Clim Dyn (2017) 48:1555–1570
DOI 10.1007/s00382-016-3160-1

Cross-validation analysis of bias models in Bayesian multi-model projections of climate

J. M. J. Huttunen[1] · J. Räisänen[3] · A. Nissinen[1] · A. Lipponen[2] · V. Kolehmainen[1]

Received: 27 April 2015 / Accepted: 28 April 2016 / Published online: 10 May 2016

Abstract Climate change projections are commonly based on multi-model ensembles of climate simulations. In this paper we consider the choice of bias models in Bayesian multimodel predictions. Buser et al. (Clim Res 44(2–3):227–241, 2010a) introduced a hybrid bias model which combines commonly used constant bias and constant relation bias assumptions. The hybrid model includes a weighting parameter which balances these bias models. In this study, we use a cross-validation approach to study which bias model or bias parameter leads to, in a specific sense, optimal climate change projections. The analysis is carried out for summer and winter season means of 2 m-temperatures spatially averaged over the IPCC SREX regions, using 19 model runs from the CMIP5 data set. The cross-validation approach is applied to calculate optimal bias parameters (in the specific sense) for projecting the temperature change from the control period (1961–2005) to the scenario period (2046–2090). The results are compared to the results of the Buser et al. (Clim Res 44(2–3):227–241, 2010a) method which includes the bias parameter as one of the unknown parameters to be estimated from the data.

Keywords Climate projections · Multi-model prediction · Bayesian estimation · Model bias · Bias change · Cross-validation · CMIP5

1 Introduction

Model-based projections of anthropogenic climate change form an important source of information for policy makers to guide in environmental decisions. Therefore, it is an important question how to best use and interpret the available model simulations.

It is widely accepted that multi-model predictions are superior to single model projections and that an ensemble of models can outperform individual ensemble members (Weigel et al. 2008; Räisänen and Ylhäisi 2012). Reviews of methods to combine multimodel ensemble predictions are given from a weather predictions perspective in Wilks (2006) and from a climate projection perspective in Tebaldi and Knutti (2007). In weather forecasts, many of the methods are based on assigning equal weights for all models and subtracting biases of each model that are determined based on past model performance. In climate projections, however, a difficulty arises due to limited knowledge of how the model biases might change between the present and future periods.

A common assumption is that the changes in model biases are small compared to the changes in climate. Several studies, however, show that biases may change as climate changes. State dependent bias models have also been proposed. For example, Buser et al. (2009) proposed a constant relation assumption. With this bias model, the bias in the mean climate changes with climate if the baseline interannual variability in the model differs from the observed variability. Furthermore, Christensen et al. (2008)

Electronic supplementary material The online version of this article (doi:10.1007/s00382-016-3160-1) contains supplementary material, which is available to authorized users.

✉ J. M. J. Huttunen
jmhuttun@gmail.com

[1] Department of Applied Physics, University of Eastern Finland, P. O. Box 1627, 70211 Kuopio, Finland

[2] Department of Physics, University of Helsinki, P. O. Box 48, 00014 Helsinki, Finland

[3] Finnish Meteorological Institute (FMI), Atmospheric Research Centre of Eastern Finland, P. O. Box 1627, 70211 Kuopio, Finland

and Boberg and Christensen (2012) demonstrate that many models overestimate warm-season temperature variability. To avoid the implicated overestimate of long-term warming, they propose temperature dependent bias correction based on quantile-quantile plots. The approach was also applied to the CMIP5 dataset in Christensen and Boberg (2012), Christensen and Boberg (2013). This approach is somewhat equivalent to the Buser's et al. constant relation assumption. Similar bias corrections were also considered in Bellprat et al. (2013), Kerkhoff et al. (2014) and Ho et al. (2012).

In this paper we consider the hybrid bias model proposed in Buser et al. (2010a), which combines the constant bias and constant relation assumptions. The hybrid bias model includes a parameter which scales the weighting between these two bias assumptions, and the parameter can either be considered as fixed or can be estimated from the data as an unknown parameter simultaneously with the other model parameters.

It is a difficult task to find the most appropriate bias model, since validation based on the future climate is impossible. In this paper, we analyse the choice of the bias model using a cross-validation approach. A key assumption in the cross-validation is that climate model outputs are random samples of possible future climates. Therefore, we can ask how well the output of a selected model can be predicted based on the data provided by all the other models in the ensemble. A similar approach known as the pseudoreality framework has been widely used in other studies (see e.g. Maraun 2012; Bellprat et al. 2013; Kerkhoff et al. 2014).

A similar cross-validation approach was used to confirm Bayesian predictions in Smith et al. (2009). However, the focus of the cross validation in this paper is the choice of the bias parameter in the hybrid bias model proposed by Buser et al. (2010a). The cross validation based selection leads to an optimal bias parameter in the sense that the distance of the model based climate prediction and the (simulated) scenario climate is minimized with respect to some metric. In this paper we mainly consider the continuous ranked probability score (CRPS), which is a widely used metric in weather and climate predictions (Hersbach 2000; Jolliffe and Stephenson 2011). The CRPS is also a strictly proper score, i.e., it is uniquely minimized by using true probabilities (see Gneiting and Raftery 2007).

We use the latest CMIP5 (World Climate Research Programme's Coupled Model Intercomparison Project phase 5) multi-model dataset, which includes a large number of general circulation model (GCM) outputs. Although several recent Bayesian multi-model methods (Buser et al. 2009, 2010a; Heaton et al. 2013) have used regional climate model (RCM) data, we use only outputs of GCMs due to the large number of available models in the CMIP5 data set. However, it is also straightforward to apply the method presented in this paper to RCM model outputs when an extensive dataset (comparable to the size of CMIP5) becomes available.

This paper is organized as follows. In Sect. 2 we describe the data and the aggregation procedure used in this study. In Sect. 3, we briefly outline the Bayesian multimodel method and the hybrid bias model presented in Buser et al. (2010a) that form the basis of our cross-validation analysis. The cross-validation approach is also presented in Sect. 3. The results are presented in Sect. 4 and conclusions are drawn in Sect. 5.

2 Data

In this section, we summarize the climate model and observational data used in this study. In the cross-validation approach presented in this paper, only simulated data (climate model data) is used for the selection of the bias parameter. However, we will also compute climate predictions using the true observational data to compare the results of the cross-validated bias model with other predictions.

The variable we consider in this study is 2m land-surface temperature and the ultimate aim of the analysis is to compute predictions for temperature change between the control period (1961–2005) and the scenario period (2046–2090). However, the methodology can also be extended to other variables. For example, the cross-validation can be connected to the predictions of both 2m -temperature and precipitation using Bayesian multimodel projections presented in Buser et al. (2010b) with a similar bias model; see also Heaton et al. (2013), Tebaldi and Sansó (2009).

In the analysis, both climate model and observational data are averaged both temporally over the summer and winter seasons and spatially over the regions introduced in the IPCC SREX report (Seneviratne et al. 2012). The regions are listed in Table 1. For each area, the spatial averages are calculated over all land grid points inside the area. The analysis is carried out separately for each season and each region.

2.1 Climate model data

This study uses data from coupled atmosphere-ocean general circulation models and Earth system models participating in the World Climate Research Programme Fifth Coupled Model Intercomparison project (CMIP5). We use 19 models for which we were able to download monthly 2 m-temperature data corresponding to the historical simulations for the recent past (the control period 1961–2005)

Table 1 The regions (SREX) used in this study

Label	Number	μ_0 (°C)	
		DJF	JJA
ALA	1	−24	10
CGI	2	−22	5
WNA	3	−4	18
CNA	4	0	24
ENA	5	−3	20
CAM	6	20	25
AMZ	7	25	25
NEB	8	25	25
WSA	9	14	9
SSA	10	23	12
NEU	11	−4	14
CEU	12	−2	18
MED	13	8	25
SAH	14	17	32
EAF	16	24	24
WAF	15	25	25
SAF	17	24	17
NAS	18	−24	14
WAS	19	8	28
CAS	20	−2	24
TIB	21	−12	15
EAS	22	−3	21
SAS	23	15	25
SEA	24	25	25
NAU	25	29	17
SAU	26	21	10

The table also includes the mean temperatures μ_0 of the regions for DJF (Northern Hemisphere winter) and JJA (Northern Hemisphere summer) seasons that are used as prior information in Bayesian multimodel analysis (see Table 3). For the coordinates of region corners, we refer to Appendix 3.A of Seneviratne et al. (2012)

and the simulations for the future (the scenario period 2046–2090) based on the Representative Concentration Pathways (RCP) 4.5 scenario (Thomson et al. 2011). The sea grid boxes are masked out by using the model-specific land-sea masks. Since in the analysis different model runs will be assumed to be independent, we chose only one model run per model family or institute. The models are summarized in Table 2.

2.2 Observational data

The observational data used in this study is the TS 3.21 high resolution monthly gridded data provided by Climate Research Unit (CRU). The data is based on station data, which have been interpolated to a regular 0.5 lon × 0.5 lat grid and can be accessed via the CRU website (http://www.cru.uea.ac.uk/data). The Land-Sea mask provided by the CRU is used as a mask for land-surface temperatures. In this study we assume that the CRU observations represent the true climate. For a detailed description of the data, see Harris et al. (2014).

3 Methods

In this paper, we consider the choice of bias model in Bayesian multi-model projection. More specifically, we apply cross-validation to find a value for the weighting parameter κ for the hybrid bias model proposed by Buser et al. (2010a). This hybrid bias model is a combination of commonly used constant bias and constant relation assumptions and the parameter κ is a weighting parameter between these bias models.

