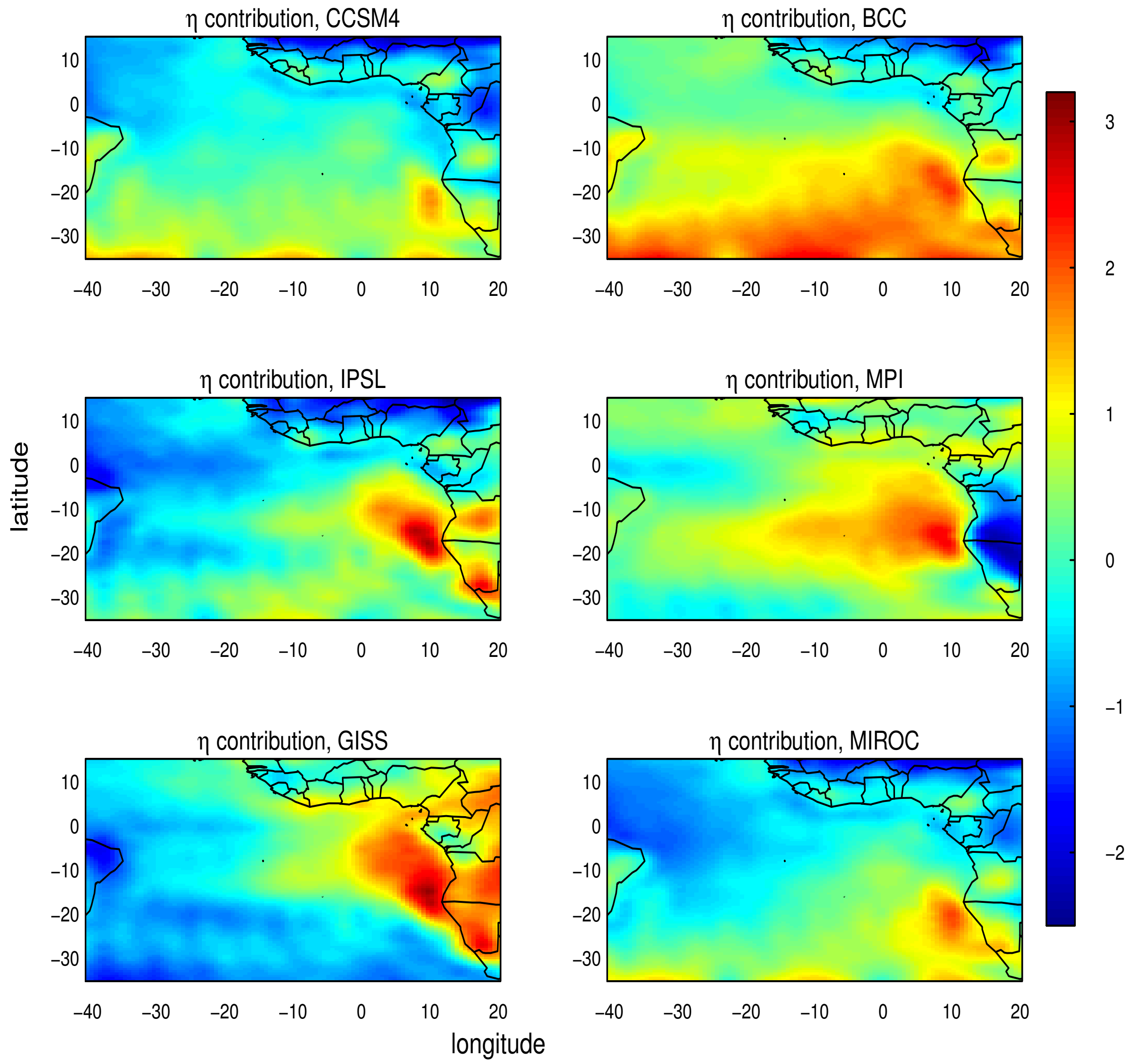

Fig. 5 Posterior means of the individual components $\{\eta_j(s), j = 1, \ldots, 6\}$ (in kelvin) associated to the six GCMs in the ensemble.

smaller Σ values lead to poorly smoothed $\mu(s)$ (panel d), while larger Σ strongly smooths $\mu(s)$ (panel f). Overall, the choice of the kernel parameters, and particularly the number of kernels p is crucial to capture the inherent spatial bias process. In fact, increasing p brings not only increased spatial details but also noticeable changes in the large scale shape of the posterior mean of the overall common bias including the location and magnitude of bias features in key locations.

