Optimal Estimation of Stochastic Energy Balance Model Parameters

DONALD P. CUMMINS, DAVID B. STEPHENSON, AND PETER A. STOTT[a]

University of Exeter, Exeter, United Kingdom

(Manuscript received 1 August 2019, in final form 6 April 2020)

ABSTRACT

This study has developed a rigorous and efficient maximum likelihood method for estimating the parameters in stochastic energy balance models (with any $k > 0$ number of boxes) given time series of surface temperature and top-of-the-atmosphere net downward radiative flux. The method works by finding a state-space representation of the linear dynamic system and evaluating the likelihood recursively via the Kalman filter. Confidence intervals for estimated parameters are straightforward to construct in the maximum likelihood framework, and information criteria may be used to choose an optimal number of boxes for parsimonious k-box emulation of atmosphere–ocean general circulation models (AOGCMs). In addition to estimating model parameters the method enables hidden state estimation for the unobservable boxes corresponding to the deep ocean, and also enables noise filtering for observations of surface temperature. The feasibility, reliability, and performance of the proposed method are demonstrated in a simulation study. To obtain a set of optimal k-box emulators, models are fitted to the 4 × CO_2 step responses of 16 AOGCMs in CMIP5. It is found that for all 16 AOGCMs three boxes are required for optimal k-box emulation. The number of boxes k is found to influence, sometimes strongly, the impulse responses of the fitted models.

1. Introduction

An energy balance model (EBM) is a simplified representation of climate where changes in global temperature are explained by imbalances in Earth's energy budget. Energy balance models are simpler than atmosphere–ocean general circulation models (AOGCMs), which explicitly describe the fluid dynamics of Earth's atmosphere and oceans. Their simplicity means that EBMs are both analytically tractable and inexpensive to simulate. Compared with purely empirical statistical models, EBMs have two distinct advantages: 1) the choice of model structure is motivated by physical reasoning, and 2) model parameters have physical interpretability. Energy balance models are therefore useful not only for climate forecasting but for making physical inferences about the climate system.

Energy balance models in the literature vary in complexity. The class of EBM considered here is the k-box model (sometimes called k-layer), which represents the atmosphere and ocean as a set of vertically stacked boxes. The simplest k-box model is the so-called one-box model, which is obtained by a linearization of the Budyko–Sellers model (Budyko 1969; Sellers 1969). The one-box model is known to insufficiently capture thermal inertia in the climate response and has been superseded by the two-box model (Gregory 2000; Held et al. 2010; Geoffroy et al. 2013a). Some recent studies have employed three-box models (Caldeira and Myhrvold 2013; Tsutsui (2016); Proistosescu and Huybers 2017; Fredriksen and Rypdal 2017). By taking the limit as $k \rightarrow \infty$ it is possible to approximate continuous vertical heat diffusion.

The k-box energy balance model (Fig. 1) used in this study is defined by the system of k linear differential equations:

$$C_1 \frac{dT_1}{dt} = F(t) - \kappa_1 T_1 - \kappa_2 (T_1 - T_2) + \xi(t), \qquad (1)$$

$$C_2 \frac{dT_2}{dt} = \kappa_2 (T_1 - T_2) - \kappa_3 (T_2 - T_3), \qquad (2)$$

$$\vdots$$

$$C_{k-1} \frac{dT_{k-1}}{dt} = \kappa_{k-1} (T_{k-2} - T_{k-1}) - \varepsilon \kappa_k (T_{k-1} - T_k), \qquad (3)$$

[a] Additional affiliation: Met Office Hadley Centre, Exeter, United Kingdom.

