Model averaging to combine simulations of future global vegetation carbon stocks

Journal:	*Environmetrics*
Manuscript ID:	env-08-0019.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Butler, Adam; Biomathematics and Statistics Scotland Doherty, Ruth; University of Edinburgh, School of Geosciences Marion, Glenn; Biomathematics and Statistics Scotland
Keywords:	Dynamic global vegetation model, Vegetation carbon, Bayesian Information Criterion, Climate change, Extrapolation

Model averaging to combine simulations of future global vegetation carbon stocks

Adam Butler†
Biomathematics & Statistics Scotland, Edinburgh, UK
Ruth M. Doherty
University of Edinburgh, UK
and Glenn Marion
Biomathematics & Statistics Scotland, Edinburgh, UK

Summary. We quantify the impact of climate model uncertainty upon predictions of future vegetation carbon stocks, for the period up to 2100, generated by a dynamic global vegetation model under a particular emissions scenario (SRES A2). Deterministic simulations are generated from the Lund-Potsdam-Jena (LPJ) model using climatic inputs derived from nine general circulation models (GCMs). Simulated changes between 1961-1990 and 2070-2099 range from +26gtC to +133gtC, and are in broadly good agreement with those obtained in a recent previous study using the LPJ model. Simulated values for the 20th century are also obtained by running LPJ with observed climate data, and this provides a baseline against which the other runs can be compared. Time series regression models are used to analyse the discrepancies between each of the GCM-based simulations and the baseline simulation, and a novel form of model averaging - in which we average not only across GCM-based simulations but also across models for each discrepancy - is then used to combine these into a single probabilistic projection for global stocks of vegetation carbon. Weights for the regression models are estimated in a simple post hoc way using BIC, and the weights for the GCMs are either estimated in the same way or else fixed to be equal. Estimating the GCM weights leads the predictions to be dominated by a single climate model and hence produces narrow predictive distributions. If GCMs are weighted equally then the predictive distributions are much more diffuse and span the full range of simulated values.

Keywords: Vegetation carbon ; Dynamic global vegetation model; General circulation model; Bayesian Information Criterion; Extrapolation; Climate change

1. Introduction

Policy makers and scientists are becoming increasingly interested in assessing the potential impacts of climate change upon physical, biological and socio-economic systems, and in using statistical methods to quantify these impacts in a probabilistic way. Probabilistic climate impact predictions enable users to account for scientific uncertainty and natural variability when making decisions about mitigation or adaptation strategies, and form a natural basis for risk assessment. They enable us, for example, to directly quantify the

†*Address for correspondence:* Adam Butler, Biomathematics & Statistics Scotland, James Clerk Maxwell Building, The King's Building, Edinburgh EH9 3JZ, United Kingdom; Telephone: +44 (0)131 650 4896; Fax: +44 (0)131 650 4901
E-mail: adam@bioss.ac.uk

probability that a particular threshold of change will be exceeded, and, when combined with an appropriate loss function, can provide a formal quantitative basis for decision making in the face of uncertainty (Räisänen and Palmer, 2001).

Vegetation plays a key role in the global carbon cycle, but the relationship between climate and vegetation carbon is complicated: vegetation carbon stocks are influenced by net primary production, heterotrophic respiration and plant mortality (Schaphoff et al., 2006; Doherty et al., 2008), which are, in turn, influenced by temperature, water stress and ambient CO_2 concentrations. In this paper we use a dynamic global vegetation model (DGVM) to assess the effect of climate uncertainty upon predictions of future global vegetation carbon stocks. Multiple simulations of past (20th century) and future (21st century) annual vegetation stocks are generated from the DGVM - these simulations are shown in Figure 1. Eighteen of the runs are generated using climate inputs derived from a set of nine different state-of-the-art General Circulation Models (GCMs) under a common emissions scenario (SRES A2). Note that the number of runs is larger than the number of GCMs because multiple runs are available for three of the GCMs (Table 1); the multiple runs ("ensembles") are obtained using different initial values or parameter values, and their inclusion is design to allow - albeit in a very limited way - for the presence of intra-GCM uncertainty. The final, "baseline", simulation is generated using gridded observational climate data (the CRU-TS-2.1 dataset; Climatic Research Unit, 2006; Mitchell and Jones, 2005), and hence only covers the 20th century.

Statistical interest lies in combining the disparate deterministic predictions of future vegetation carbon into a single probabilistic prediction. We know that GCMs provide a biased and noisy representation of the real climate, and so we begin our analysis by analysing the dynamic statistical properties of the bias associated with each of the GCM-based simulations. We do this by assuming - in the absence of observational data at appropriate temporal and spatial scales - that the baseline simulation provides the best available description of true vegetation carbon stocks for the 20th century, and using time series regression models to analyse the discrepancy between this run and each of the GCM-based simulations. The assumption that the baseline simulation provides an accurate description of 20th century vegetation carbon stocks depends upon (a) the accuracy of observational data for climate, soil type and atmospheric concentration of carbon dioxide during the 20th century and (b) the ability of the LPJ model to accurately encapsulate the processes that determine levels of global vegetation carbon; both of these issues have been extensively considered by previous authors, and in Section 2 we briefly review these studies.

For future years, throughout the 21st century, we do not know

(a) which GCM to use as a basis for prediction; or
(b) which regression model to use in describing the relationship between this GCM-based simulation and reality,

and both of these choices impact upon the predicted changes in vegetation carbon stocks. In Section 3 we therefore use a relatively simple *post hoc* form of likelihood-based model averaging (Buckland et al., 1997) to deal with these two sources of uncertainty. General statistical methodology for model averaging is well developed (Hoeting et al., 1999), and recent papers (Raftery et al., 2005; Berrocal et al., 2007; Sloughter et al., 2007) have shown that it can be used to provide a rigorous probabilistic framework for combining the predictions associated with a set of distinct deterministic models. The idea of averaging across both deterministic and statistical models appears to be novel, however, and through this

development we are able to simultaneously quantify two important elements of predictive uncertainty.

The same methodological approach could also potentially be used in other environmental and ecological applications - to combine predictions of climate variables generated directly by GCMs, for example, or to assess the impact of climate uncertainty upon the output from a rainfall-runoff model. Estimating weights *post hoc* can be less efficient than estimating parameters and weights simultaneously (as in, for example, Raftery et al., 2005), but has the substantial practical advantage that it extends easily to more complicated situations - making it, for example, straightforward to deal with spatial, spatio-temporal or multivariate data, and making it trivial to incorporate additional deterministic simulations into the analysis.

2. Simulations of vegetation carbon

We use a dynamic global vegetation model to simulate values of global annual vegetation carbon throughout the 21st century. We focus upon a specific socio-economic scenario, SRES A2 (Nakicenovic and Swart, 2000), in which slow technological change, high population growth and regionally orientated economic growth result in a large increase in anthropogenic CO_2 emissions.

Stocks of vegetation carbon - the amount of above ground carbon, excluding litter - reflect the annual assimilation of Net Primary Productivity (NPP) by different plant types. Levels of NPP are, in turn, influenced by atmospheric CO_2 concentrations and temperature, with rising temperatures and CO_2 concentrations promoting plant growth. Rising temperatures also increase evapotranspiration and reduce soil water availability, however, and can thereby induce plant water stress (excessive atmospheric demand and/or low soil water availability). Plant uptake of CO_2 and loss of water are regulated through the leaf stomata, and in times of water stress plants will act to prevent water loss by reducing stomatal opening, thereby reducing rates of photosynthesis and (hence) plant growth. Changes in assimilated vegetation carbon in response to climate change are therefore potentially complicated, since they depend upon the combined effects of temperature change, precipitation change and baseline climate.

2.1. The LPJ model: description and validation

The Lund-Potsdam-Jena (LPJ) dynamic global vegetation model (Sitch et al., 2003; Gerten et al., 2004) is a process-based biogeography-biogeochemistry model which simulates the spatio-temporal dynamics of terrestrial vegetation, together with land-atmosphere carbon and water exchanges. We use Version 1.2 of LPJ, which simulates potential natural vegetation and does not account for agricultural use.

The LPJ model inevitably provides a simplified and imperfect representation of the global ecosystem (in particular it is known to have difficulties in describing the roles of disturbance and nitrogen deposition; Magnani et al., 2007). This version of the LPJ model has, however, already been extensively validated for terrestrial carbon and hydrological exchanges and vegetation distribution (e.g. Sitch et al., 2003, Gerten et al., 2004, Zaehle et al., 2005, Hickler et al., 2006, Schaphoff et al., 2006), and the effects of structural uncertainty (Cramer et al., 2001; Smith et al., 2001) and parameter uncertainty (Zaehle et al., 2005) have also been explored. Comparisons of net ecosystem exchange against observational data from EUROFLUX sites have been found to give reasonable agreement (Sitch et al., 2003;

Zaehle et al., 2005). Hickler et al. (2006) found that modelled variation in NPP across a large number of sites - spanning several biomes - showed a strong correlation with estimates obtained from field measurements, although LPJ tended to simulate higher NPP than given in the EUROFLUX EMDI dataset for grasslands. Crucially, LPJ was found to realistically simulate the dominant plant functional types in most regions (Sitch et al., 2003; Schaphoff et al., 2006), and simulations of total vegetation carbon are therefore based on a realistic global distribution of vegetation types. Historical data on global vegetation carbon stocks simply do not exist - the only available data are of limited temporal and spatial extent - so it is not possible to directly assess the ability of LPJ to accurately reproduce real trends in vegetation carbon stocks at the global level.