Before going to the details of our cross-validation approach, we briefly outline the Bayesian multi-model projection methodology by Buser et al. (2010a), which forms the basis of the cross-validation approach.

3.1 Notations

We follow the representation and notations of Buser et al. (2010a): $X_{0,t}$ denotes temperatures during the control period of the chosen reference model in year $1960 + t$, $t = 1, \ldots, T$ $(T = 45)$ and $Y_{0,t}$ denotes the corresponding scenario temperatures in year $2045 + t$, $t = 1, \ldots, T$. For the i'th model $(i = 1, \ldots, N_{\mathrm{m}})$, the model outputs for the control period are denoted as $X_{i,t}$ and for the scenario period as $Y_{i,t}$.

3.2 Bayesian multi model projections

As in Buser et al. (2010a), the multi-model climate predictions are made using a Bayesian framework. For other Bayesian multi-model approaches, see Tebaldi et al. (2005), Tebaldi and Sansó (2009), Heaton et al. (2013). The idea is to construct a probability distribution for the scenario climate given all the available data:

$$\mathcal{D} = \{X_{0,t}, X_{i,t}, Y_{i,t}; t = 1, \ldots, T,\ i = 1, \ldots, N_{\mathrm{m}}\}.$$

In this approach, we specify the *likelihood density* $p(\mathcal{D}|\Theta)$ where Θ is a set of model parameters (specified below). In Bayesian formalism, the model parameters Θ are also considered as random quantities and the *prior distribution* $p(\Theta)$ incorporates our prior beliefs about the parameters (also specified below).

Then the posterior distribution, given by the Bayes formula

$$p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)p(\Theta), \tag{1}$$

Table 2 The CMIP5 climate models used in this study

i	Model	Institution
1	CSM1.1(m)	Beijing Climate Center, China Meteorological Administration
2	BNU-ESM	Beijing Normal University, China
3	CanESM2	Canadian Centre for Climate Modelling and Analysis
4	CMCC-CM	Centro Euro-Mediterraneo per I Cambiamenti Climatici, Italy
5	CNRM-CM5	Centre National de Recherches Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique
6	ACCESS-1.3	Commonwealth Scientific and Industrial Research Organisation (CSIRO), and Bureau of Meteorology (BoM), Australia
7	CSIRO-Mk3	CSIRO Atmospheric Research and Queensland Climate Change Centre of Excellence in Brisbane, Australia
8	FIO-ESM	The First Institute of Oceanography, SOA, China
9	EC-EARTH	EC-EARTH consortium
10	IPSL-CM5A-MR	Institut Pierre-Simon Laplace, France
11	FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences, and CESS, Tsinghua University
12	MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies, Japan
13	HadGEM2-ES	Met Office Hadley Centre, UK
14	MPI-ESM-MR	Max Planck Institute for Meteorology (MPI-M), Germany
15	CGCM3	Meteorological Research Institute, Japan
16	GISS-E2-H-CC	NASA Goddard Institute for Space Studies, USA
17	CCSM4	National Center for Atmospheric Research (NCAR), USA
18	NorESM1-ME	Norwegian Climate Centre
19	CESM1-CAM5	National Science Foundation, Department of Energy, National Center for Atmospheric Research, USA

gives a probability distribution for the parameters given the data.

Given the posterior probability density of the parameters, the conditional probability distribution for the scenario climate is given by

$$p(Y_{0,t}|\mathcal{D}) = \int p(Y_{0,t}|\Theta)p(\Theta|\mathcal{D})\,\mathrm{d}\Theta. \quad (2)$$

The first task is to specify the distributions for historical and scenario temperatures given the parameters Θ, and also distributions for model predictions $X_{i,t}$ and $Y_{i,t}$ given Θ.

3.3 *Distribution of data*

The distribution of data is chosen in the same manner as in Buser et al. (2010a). All data is assumed to be normally distributed and mutually independent. The chosen statistical models for $X_{0,t}$, $X_{i,t}$ and $Y_{0,t}$ are:

$$X_{0,t} \sim \mathcal{N}(\mu + \gamma(t - T_0), \sigma^2) \quad (3)$$

$$X_{i,t} \sim \mathcal{N}(\mu + \beta_i + (\gamma + \gamma_i)(t - T_0), \sigma^2 b_i^2) \quad (4)$$

$$Y_{0,t} \sim \mathcal{N}(\mu + \Delta\mu + (\gamma + \Delta\gamma)(t - T_0), \sigma^2 q^2) \quad (5)$$

where $T_0 = (T + 1)/2$. Centering the time around T_0 allows us interpret the parameter μ as the mean value of the temperature during the control period and the parameter γ represents the linear trend of temperature during the control period. The parameter σ represents the interannual standard deviation of the temperature during the control period. For the i'th model, the parameters β_i and γ_i are additive biases during the control period and b_i is the multiplicative bias in the interannual variation. The parameter $\Delta\mu$ represents mean temperature change between the control and scenario periods, q is the change in the interannual variability and $\Delta\gamma$ is the change in the trend.

In this paper we consider the choice of the model for the bias between the control and scenario period. A common choice for the change of bias is the *constant bias* assumption (Buser et al. 2009, 2010b):

$$Y_{i,t} \sim \mathcal{N}(\mu + \Delta\mu + \beta_i + \Delta\beta_i + (\gamma + \Delta\gamma + \gamma_i + \Delta\gamma_i)(t - T_0), \sigma^2 q^2 b_i^2 q_{b_i}^2) \quad (6)$$

where $\Delta\beta_i$ and $\Delta\gamma_i$ are additive changes in the biases between the control and scenario periods and q_{b_i} is the change in the multiplicative bias. In addition, $\Delta\beta_i$ and $\Delta\gamma_i$ are assumed to be close to zero, which means that the

Table 3 Parameters for the prior distribution $p(\Theta)$

Parameter	Unit	Distribution	95 % confidence interval
μ	°C	$\mathcal{N}(\mu_0, 25)$	μ_0-9.8, μ_0+ 9,8
$\Delta\mu$	°C	$\mathcal{N}(0, 16)$	−7.8, 7,8
β_i	°C	$\mathcal{N}(0, 16)$	−7.8, 7,8
$\Delta\beta_i$	°C	$\mathcal{N}(0, 0.5)$	−1.4, 1.4
γ	°C/yr	$\mathcal{N}(0, 0.1)$	−0.6, 0.6
$\Delta\gamma$	°C/yr	$\mathcal{N}(0, 0.1)$	−0.6, 0.6
γ_i	°C/yr	$\mathcal{N}(0, 0.1)$	−0.6, 0.6
$\Delta\gamma_i$	°C/yr	$\mathcal{N}(0, 0.0005)$	−0.045, 0.045
σ^{-2}	°C$^-$2	$\mathcal{G}(0.1, 0.1)$	0,9.8
q^{-2}		$\mathcal{G}(0.1, 0.1)$	0,9.8
b_i^{-2}		$\mathcal{G}(0.1, 0.1)$	0,9.8
$q_{b_i}^{-2}$		$\mathcal{G}(3, 3)$	0.2,2.4

The mean temperatures μ_0 for each region are listed in Table 1. $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with the mean μ and the variance σ^2 and $\mathcal{G}(\alpha, \beta)$ is the Gamma distribution with the shape α and rate β

bias in the models is assumed to remain relatively stable between the control and scenario period. In other words, with the constant bias assumption, the models are assumed to predict the climate shift between the control and scenario period accurately. Another common bias model is the constant relation assumption (Buser et al. 2009, 2010b) given by

$$Y_{i,t} \sim \mathcal{N}(\mu + b_i\Delta\mu + \beta_i + \Delta\beta_i + (\gamma + b_i\Delta\gamma + \gamma_i + \Delta\gamma_i)(t - T_0), \sigma^2 q^2 b_i^2 q_{b_i}^2). \quad (7)$$

With the constant relation bias model, a model which overestimates (or underestimates, resp.) the difference between a warm and a cold year (as characterized e.g. by the standard deviation) in the control period by the factor b_i, is also assumed to overestimate (or underestimate) the climate change by the same factor. For more details about these bias models, we refer to Buser et al. (2009).

We adopt the hybrid bias model for the climate model outputs introduced in Buser et al. (2010a):

$$Y_{i,t} \sim \mathcal{N}(\mu + \Delta\mu + \beta_i + \Delta\beta_i + \kappa(b_i - 1)\Delta\mu + (\gamma + \Delta\gamma + \gamma_i + \Delta\gamma_i + \kappa(b_i - 1)\Delta\gamma)(t - T_0), \sigma^2 q^2 b_i^2 q_{b_i}^2) \quad (8)$$

where the parameter κ takes values between 0 and 1. For $\kappa = 0$, the model corresponds to the constant bias model and for $\kappa = 1$, it corresponds to the constant relation model. In this paper, the task is to select κ by applying a cross-validation approach.