Corresponding author: Donald P. Cummins, dc533@exeter.ac.uk

DOI: 10.1175/JCLI-D-19-0589.1

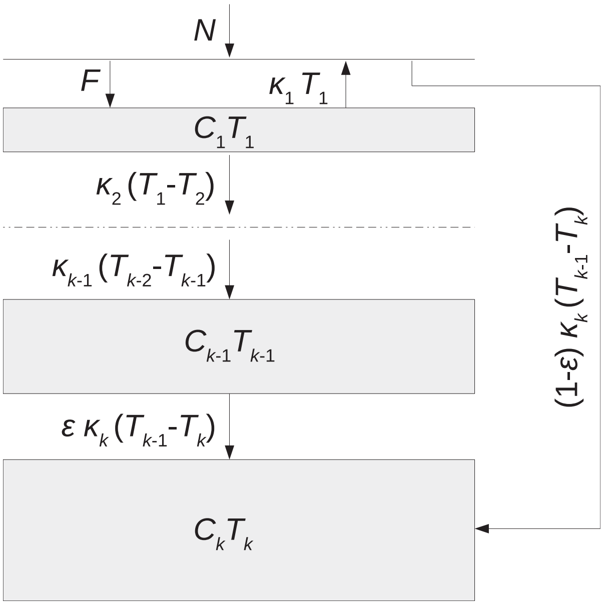
N
F
$\kappa_1 T_1$
$C_1 T_1$
$\kappa_2 (T_1 - T_2)$
$\kappa_{k-1} (T_{k-2} - T_{k-1})$
$C_{k-1} T_{k-1}$
$\varepsilon \kappa_k (T_{k-1} - T_k)$
$C_k T_k$
$(1-\varepsilon) \kappa_k (T_{k-1} - T_k)$

FIG. 1. Vertical layout of the boxes in the k-box energy balance model. The thickness of each box indicates its heat capacity, and the arrows represent the flow of heat between adjacent boxes. The top of the atmosphere has no heat capacity and so is represented by a horizontal line. The dashed line in the middle is an abbreviation of the intervening boxes.

$$C_k \frac{dT_k}{dt} = \kappa_k (T_{k-1} - T_k). \qquad (4)$$

The first box represents the atmosphere and uppermost layer of the ocean, while boxes 2 to k together represent the deep ocean. Each box i has a temperature T_i and heat capacity C_i and is coupled to adjacent boxes above and below; T_1 is defined to be global mean surface temperature (GMST) anomaly relative to preindustrial. Heat transfer coefficients $\kappa_i > 0$ determine the strength of thermal coupling between boxes i and $i - 1$. In the literature κ_1 is often written as λ and is referred to as the climate feedback parameter (e.g., Geoffroy et al. 2013a). We follow the convention of Fredriksen and Rypdal (2017) and use the letter κ for both climate feedback and heat uptake by the deep ocean. The heat transfer coefficient κ_k in the equation for box $k - 1$ is multiplied by a so-called efficacy factor $\varepsilon > 0$, introduced by Held et al. (2010), to simulate variation in the effective strength of κ_1 during periods of transient (nonequilibrium) warming. The term $F(t)$ denotes radiative forcing measured at the top of the atmosphere and $\xi(t)$ is a stochastic disturbance (see below). Table 1 contains physical units and a brief description of each parameter.

Natural variability in GMST can be partially explained within the EBM framework using a stochastic process in the radiative forcing term (Hasselmann 1976). To enforce continuity of F in time we model $F(t)$ as a red-noise process:

TABLE 1. Parameters of k-box model with physical units and description.

Parameter	Unit	Description
γ	Dimensionless	Stochastic forcing continuous-time autocorrelation parameter
C_i	W yr m^{-2} K^{-1}	Total heat capacity of box i
κ_i	W m^{-2} K^{-1}	Heat transfer coefficient; controls heat flux across upper boundary of box i.
ε	Dimensionless	Deep ocean heat uptake efficacy factor
σ_η	W m^{-2}	Standard deviation of TOA stochastic forcing component
σ_ξ	W m^{-2}	Standard deviation of stochastic disturbance applied to surface box
$F_{4\times CO_2}$	W m^{-2}	Effective radiative forcing after quadrupling preindustrial atmospheric CO$_2$
τ_i	Yr	The ith characteristic time scale of the k-box model
a_i	Dimensionless	Weighting of ith exponential basis function in step response of surface temperature
ECS	K	Equilibrium climate sensitivity; final temperature after doubling atmospheric CO$_2$
TCR	K	Transient climate response; surface temperature after 70 years of 1% yr^{-1} CO$_2$ increase.