To assess how the choice of kernel locations or centers of the kernels influences the results, we compare three different sets of kernels that only differ for the location of the centers while using the same number of kernels $p = 64$ and scale matrix $\Sigma = 0.5\mathbf{I}_2$. Figure 7 shows the three sets of kernels, along with the corresponding surface plots of the posterior mean of the overall bias $\mu(s)$. The three different sets of kernels are shown in column (a). In the upper row the centers of the kernels are equally-spaced. The middle and the lower rows feature two different sets of unequally-spaced kernel centers. The different kernel locations yield noticeable differences in the large scale shape of $\mu(s)$ (column b) including the location and magnitude of the bias. The most prominent feature is that equally-spaced kernel locations produce a stronger and more extensive warm bias over the Angola-Benguela front region compared to unequally-spaced kernel setups, which is also closer to the bias estimates shown in Figure 4. The unequally-spaced kernels lead to reduced bias in both warm and cold bias regions.

We performed a further sensitivity analysis to assess the sensitivity of the results to the choice of the parameter θ of the Half-Cauchy (HC) prior for $\tau_1, \ldots, \tau_6$. While the sensitivity analysis could be performed for all prior choices, we only focus on θ as hyperparameters of variance components are

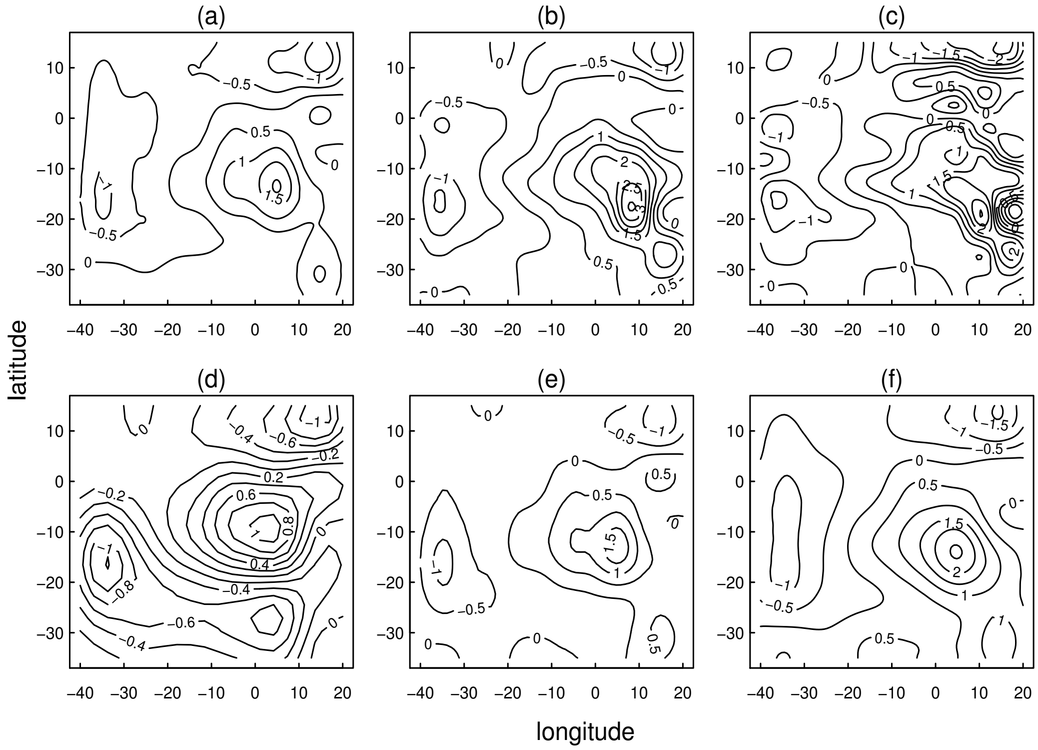

Fig. 6 Assessing the influence of choice of the number p and scale Σ of Gaussian kernels. The upper row shows contour plots of the overall common bias $\mu(s)$ fixing $\Sigma = 0.5\mathbf{I}_2$ while p varies: (a) $p = 15$; (b) $p = 28$; (c) $p = 48$. The lower row shows contour plots of $\mu(s)$ fixing $p = 15$ while Σ varies: (d) $\Sigma = 0.1\mathbf{I}_2$; (e) $\Sigma = 1.2\mathbf{I}_2$; (f) $\Sigma = 5\mathbf{I}_2$.