Due to the independency assumption, the likelihood $p(\mathcal{D}|\Theta)$ is the product of the Gaussian distributions (3)–(5), (8),

$$p(\mathcal{D}|\Theta) \propto \prod_{t=1}^{T} \frac{1}{\sigma} e^{-\frac{[X_{0,t} - \mu - \gamma(t - T_0)]^2}{2\sigma^2}} \times \prod_{t=1}^{T}\prod_{i=1}^{N_m} \frac{1}{\sigma b_i} e^{-\frac{[X_{i,t} - \mu - \beta_i - (\gamma + \gamma_i)(t - T_0)]^2}{2\sigma^2 b_i^2}} \times \prod_{t=1}^{T}\prod_{i=1}^{N_m} \frac{1}{\sigma q b_i q_{b_i}} e^{-\frac{[Y_{i,t} - \mu - \Delta\mu - \beta_i - \Delta\beta_i - \kappa(b_i - 1)\Delta\mu - \ldots]^2}{2\sigma^2 q^2 b_i^2 q_{b_i}^2}}. \quad (9)$$

3.4 *Prior distributions*

We need to specify a prior distribution for all unknown parameters in the models (3)–(5) and (8). The prior distribution for the parameters is specified as in Buser et al. (2010a). There are two types of parameters. The parameters μ, $\Delta\mu$, β_i, $\Delta\beta_i$, γ, $\Delta\gamma$, γ_i and $\Delta\gamma_i$ are related to the means of the normal distributions. It is a common practice to assume normal distributions for such parameters since this simplifies the computations (e.g. see Gelman et al. 2003). The other parameters σ^2, q^2, b_i^2, and $q_{b_i}^2$ are related to the variances or multiplicative changes of the variances. It is common practice to work with the precisions (the inverses of the variances) and choose Gamma distribution as the prior distributions for such precision parameters (Gelman et al. 2003). Thus, as in Buser et al. (2009, 2010a), we consider the precision σ^{-2} as unknown. The same approach is also taken for the multiplicative factors q, b_i and q_{b_i}.

The vector of the unknown parameters is

$$\Theta = (\mu, \Delta\mu, \beta_i, \Delta\beta_i, \gamma, \Delta\gamma, \gamma_i, \Delta\gamma_i, \sigma^{-2}, q^{-2}, b_i^{-2}, q_{b_i}^{-2}; i = 1, \ldots, N_m).$$

The parameter κ could also be added to Θ and estimated from the data; see Buser et al. (2010a). However, for the cross-validation approach, the bias parameter κ is considered as known (auxiliary) parameter and not included to Θ.

All of these parameters are assumed to be mutually independent. Therefore only the marginal prior distributions, or more precisely the parameters of the Gaussian and Gamma distributions, have to be specified. The parameters are presented in Table 3. For μ, $\Delta\mu$, β_i, γ, $\Delta\gamma$, γ_i, q and b_i, the parameters are chosen such that the distributions are almost flat and only values that are very far from the physical plausibility are excluded. Thus the posterior distribution for these parameters is mainly determined by the likelihood (i.e., the data). However, the parameters $\Delta\beta_i$, $\Delta\gamma_i$ and q_{b_i} are different due to an identifiability problem. For example, if these parameters are allowed to vary significantly, a large value in $\Delta\mu$, $\Delta\gamma$ and σ could be compensated by opposite model bias changes $\Delta\beta_i$ and q_{b_i}. To overcome the issue, the values of $\Delta\beta_i$ and $\Delta\gamma_i$ are assumed to be small

and the values of the q_{b_i} to be close to unity. Therefore, narrow (more informative) distributions are chosen for the bias change terms.

The assumption that $\Delta\beta_i$ and $\Delta\gamma_i$ are Gaussian and with a small variance is equivalent to the assumption that $\sum_i \Delta\beta_i^2$ and $\sum_i \Delta\gamma_i^2$ are small, conditions that are commonly used for regularisation in over-parameterized problems. On the other hand, we could consider parameters $v_i = \Delta\mu + \beta_i$ which would be identifiable. The Gaussian assumption given for β_i corresponds to the priori assumption that all v_i's are similar (highly correlated). See Buser et al. (2009) for more details.

3.5 *Integration of the posterior distribution*

An explicit or numerical integration of the high dimensional distributions is usually not possible. Therefore, as in Buser et al. (2010a), we use Markov Chain Monte Carlo (MCMC) to compute approximations for the densities $p(\Theta|\mathcal{D})$ and $p(Y_{0,t}|\mathcal{D})$. More specifically, we use the Gibbs sampler (see e.g. Gilks et al. 1996) to compute samples from the posterior distribution $p(\Theta|\mathcal{D})$. With the Gibbs sampler, a set of samples is generated as follows. We start with some initial value

$$\Theta^{(0)} = (\mu^{(0)}, \Delta\mu^{(0)}, \beta_i^{(0)}, \Delta\beta_i^{(0)}, \gamma^{(0)}, \ldots)$$

and set $s = 0$. First, we draw a sample $\mu^{(s+1)}$ from the full conditional of μ (the distribution of μ conditioning all available information except the parameter itself):

$$p(\mu|\mathcal{D}, \Delta\mu^{(s)}, \beta_1^{(s)}, \beta_2^{(s)}, \ldots) \propto p(\mu, \Delta\mu^{(s)}, \beta_1^{(s)}, \beta_2^{(s)}, \ldots|\mathcal{D}). \tag{10}$$

We continue to the next parameter and draw $\Delta\mu^{(s+1)}$ from the full conditional of $\Delta\mu$

$$p(\Delta\mu|\mathcal{D}, \mu^{(s+1)}, \beta_1^{(s)}, \beta_2^{(s)}, \ldots) \propto p(\mu^{(s+1)}, \Delta\mu, \beta_1^{(s)}, \beta_2^{(s)}, \ldots|\mathcal{D}). \tag{11}$$

The procedure is continued until all parameters have been updated one at a time and we have $\Theta^{(s+1)}$. We increase $s \leftarrow s+1$ and repeat the sample generation N_s times to obtain a set of samples $\{\Theta^{(s)}\}_{s=1}^{N_\mathrm{s}}$. It is common to ignore a number of samples at the beginning (burn-in period) due to the fact that it takes time to converge to the stationary distribution $p(\Theta|\mathcal{D})$. For more details, see e.g. Gilks et al. (1996).

The generated samples give a discrete approximation for the posterior distribution

$$p(\Theta|\mathcal{D}) \approx \frac{1}{N_\mathrm{s} - N_\mathrm{b}} \sum_{s=N_\mathrm{b}+1}^{N_\mathrm{s}} \delta(\Theta - \Theta^{(i)}) \tag{12}$$

where N_b is the length of the burn-in period and δ is the Dirac delta distribution. By Eqs. (12) and (2), the distribution of $Y_{0,t}$ given the data $\mathcal{D}$ has an approximation which can be formally written as

$$\begin{aligned} p(Y_{0,t}|\mathcal{D}) &\approx \frac{1}{N_\mathrm{s} - N_\mathrm{b}} \sum_{s=N_\mathrm{b}+1}^{N_\mathrm{s}} \int p(Y_{0,t}|\Theta)\delta(\Theta - \Theta^{(i)})\,\mathrm{d}\Theta \\ &= \frac{1}{N_\mathrm{s} - N_\mathrm{b}} \sum_{s=N_\mathrm{b}+1}^{N_\mathrm{s}} p(Y_{0,t}|\Theta^{(s)}). \end{aligned} \tag{13}$$

By (5), the distributions $p(Y_{0,t}|\Theta^{(s)})$ are Gaussian with the mean $\mu^{(s)} + \Delta\mu^{(s)} + (\gamma^{(s)} + \Delta\gamma^{(s)})(t - T_0)$ and the variance $\sigma^{(s)}q^{(s)}$. Hence, the conditional density of $Y_{0,t}$ can be approximated with a sum of the Gaussian densities $p(Y_{0,t}|\Theta^{(s)})$.

In general, the sampling from the full conditionals of $p(\Theta|\mathcal{D})$ is carried out numerically by evaluating the full conditionals on a grid and then sampling from this discrete distribution. The numerical sampling requires a large number of evaluations of the posterior distribution and a computational implementation can be very slow. However, fortunately in our case the distributions are rather simple and it is a straightforward (but tedious) task to check that the full conditionals, except the conditional of b_i^{-2}, are either Gaussian or Gamma distributions.[1] The sampling from these distributions can be carried out directly using existing random number generators, which decreases the computational costs significantly. Sampling from the full conditional of b_i^{-2} is carried out numerically in our implementation, but other approaches such as accept-reject sampling or a Metropolis update for b_i^{-2} can also be used (see e.g. Gilks et al. 1996).

3.6 Cross-validation approach

The cross-validation is a model validation technique, in which the basic idea is to partition the data into two subsets: *training* set and *testing* set. In the validation, the training set is considered as *known*, available data and it is used to train the model (basically this may involve estimation of some parameters in the model using the data in the training set). The data in the testing set is considered as *unknown* scenario (future) data, which we wish to predict using the trained model. The performance of the model can be measured by comparing the model predictions to the testing data by using some metric such as the mean square error.

[1] By the Bayes formula (1), the posterior distribution $p(\Theta|\mathcal{D})$ is the product of the likelihood distribution (9) and the prior distributions given in Table 3. For example, one can see that (10) and (11) are Gaussian densities by re-organizing the terms.

Commonly, to reduce variability, the training and testing is repeated using different partitioning.

In this paper, we use so called leave-one-out cross-validation for the selection of the bias parameter κ (see e.g. McQuarrie and Tsai 1998). A similar approach was also used in Räisänen et al. (2010) and Räisänen and Ylhäisi (2012) to evaluate the potential effects of non-uniform climate model weighting on the quality of climate change projections. The cross-validation is used to find an optimal value of the κ parameter in the hybrid model (8) in the sense that the distance between the climate model predictions and the (simulated) scenario climate is minimized with respect to a chosen metric. The analysis will be carried out separately for each SREX region and for summer and winter seasons.