$$\frac{dF}{dt} = -\gamma [F - F_{\text{det}}(t)] + \eta(t), \qquad (5)$$

where $F_{\text{det}}(t)$ and $\eta(t)$ are the respective deterministic and stochastic forcing components. Here we assume $\eta(t)$ to be a Gaussian white-noise (WN) process with mean zero and standard deviation σ_η. In the limit as $\gamma \to \infty$ the stochastic forcing becomes white noise, whereas if $\gamma \to 0$ we have a random walk. Interannual variation in radiative forcing is insufficient to explain all of the natural variability in surface temperature. Residual surface temperature variability is explained here by a Gaussian WN disturbance $\xi(t)$ with mean zero and standard deviation σ_ξ. The term $\xi(t)$ functions like an external forcing but is not measurable at the top of the atmosphere since it represents dynamic variability, which is generated internally.

As parameters of the k-box EBM do not correspond to well-defined physical quantities in the real world, it is not possible to calculate realistic parameter values directly from first principles. Parameter values must instead be estimated empirically from data. In this paper a maximum likelihood method is presented for estimating parameters of k-box models. The structure of the paper is as follows: section 2 provides a summary and critique of some methods previously employed to fit box models;

section 3 describes data requirements for successful parameter estimation and the specific data from phase 5 of the Coupled Model Intercomparison Project (CMIP5) used in this study; section 4 outlines the proposed maximum likelihood framework; section 5 describes a software tool created for applying the method described in this paper; section 6 evaluates the robustness of the proposed method in a simulation study; section 7 explains how the method was applied to climate model data from CMIP5 and presents an analysis of the results; and the content of the paper is summarized in section 8.

2. Methods for fitting *k*-box energy balance models

Maximum likelihood estimation is simple for one-box model parameters: given uniformly sampled data, estimation reduces to an ordinary least squares problem with a closed-form solution (Rypdal and Rypdal 2014). When more boxes are added, latent variables appear and the estimation problem becomes more difficult. Several methods have been proposed in the literature for estimating parameters of box models with $k \geq 2$, including least squares curve fitting (Geoffroy et al. 2013a; Caldeira and Myhrvold 2013), frequency-domain regression (Fredriksen and Rypdal 2017), and Bayesian estimation (Proistosescu and Huybers 2017; Jonko et al. 2018). Box models have previously been fitted to the historical record, to paleoclimate reconstructions, and to data from general circulation model experiments. Three examples of existing methods are described below. The first method described, proposed by Geoffroy et al. (2013a), is compared in section 7 with the new method proposed in this paper.

Geoffroy et al. (2013a) derived explicit time-dependent solutions for the two-box model under purely deterministic forcing scenarios. They proposed a procedure for estimating model parameters using measurements of GMST and top-of-the-atmosphere (TOA) net downward radiative flux (see section 3) from the step responses of AOGCMs in CMIP5. Their method uses prior information about characteristic time scales to estimate the model parameters in sequence, with the sum of squared residuals as the criterion to be minimized. The time-dependent solution of the two-box model is a sum of saturating exponentials and so estimating parameters in parallel by nonlinear least squares can be a notoriously difficult problem (Kaufmann 2003), which is avoided by estimating parameters sequentially. In a companion paper Geoffroy et al. (2013b) added a deep ocean heat uptake efficacy factor ε to their model, requiring the use of iteration in their fitting procedure. Geoffroy et al. (2013a) did not specify an error model; however, their least squares fitting criterion would correspond to maximum likelihood estimation under an assumption of errors which are independent and identically distributed (i.i.d.) and Gaussian. We have found this assumption to be inconsistent with time series of residuals obtained by subtracting fitted two-box model trajectories from AOGCM step responses: such residual time series exhibit strong autocorrelation. Without specifying an error model it is also impossible to correctly construct confidence intervals for parameter estimates.