more sensitive than the hyperparameters of other forms such as Gaussian prior for regression coefficients (Gelman 2006). Figure 8 illustrates the posterior distributions of τ_j for the three choices $\theta \in \{20, 35, 40\}$. The three choices produce slightly different posterior distributions, but they reflect the same general pattern. The sensitivity to θ differs slightly across the ensemble members. For instance, CCSM4 provides the smallest variation across all three choices and IPSL produces the largest variation. Thus, we consider the results to be robust against the specific choice of θ.

5 Discussion

We have proposed a Bayesian hierarchical method for the probabilistic assessment and quantification of spatially referenced climate model biases in a multi-model ensemble. The approach synthesizes an overall shared bias as a non-stationary spatial field and quantifies the associated uncertainty. The approach optimizes the way information about the bias is combined within the ensemble. Specifically, the presented model accounts for the variability of the bias across ensemble members, and the contribution of each member to the overall common bias is determined based on the posterior inferences on each model's variability parameter.

Application of the model to the case of tropical Atlantic near-surface air temperature from an ensemble of six historical simulations contributing to CMIP5 exemplified how the proposed approach allows to gain deeper insights into climate model bias compared to more traditional assessments: Known common features of the bias in this region are well captured by our statistical model, such as the warm bias over the Angola-Benguela front region. But, our model further reveals that the different GCMs unequally contribute to determining this bias, which also results in a variety of model-specific features of the bias over the same area. The proposed statistical decomposition of each model's bias into a shared/common and a model-specific component stimulates additional investigation of the underlying physical processes as well. In our application, for instance, the errors of opposite sign emerging in the model-specific components of the bias over the near-coastal oceanic waters of equatorial Africa deserve further analysis.