2.2. Generation of simulated runs

The LPJ model requires input data on atmospheric CO_2 concentrations and soil texture, together with climatological data on temperature, precipitation and fractional cloud cover. In this paper we use the model to generate nineteen sets of simulations; each of these runs is generated using a different set of climate inputs, but the same CO_2 and soil texture data are used in all cases. Annual CO_2 concentrations are based on the Mauna Loa observational records up to the year 1990 (as in Schaphoff et al., 2006) and on predictions generated by the Bern-CC global carbon model under scenario SRES A2 for the period 1990-2100 (Houghton et al., 2001, Appendix II). Soil texture data are grouped into eight discrete classes, as in Gerten et al. (2004).

Eighteen of the runs are generated using nine different state-of-the-art General Circulation Models (GCMs; Table 1). GCMs are based on a fundamental set of physical equations, and represent current best understanding of the physical behaviour of the coupled atmosphere-ocean system. The simulations which we use were created as part of the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2007) and are available through the Program for Climate Model Diagnosis and Intercomparison (PCMDI, 2006). We consider only simulations generated under scenario SRES A2‡; under this scenario CO_2 concentrations are 836 ppmv in 2100 and mean global annual temperatures for the period 2070-2099 are equal to 16.2^{o}C (based on averaging across the nine GCMs; Table 2). Multiple runs are available for some of the GCMs: each of these so-called ensemble runs is based on a different set of initial conditions and/or parameter values, and the use of ensembles is designed to quantify the effects of intra-model uncertainty (which is often called "natural variability" or "internal model variability" in the climate literature). The number of available ensemble runs is small, however, so we are only able to do this in a very limited way. Soil and CO_2 data are interpolated to the appropriate spatial resolution for each GCM, and LPJ simulations are thereby generated at the native spatial scale associated with that climate model.

The remaining ("baseline") run from the LPJ model is generated, for the period 1900-2001 only, using observational climate data interpolated from individual weather station onto a $0.5^o \times 0.5^o$ grid. These gridded data come from the CRU-TS-2.1 dataset (produced by the Climatic Research Unit; Climatic Research Unit, 2006; Mitchell and Jones, 2005;

‡Note that GCM simulations of future climate are contingent on a particular scenario for future emissions of greenhouse gases and are, consequently, referred to as "projections" rather than "predictions" in the applied literature. We use the term "prediction" throughout this paper, in keeping with the standard statistical terminology, but stress the importance of appreciating that these predictions are conditional upon a particular set of socio-economic assumptions.

New et al., 2001); previous versions of the same dataset were used by previous authors in assessing the validity of simulations generated by LPJ (Sitch et al., 2003; Zaehle et al., 2005; Schaphoff et al., 2006; Hickler et al., 2006). A number of studies have been concerned with quantifying the accuracy of gridded climate data, but this is not a straightforward task because coverage varies substantially and irregularly across both space and time (e.g. Jones et al., 1997; Folland et al., 2001).

For each run of the LPJ model, a spin-up period of one thousand years is used to ensure that the initial values are in equilibrium; climatic inputs for this spin-up period are based on repeated use of climate data/simulations for the period 1850-1880. The LPJ model is run on a daily basis: vegetation carbon stocks represent the accumulated sum of daily NPP, and the modelled values therefore need to accurately reflect day-to-day and month-to-month variations in productivity - the ability to capture these variations will be important even when, as here, interest is solely in aggregated annual values. Daily values of the climate variables are generated from monthly values via linear interpolation; alternative, more realistic approaches, to temporal disaggregation could also be used (e.g. Gerten et al., 2004, Fowler et al., 2007).

2.3. Key features of the simulations

The nineteen simulations of global annual vegetation carbon stocks are shown in Figure 1. The global mean values are calculated by averaging across space, using weights proportional to the cosine of the latitude of each grid cell, after restricting attention only to those cells that contain at least 50% land. Table 2 summarises Figure 1 by comparing mean vegetation carbon stocks between the present day (as represented by a thirty period from 1961-90) and the end of the 21st century (2070-2099) for each GCM, and comparing these changes against corresponding changes in mean temperature and daily precipitation.

Using the gridded CRU climate data we simulate the mean level of global vegetation carbon to be 789.9 gtC over the thirty year period 1961-90 (Table 2). This is in agreement with the value (779 gtC) obtained by Schaphoff et al. (2006) for 1971-2000 using the same climate data, although both of these values are higher than those suggested by earlier studies: 466-654 gtC (Houghton et al., 2001) and 640 gtC (Cao and Woodward, 1998). The true global value is not very well known (Benjamin Smith, Pers. Comm, 2008).

It can clearly be seen (Figure 1) that predicted future trends in vegetation carbon stocks differ in both direction and magnitude between the GCMs. All of the simulation runs exhibit increases in the period up to 2050, but two of the GCMs (HadCM3 and HadGEM1) show subsequent decreases during the second half of the 21st century whilst simulated values from another (CNRM-CM3) are relatively stable from around 2080 onwards. Simulated changes in global mean vegetation carbon stocks from 1961-1990 to 2070-2099 range between +26gtC and +133 gtC.

In a previous study, Schaphoff et al. (2006) generated simulations from LPJ using five GCMs. All simulations were generated under the Is92a emissions scenario, in which ambient CO2 concentrations reach 703 ppmv in 2100 (Houghton et al., 2001) and the average global annual mean temperature in 2071-2100 is equal to 17.9^{o}C (based on averaging across the five GCMs that they consider). Schaphoff et al. (2006) ran a similar version of LPJ to that used in this paper, and four of the same GCMs were used in both studies (although Schaphoff et al., 2006 generally used earlier versions of these models). They reported changes in global mean vegetation carbon stocks from 1971-2000 to 2071-2100 that ranged between -8gtC and +151 gtC. Of the GCMs that were common to both studies Schaphoff et al.

(2006) reported changes of +151 gtC (CSIRO), +51 gtC (ECHAM5), +5 gtC (HadCM3) and -8 gtC (CGCM1), whilst we have identified changes of +108 gtC (CGCM3), +107 gtC (CSIRO), +89 gtC (ECHAM5) and +29 gtC (HadCM3). Previous studies using different DGVMs have suggested vegetation increases of 150-340 gtC between the present day and 2100 (Cramer et al., 2001; White et al., 1999). Finally, mean vegetation carbon totals for the period 1961-90 are rather lower that the 1971-2000 values given by Schaphoff et al. (2006).

3. Model averaging

Let the vector $\mathbf{y} = (\mathbf{y}_O, \mathbf{y}_P)$ denote the baseline simulation of annual global vegetation carbon. The values, $\mathbf{y}_O$, for the period of observation (1900-2001) are known (non-missing), whilst the values $\mathbf{y}_P$ for the period of prediction (2002-2100) are unknown (missing). Let $\mathbf{Y}$ denote the corresponding multivariate random variable, from which $\mathbf{y}$ is assumed to be a realisation.

The aim of our analysis will be to draw inferences about the predictive distribution of $\mathbf{Y}_P|(\mathbf{Y}_O = \mathbf{y}_O)$. It is important to note that we are therefore concerned with predicting the level of vegetation carbon that the LPJ model will simulate given future climate (or, to be more precise, given imperfect observations of future climate), rather than with predicting the actual level of vegetation carbon. It is only possible to draw inferences about the latter quantity if we are prepared to make the additional assumption that the LPJ model provides an accurate representation of the processes that determine levels of global vegetation carbon.

3.1. Methodology

3.1.1. Prediction using a single GCM-based simulation

Let $\mathbf{f} = (\mathbf{f}_O, \mathbf{f}_P)$ denote a single GCM-based simulation of annual global vegetation carbon, covering the same period as above. All values of $\mathbf{f}$ are known (non-missing).

We assume that $\mathbf{f}$ is related to the expected value of $\mathbf{Y}$ through the formula

$$\boldsymbol{\mu} := \mathbb{E}(\mathbf{Y}) = \mathbf{f} + \epsilon(\mathbf{x}; \boldsymbol{\theta})$$

where ϵ is a vector-valued function whose output depends upon the values of a known covariate $\mathbf{x}$ and one or more unknown parameters $\boldsymbol{\theta}$. The vectors $\mathbf{x}$ and $\boldsymbol{\mu}$ are of the same length as $\mathbf{f}$ and $\mathbf{y}$, and we assume here, and throughout this article, that all arithmetic operations of vectors are performed pointwise.

The function ϵ quantifies the systematic bias between the GCM-based simulation run $\mathbf{f}$ and the baseline simulation run $\mathbf{y}$. If the GCM-based run were unbiased, relative to the baseline run, then we would have $\epsilon(\mathbf{x}; \boldsymbol{\theta}) = 0$ and $\mathbb{E}(\mathbf{Y}) = \mathbf{f}$. An assumption of constant but non-zero bias would imply that $\epsilon(\mathbf{1}; \alpha) = \alpha\mathbf{1}$, where α is an unknown parameter whose value determines the sign and magnitude of the bias term. Alternatives might be to assume that the bias is a linear function of the model signal $\mathbf{f}$ (Raftery et al., 2005) or of time $\mathbf{t}$, or that it has a more complicated parametric form.