Fredriksen and Rypdal (2017) estimated parameters of a three-box model with natural variability driven by a Gaussian WN process in the forcing term. They proposed an iterative least squares-based fitting algorithm to estimate the model parameters. Their method alternates between fitting the signal (expected temperature series for the first box) in the time domain and fitting the noise (time series of residuals) in the frequency domain. Fredriksen and Rypdal (2017) estimated model parameters using estimates of GMST from HadCRUT4 (Morice et al. 2012) and the Moberg et al. (2005) paleoclimate reconstructions, and forcing estimates from Crowley (2000) and Hansen et al. (2011). Unlike the other studies cited in this section, Fredriksen and Rypdal (2017) estimated parameters of box models ($k \geq 2$) without access to measurements of TOA net downward radiative flux, as they were fitting to historical datasets. Only a subset of the model parameters was estimated from data since, without radiative flux measurements, a wide range of possible values for the three characteristic time scales τ_1, τ_2, and τ_3 was found to be equally compatible with the observations. In their analysis three candidate time scale configurations were chosen and the remaining parameters estimated. An important result of Fredriksen and Rypdal (2017) is that the stochastically forced three-box model produces a similar noise spectrum to so-called scale-invariant models, a related class of simple climate model. Parameters of scale-invariant models have been estimated by maximum likelihood (Rypdal and Rypdal 2014) and more recently using Bayesian inference (Rypdal et al. 2018). The method of Rypdal et al. (2018) is generally applicable to linear response models, including box models, although the authors only present results for the scale-invariant model.

Jonko et al. (2018) estimated parameters of the two-box model using Bayesian hierarchical methods. In their model likelihood the variability in observed temperatures $T_1(t)$ and TOA net downward radiative flux $N(t)$ are jointly modeled as a vector autoregressive process of order one [VAR(1)]. All VAR(1) correlations are considered free parameters, not constrained by the physical parameters of the EBM. Given prior distributions for parameters to be estimated, Markov chain Monte Carlo (MCMC) is used to form an approximation to the posterior distribution. Jonko et al. (2018) used their method to pool information from 24 AOGCM step responses and produce a joint posterior for equilibrium climate sensitivity (ECS). They also included time series of historical temperature observations in their model likelihood to further constrain estimates of future warming. By opting not to include a stochastic forcing term Jonko et al. (2018) increase the number of parameters to estimate and lose physical motivation for the natural temperature variability in their model.

Of the approaches considered above, none can be considered optimal in the sense of maximum likelihood or sampling from the posterior distribution of the full, stochastic k-box energy balance model. We therefore propose to develop a maximum likelihood method for estimating stochastic k-box models with $k \geq 2$. Maximum likelihood estimators are widely used and have known asymptotic sampling properties allowing for simple quantification of uncertainty. Furthermore, optimal complexity of maximum likelihood models can be identified using information criteria.

3. Step response and CMIP5 data

The k-box model is a linear time-invariant system and is therefore completely characterized by its impulse response or alternatively its step response (of which the impulse response is the time derivative). The step response contains information about model behavior on all relevant time scales. The CMIP5 archive includes experiments (Taylor et al. 2012) designed to elicit the step response of AOGCMs by subjecting them to a step forcing of the form

$$F(t) = \begin{cases} F_{4\times\mathrm{CO}_2} & \text{if} \quad t \geq 0, \\ 0 & \text{otherwise}. \end{cases} \tag{6}$$

The forcing is achieved by an instant quadrupling of atmospheric carbon dioxide (CO_2) concentration. The reasoning behind this choice of forcing is that the amplitude should be large enough that the signal-to-noise ratio is high, but small enough not to induce strongly nonlinear behavior such as tipping points. Ideally the step-forcing experiment would be long enough for the system to stabilize at a new equilibrium temperature and multiple ensemble runs would be available for each AOGCM. However, since Earth system models (ESMs) are expensive to run, the step-forcing experiments in CMIP5 are typically 150 years in length and consist of a single ensemble member. These experiments nevertheless constitute the most information-rich datasets from which to infer the parameters of k-box models and simple climate models in general. The output of an AOGCM step-forcing experiment can even be used on its own to make climate predictions by convolving it with a forcing signal of interest (Good et al. 2011; Lucarini et al. 2017).