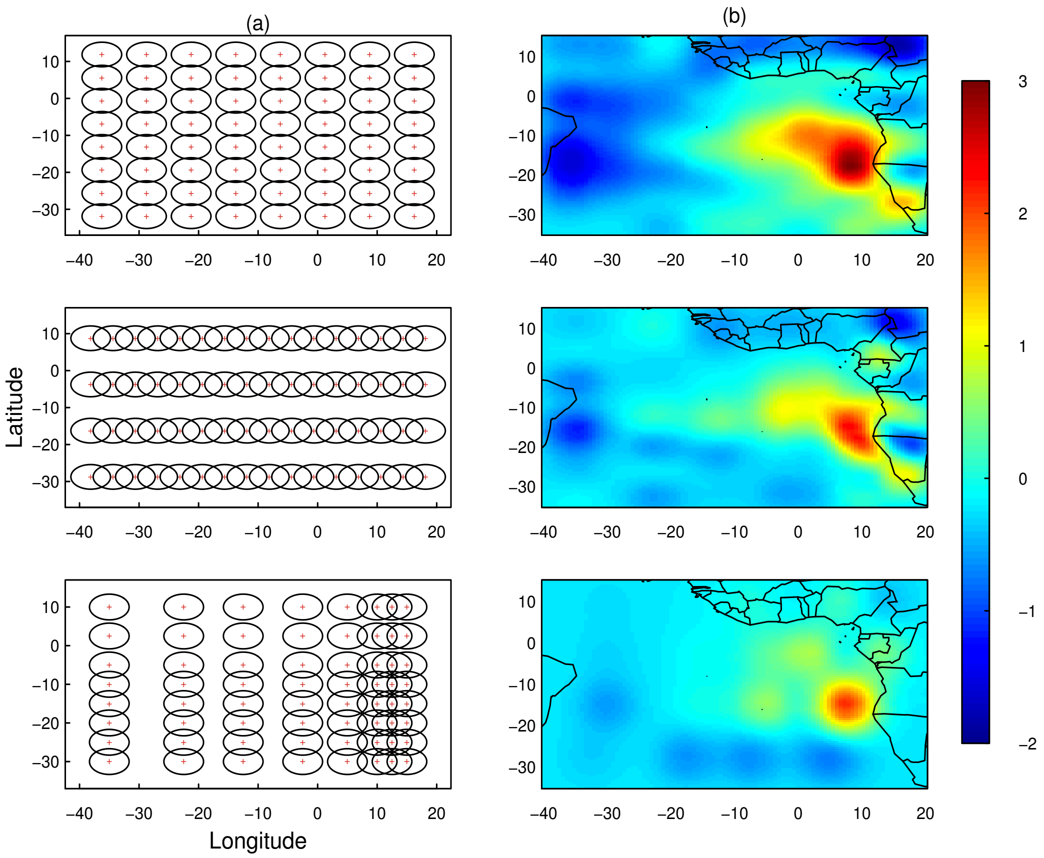

Fig. 7 Comparison of the posterior mean of the overall bias $\mu(s)$ for three different choices of kernel locations: (a) Gaussian weighting kernel locations; (b) the posterior mean surfaces of the overall common bias.

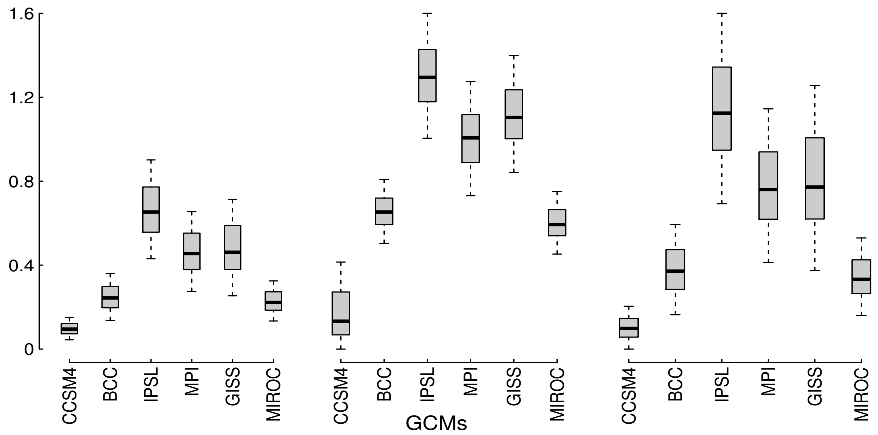

Fig. 8 Boxplots of MCMC draws associated to τ_j for different choices of the hyperparameter θ for the Half-Cauchy prior. Panels are for $\theta \in \{20, 35, 40\}$, respectively, from left to right.

The basic idea underlying our statistical model is generic and could be applied to a wider range of climate models, geographical locations and geophysical variables. Indeed this will be included in our future work that considers an extension to a spatiotemporal model involving a larger set of GCM simulations. The challenge will be to formulate a computationally efficient method for such an extensive approach taking into account the spatial and temporal features simultaneously. Another future focus is to consider biases of multivariate outputs from GCMs such as temperature and precipitation which may provide a broader assessments of climate model uncertainties.

Finally, in section 2.1 we have mentioned that we interpolated the outputs from the six GCMs to the same observational grid to resolve the misalignment between observations and model outputs before fitting the Bayesian hierarchical model. The uncertainty associated to the interpolation can affect the bias estimation in case of strong spatial misalignment. In our case study, both reanalysis and climate model outputs feature high spatial resolution over the investigated domain. We therefore expected interpolation to only minimally influence the results, and hence did not explicitly accounted for it in our model. Nonetheless, when there is concern of substantial uncertainty due to interpolation, it may be desirable to build a model that is able to handle such spatial misalignment directly. One possible approach is the Bayesian hierarchical method for nested block-level realignment (e.g., Banerjee et al. 2014), but this method requires the model output to be nested in the observational grid (Mugglin and Carlin 1998). A simpler solution is, once the outputs are firstly predicted to the observational grid using a stochastic model based approach such as the kriging method, to rectify the uncertainty that has been introduced by inflating the variance of the error $\varepsilon_j(s)$ in model (2). We denote the predicted value from climate model j at spatial location s by $\widehat{X}_j(s)$. Its variance, $\delta_j^2(s) = \mathrm{var}(\widehat{X}_j(s))$ is zero if the output grid and observation grid coincide in s, otherwise it will be positive. Thus we specify

$$\mathrm{var}(\varepsilon_j(s)) = \sigma_j^2 + \gamma_j \delta_j^2(s),$$

where the modulating parameter γ_j is positive. This slight modification adds further parameters to the Bayesian hierarchical model for which we can assign prior distributions similarly to σ_j^2.