We further assume that the joint distribution of the residuals,

$$\mathbf{Z} := \mathbf{Y} - \mathbb{E}(\mathbf{Y}) = \mathbf{Y} - \boldsymbol{\mu} = \mathbf{Y} - \mathbf{f} - \epsilon(\mathbf{x}; \boldsymbol{\theta})$$

can be described by a probability density function g whose form depend upon the values of one or more unknown parameters $\boldsymbol{\psi}$, so that

$$\mathbf{Z} \sim g(\bullet; \boldsymbol{\psi}),$$

and $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$. Under this model, the predictive distribution for $\mathbf{Y}_P|(\mathbf{Y}_O = \mathbf{y}_O)$ is equal to

$$\begin{aligned}\mathbb{P}(\mathbf{Y}_P = \mathbf{y}_P|\mathbf{Y}_O = \mathbf{y}_O) &= \mathbb{P}(\mathbf{Z}_P + \boldsymbol{\mu}_P = \mathbf{y}_P|\mathbf{Z}_O + \boldsymbol{\mu}_O = \mathbf{y}_O) \\ &= \mathbb{P}(\mathbf{Z}_P = \mathbf{y}_P - \boldsymbol{\mu}_P|\mathbf{Z}_O = \mathbf{y}_O - \boldsymbol{\mu}_O) \\ &= g_{P|O}\left(\mathbf{y}_P - \mathbf{f}_P - \epsilon(\mathbf{x}_P;\boldsymbol{\theta})|\ \mathbf{y}_O - \mathbf{f}_O - \epsilon(\mathbf{x}_O;\boldsymbol{\theta})\right),\end{aligned}$$

where $g_{P|O}$ denotes the relevant conditional distribution of g.

In many situations it will be appropriate to assume that the residuals $\mathbf{Z}$ have a multivariate normal distribution,

$$\mathbf{Z} \sim \mathrm{MVN}(\mathbf{0}, \Sigma),$$

whose covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{OO} & \Sigma_{OP} \\ \Sigma_{PO} & \Sigma_{PP} \end{bmatrix}$$

depends upon the values of unknown parameters $\boldsymbol{\psi}$. In this case, standard Gaussian theory tells us that

$$\mathbf{Y}_P|(\mathbf{Y}_O = \mathbf{y}_o) \sim \mathrm{MVN}(\boldsymbol{\mu}_P + \Sigma_{PO}\Sigma_{OO}^{-1}(\mathbf{y}_O - \boldsymbol{\mu}_O), \Sigma_{PP} - \Sigma_{PO}\Sigma_{OO}^{-1}\Sigma_{OP}).$$

The simplest special case occurs if the components of Z are independent and normally distributed with common variance σ^2, so that $\Sigma = \sigma^2 I$. Often, however, we might expect the residuals to exhibit temporal dependence; for an autoregressive time series model of order one - an AR(1) model - the covariance matrix will have elements of the form $\Sigma_{kl} = \sigma^2\rho^{|k-l|}$ and depend upon the values of two unknown parameters $\boldsymbol{\psi} = (\sigma^2, \rho)$.

Overall, the selection of an appropriate statistical model for the discrepancy term $\mathbf{Y} - \mathbf{f}$ consists of choosing a parametric form for ϵ and a joint distribution for $\mathbf{Z}$. The discrepancy term is driven by all processes that are *not* contained in the simulation run $\mathbf{f}$ - that is, by the climatic processes that are not included within the GCM used to create $\mathbf{f}$. As such, there will typically be little prior information regarding the properties that this term should have, and so we suggest selecting an appropriate model using statistical rather than mechanistic criteria. Statistical model selection can proceed along standard lines by, for example, selecting the model with the lowest value of the Bayesian Information Criterion (BIC; Schwarz, 1978) or Akaike Information Criterion (AIC; Akaike, 1973). Both criteria are defined to be equal to the deviance minus a penalty term: the penalty for BIC is equal to the product of the number of unknown parameters and the log of the number of datapoints, whereas for AIC it is equal to twice the number of unknown parameters.

3.1.2. Model averaging across simulation runs

Recall that in our application there are eighteen GCM-based simulation runs, taken from nine different GCMs. It is not clear which of the simulation runs should be used as a basis for prediction; model averaging provides a formal statistical mechanism to account for this uncertainty. This approach is also able to deal in a balanced way with the fact that some GCMs have ensemble runs whilst other do not, if the prior weights are chosen in such a way that each of the GCMs is given equal weight *a priori* (1/9, in our case).

Let $\mathbf{f}_i$ denote the LPJ simulation of vegetation carbon that was generated using the i-th GCM run, where $i \in \Omega$ and where Ω denotes the set of all GCM runs that have been used. For each $i \in \Omega$ we assume, as above, that

$$\mathbf{Z}_i := \mathbf{Y} - \boldsymbol{\mu}_i = \mathbf{Y} - \mathbf{f}_i - \epsilon(\mathbf{x}_i, \boldsymbol{\theta}_i) \sim g(\bullet; \boldsymbol{\psi}_i)$$

where $\mu_i := \mathbb{E}(\mathbf{Y}) = \mathbf{f}_i + \epsilon(\mathbf{x}_i, \boldsymbol{\theta}_i)$ (and hence $\mathbb{E}(\mathbf{Z}_i) = \mathbf{0}$).

Any of the simulation runs $i \in \Omega$ could, potentially, be used as a basis for drawing inferences about the missing (future) values of $\mathbf{Y}$, and each would lead to a different predictive distribution. Assume that one of these runs, $I \in \Omega$, leads to the best predictive distribution for $\mathbf{Y}$, but that the value of I is unknown. It follows (Raftery et al., 2005) that

$$\begin{aligned}\mathbb{P}(\mathbf{Y} = \mathbf{y}) &= \sum_{i\in\Omega} \mathbb{P}(I = i)\mathbb{P}(\mathbf{Y} = \mathbf{y}|I = i) = \sum_{i\in\Omega} \mathbb{P}(I = i)\mathbb{P}(\mathbf{Z}_i = \mathbf{y} + \boldsymbol{\mu}_i|I = i) \\ &= \sum_{i\in\Omega} w_i g(\mathbf{y} - \boldsymbol{\mu}_i; \boldsymbol{\psi}_i) = \sum_{i\in\Omega} w_i g\left(\mathbf{y} - \mathbf{f}_i - \epsilon\left(\mathbf{x}_i, \boldsymbol{\theta}_i\right); \boldsymbol{\psi}_i\right),\end{aligned}$$

where $w_i := \mathbb{P}(I = i)$ denotes the probability that simulation run i provides the best basis for prediction. The expected value of $\mathbf{Y}$ is equal to a weighted sum of the bias-corrected simulation runs,

$$\mathbb{E}(\mathbf{Y}) = \sum_{i\in\Omega} w_i \boldsymbol{\mu}_i = \sum_{i\in\Omega} w_i \{\mathbf{f}_i + \epsilon(\mathbf{x}_i, \beta_i)\}.$$

This highlights the fact that the interpretation of w_i is subtle; the value of w_i is not directly related to the distance between the GCM-based simulation $\mathbf{f}_i$ and the baseline simulation $\mathbf{y}$, but rather to the distance between the *bias-corrected* GCM-based simulation $\mathbf{f}_i + \epsilon(\mathbf{x}_i, \beta_i)$ and $\mathbf{y}$. As such, the relative weights $w_1, \ldots, w_{|\Omega|}$ that are allocated to different GCMs are contingent upon the parametric form that is used to describe the bias term ϵ.

3.1.3. Model averaging across parametric forms for the bias term

In many situations, including our application to vegetation carbon simulations generated by LPJ, it will unclear which parametric form should be used to describe the bias term $\epsilon(\mathbf{x}, \boldsymbol{\theta})$. We therefore propose using model averaging to account for the uncertainty associated with this choice, as well as the choice of simulation run; this appears to be a novel methodological development.

Let ϵ_j denote the j-th possible parametric form for the bias term, where $j \in B$ and where B denotes the set of all parametric forms that are under consideration (in our application there will be four such forms: constant, linear, quadratic and sinusoidal). Also let

$$\mathbf{Z}_{ij} := \mathbf{Y} - \mathbf{f}_i - \epsilon_j(\mathbf{x}_i, \boldsymbol{\theta}_i) = \mathbf{Y} - \boldsymbol{\mu}_{ij},$$

and assume that $\mathbf{Z}_{ij} \sim g(\bullet; \boldsymbol{\psi}_{ij})$ and $\mathbb{E}(\mathbf{Z}_{ij}) = \mathbf{0}$.