The models in CMIP5 have equilibration times in the thousands of years, meaning that a 150-yr time series of temperatures contains insufficient information to identify all model parameters. Attempting to fit to such datasets results in massively correlated parameter estimates with correspondingly large uncertainty. This difficulty can be overcome by using measurements of net downward radiative flux at the top of the atmosphere (TOA) to constrain κ_1. Using Eqs. (1) and (3) we extract the relation

$$N(t) = F(t) - \kappa_1 T_1(t) + (1 - \varepsilon)\kappa_k[T_{k-1}(t) - T_k(t)], \tag{7}$$

where $N(t)$ denotes the TOA net downward radiative flux. If the system is in equilibrium at time t, that is, $T_{k-1}(t) = T_k(t)$, and/or if $\varepsilon = 1$, Eq. (7) reduces to the traditional Gregory relation $N(t) = F(t) - \kappa_1 T_1(t)$ (Gregory et al. 2004). Note that, since fitting to 4 × CO_2 experiments is essentially not feasible without measurements of $N(t)$, fitting to historical temperature observations with all parameters free is unlikely to produce meaningfully constrained estimates.

4. Maximum likelihood framework

Computing the likelihood function for the k-box model is nontrivial. We typically observe the temperature of only the first box and hence for $k \geq 2$ at least half of the model state variables are unobserved (latent). In this section we start by obtaining a rigorous state-space formulation of the k-box model. We then show how the likelihood of this state-space representation can be evaluated recursively using the Kalman filter. Numerical maximization of the likelihood is briefly described and a method for constructing confidence intervals given. Finally, we explain how optimal model complexity can be identified using information criteria.

a. Matrix representation

The purely deterministic, homogeneous (externally and internally unforced) k-box model with $\varepsilon = 1$ can be written in matrix form:

$$\dot{\mathbf{x}}_h(t) = \mathbf{A}\mathbf{x}_h(t), \tag{8}$$

where

$$\mathbf{x}_h(t) = [T_1(t), \ldots, T_k(t)]', \tag{9}$$

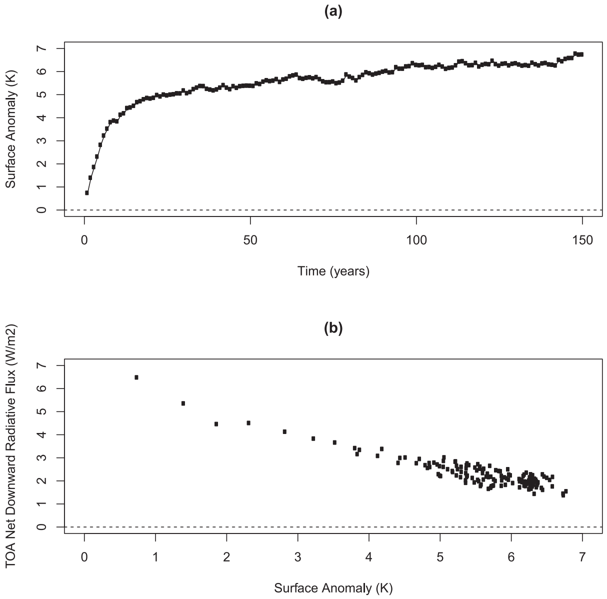

FIG. 2. Example simulated dataset from a two-box model with parameters: $\gamma = 1.58$; C_1, C_2 = 7.73, 89.3 W yr m^{-2} K^{-1}; κ_1, κ_2 = 0.632, 0.522 W m^{-2} K^{-1}; $\varepsilon = 1.52$; σ_η, σ_ξ, $F_{4\times\mathrm{CO_2}}$ = 0.428, 0.643, 6.86 W m^{-2}. (a) Increasing surface temperatures during the first 150 years after CO$_2$ quadrupling. (b) Values of TOA net downward radiative flux in each year plotted against the corresponding surface temperature.