The quality of a chosen bias model (or a choice of the parameter κ) can be measured using the cross-validation approach as follows. We choose the ℓ'th model to represent "the reality" such that temperatures during the control period are considered as the observational data of the historical period and the predicted temperatures for the scenario period are considered as "unknown data" of future temperatures. The Bayesian analysis is used to predict future temperatures that are then compared to the actual future temperatures of the ℓ'th model. For our primary measure of the distance between the predictions and the true climate, we use the Continuous Ranked Probability Score (CRPS) (see "Appendix"). This procedure is repeated by choosing all models as "the reality" one by one.

More specifically, the approach used in this paper can be presented as the following algorithm:

0. Fix the bias parameter κ.
1. Choose testing model $\ell \in \{1, 2, ..., N_{\mathrm{m}}\}$ and substitute the observation and the future (unknown) data to be

   $$X_{0,t} \leftarrow X_{\ell,t}, \quad Y_{0,t} \leftarrow Y_{\ell,t}$$

   and exclude $X_{\ell,t}$ and $Y_{\ell,t}$ from the set of climate model data $\mathcal{D}$.
2. Calculate the Bayesian multi-model predictions for $Y_{0,t}$ as described in Sect. 3.2. The resulting Markov chain approximates $p(Y_{0,t}|\mathcal{D})$ by Eq. (13).
3. Calculate $\mathrm{CRPS}_\ell(\kappa)$ (see "Appendix" for details).
4. Repeat steps 1-3 until all of the N_{m} models have been used as the testing model.
5. Compute the mean of CRPSs:

   $$\overline{\mathrm{CRPS}}(\kappa) = \frac{1}{N_{\mathrm{m}}} \sum_{\ell=1}^{N_{\mathrm{m}}} \mathrm{CRPS}_\ell(\kappa).$$

   The mean $\overline{\mathrm{CRPS}}(\kappa)$ can be considered as a measure of the quality of the bias model (or the bias parameter κ). The optimal κ, in the sense of cross-validation, can be chosen by minimizing $\overline{\mathrm{CRPS}}(\kappa)$ with respect to κ or, in practice, by repeating the above procedure for a discrete set of κ's and choosing the κ with the smallest $\overline{\mathrm{CRPS}}(\kappa)$. Later we call this κ as the cross-validated κ.

The cross-validation approach can also be applied using other scores. In this study, we have also carried out computations using the logarithmic score $\mathrm{LOG} = -\log p(Y_{0,t}^{\mathrm{obs}}|\mathcal{D})$. The logarithmic score is another strictly proper score; see Gneiting and Raftery (2007). A comparison between these two alternative scores (CRPS and LOG) will be presented in Sect. 4.

4 Results

The results are based on the Bayesian analysis described in Sect. 3. As in Buser et al. (2009), the length of Markov chains was chosen to be $N_{\mathrm{s}} = 500{,}000$ samples where the first $N_{\mathrm{b}} = 100000$ were discarded as the burn-in period. Then the sets of samples were thinned by taking only every tenth sample of the generated chain, i.e., the lengths of the final sample sets $\{\Theta^{(s)}\}$ were 40000.

To check the convergence of the chains, we calculated the effective sample sizes of the chains that are based on approximative autocorrelation functions and represent the number of (effectively) independent samples in the chains. All of the chains have at least 200 effective samples and only 0.04 % of the all (more than 1.2 million) chains have less than 500 effective samples. We also studied convergence by computing several chains for the same problems. The estimates of the parameters (the mean of the chains) were altered only slightly between subsequent MCMC runs. For example, the estimate of $\Delta\mu$ for NEU/DJF was altered less than 0.2 % between subsequent runs.

4.1 Cross-validation

In the cross-validation, CRPSs were computed for $\kappa = 0, 0.1, 0.2, \ldots, 0.9, 1$. The analysis was carried out separately for each SREX region and for the DJF (Northern Hemisphere winter) and JJA (Northern Hemisphere summer) seasons. The cross-validated values of κ for each region are shown in Figs. 1 (DJF) and 2 (JJA). Figure 3 shows $\mathrm{CRPS}_\ell(\kappa)$ for each "reality model" as a function of κ including also the mean values $\overline{\mathrm{CRPS}}(\kappa)$ (shown for the selected regions, see Fig. S1–S7 in the supplementary material for all regions).

The results show that the cross-validated κ can differ significantly between the summer and winter season, as well as between the regions.

However, regions with similar climate also have quite similar values for the bias parameter κ. Importantly, CRPS varies substantially between individual verifying models, as

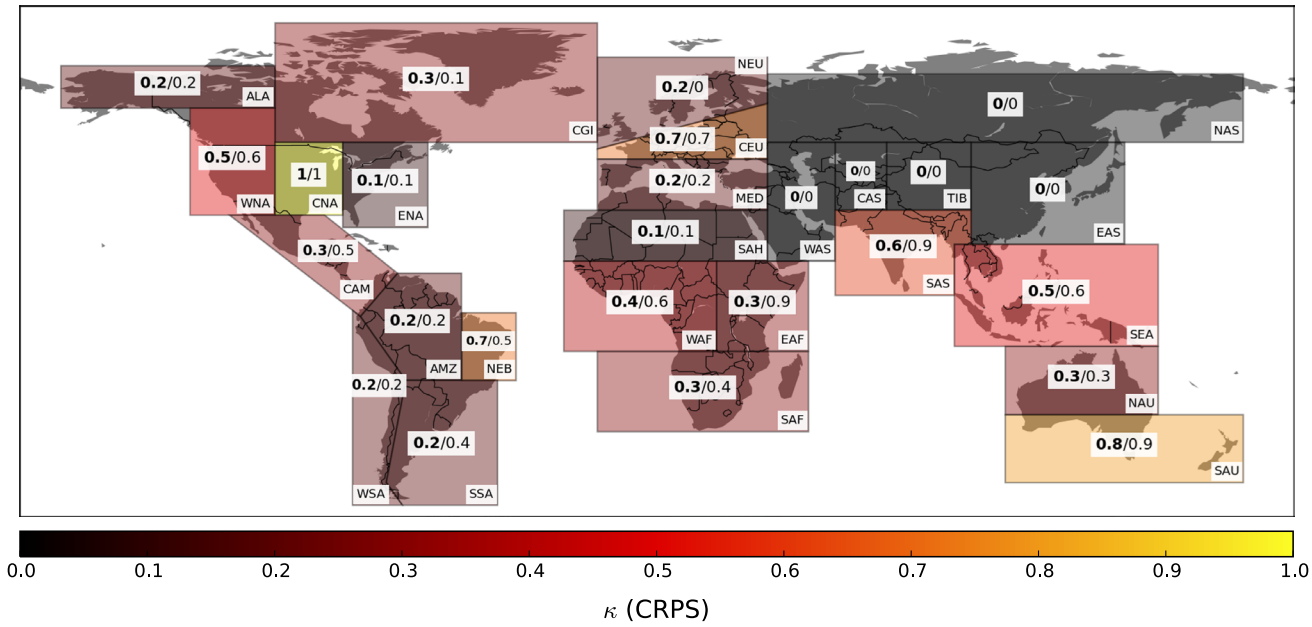

Fig. 1 The cross-validated parameters κ that minimize CRPS (the *boldface font*) and LOG (the *regular font*) for DJF (Northern Hemisphere winter). The *color coding* corresponds to the CRPS values

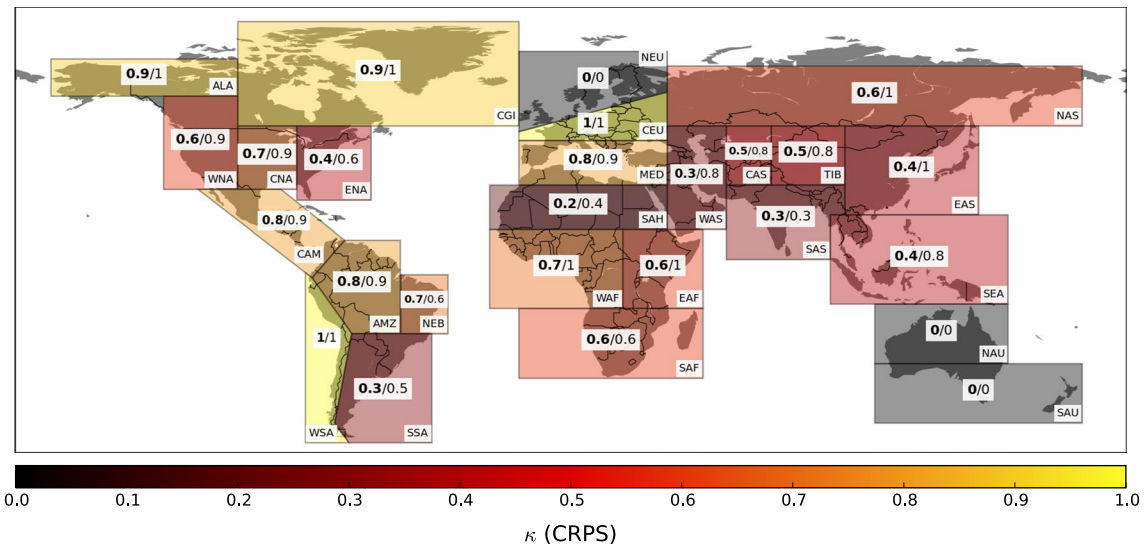

Fig. 2 The cross-validated bias parameters κ that minimize CRPS (the *boldface font*) and LOG (the *regular font*) for JJA (Northern Hemisphere summer). The *color coding* corresponds to the CRPS values

can be seen from Fig. 3. In general, it is largest for models which are outliers in terms of the simulated climate change and whose future climate is therefore difficult to predict with our statistical model. Furthermore, the value of κ that is optimal in the ensemble mean sense does not always minimize CRPS for the individual models. This variation is important because only one realization of the future climate will be observed in the real world. Comparing with this large inter-model variation, the mean CRPS changes in some cases (e.g., NEU in DJF) negligibly with κ. This indicates that the ensemble gives no guidance on the choice of κ in such cases, but does not unfortunately guarantee that the actual projection would be insensitive to κ (see NEU, DJF in Fig. 5 below). In other cases (e.g., CEU in JJA), the variation of the mean CRPS with κ might still have some practical significance, suggesting that values of κ that are close to the cross-validated optimum are more likely to lead to good climate projections than values far from this optimum.