It follows that

$$\begin{aligned}\mathbb{P}(\mathbf{Y} = \mathbf{y}) &= \sum_{i\in\Omega}\sum_{j\in B} \mathbb{P}(I = i, J = j)\mathbb{P}(\mathbf{Y} = \mathbf{y}|I = i, J = j) \\ &= \sum_{i\in\Omega}\sum_{j\in B} w_{ij} g\left(\mathbf{y} - \boldsymbol{\mu}_{ij}; \boldsymbol{\psi}_{ij}\right) = \sum_{i\in\Omega}\sum_{j\in B} w_{ij} g\left(\mathbf{y} - \mathbf{f}_i - \epsilon_j\left(\mathbf{x}_{ij}, \boldsymbol{\theta}_{ij}\right); \boldsymbol{\psi}_{ij}\right),\end{aligned}$$

where $w_{ij} := \mathbb{P}(I = i, J = j)$. If we let $w_i := \mathbb{P}(I = i) = \sum_{j\in B} w_{ij}$ denote the marginal probability that simulation run I provides the best basis for prediction, then

$$\begin{aligned}\mathbb{E}(\mathbf{Y}) &= \sum_{i\in\Omega}\sum_{j\in B} w_{ij}\boldsymbol{\mu}_{ij} = \sum_{i\in\Omega}\sum_{j\in B} w_{ij}\{\mathbf{f}_i + \epsilon_j(\mathbf{x}_{ij}; \boldsymbol{\theta}_{ij})\} \\ &= \sum_{i\in\Omega} w_i \left(\mathbf{f}_i + \frac{\sum_{j\in B} w_{ij}\epsilon_j(\mathbf{x}_{ij}; \boldsymbol{\theta}_{ij})}{\sum_{j\in B} w_{ij}}\right).\end{aligned}$$

3.1.4. Estimation of parameters and weights

For each choice of simulation run $i \in \Omega$ and bias model $j \in B$, we (simultaneously) estimate the parameters $(\boldsymbol{\theta}_{ij}, \boldsymbol{\psi}_{ij})$ by numerical maximum likelihood estimation; let $(\hat{\boldsymbol{\theta}}_{ij}, \hat{\boldsymbol{\psi}}_{ij})$ denote the resulting estimators. The variance of $(\hat{\boldsymbol{\theta}}_{ij}, \hat{\boldsymbol{\psi}}_{ij})$ is approximately equal to the inverse of the observed information matrix (which is, in turn, equal to the negative of the hessian of the log-likelihood function evaluated at the maximum likelihood estimate), and an approximation to $\mathrm{var}(\hat{\boldsymbol{\mu}}_{ij})$ can then easily be computed via the delta method.

We consider the effects of increasing the variance of g from $\mathrm{var}(\mathbf{Z}_{ij})$ to $\mathrm{var}(\mathbf{Z}_{ij})+\mathrm{var}(\hat{\boldsymbol{\mu}}_{ij})$, as an approximate (conservative) means of assessing the degree of uncertainty associated with the estimation of $\boldsymbol{\mu}_{ij}$ when drawing predictive inferences. Note, however, that this does not account for the uncertainty associated with estimating $\hat{\boldsymbol{\psi}}_{ij}$.

The relative performance of different simulation runs and bias models is assessed using either BIC or AIC; let Λ_{ij} denote the BIC (or AIC) values associated with simulation run $i \in \Omega$ and bias model $j \in B$. We can use these to obtain approximate estimates for the weight w_{ij}, through Equation 18 of Buckland et al. (1997):

$$w_{ij} = \frac{\phi_{ij} \exp(-\Lambda_{ij}/2)}{\sum_{l \in B} \sum_{k \in \Omega} \phi_{kl} \exp(-\Lambda_{kl}/2)}. \tag{1}$$

ϕ_{ij} denotes the prior weight associated with fitting bias model $j \in B$ to simulation run $i \in \Omega$. Note that the basic weighting scheme is also similar to that used by Murphy et al. (2004) in averaging across climate predictions, except that they use a statistic known as the "Climate Prediction Index" (CPI) in place of BIC.

Equation 1 allocates weight to simulation run $i \in \Omega$ based on the performance of that run during the 20th century, but this approach is open to criticism on the grounds that the past performance of climate models does not necessarily provide a reliable basis for determining their predictive ability (see Section 4). An alternative approach is to keep the weights associated with simulations $i \in \Omega$ fixed at their prior values, so that $\sum_{j \in B} w_{ij} = \sum_{j \in B} \phi_{ij}$. This leads to the formula

$$w_{ij} = \frac{\phi_{ij} \exp(-\Lambda_{ij}/2)}{\sum_{l \in B} \exp(-\Lambda_{il}/2)(\sum_{k \in \Omega} \phi_{kl})}, \tag{2}$$

in which the BIC/AIC values are only used to determine the weights associated with potential models $j \in B$ for the bias *within* the context of a particular deterministic simulation run i.

3.2. Application to LPJ simulations of vegetation carbon

Explanatory analyses motivated us to consider four possible models for the bias term,

$$\begin{aligned}
\epsilon(\mathbf{1}; \alpha) &= \alpha\mathbf{1}; \\
\epsilon(\mathbf{t}; (\alpha, \beta)) &= \alpha\mathbf{1} + \beta\mathbf{t}; \\
\epsilon(\mathbf{t}; (\alpha, \beta, \gamma)) &= \alpha\mathbf{1} + \beta\mathbf{t} + \gamma\mathbf{t}^2; \text{ and} \\
\epsilon(\mathbf{t}; (\alpha, \beta, a, b)) &= \alpha\mathbf{1} + \beta \sin(a\mathbf{1} + b\mathbf{t}),
\end{aligned}$$

which correspond, respectively, to constant, linear, quadratic and sinusoidal trends over time.

All unknown parameters are estimated via numerical maximum likelihood, and multiple sets of overdispersed initial values are used to ensure that the optimisation algorithm has converged to a global maximum.

For each of these choices of ϵ, and for each simulation run $i \in \Omega$, there is strong evidence (in terms of BIC/AIC) against the use of uncorrelated errors $\mathbf{Z}_{ij}$ and in favour of models that include autocorrelation. Different time series models for the correlation structure - AR(1), AR(2), AR(3) and ARMA(1,1) models - had broadly similar performance, with the AR(1) generally having similar or better (lower) BIC values than those associated with the other three models. We therefore proceeded under the assumption that the errors were described by an AR(1) model with unknown variance and correlation.

The BIC values associated with the different simulation runs and choice of bias model ϵ are shown in Table 3; note that low values of BIC indicate good performance. The overall differences in performance between deterministic models (GCMs) are markedly greater than those between models for the bias term, but the relative performances of the four bias models still vary substantially: the no trend model has lowest BIC for eight of the GCM-based simulations, the linear model has lowest BIC for six, and the sinusoidal model has lowest BIC for four.

We also explored the performance of a wider set of models for ϵ, before deciding upon the set of four that we have shown here. Models which included higher order polynomial trends in $\boldsymbol{\mu}_{it}$, or which included the model signal f_{it} as a covariate rather than an offset, generally did perform better, in terms of AIC/BIC, than any of the models that we have used. We chose to exclude these models from our analyses on the grounds that they seem to provide a particularly unreliable basis for long-range extrapolation - levels of estimation uncertainty associated with the parameters of cubic or higher order polynomial models appeared to be extremely high, whilst LPJ simulations of vegetation carbon for the 20th century (expressed as anomalies relative to the reference period 1961-1990) cover a very narrow range of values relative to those that are simulated for the 21st century. Note that the decision of which statistical models to include in the set B is fairly subjective, and can have a substantial impact upon the results of the model averaging; there is also subjectivity in the selecting the set of GCMs and GCM ensembles Ω.

3.3. Results of model averaging

Model weights w_{ij} are calculated using Equation 1 or 2. In the absence of any other information, we assign equal prior weight to each of the nine GCMs, and, for GCMs with more than one ensemble run, assign equal prior weight to each member of the ensemble. This allows us to deal with the nine GCMs in a balanced way, despite the fact that some GCMs have multiple ensembles and others do not. We assume that the four bias models are all equally plausible *a priori*. Hence $\phi_{ij} = 1/36m_I$, where m_i denotes the number of ensembles associated with the GCM that was used to generate run $i \in \Omega$.

We find that Equation 1 lends virtually all weight (more than 99.9%) to be concentrated on just two of the GCM runs, both of which originate from the same GCM (ensembles 3 and 5 from CGCM). The associated predictive distributions for future vegetation carbon stocks are consequently very precise (Figure 2; top), but will also be highly sensitive to the assumption that past performance of a GCM is indicative of future performance. In contrast, Equation 2 produces much more diffuse predictive distributions (Figure 2; bottom) that span the full range of responses associated with the different GCMs, and so lead to less precise but perhaps more robust predictions of future change.

Predictive distributions from Equation 2 are actually moderately precise in the years up to 2050, but become much more diffuse during the second half of the century. In particular, predictions from 2050 onwards reflect the possibility that the trend in carbon may be either increasing or decreasing. From about 2080 onwards there is a non-negligible probability that vegetation carbon stocks will be lower than in 2002, and by the end of the twenty-first century the predictions of vegetation carbon stocks are, unsurprisingly, highly uncertain.

In Figure 3 we compare the predictive distributions (PDFs) for 2095 that are generated either by model averaging across both bias models and GCMs, or by averaging across GCMs using a particular bias model (e.g. sinusoidal or "no trend"). Model averaging using BIC gives a median predicted anomaly in 2095 of 116.5 gtC, and an associated 95% credible interval of [0.4,161.8] gtC. Note that this interval spans the range of the GCM-driven simulations from LPJ, reflecting the fact that Equation 2 applies equal weight to each GCM. The predictive distributions obtained by averaging across bias models using BIC or AIC are fairly similar to those obtained by selecting the model with the lowest BIC/AIC, but do show differences from those obtained using by the same bias model for each GCM-based simulation. Most notably, the predictive distribution associated with the no trend model differs markedly from the other three models models in which the discrepancy $\mathbf{Y} - \mathbf{f}$ is allowed to show systematic variation over time. These results suggest that, for our application, it is crucial to allow the parametric form for ϵ to depend upon which simulation run i is being considered is, but that it is less important to account for the uncertainty associated with the choice of parametric form (i.e. model selection and model averaging over candidate models $j \in B$ give similar results; note that the same is emphatically *not* true in the case of the simulation runs $i \in \Omega$).