and

$$A_{i,j} = \begin{cases} -(\kappa_i + \kappa_{i+1})/C_i, & \text{if} \quad i = j \neq k, \\ \kappa_j/C_i, & \text{if} \quad j = i+1, \\ \kappa_i/C_i, & \text{if} \quad j = i-1, \\ -\kappa_i/C_i, & \text{if} \quad i = j = k, \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

For $\varepsilon \neq 1$ two entries in the penultimate row of $\mathbf{A}$ must be changed to match Eq. (3). The matrix $\mathbf{A}$ is tridiagonal because each box is coupled only to its immediate neighbors above and below. The ith eigenvalue of $\mathbf{A}$ is $-1/\tau_i$ where τ_i is the ith characteristic time scale of the linear system. An analytical expression for τ_i is given in Geoffroy et al. (2013a) for the case $k = 2$. See appendix A for a proof that $\mathbf{A}$ has real and nonpositive eigenvalues for any k when $\varepsilon = 1$.

Analysis of the full inhomogeneous, stochastic k-box model is simplified by the inclusion of radiative forcing F as a state variable. Defining the state vector

$$\mathbf{x}(t) = [F(t), T_1(t), \ldots, T_k(t)]', \tag{11}$$

we can write the full model

$$\dot{\mathbf{x}}(t) = \mathbf{A}^+\mathbf{x}(t) + \mathbf{b}u(t) + \mathbf{w}(t), \tag{12}$$

where $\mathbf{A}^+$ is simply matrix $\mathbf{A}$ augmented with one additional row–column pair (above and to the left) to account for Eq. (5):

$$A^+_{1,1} = -\gamma, \tag{13}$$

$$A^+_{2,1} = 1/C_1; \tag{14}$$

and where

$$\mathbf{b} = (\gamma, 0, \ldots, 0)', \tag{15}$$

$$u(t) = F_{\text{det}}(t), \tag{16}$$

and

$$\mathbf{w}(t) \sim N(\mathbf{0}, \mathbf{Q}) \tag{17}$$

with

$$Q_{i,j} = \begin{cases} \sigma_\eta^2, & \text{if} \quad i = j = 1, \\ (\sigma_\xi/C_1)^2, & \text{if} \quad i = j = 2, \\ 0, & \text{otherwise}. \end{cases} \tag{18}$$

b. Discretization scheme

The continuous-time model is a system of stochastic differential equations and may be analyzed using the tools of stochastic calculus. However if observations consist of uniformly spaced discrete samples then it makes sense to discretize the model (see section 7 for details of sampled data used in this study). Assuming constancy of the deterministic forcing input $u(t) = F_{\text{det}}(t)$ between samples, the model can be discretized exactly (see appendix B)

$$\mathbf{x}(t) = \mathbf{A}_d\mathbf{x}(t-1) + \mathbf{b}_d u(t-1) + \mathbf{w}_d(t), \tag{19}$$

where

$$\mathbf{A}_d = e^{\mathbf{A}^+}, \tag{20}$$

$$\mathbf{b}_d = (\mathbf{A}^+)^{-1}(\mathbf{A}_d - \mathbf{I})\mathbf{b}, \tag{21}$$

$$\mathbf{w}_d(t) \sim N(\mathbf{0}, \mathbf{Q}_d), \quad \text{and} \tag{22}$$

$$\mathbf{Q}_d = \int_{s=0}^{1} e^{\mathbf{A}^+ s}\mathbf{Q}e^{\mathbf{A}^{+\prime} s}\, ds, \tag{23}$$

with subscript d denoting discretization. The integral in Eq. (23) can be evaluated via the matrix exponential method described in section 1 of Van Loan (1978).