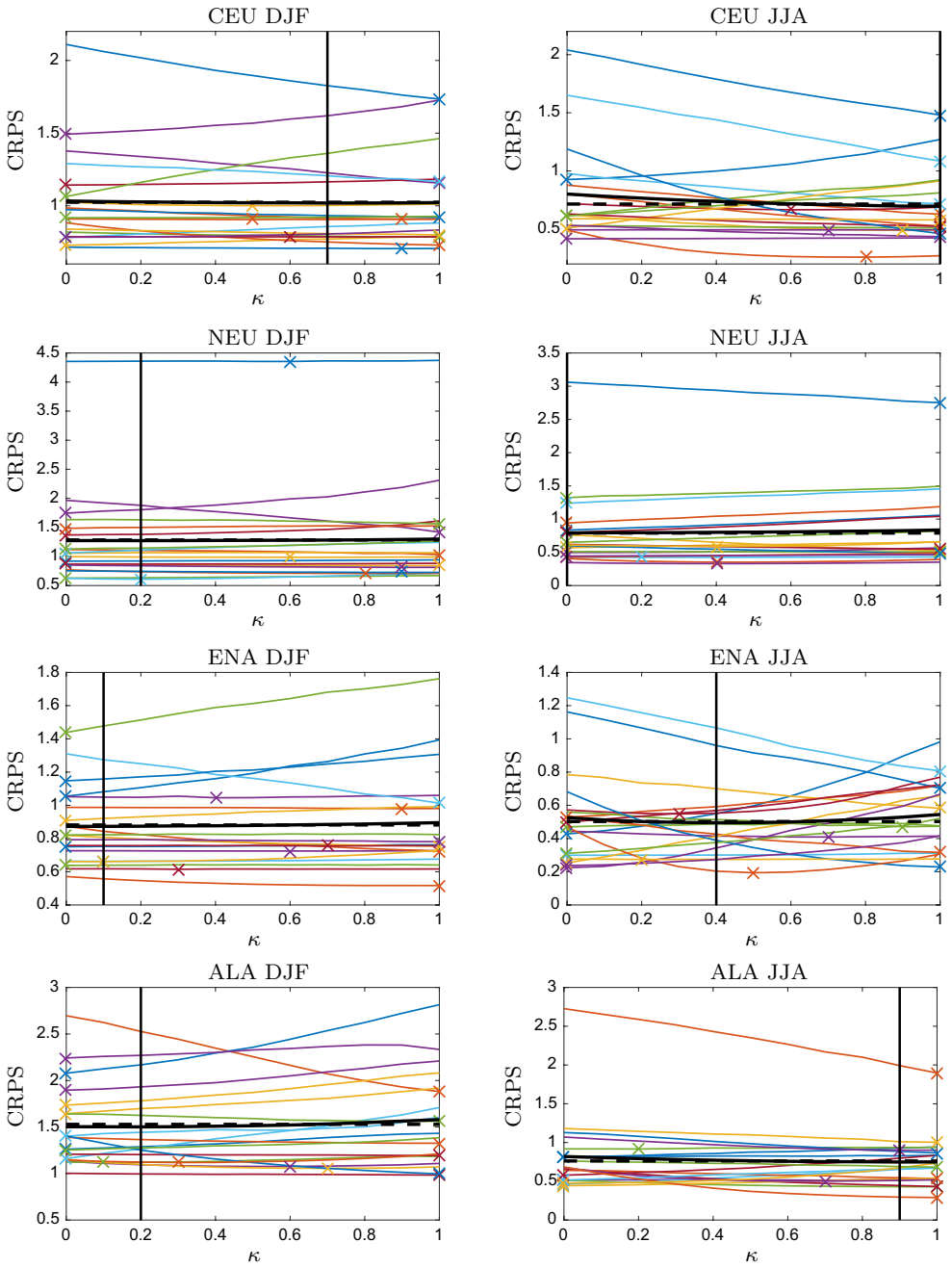

Fig. 3 CRPS values as a function of κ. *Coloured solid lines* are $\mathrm{CRPS}_\ell(\kappa)$ for each "reality model" ℓ. The *thicker black lines* correspond to the mean of the values. The *crosses mark* the minimum value for the each "reality model" ℓ and the *vertical black line* marks the cross-validated κ. The *dashed lines* corresponds to the mean of the CRPS value when the predictions are computed using the approach of Buser et al. (2010a) in which κ is also estimated

We also studied the uncertainty in the MCMC approximations by repeating the cross-validation approach several times. The mean CRPS curves for regions CEU, NEU and CGI in DJF for subsequent MCMC runs are shown in Fig. S8 in the supplementary material. We note that the optimal values may vary a step (± 0.1) between subsequent MCMC runs, or even two steps (± 0.2) if the mean CRPS is very flat (e.g. CGI in DJF).

To compare different metrics, we have also carried the cross-validation analysis using the logarithmic score. The values of κ that minimize the CRPS and the logarithmic score are both shown in Figs. 1 and 2. Compared to CRPS, the logarithmic score tends to favor higher values of κ. This may be because (i) the logarithmic score is less tolerant against verifying observations that fall far in the

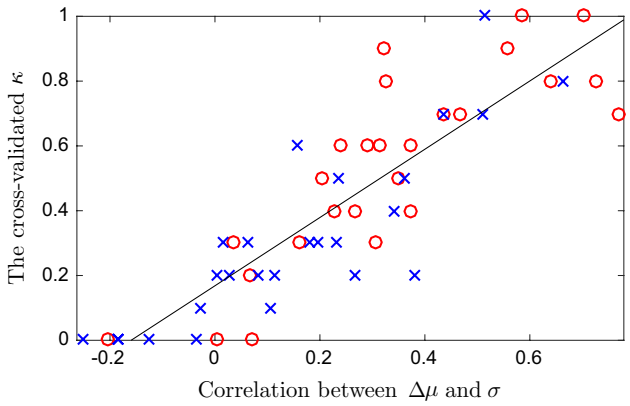

Fig. 4 The cross-validated κ versus the correlation between temperature change $\Delta\mu$ and variability σ for each region. The points corresponding to DJF are marked with *crosses* (x) and JJA with *circles* (o). The *black line* is a linear fit to all of the points. The correlation coefficient $r = 0.87$

tails of the predicted distribution, and (ii) the frequency of such cases tends to be reduced by increasing κ because the predicted distributions become wider, as can be seen from Fig. 5 below (to be discussed in more depth in Sec. 4.2).

Based on the assumptions behind the constant bias and constant relation models, one could hypothesise that large values of κ would correspond to regions and seasons in which there is a strong correlation between the temperature increase between the control and scenario periods and the variability of the simulated temperatures during the control period. To test this hypothesis, we calculated (rough) estimates for correlations from the raw climate model output data using the following simple procedure. The temperature increase in the i'th model $\Delta\mu_i$ is estimated as the difference between the means of the temperatures of the scenario period $Y_{i,1}, \ldots, Y_{i,45}$ and the control period $X_{i,1}, \ldots, X_{i,45}$. Furthermore, we calculate interannual variability in the i'th model by removing a linear trend from the temperatures $X_{i,1}, \ldots, X_{i,45}$ (using a linear least-squares fit) and by calculating σ_i as the ensemble standard deviation of the detrended data. This gives a set of pairs (μ_i, σ_i) for each model $i = 1, \ldots, 19$ from which the correlation can be calculated. The procedure is carried out separately both for each region and for the summer and winter seasons. Figure 4 shows the correlations for each region as a function of the cross-validated κ of the region. The figure shows that there is a linear dependency between the correlation and the cross-validated κ, as suspected.

The cross-validation approach can also be carried out for the approach proposed in Buser et al. (2010a), in which the parameter κ is included to the model parameters Θ and estimated from the data $\mathcal{D}$ along with all other parameters. Thus, instead of fixing κ, we compute the prediction $p(Y_{0,t}|\mathcal{D})$ using the approach of Buser et al. (2010a) and compute CRPS values for these predictions. Figure 3 and Figs. S1–S7 in the supplementary material also include the means of the CRPS values computed using this approach. The average cross-validated CRPS for Buser's approach tends to be slightly above the corresponding value for the cross-validated "optimal" κ. However, the difference is generally small. Also note that this comparison is not fully fair because the optimal value was chosen "after the fact", i.e. after the cross-validation. For some regions and seasons, the use of the cross-validated κ can produce better predictions in terms of CRPS, but the difference may not be significant.