3.4. Assessment of past performance

We can partially assess the performance of our model averaging approach by predicting values of $\mathbf{Y}$ for periods in which the baseline simulation is already available - in Figure 4, for example, we compare the predictive distributions for the year 2001 that we obtain using Equations 1 and 2. These predictions are based on estimating the values of the parameters and weights (θ_{ij}, ψ_{ij} and w_{ij}) using values from a restricted period only - either 1900-60, 1900-70, 1900-80 or 1900-90.

We see that predictive distributions based on values from the shortest period (1900-60) are highly uncertain, whether we use Equation 1 or 2. Equation 1 yields a distribution with two widely separated modes, with all intervening values - including the true value - having low or very low probability densities. Equation 2, in contrast, generates a highly diffuse prediction, and assigns non-negligible densities to a much wider set of values - including the true value. Similar but less diffuse results are obtained using data from 1900-1970, with the crucial difference that one of the modes in the predictive distribution generated by Equation 1 now lies close to the actual value of y, so that a relatively high density is assigned to this value. Predictive distributions based on the longest periods (1900-80 or 1900-90) are substantially more precise - less diffuse - than those based on 1900-60 and 1900-70. Crucially, the distributions obtained using Equation 1 no longer appear to be bimodal, and now assign a relatively high probability density to the true value. Qualitatively similar results are obtained by using other periods (1900-1965, 1900-1970 and 1900-1975; not shown), and by considering years other than 2001 - the predictive distributions obtained via Equation 1 are always relatively precise but sometimes highly inaccurate (i.e. assign a very low density to the true value of $\mathbf{y}$), whereas the predictions obtained using Equation

2 are relatively diffuse but always assign a moderate density to the true value of $\mathbf{y}$.

4. Discussion

The substantial differences between the simulated levels of vegetation carbon obtained using different GCMs (in Section 2) reflect the sensitive balance between the effects of temperature, precipitation and ambient CO_2 concentration: concomitant increases in temperature and precipitation will lead to increased vegetation growth, whereas increases in CO_2 concentrations and temperature that are not associated with any substantial increase in precipitation will lead to declining stocks of vegetation carbon.

The largest relative temperature increases between 1961-90 and 2070-2099 are shown by the HadCM3, HADGEM1 and CNRM-CM3 models (39%, 44% and 39% respectively). For the HadCM3 and HadGEM1 models these increases are not accompanied by any substantial change in precipitation (+1.5% and -0.4% respectively), with the result that plant water stress is the dominant effect on vegetation carbon stocks after 2050 and that there are consequently only minimal increases in overall vegetation carbon over the course of the 21st century (+3.3% and +3.7%, respectively, between the periods 1961-90 and 2070-2099). Precipitation increases in CNRM-CM3, in contrast, are moderately large (+5.3%), leading to a more prolonged, but still relatively slow, increase in vegetation carbon stocks (+7.6%): in this simulation the effects of increasing temperatures and CO2 on plant growth are presumably balanced by reduced growth resulting from plant and soil water stress. The NCAR-CCSM3 and CCCMA-CGCM3.1 models are associated with relatively large increases in both temperature (+35% and +38%) and rainfall (+13.3% and +8.1%), leading to sustained and relatively large increases in levels of vegetation carbon (+17.1% and +15.8%).

The results for the remaining models are somewhat less straightforward to interpret. The overall temperature and precipitation changes for ECHAM5 are similar to those for the NCAR-CCSM3 and CCCMA-CGCM3.1 models, for example, but this model shows only relatively modest increases in vegetation carbon (+13.0%). This may relate to the fact that ECHAM5 exhibits the lowest levels of baseline rainfall, during 1961-1990, and so may experience relatively high levels of water stress; however, the CNRM-CM3 model exhibits even smaller increases in vegetation carbon despite having the highest level of baseline rainfall. A more detailed regional evaluation may be required in order to fully understand the causes of inter-GCM differences in simulated values of global vegetation carbon.

There is generally good agreement between the simulated values of vegetation carbon that we have presented and those that were reported by Schaphoff et al. (2006). Values for three of the GCMs that were used in both studies are similar (ECHAM, CSIRO and HadGCM3), and exhibit the same ordering, although the results that we obtain using CGCM3 are very different to those which were obtained by Schaphoff et al. (2006) using CGCM1. We might actually have expected to see larger increases in vegetation carbon over the 21st century than those reported by Schaphoff et al. (2006), since atmospheric CO_2 concentrations for the end of the 21st century are much higher under the SRES A2 emissions scenario than under the Is92a scenario (19% higher). Despite the higher CO_2 concentrations, however, the overall mean global annual temperature for 2070-2099 is substantially lower - by 1.7^{o}C - for the set of GCMs that we used than for the set used by Schaphoff et al. (2006). This difference may be due to a reduced range of variability amongst the newer versions of GCMs (as noted by Meehl et al., 2007), and probably compensates for the differences in

CO_2 concentrations.

Note that the total amount of stored carbon will depend upon levels of soil and litter carbon as well as the level of vegetation carbon. Higher temperatures increase heterotrophic respiration, and soil carbon stocks therefore have the potential to decrease in the future. Several of the GCM-driven simulations in Schaphoff et al. (2006) simulate soil carbon decreases that are equal in magnitude to vegetation carbon increases. Note that our results also do not account for future changes in land-use, which could potentially alter distributions of Plant Functional Types.

We have combined the simulations of vegetation carbon into a single predictive distribution using a form of model averaging in which the model weights are estimated *post hoc*. We have found that the results of the model averaging procedure are strongly dependent upon the procedure used to estimate the weights (Section 3.3). When the weights associated with the different GCMs were estimated based on past performance, using Equation 1, then we found that a large amount of weight tended to be allocated to a small number of the climate models (one, in this case). Similar, although less extreme, phenomenon were reported by Min et al. (2007) in the context of surface air temperatures, by Fowler et al. (2007) in an analysis of temperature and precipitation at the catchment scale, and by Fowler and Ekström (2008) in an analysis of UK precipitation. GCM predictions of the far future relate to climatic conditions that have no analogue in the observational record, so it seems questionable as to whether we can robustly assess the relative accuracy of future GCM-based predictions solely on the basis of their performance during the historical period. One approach to dealing with these difficulties is to keep the marginal weights associated with the different GCMs fixed at their prior values, using Equation 2, allowing us to deal with the effect of predictive uncertainty in a more conservative - and, in this particular application, probably more plausible - fashion. Fixed model weights have been adopted by a number of authors in the context of climate prediction (e.g. Palmer and Räisänen, 2002, Räisänen and Palmer, 2001).

The key advantage of estimating the weights *post hoc*, using BIC/AIC, lies in the fact that this approach is straightforward and quick to implement: we can fit a separate model ϵ_j to each simulation run $\mathbf{f}_i$ using maximum likelihood, and then combine the results in a trivial and instantaneous way. The current methodology could, therefore, easily be extended to more complicated situations in which the predictions have spatial, spatio-temporal or multivariate structure. There are some limitations, however. Equations 1 and 2 only provide approximate estimators for the weights w_{ij}, and the fact that we fit separate models for each simulation run $\mathbf{f}_i$ prevents us from obtaining more efficient inferences by pooling some elements of the parameter vector $\boldsymbol{\theta}_{ij}$ across models (as in Raftery et al., 2005). Some of these difficulties could be avoided through the use of a fully Bayesian approach (using Reversible Jump Markov chain Monte Carlo; Green, 1995), or by estimating parameters and weights simultaneously using the EM-algorithm (Raftery et al., 2005), but, for computational reasons, these approaches would generally be less straightforward to generalise to more complicated situations.

Model averaging does not provide the only statistical methodology for combining a set of deterministic predictions into a single probabilistic prediction - see Tebaldi and Knutti (2007) for a recent review of methods that have been used in the context of climatology and Fowler et al. (2007) for a review of methods used in climate impact prediction. Alternative approaches for analysing runs drawn from a set of different models - so-called "ensembles of opportunity" - involve treating the predictive runs $\{f_i : i \in \Omega\}$ as explanatory variables and the baseline run y as the response variable in the context of a regression model (e.g. Allen

and Stott, 2003 - optimal fingerprinting; Gneiting et al., 2005 - EMOS; Greene et al., 2006), or treating y and $\{f_i : i \in \Omega\}$ as mutually independent sources of data about the true, latent, process in the context of Bayesian hierarchical model (Tebaldi et al., 2004, 2005; Smith et al., 2008; Furrer et al., 2007). Lopez et al. (2006) compared the results obtained using these approaches when analysing global temperature. It is important to be aware that there are fundamental difficulties associated with generating probabilistic predictions using ensembles of opportunity (Stainforth et al., 2007), since the models are not drawn from any meaningful population, are not selected either systematically or at random, and are far from independent (Knutti et al., 2008). These issues inevitably afflict all studies that use outputs obtained from multiple GCMs, and cannot - with the possible exception of the final issue (dependence) - be corrected for within the statistical analysis. Analyses of ensemble runs that have been generated from a single common model (e.g. a single GCM) rest on a firmer conceptual basis, and statistical methods for this purpose are relatively well-developed (SACCO methods: "statistical analysis of computer code output"; e.g. Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2006).