c. State-space representation

As a linear time-invariant system the k-box model is amenable to powerful numerical techniques from

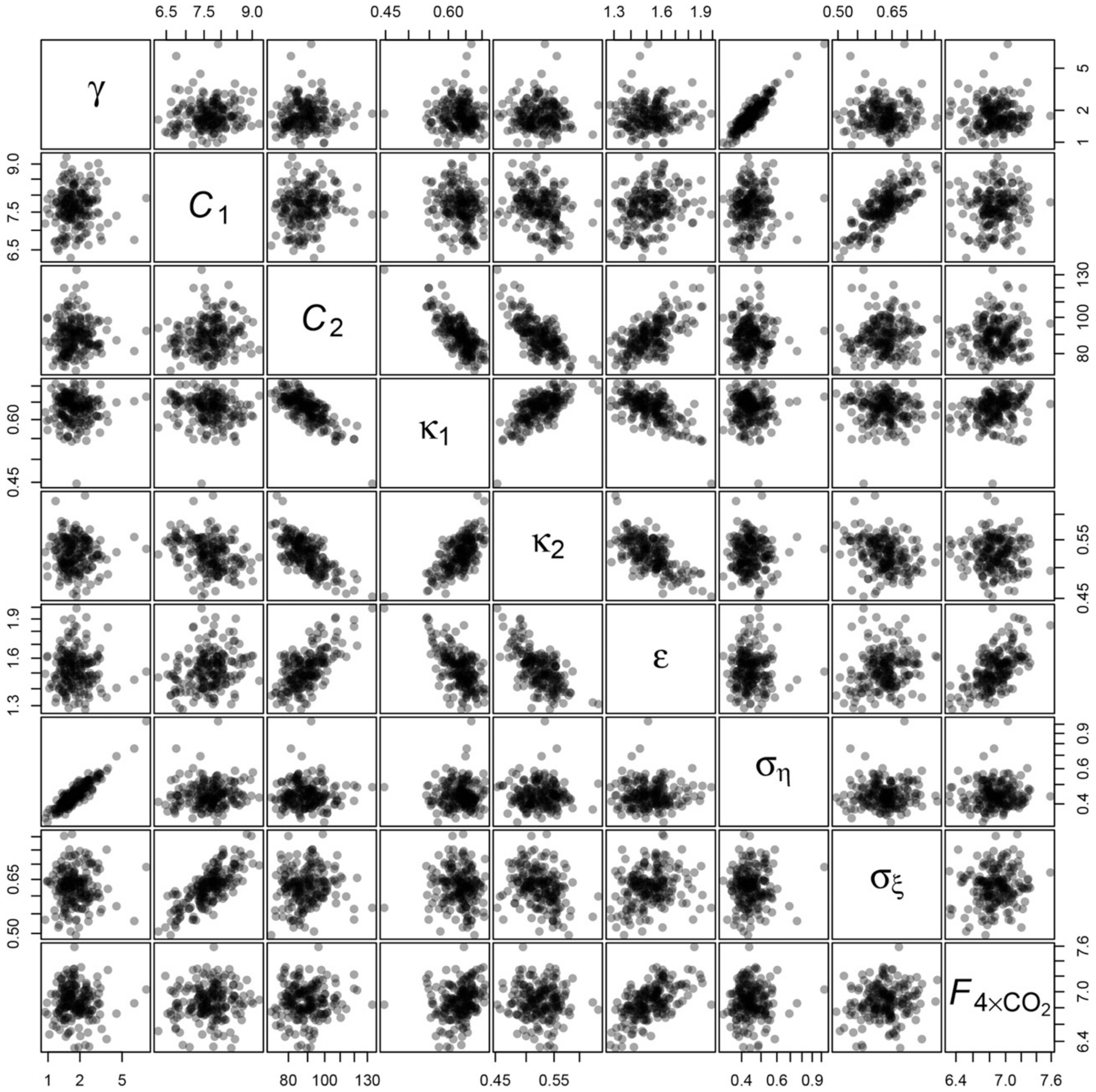

FIG. 3. Pairs plot showing approximate sampling distribution of the maximum likelihood estimator. Each point represents a model fitted to a simulated dataset. Simulated datasets are from a two-box model with parameters: $\gamma = 1.58$; C_1, $C_2 = 7.73, 89.3$ W yr m^{-2} K^{-1}; κ_1, $\kappa_2 = 0.632, 0.522$ W m^{-2} K^{-1}; $\varepsilon = 1.52$; σ_η, σ_ξ, $F_{4\times\mathrm{CO}_2} = 0.428, 0.643, 6.86$ W m^{-2}. Plot axes are logarithmic to increase visibility of parameter correlations.

control theory, in particular the Kalman filter (Kalman 1960). By choosing a model for our observation process $\mathbf{y}(t)$ we can write the k-box model in state-space form

$$\mathbf{x}(t) = \mathbf{A}_d\mathbf{x}(t-1) + \mathbf{b}_d u(t-1) + \mathbf{w}_d(t), \qquad (24)$$

$$\mathbf{y}(t) = \mathbf{C}_d\mathbf{x}(t) + \mathbf{v}_d(t), \qquad (25)$$

where matrix $\mathbf{C}_d$ is our observation operator and $\mathbf{v}_d(t)$ is an (optional) additive observation error. If we observe TOA net downward radiative flux $N(t)$ and surface temperature $T_1(t)$ at each time t, both without error, then

$$\mathbf{y}(t) = [T_1(t), N(t)]' = \mathbf{C}_d\mathbf{x}(t), \qquad (26)$$

where entries of $\mathbf{C}_d$ are determined by Eq. (7). In the general case (e.g., the historical record) our observations might be contaminated by errors $\mathbf{v}_d(t)$ such that

$$\mathbf{v}_d(t) \sim N(\mathbf{0}, \mathbf{\Sigma}_t), \tag{27}$$

but for climate model experiments we assume $\mathbf{v}_d(t) = \mathbf{0}$ for all t.