4.2 Predictions from observational CRU data

To study the effect of the bias model to the predictions, we also computed multimodel predictions for different values of κ using the real CRU observational data (see Sect. 2.2). Bayesian multimodel predictions are computed as described in Sect. 3.2. Figure 5 shows the predictions for the mean temperature change $\Delta\mu$ between the control and scenario periods $\Delta\mu$ with 90 % confidence intervals (the intervals are estimated using the samples $\Delta\mu^{(i)}$) for a selection of areas as a function of κ (for the complete set of the estimates, see Figs. S9–S35 in the supplementary material). As can be seen, the bias model may have a significant effect on the predictions and also on the uncertainty intervals. However, the bias parameter κ does not have a significant effect to the estimates of the internal variability parameters σ and q and the trend parameters γ and $\gamma + \Delta\gamma$. Similar observations were also made by Buser et al. (2010a).

In some cases, $\Delta\mu$ either increases or decreases systematically with increasing κ, but this is not always the case. This likely depends on whether or not there is a systematic bias in the interannual variability in the models. For example, if the models simulate too strong interannual variability in an area, the constant relation framework ($\kappa = 1$) indicates that the models will also overestimate the long-term climate change. Therefore, $\Delta\mu$ becomes smaller than the temperature change simulated by the models (which is naturally close to $\Delta\mu$ for the constant bias case ($\kappa = 0$)). However, when the simulated interannual variability is close to that observed, $\Delta\mu$ remains close to the value directly simulated by the models even for large κ. As proposed in Buser et al. (2010a), the parameter κ can also be included to the model parameters Θ and estimated from the data $\mathcal{D}$ along with all other parameters. To compare the estimated κ with the cross-validated κ, we have computed Bayesian multi-model predictions using the approach described in Buser et al. (2010a) (the prior model for κ is chosen to be uniform on the unit interval). Figure 6 shows the probability density functions (PDF) $p(\kappa|\mathcal{D})$ (histograms) for the selected areas.[2] In most cases (exceptions

[2] Figures S35−S61 in the supplementary material show the estimates for each region and also estimates of the 2D joint histogram of the parameters $\Delta\mu$ and κ for each region.

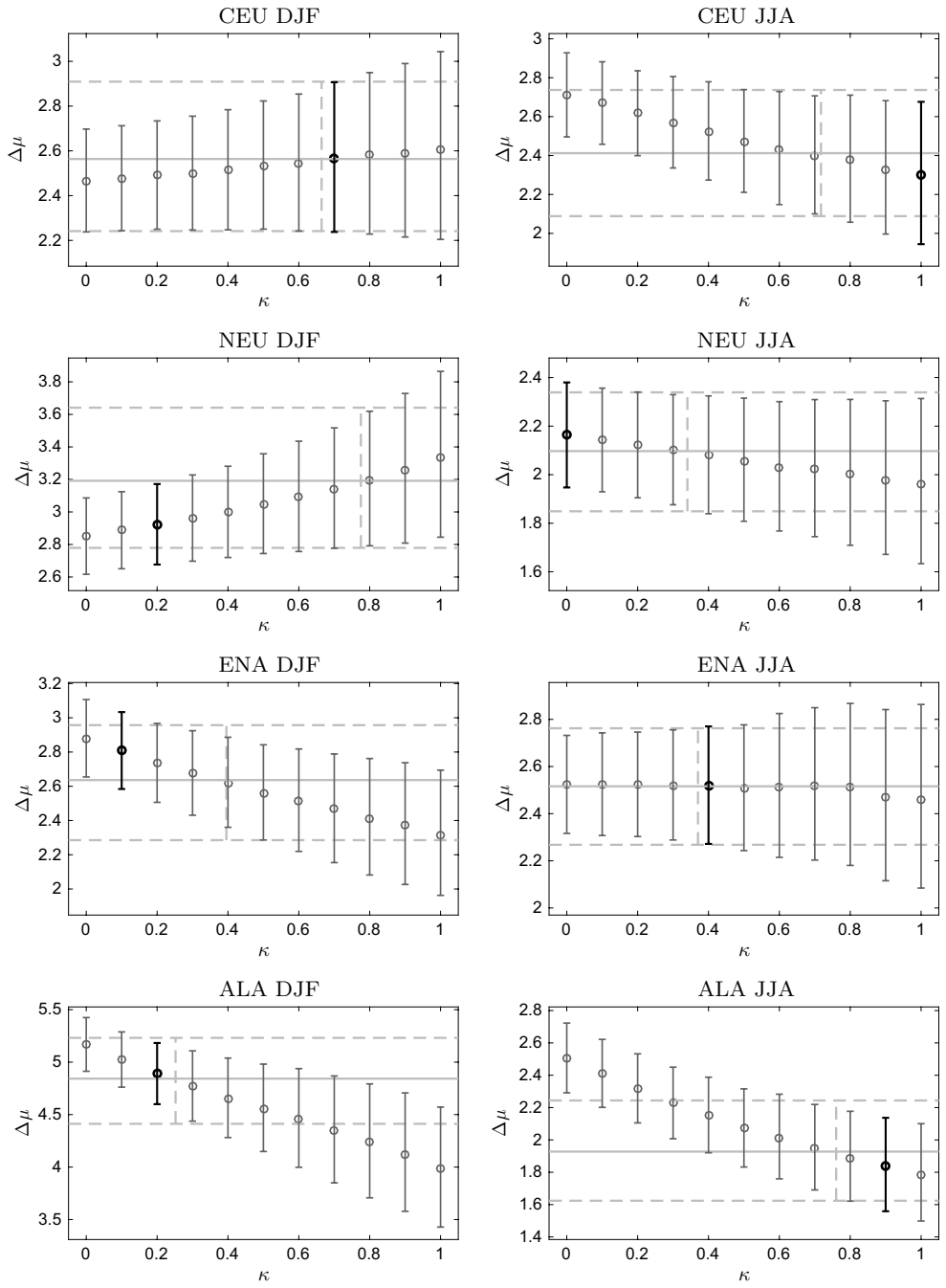

Fig. 5 The estimates of $\Delta\mu$ as a function of the bias parameter κ for a selection of areas (all areas are shown in the supplementary material). The *circle* corresponds to the means of the Markov chains $\{\Delta\mu^{(i)}\}$ and the *error bars* correspond to 90 % confidence limits. The estimates corresponding to the parameter κ obtained through the cross-validation are marked with *bold lines*. The figure also includes estimates of $\Delta\mu$ when κ is included as a parameter in the Bayesian approach: the *gray solid horizontal line* corresponds to the mean of $\Delta\mu$ and the *dashed lines* are 90 % confidence limits. The *dashed vertical gray line* marks the estimate of κ obtained as the mean of the chain

are discussed below), the cross-validated κ is reasonably close to either the mean or the maximum point of $p(\kappa|\mathcal{D})$. This can also be seen from Fig. 7 which shows a significant linear correlation between the estimated and the cross-validated κ. The parameters κ given by the cross-validation approach are in most cases slightly larger than the estimated parameters κ (both the mean and maximum points). In other words, the cross-validation approach has a slightly higher tendency towards the constant relation assumption.