5. Conclusions

Using state-of-the-art climate models and a dynamic global vegetation model we have simulated trends in future global vegetation carbon stocks under a particular emissions scenario, SRES A2. The simulated values suggest a reasonable consensus amongst GCMs about both the direction and magnitude of change in the period up to 2050, but indicate substantial uncertainty beyond that point; they are generally in good agreement with those obtained in an earlier study by Schaphoff et al. (2006). Simulated values for vegetation carbon stocks in 2095 (relative to 1961-1990) range from +0.7gtC to +151.4gtC.

There are substantial advantages in using statistical approaches to combine these deterministic simulations into a single probabilistic prediction/projection (Räisänen and Palmer, 2001; Collins, 2007), but there is currently little agreement on how this should be done (Fowler et al., 2007; Tebaldi and Knutti, 2007). In this paper we have adopted a simple form of model averaging in which the model weights are estimated *post hoc* using BIC or AIC values (Buckland et al., 1997). The novel aspect of our statistical approach is that we account for two distinct sources of uncertainty: lack of knowledge about which GCM provides the best basis for prediction, and lack of knowledge about the form of the relationship between each of these GCM-based simulations and the simulation run obtained using observed climate data (which we treat as being equivalent to the "truth", given the lack of any actual data on global vegetation carbon stocks). The statistical methodology is generic, so that it could be used to combine long-term deterministic simulations generated by other environmental models, and ought to generalise easily to situations in which these simulations are multivariate or exhibit spatial structure.

The BIC/AIC values can either be using to estimate the weights associated with both the simulation runs themselves and with the models that we use to describe the discrepancy associated with each of these runs, using information obtained from the period for which observational climate data are available, or can just be used to estimate the weights associated with the discrepancy models whilst assign equal weight to each GCM. The former approach leads the vast majority of weight to be attributed to a single GCM (CCCMA-CGCM3.1), and therefore produces narrow predictive distributions for the change in vegetation carbon stocks from 1961-1990 until 2071-2099: a median of 118.6 gtC and a 95% credible interval

of [115.0,122.5] gtC. These results seem implausibly precise, and, because they concentrate so much weight on a single GCM, will be highly sensitive to the assumption that past performance provides a good indicator of future performance. The latter approach produces a much more diffuse predictive distribution that spans the full range of simulated responses - a median of 116.5 gtC, a 95% credible interval of [+0.4,+161.8] gtC, and a probability of 0.022 that the values will be lower than the average for 1961-90. The latter values suggest that, under the SRES A2 emissions scenario, (a) vegetation has the potential to sequester more carbon in the future, as indicated by previous studies; (b) the quantity of carbon that could be stored in this way is highly uncertain; and (c) there is a small but non-negligible probability that vegetation carbon stocks will actually fall slightly over the course of the 21st century. These findings are contingent upon the set of GCMs that were selected, the set of statistical models for ϵ that were considered, and the dynamic global vegetation model that was used, and should be interpreted alongside concurrent changes in soil and litter carbon.

Acknowledgements

This work was funded by European Commission Framework 6 Integrated project ALARM (Assessing LArge scale environmental Risks for biodiversity with tested Methods) (GOCE-CT-2003-506675) and by the Scottish Government. Jonathan Rougier provided extensive advice regarding the statistical methodology and notation. Ben Smith, Stephen Sitch, Thomas Hickler, Sybil Schaphoff and Dieter Gerten provided LPJ model code and helpful discussions regarding Section 2. Useful comments on the manuscript were provided by Chris Glasbey and Clive Anderson. We acknowledge the GCM modelling groups for providing their data for analysis, and the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model output.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Budapest, pp. 267–281. Akademiai Kiado.

Allen, M. R. and P. A. Stott (2003). Estimating signal amplitudes in optimal fingerprinting, part I: theory. *Climate Dynamics 21*, 477–491.

Berrocal, V. J., A. E. Raftery, and T. Gneiting (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review 135*(4), 1386–1402.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*(2), 603–618.

Cao, M. and F. I. Woodward (1998). Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature 393*, 249–252.

Climatic Research Unit (2006). CRU-TS-2.1 gridded global climate dataset. http://www.cru.uea.ac.uk/∼timm/grid/CRU_TS_2_1.html (accessed April 2006).

Collins, M. (2007). Ensembles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A 365*(1857), 1957–1970.

Cramer, W., B. F. Alberte, I. Woodward, C. I. Prentice, R. A. Betts, V. Brovkin, P. M. Cox, V. Fisher, J. A. Foley, A. D. Friend, C. Kucharik, M. R. Lomas, N. Ramankutty, S. Sitch, B. Smith, A. White, and C. Young-Molling (2001). Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. *Global Change Biology 7*(4), 357–373.

Doherty, R., S. Sitch, B. Smith, S. L. Lewis, and P. Thornton (submitted, 2008). Implications of uncertainty in multi-model simulations of future climate for the carbon cycle and biogeography in East Africa. *Glob. Ecol. and Biogeography*.

Folland, C. K., N. A. Rayner, S. J. Brown, T. M. Smith, S. S. P. Shen, D. E. Parker, I. Macadam, P. D. Jones, R. N. Jones, N. Nicholls, and D. M. H. Sexton (2001). Global temperature change and its uncertainties since 1861. *Geophysical Research Letters 28*, 2621–2624.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology 27*(12), 1547–1578.

Fowler, H. J. and M. Ekström (in revision, 2008). Multi-model ensemble estimates of climate change impacts on UK seasonal rainfall extremes. *International Journal of Climatology*.

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

Gerten, D., S. Schaphoff, U. Haberlandt, W. Lucht, and S. Sitch (2004). Terrestrial vegetation and water balance - hydrological evaluation of a dynamic global vegetation model. *Journal of Hydrology 286*, 249–270.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CPRS estimation. *Monthly Weather Review 133*, 1098–1118.

Goldstein, M. and J. Rougier (2006). Bayes linear calibrated prediction for complex systems. *J. Amer. Statist. Assoc. 101*(475), 1132–1143.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Greene, A. M., L. Goddard, and U. Lall (2006). Probabilistic multimodel regional temperature change projections. *Journal of Climate 19*(17), 4326–4343.

Hickler, T., C. Prentice, B. Smith, M. T. Sykes, and S. Zaehle (2006). Implementing plant hydraulic architecture within the LPJ Dynamic Global Vegetation model. *Global Ecology and Biogeography 15*(6), 567–577.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science 14*(4), 382–417.

Houghton, J. T., Y. Ding, M. Griggs, M. Noguer, P. J. van der Linder, and D. Xiaosu (Eds.) (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge University Press.

IPCC (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC. Eds.: Core Writing Team, Pacjauri, R. K. and Reisinger, A.

Jones, P. D., T. J. Osborn, and K. R. Briffa (1997). Estimating sampling errors in large-scale temperature averages. *Journal of Climate 10*, 2548–2568.

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

Knutti, R., M. R. Allen, P. Friedlingstein, J. M. Gregory, G. C. Hegerl, G. A. Meehl, M. Meinshausen, J. M. Murphy, G.-K. Plattner, S. C. B. Raper, T. F. Stocker, P. A. Stott, H. Teng, and T. M. L. Wigley (2008). A review of uncertainties in global temperature projections over the twenty-first century. *Journal of Climate 21*, 2651–2663.

Lopez, A., C. Tebaldi, M. New, D. Stainforth, M. Allen, and J. Kettleborough (2006). Two approaches to quantifying uncertainty in global temperature changes. *Journal of Climate 19*(19), 4785–4796.

Magnani, F., M. Mencuccini, M. Borghetti, P. Berbigier, F. Berninger, S. Delzon, A. Grelle, P. Hari, P. G. Jarvis, P. Kolari, A. S. Kowalski, H. Lankreijer, B. E. Law, A. Lindroth, D. Loustau, G. Manca, J. B. Moncrieff, M. Rayment, V. Tedeschi, R. Valentini, and J. Grace (2007). The human footprint in the carbon cycle of temperate and boreal forests. *Nature 447*(7146), 848–852.

Meehl, G. A., T. F. Stocker, W. D. Collins, P. Friedlingstein, A. T. Gaye, J. M. Gregory, A. Kitoh, R. Knutti, J. M. Murphy, A. Noda, S. C. B. Raper, I. G. Watterson, A. J. Weaver, and Z. C. Zhao (2007). Global climate projections. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller (Eds.), *Climate Change 2007: The physical Science Basis. Contribution of Working Group I to the Fourth Assessment of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Min, S. K., D. Simonis, and A. Hense (2007). Probabilistic climate change predictions applying Bayesian model averaging. *Phil. Trans. R. Soc. A 365*(1857), 2103–2116.

Mitchell, T. D. and P. D. Jones (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol. 25*, 693–712.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature 430*, 768–772.

Nakicenovic, N. and R. Swart (2000). *Special Report on Emissions Scenarios*. Cambridge University Press (UK).