Acknowledgements The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7/2007-2013) under Grant agreement n 603521 - PREFACE. The authors would like to thank the two anonymous reviewers and the associate editor for helpful comments on the manuscript.

6 Appendix: choice of priors

In this section, we provide details of prior and hyperparamters choices. All priors are approximately non-informative. For the error variances $\{\sigma_{\varepsilon,j}^2 : j = 1\ldots,6\}$, we assign the uniform prior on the standard deviation scale $\sigma_{\varepsilon,j} \sim \mathrm{Unif}(a,b)$ by choosing $a = 0$ and $b = 10^2$ for each j independently. Accordingly, the error variances, which are proportional to $(b-a)^2$, are very large so that the priors are approximately non-informative. For $\{\tau_j^2 : j = 1\ldots,6\}$, we use a Half-Cauchy (HC) prior, which is a conditionally conjugate family of a half t distribution (Gelman 2006). The Half t distribution corresponds to the absolute value of a Student-t distribution centered at zero, whose probability distribution is proportional to

$$\big(1+\frac{1}{\mathrm{df}}(\frac{\tau_j}{\theta})^2\big)^{-(\mathrm{df}+1)/2} \tag{8}$$

with two parameters: degrees of freedom df and scale parameter θ. We obtain the proper HC probability distribution for τ_j as a special case of (8) by setting $\mathrm{df} = 1$,

$$p(\tau_j) \propto \big(\theta^2 + \tau_j^2\big)^{-1}, \quad j = 1,\ldots,6$$

we specify priors for τ_j as $\tau_j \sim \mathrm{HC}(\theta)$, independently for each j. Large but finite value of the scale parameter θ represents an approximately non-informative prior distribution. In the limit $\theta \to \infty$ this becomes a uniform prior density on $p(\tau_j)$. For our analysis, we set $\theta = 30$. To choose a prior for the $p \times p$ covariance matrix $\mathbf{G}$, the variances $G_1,\ldots,G_p$ and the pair-wise covariances $G_{kl} : k,l = 1,\ldots,p$ must be explicitly specified. One way to achieve this is to use the separation technique (Gelman and Hill 2006; O'Malley et al. 2008)

$$\mathbf{G} = \Gamma \mathbf{Q} \Gamma$$

where Γ is the diagonal matrix with diagonal elements $\omega_1^2,\ldots,\omega_p^2$ and $\mathbf{Q}$ is new $p \times p$ covariance matrix. The role of the new parameters ω_k^2 and $\mathbf{Q}$ is to derive appropriately scaled priors for the variances and pair-wise covariances related to $\mathbf{G}$. We assign proper uniform prior on $\omega_k^2 \sim \mathrm{Unif}(0,10^2)$ independently for each k. The covariance component $\mathbf{Q}$ is given the inverse Wishart distribution $\mathrm{IW}(p+1,\mathbf{I}_p)$. The two parameters degrees of freedom $p+1$ and the identity matrix $\mathbf{I}_p$ fully determine the distribution. The variances and pair-wise covariances associated to $\mathbf{G}$ are then obtained as $G_k = \omega_k^2 Q_p$ and $G_{kl} = \omega_k \omega_l Q_{kl}$. To make inference, we require the standard deviations $|G_k|^{1/2}$ and correlations ρ_{kl}

$$|G_k|^{1/2} = |\omega_k|\sqrt{Q_k} \quad \text{and} \quad \rho_{kl} = \frac{G_{kl}}{|G_k|^{1/2}|G_l|^{1/2}}, \quad k,l = 1,\ldots,p$$

References

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data. CRC Press, New York

Berliner LM (2003) Physical-statistical modeling in geophysics. Journal of Geophysical Research: Atmospheres 108:8776. doi:10.1029/2002JD002865

Boberg F, Christensen JH (2012) Overestimation of Mediterranean summer temperature projections due to model deficiencies. Nature Climate Change 2:433-436