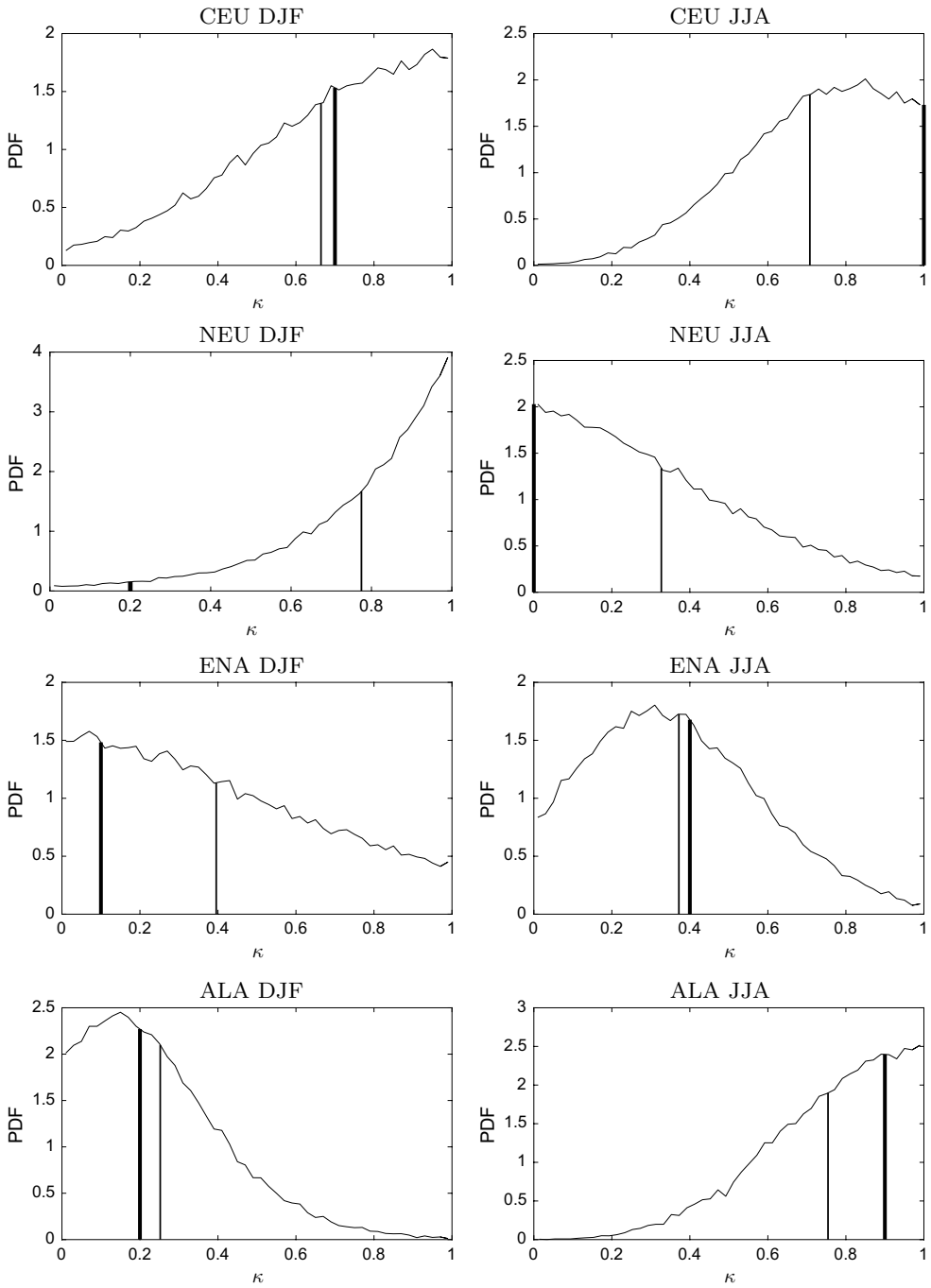

Fig. 6 The forecasted probability density function of κ when κ is included to the set of the model parameters Θ in the Bayesian analysis. The mean is shown with a *thin line* and the cross-validated κ with a *thick line*

The reason for the larger values might be the following. First, the prediction method gives too narrow predictive distributions, in the sense that too many verifying observations fall in the tails. For example, this can be seen from the rank histograms which show a significant accumulation of observations to the tails of the predicted distributions (see Fig. S62 in the supplementary material). This problem is reduced by increasing κ, which increases the width of the predictive distributions. On the other hand, the Buser et al. method tends to give wider predicted distributions by allowing for the uncertainty in κ. Thus, when κ is fixed without uncertainty, slightly higher values of κ are needed in our method to reduce the underdispersion of the predicted distribution. This underdispersion might also depend on the shape of the distributions assumed by the statistical model. Thus, it might be potentially alleviated by replacing the normal distributions in (3)–(8) with a distribution with heavier tails.

To compare the estimates of the temperature change $\Delta\mu$, Fig. 5 includes also the estimates of $\Delta\mu$ with 90 %

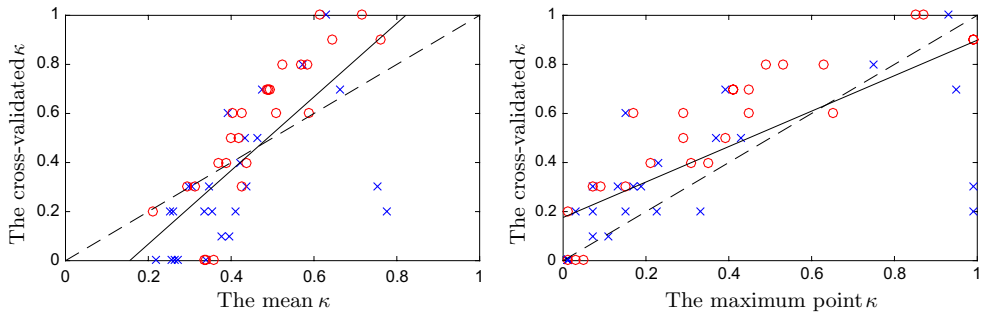

Fig. 7 *Left*: the estimated κ (the mean of the MC chain) compared to the cross-validated κ from the cross-validation analysis for each region. *Right*: the maximum points of forecasted PDFs $p(\kappa|\mathcal{D})$ (estimated using a histogram from the MC chain) versus the cross-validated κ from the cross-validation. The points corresponding to DJF are marked with *crosses* (x) and JJA with *circles* (o). The *solid lines* are linear fits to the points and the *dashed line* is $y = x$. The correlation coefficients are $r = 0.72$ (mean) and $r = 0.76$ (max)

uncertainty range when κ is included to the unknown parameters Θ. The means of $\Delta\mu$ are similar to the estimates of $\Delta\mu$ for a fixed κ which is near the mean of κ. As expected, however, the uncertainty interval of $\Delta\mu$ becomes slightly wider when κ is included as an unknown parameter. This is the case particularly when $\Delta\mu$ is substantially affected by κ.

There are two notable exceptions, NEU and CGI in winter (DJF), for which the cross-validation analysis favours much smaller values of κ than Buser's et al method. However, the mean $\overline{\text{CRPS}}(\kappa)$ is very flat in these cases, meaning that the cross-validated κ has a significant uncertainty. The forecasted PDFs for κ given by the Buser et al. (2010a) method are significantly concentrated on large values of κ. However, we found out that the estimates of κ given by the Buser et al. (2010a) method may depend substantially on the selection of the models that are included to the climate model outputs. For example, for NEU in DJF, the PDF of κ for the Buser et al. (2010a) method becomes flat if a single model (CMCC-CM) is excluded (Fig. S63). When this model is included (Fig. 6), by contrast, the method strongly favors large values of κ. This is because much larger warming is simulated by the CMCC-CM model than the others in NEU in DJF, and this large warming is much more difficult to reconcile with small than large values of κ (note the increase in $\Delta\mu$ with κ in this case in Fig. 5). The cross validation approach seems to be less sensitive to the CMCC-CM model: although the cross-validated κ is reduced from 0.2 to 0 when this model is excluded (Fig. S63), $\overline{\text{CRPS}}(\kappa)$ still remains flat. For similar reasons, the Buser et al. method strongly favours large κ for CGI in DJF when including all models (Figs. S40), but not when excluding CanESM2 (Fig. S63). On the other hand, the cross-validated κ can also be sensitive to the selection of the models: for example, if FGOALS-g2 (the ascending curve in Fig. 3) is excluded from the analysis for NEU DJF, the cross validated κ increased from 0.2 to 0.7 (although the

Table 4 The first and second columns list the region and its index

Area	#	DJF %	JJA %	Christensen and Boberg (2013) (%)
ALA	1	5.4	**26.7**	7.7
CGI	2	2.9	**−15.3**	5.2
WNA	3	2.5	**2.6**	19.4
CNA	4	6.0	**11.7**	9.1
ENA	5	2.5	**0.3**	6.8
CAM	6	5.0	**0.5**	1.6
AMZ	7	5.7	21.0	16.1
NEB	8	14.9	15.2	22.4
WSA	9	**0.2**	−10.8	−1.2
SSA	10	**4.0**	3.3	−7.2
NEU	11	0.0	**0.0**	12.1
CEU	12	−4.6	**15.2**	19.4
MED	13	−0.8	**−11.9**	15.6
SAH	14	−0.3	**5.1**	18.2
WAF	15	3.9	3.8	−7.3
EAF	16	−1.0	20.8	0.3
SAF	17	**4.7**	12.5	−0.1
NAS	18	0.0	**11.5**	7.0
WAS	19	0.0	**0.9**	19.2
CAS	20	0.0	**8.6**	11.0
TIB	21	0.0	**8.1**	17.4
EAS	22	0.0	**2.2**	14.2
SAS	23	18.2	**17.0**	20.8
SEA	24	−2.3	4.2	5.2
NAU	25	**10.4**	0.0	13.4
SAU	26	**−8.7**	0.0	3.6

The third and fourth columns show the relative decrease $1 - \Delta\mu_{\kappa_{\text{CV}}}/\Delta\mu_{\kappa=0}$ (in percent) for the northern winter (DJF) and the northern summer (JJA), respectively. The boldface font indicates the warmer season in the area, except when the difference between DJF and JJA is very small. The last column represents values of the relative change presented in Christensen and Boberg (2013) for the warmest 50 % of months during 2071–2100

mean $\overline{\mathrm{CRPS}}(\kappa)$ still remains flat). Thus, the largest discrepancies between our cross-validation method and the Buser et al. method appear to be associated with different sensitivities to outlying models.

In an earlier study, Christensen and Boberg (2012, 2013) used regression between simulated present-day temperature variability and future temperature changes to infer how multi-model mean, constant-bias-model temperature change estimates should be adjusted to account for biases in simulated variability. In most regions, the adjustment reduced the estimated warm-season (warmest 50 % of months) warming. We conducted a similar analysis, comparing $\Delta\mu$ between the cross-validated κ and the constant-bias model ($\kappa = 0$) (Table 4). We also find that the warming is commonly reduced, particularly in JJA. The magnitude of this change, in some cases up to over 20 %, is similar to that reported by Christensen and Boberg (2013). However, at the level of individual regions, there is no detailed agreement with their study. This relates probably both to differences in the ensembles used and those in methodology. In particular, Christensen and Boberg used the data for the warmest 50 % of months (for 2071-2100) simultaneously, thus including in their variability analysis a contribution from the seasonal cycle in addition to interannual variability. Here, only interannual variability is considered.