New, M., M. Todd, M. Hulme, and P. Jones (2001). Precipitation measurements and trends in the twentieth century. *International Journal of Climatology* (15), 1889–1922.

Palmer, T. N. and J. Räisänen (2002). Quantifying the risk of extreme seasonal precipiation in a changing climate. *Nature 415*, 512–514.

PCMDI (2006). CMIP3 multi-model dataset archive. http://www.pcmdi.llnl.gov (accessed April 2006).

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(5), 1155–1174.

Räisänen, J. and T. N. Palmer (2001). A probability and decision-model analysis of a multi-model ensemble of climate change simulations. *Journal of Climate 14*, 3212–3226.

Schaphoff, S., W. Lucht, D. Gerten, S. Sitch, W. Cramer, and I. C. Prentice (2006). Terrestrial biosphere carbon storage under alternative climate projections. *Climatic Change 74*, 97–122.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Sitch, S., B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. Sykes, K. Thonicke, and S. Venevski (2003). Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ Dynamic Vegetation Model. *Global Change Biology 9*, 161–185.

Sloughter, J. M., A. E. Raftery, and T. Gneiting (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review 135*, 3209–3220.

Smith, B., C. I. Prentice, and M. T. Sykes (2001). Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within european climate space. *Global Ecology & Biogeography 10*, 621–637.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (forthcoming, 2008). Bayesian modeling of uncertainty in ensembles of climate models. *J. Amer. Statist. Assoc.*.

Stainforth, D., M. R. Allen, E. R. Tredger, and L. A. Smith (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. A 365*(1857), 2163–2177.

Tebaldi, C. and R. Knutti (2007). The use of the multimodel ensemble in probabilistic climate projections. *Phil. Trans. Roy. Soc. A. 365*(1857), 2053–2075.

Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith (2004). Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters 31*.

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

White, A., M. G. R. Cannell, and A. D. Friend (1999). Climate change impacts on ecosystems and the terrestrial carbon sink: a new analysis. *Global Environmental Change 9*, S21–S30.

Zaehle, S., S. Sitch, B. Smith, and F. Hatterman (2005). Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics. *Global Biogeochemical Cycles 19*(GB3020).

Table 1. Description of the nine GCMs which are used to provide climate inputs to the LPJ model

Model name	Institution	Spatial resolution	# runs
NCAR-CCSM3	National Centre for Atmospheric Research, USA	$1.4065^o \times 1.4065^o$	5
CCCMA-CGCM3.1	Canadian Centre for Climate Modelling and Analysis, Canada	$3.7500^o \times 3.7500^o$	4
CNRM-CM3	Centre National de Recherches Meteorologiques, France	$2.8125^o \times 2.8125^o$	1
CSIRO-MK3.0	CSIRO Atmospheric research, Australia	$1.8750^o \times 1.8750^o$	1
ECHAM5	Max Planck Institute for Meteorology, Germany	$1.8750^o \times 1.8750^o$	3
GFDL-CM2.1	Geophysical Fluid Dynamics Laboratory, USA	$2.5000^o \times 2.0000^o$	1
HadCM3	Hadley Centre for Climate Prediction and Research, UK	$3.7500^o \times 2.5000^o$	1
HadGEM1	Hadley Centre for Climate Prediction and Research, UK	$1.8750^o \times 1.2410^o$	1
MRI-CGCM2_3	Meteorological Research Institute, Japan	$2.8125^o \times 2.8125^o$	1

Table 2. Comparison of daily mean global surface temperature and precipitation values generated by different GCMs for 1961-1990 and 2070-2099, together with % change between these two periods. Observational climate data (CRU) are shown for the period 1961-1990. Corresponding results are also shown for simulations of global vegetation carbon stocks from the LPJ model. For NCAR-CCSM3, CCCMA-CGCM3.1 and ECHAM5 values are obtained by averaging across the available ensembles (but with ensemble 5 for NCAR-CCSM3 excluded, since this run terminates in 2089).

GCM	Surface temperature (daily mean, ^{o}C)			Precipitation (daily, mm)			LPJ vegetation carbon (annual, gtC)		
	1961-1990	2070-2099	Change (%)	1961-1990	2070-2099	Change (%)	1961-1990	2070-2099	Change (%)
CRU data	13.5	•	•	2.14	•	•	790	•	•
NCAR-CCSM3	12.9	17.4	+35	2.25	2.55	+13.3	780	913	+17.1
CCCMA-CGCM3.1	11.5	15.8	+38	2.04	2.21	+8.1	681	789	+15.8
CNRM-CM3	11.8	16.3	+39	2.57	2.71	+5.3	763	821	+7.6
CSIRO-MK3.0	12.1	15.6	+29	2.01	2.02	+0.6	618	725	+17.3
ECHAM5	13.2	18.1	+37	2.00	2.15	+7.4	692	781	+13.0
GFDL-CM2.1	12.2	16.5	+35	2.26	2.24	-0.6	580	649	+12.0
HadCM3	12.4	17.3	+39	2.27	2.31	+1.5	798	827	+3.7
HadGEM1	11.4	16.4	+44	2.40	2.40	-0.4	798	824	+3.3
MRI-CGCM2_3	12.4	15.6	+26	2.08	2.17	+4.1	886	1013	+14.3

Table 3. Relative BIC values associated with each combination of GCM-based simulations $\mathbf{f}_i$ and statistical models for the bias term $_j$, based on model fit during the period 1900-2001. Values are relative to those of the combination with lowest BIC. For each GCM we also list the associated bias model $_j$ with best fit (lowest BIC).

Simulation run, $\mathbf{f}_i$		Model for bias, $_{ij}$				
GCM	Ensemble	No trend	Linear	Quadratic	Sinusoidal	Best fit
NCAR-CCSM3	1	69.07	64.23	70.85	67.00	Linear
NCAR-CCSM3	2	49.39	50.51	54.99	54.93	No trend
NCAR-CCSM3	3	39.27	38.87	45.46	41.61	Linear
NCAR-CCSM3	4	64.08	60.18	65.49	72.08	Linear
NCAR-CCSM3	5	65.35	63.35	68.37	69.63	Linear
CCCMA-CGCM3.1	1	40.49	45.10	44.78	58.38	No trend
CCCMA-CGCM3.1	3	25.58	28.36	31.03	9.79	Sinusoidal
CCCMA-CGCM3.1	4	46.13	48.16	53.05	59.66	No trend
CCCMA-CGCM3.1	5	8.40	11.42	17.40	0.00	Sinusoidal
CNRM-CM3		111.13	112.77	118.96	120.52	No trend
CSIRO-MK3.0		51.67	56.88	59.00	54.67	No trend
ECHAM5	1	50.24	53.07	54.86	51.37	No trend
ECHAM5	2	130.23	128.40	134.49	130.18	Linear
ECHAM5	3	73.91	79.76	86.32	80.91	No trend
GFDL-CM2.1		116.52	112.19	117.23	123.89	Linear
HadCM3		85.98	89.75	95.32	91.15	No trend
HadGEM1		78.51	80.85	87.11	75.53	Sinusoidal
MRI-CGCM2_3		55.47	50.01	55.80	46.96	Sinusoidal

Fig. 1. LPJ simulations of global annual vegetation carbon stocks for the 20th and 21st centuries. Carbon stocks are measured in gigatonnes of carbon (gtC), and are reported as anomalies relative to the mean value for a thirty year reference period (1961-1990). Climate inputs to the baseline simulation run are based on the CRU-TS-2.1 gridded observational climate dataset, with inputs to the remaining eighteen runs provided by outputs from nine different General Circulation Models. For GCMs with more than one ensemble run these are shown as dotted lines.

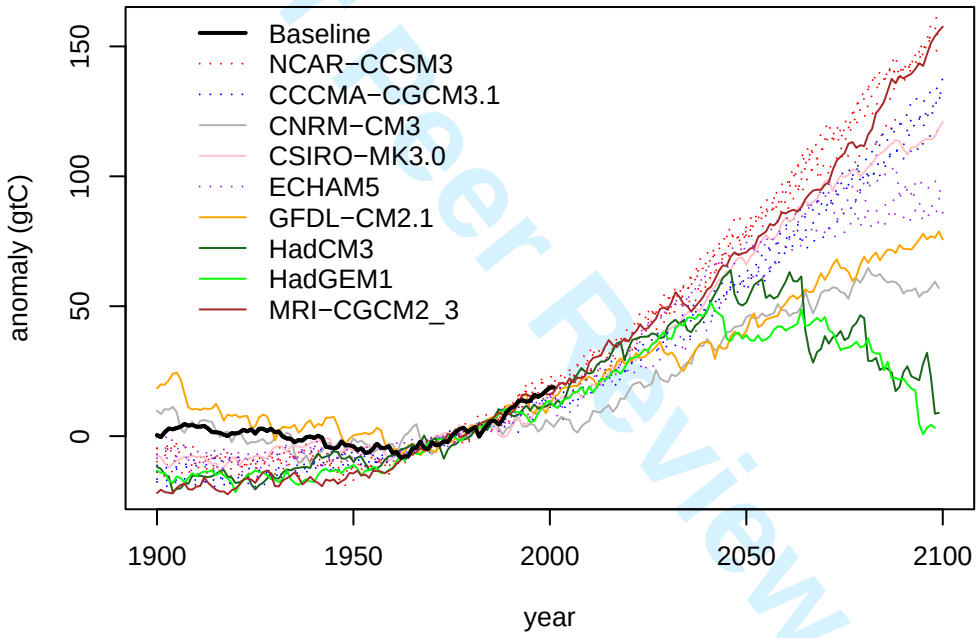

Fig. 2. Predictive distributions for global vegetation carbon stocks during the twenty-first century, based on model averaging using Equation 1 (top) or Equation 2 (bottom). 2.5%, 50% and 97.5% quantiles of the predictive distribution are shown (thin black lines), together with the baseline simulation (thick black) and GCM-based runs (grey). Stocks are reported as anomalies relative to the mean value for the period 1961-1990. We account for estimation error by replacing var($\mathbf{Z}_{ij}$) with var($\mathbf{Z}_{ij}$) + var($\hat{\boldsymbol{\mu}}_{ij}$).