Brohan P, Kennedy JJ, Harris I, Tett SF, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. Journal of Geophysical Research: Atmospheres 111. doi:10.1029/2005JD006548

Buser CM, Knsch HR, Lthi D, Wild M, Schr C (2009) Bayesian multi-model projection of climate: bias assumptions and interannual variability. Climate Dynamics 33:849-868

Christensen JH, Boberg F, Christensen OB, LucasPicher P (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. Geophysical Research Letters 35. doi:10.1029/2008GL035694

Cowles MK, Carlin BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91:883-904

Cressie N (1993) Statistics for spatial data. Wiley-Interscience, New York

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen, 2013: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge,United Kingdom and New York, NY, USA

Furrer R, Sain SR, Nychka D, Meehl GA (2007) Multivariate Bayesian analysis of atmosphereocean general circulation models. Environmental and Ecological Statistics 14:249-266

Gelfand AE, Diggle P, Fuentes M, Guttorp P (2010) Handbook of Spatial Statistics. Chapman and Hall / CRC, Florida

Gelman A (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1:515-534

Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London

Higdon D (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environmental and Ecological Statistics 5:173-190

Jun M, Knutti R, Nychka DW (2008) Spatial analysis to quantify numerical model bias and dependence: how many climate models are there?. Journal of the American Statistical Association 103:934-947

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y (1996) The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77:437-471

Kang EL, Cressie N, Sain SR (2012) Combining outputs from the North American regional climate change assessment program by using a Bayesian hierarchical model. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61:291-313

Kaufman CG, Sain SR (2010) Bayesian functional ANOVA modeling using Gaussian process prior distributions, Bayesian Analysis 5:123-149

Keller CF (2009) Global warming: a review of this mostly settled issue. Stochastic Environmental Research and Risk Assessment 23:643676

Kistler R, Collins W, Saha S, White G, Woollen J, Kalnay E, Chelliah M, Ebisuzaki W, Kanamitsu M, Kousky V, van den Dool H (2001) The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. Bulletin of the American Meteorological society 82:247-267

Knutti R (2010) The end of model democracy? An editorial comment. Climatic Change 102:395404

Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of coupled climate models. Climate Dynamics 17:83-106

Leith NA, Chandler RE (2010) A framework for interpreting climate model outputs. Journal of the Royal Statistical Society: Series C (Applied Statistics) 59:279-296

Milinski, S., J. Bader, H. Haak, A. C. Siongco, J. H. Jungclaus (2016) High atmospheric horizontal resolution eliminates the wind-driven coastal warm bias in the southeastern tropical Atlantic. Geophysical Research Letter, accepted for publication

Mugglin AS, Carlin BP (1998) Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. Journal of Agricultural, Biological and Environmental Statistics 3: 111130

Neuman SP (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environ-

mental Research and Risk Assessment 17: 291-305
O'Malley AJ, Zaslavsky AM (2008) Domain-level covariance analysis for multilevel survey data with structured nonresponse. Journal of the American Statistical Association 103:1405-1418
Polson NG, Scott JG (2012) On the half-Cauchy prior for a global scale parameter. Bayesian Analysis 7:887-902
Sain SR, Furrer R (2010) Combining climate model output via model correlations. Stochastic Environmental Research and Risk Assessment 24:821-829
Stroud JR, Mller P, Sansó B (2001) Dynamic models for spatiotemporal data. Journal of the Royal Statistical Society. Series B, Statistical Methodology 63: 673-689
Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society 93:485-498
Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles. Journal of Climate 18:1524-1540
Toniazzo T, Woolnough S (2014) Development of warm SST errors in the southern tropical Atlantic in CMIP5 decadal hindcasts. Climate Dynamics 43:2889-2913
Wahl, S., M. Latif, W. Park, N. Keenlyside (2015) On the Tropical Atlantic SST warm bias in the Kiel Climate Model. Climate Dynamics doi:10.1007/s00382-009-0690-9
Wang C, Zhang L, Lee SK, Wu L, Mechoso CR (2014) A global perspective on CMIP5 climate model biases. Nature Climate Change 4: 201-205
Zanchettin D, Bothe O, Lehner F, Ortega P, Raible CC, Swingedouw D (2015) Reconciling reconstructed and simulated features of the winter Pacific/North American pattern in the early 19th century. Climate of the Past 11:939-958