5 Conclusions and discussion

This paper considered Bayesian multi-model predictions or computation of the predictions for temperature change between control and scenario periods using an ensemble of model outputs. We have developed a cross-validation approach to find, in a specific sense, an optimal value for the parameter κ in the bias model proposed by Buser et al. (2010a). The key of the approach is to select one of the model outputs as "the reality" and predict the output using other climate model output. The predictions are then compared to the actual outputs of the "reality" model and the difference is measured using the CRPS. The procedure is repeated by selecting each climate model output as "the reality" one by one. The approach can also be applied to predictions of other variables such as precipitation.

The cross-validation approach was applied to the CMIP5 dataset by considering separately all IPCC SREX regions (Seneviratne et al. 2012) and summer and winter seasons. The results show that the cross-validated bias parameter can vary significantly between the regions and seasons. This gives an indication that the pre-specification of a fixed bias model such that the commonly known constant bias assumption (corresponds to $\kappa = 0$) or the constant relation assumption proposed by Buser et al. (2009) ($\kappa = 1$) should be in principle avoided.

Buser et al. (2010a) proposed an approach to estimate the bias parameter κ by including the parameter as one of the unknown parameters in the Bayesian multi-model approach. Our results show that there is a significant correlation between the estimated κ and the cross-validated κ calculated using the proposed cross-validation approach. However, comparing to the estimated values of κ, the cross-validated parameters κ are slightly larger favouring the constant relation assumption. These slightly larger values could perhaps be caused by too narrow predictive distributions, which are compensated by increasing κ in the cross-validation analysis. This may indicate that the uncertainty is underestimated in the estimation which may be caused by, for example, too narrow prior distribution.

For several regions, the mean CRPS in cross validation depends only very weakly on κ. This indicates that the cross-validated κ may be very uncertain. This could indicate that there is no single value for κ that would be suitable to model bias changes with all of the climate models included to the inference.

There were also two notable exceptions for which our method and Buser et al's method give significantly different results. Namely, for the regions NEU and CGI in winter, our cross-validation analysis results in very small values of κ preferring constant bias assumption, but Buser's et al method prefers large values of κ. On the other hand, we found out that the results of Buser's method are significantly changed if we exclude climate models predicting large winter warming. When such models are excluded, the difference of the approaches is greatly reduced.

Due to all these complications, our general conclusion is that predictions of future climate change should be preferably computed using all approaches available (e.g. the method by Buser et al. (2010a) and our cross-validation method). If all of the methods give similar predictions, the predictions could be trusted with more support. However, if the approaches yield significantly different predictions, the causes of the discrepancies should be investigated and studied further.

Finally we note that the CMIP5 dataset was chosen due to the large number of output of different models. However, due to the relatively low spatial resolution of many of these models, we found it prudent to only present the projections at the scale of the SREX regions. To obtain more spatially detailed predictions of temperature (or e.g. precipitation) change, the approach could be applied to regional climate model data (as in Buser et al. 2009, 2010a) or high-resolution general circulation model data if a large ensemble of such model output becomes available.

Acknowledgments This work has been supported by strategic funding of the University of Eastern Finland and funding of Academy of Finland (application numbers 213476, 250215 AND 272041, Finnish Programme for Center of Excellence in Research 2006–2011, 2012–2017 AND 2014–2019). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 2 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Appendix

Continuous ranked probability score

Let x be a scalar variable (e.g. 2 m-temperature). Suppose that a probability function forecast of x is given by $p(x)$ and we have also an observation x^{obs} of the x. The Continuous Ranked Probability Score (CRPS) (Stanski et al. 1989; Hersbach 2000; Candille and Talagrand 2005; Grimit et al. 2006) is defined as

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[P^{\mathrm{pred}}(x) - P^{\mathrm{obs}}(x) \right]^2 \mathrm{d}x \tag{14}$$

where $P^{\mathrm{pred}}(x) = \int_{-\infty}^{x} p(x')\,\mathrm{d}x'$ is the cumulative distribution function of $p(x)$ and P^{obs} is the cumulative distribution function for the observation:

$$P^{\mathrm{obs}}(x) = \begin{cases} 0, & \text{if } x < x^{\mathrm{obs}} \\ 1, & \text{if } x \geq x^{\mathrm{obs}} \end{cases} = H(x - x^{\mathrm{obs}}).$$

where H is the Heaviside function ($H(x) = 1$ if $x \geq 0$ and 0 otherwise).

In this paper, the "distance" between the predictions of the future temperatures $Y_{0,t}$ (when ℓ'th model is taken as the "truth" in the cross-validation) and the actual future temperatures $Y_{\ell,t}$ is measured using the CRPS. The (total) CRPS score is calculated as the mean of CRPSs calculated for every year $t = 1, \ldots, T$. By Eq. (13), the prediction distribution is a (weighted) sum of Gaussian distributions and the CRPS for such Gaussian mixture model can be calculated in closed form using an expression given in Grimit et al. (2006).

References

Bellprat O, Kotlarski S, Lüthi D, Schär C (2013) Physical constraints for temperature biases in climate models. Geophys Res Lett 40:4042–4047

Boberg F, Christensen J (2012) Overestimation of mediterranean summer temperature projections due to model deficiencies. Nat Clim Change 2:433–436

Buser C, Künsch H, Lüthi D, Wild M, Schär C (2009) Bayesian multi-model projection of climate: bias assumptions and interannual variability. Clim Dyn 33(6):849–868. doi:10.1007/s00382-009-0588-6

Buser C, Künsch H, Schär C (2010a) Bayesian multi-model projections of climate: generalization and application to ensembles results. Clim Res 44(2–3):227–241

Buser C, Künsch H, Weber A (2010b) Biases and uncertainty in climate projections. Scand J Stat 37:179–199

Candille G, Talagrand O (2005) Evaluation of probabilistic prediction systems for a scalar variable. Q J R Meteorol Soc 131:2131–2150

Christensen J, Boberg F (2012) Temperature dependent climate projection deficiencies in CMIP5 models. Geophys Res Lett 39(L24):705

Christensen J, Boberg F (2013) Correction to temperature dependent climate projection deficiencies in CMIP5 models. Geophys Res Lett 40:2307–2308

Christensen J, Boberg F, Christensen O, Lucas-Picher P (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. Geophys Res Lett 35(L20):709

Gelman A, Carlin J, Stern H, Rubin D (2003) Bayesian data analysis, 2nd edn. Chapman & Hall/CRC, Boca Raton

Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice. Chapmann & Hall, Boca Raton

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. Am Stat Assoc 102:359–378

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q J R Meteorol Soc 132:2925–2942

Harris I, Jones P, Osborn T, Lister D (2014) Updated high-resolution grids of monthly climatic observations the CRU TS3.10 dataset. Int J Climatol 34(3):623–642

Heaton M, Greasby T, Sain S (2013) Modeling uncertainty in climate using ensembles of regional and global climate models and multiple observation-based data sets. SIAM/ASA J Uncertain Quantif 1:535–559

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570

Ho C, Stephenson D, Collins M, Ferro C, Brown S (2012) Calibration strategies: a source of additional uncertainty in climate change projections. Bull Amer Meteor Soc 93:21–26

Jolliffe I, Stephenson D (eds) (2011) Forecast verification: a practitioner's guide in atmospheric science, 2nd edn. Wiley, New York

Kerkhoff C, Künsch H, Schär C (2014) Assessment of bias assumptions for climate models. J Clim 27(17):6799–6918

Maraun D (2012) Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. Geophys Res Lett 39(L06):706

McQuarrie A, Tsai CL (1998) Regression and time series model selection. World Scientific, Singapore

Räisänen J, Ylhäisi J (2012) Can model weighting improve probabilistic projections of climate change? Clim Dyn 39:1981–1998

Räisänen J, Ruokolainen L, Ylhäisi J (2010) Weighting of model results for improving best estimates of climate change. Clim Dyn 35:407–422

Seneviratne S, Nicholls N, Easterling D, Goodess C, Kanae S, Kossin J, Luo Y, Marengo J, McInnes K, Rahimi M, Reichstein M, Sorteberg A, Vera C, Zhang X (2012) Changes in climate extremes and their impacts on the natural physical environment. In: IPCC (ed) Managing the risks of extreme events and disasters to advance climate change adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), pp 109–230

Smith R, Tebaldi C, Nychka D, Mearns L (2009) Bayesian modeling of uncertainty in ensembles of climate models. J Am Stat Assoc 104:97–116

Stanski H, Wilson L, Burrows W (1989) Survey of common verification methods in meteorology. Research report 89-5, Atmospheric Environment Service Forecast Research Division, Canada

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 365:2053–2075

Tebaldi C, Sansó B (2009) Joint projections of temperature and precipitation change from multiple climate models: a hierarchical bayesian approach. J R Stat Soc A 172:83–106

Tebaldi C, Smith R, Mearns DNL (2005) Quantifying uncertainty in projection of regional climate change: a bayesian approach to the analysis of multimodel ensembles. J Clim 18:1524–1540

Thomson AM, Calvin KV, Smith SJ, Kyle GP, Volke A, Patel P, Delgado-Arias S, Bond-Lamberty B, Wise MA, Clarke LE, Edmonds JA (2011) RCP4.5: a pathway for stabilization of radiative forcing by 2100. Clim Change 109(1–2):77–94. doi:10.1007/s10584-011-0151-4

Weigel A, Liniger M, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260

Wilks D (2006) Comparison of ensemble-MOS methods in the Lorenz 96 setting. Meteorol Appl 13:243–256