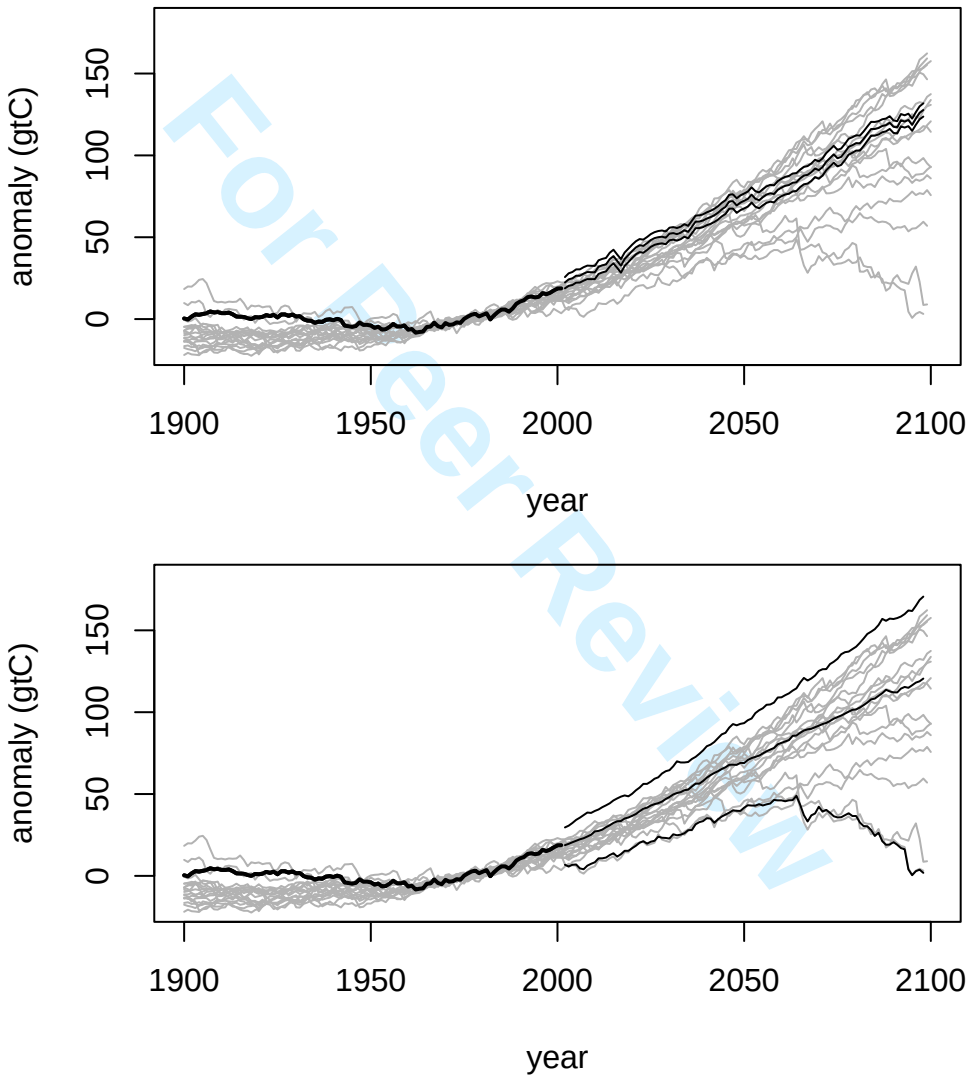

Fig. 3. Predictive distributions for global vegetation carbon stocks in 2095, based on averaging across simulation runs using fixed weights $_i$ (i.e. using Equation 2). The bias term is dealt with in six different ways: a) always using a sinusoidal model (whichever GCM run i $\in \Omega$ is being considered); b) always assuming that the bias is constant over time; c) by selecting the model for that has lowest AIC for each i $\in \Omega$; d) by selecting the model for that has lowest BIC; e) by averaging across the set of possible models for using AIC and f) by averaging across this set using BIC. Results are shown with (black) and without (grey) accounting for uncertainty in the estimation of $\hat{\boldsymbol{\mu}}_{ij}$.

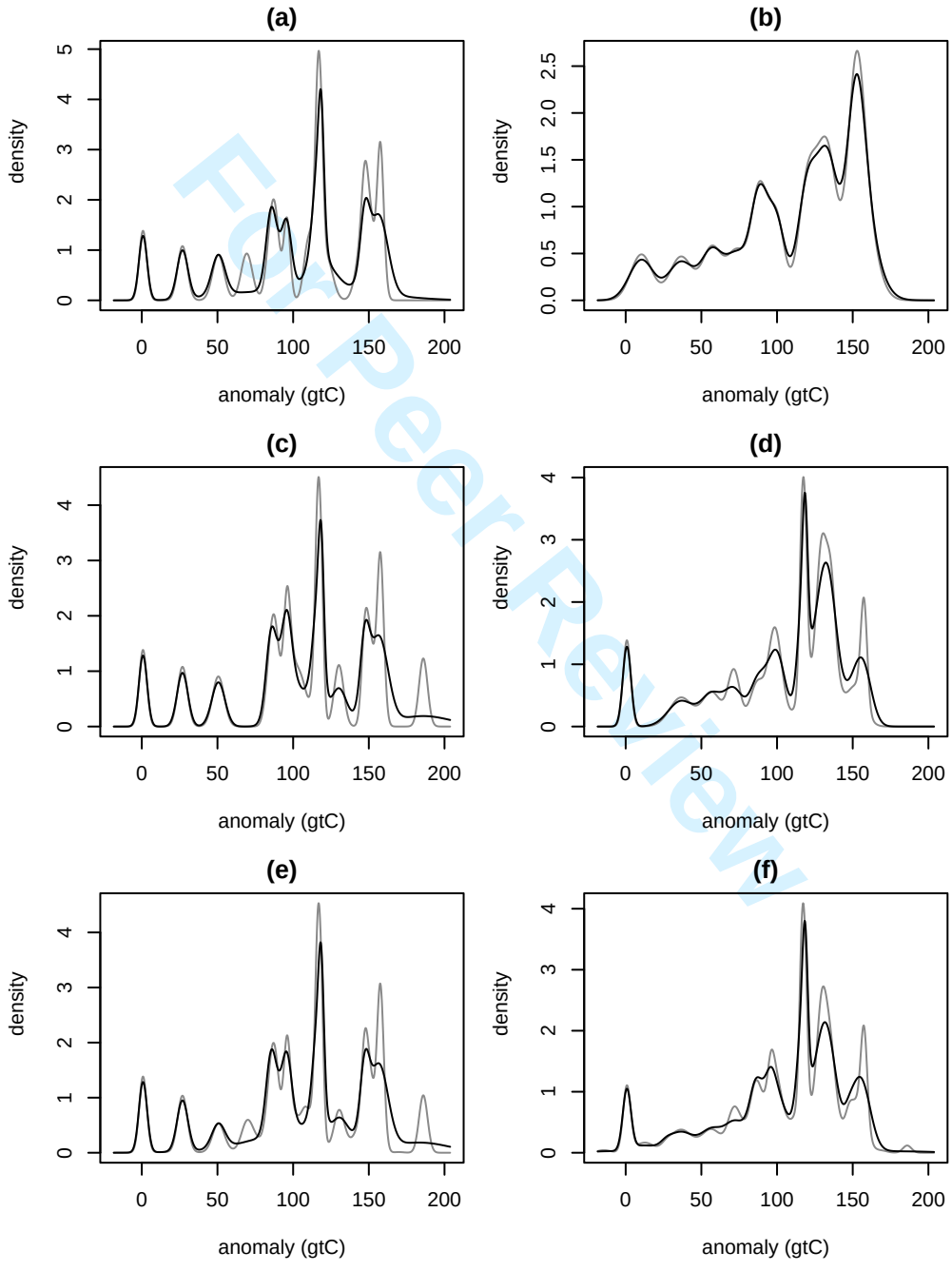

Fig. 4. Predictive distributions for global vegetation carbon in the year 2001, based only on data up to 1960 (dotted), 1970 (dashed), 1980 (thin solid) and 1990 (thick solid). Predictions are produced by model averaging across simulation runs and parametric forms for the bias using either Equation 1 (top) or Equation 2 (bottom). We account for estimation error by replacing var($\mathbf{Z}_{ij}$) with var($\mathbf{Z}_{ij}$) + var($\hat{\boldsymbol{\mu}}_{ij}$). The actual value of the baseline run in 2001 is also shown, for comparison (large circle).