d. Kalman filter

The Kalman filter was originally developed as a minimum mean-square-error (MMSE) estimator of state variables in a noisy linear dynamic system (Kalman 1960). It may also be used to recursively calculate the likelihood of a time series of observations from this class of model (Tusell 2011). The Kalman filter estimates the system state at time t using the information contained in all previous observations up to and including time t, through a recursive procedure iterating over two steps: a prediction step and an update step. For the k-box model in state-space form we can write the Kalman recursions as follows, using the hat/subscript notation of Reid (2001).

1) Prediction step

Given $\hat{\mathbf{x}}_{t-1|t-1}$, our best estimate of the system state at time $t-1$ given data, the predicted state $\hat{\mathbf{x}}_{t|t-1}$ at time t is

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{A}_d \hat{\mathbf{x}}_{t-1|t-1} + \mathbf{b}_d u_{t-1}. \tag{28}$$

The predicted error covariance of this a priori estimate is

$$\mathbf{P}_{t|t-1} = \mathbf{A}_d \mathbf{P}_{t-1|t-1} \mathbf{A}'_d + \mathbf{Q}_d, \tag{29}$$

where $\mathbf{P}_{t-1|t-1}$ is the covariance of the estimated state at time $t-1$.

2) Update step

Having then observed $\mathbf{y}_t$ we update our a priori estimate of $\mathbf{x}_t$ with this new information to obtain an a posteriori state estimate $\hat{\mathbf{x}}_{t|t}$ with corresponding covariance $\mathbf{P}_{t|t}$. Our measurement prefit residual is

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{C}_d \hat{\mathbf{x}}_{t|t-1} \tag{30}$$

which has covariance

$$\mathbf{S}_t = \sum\nolimits_t + \mathbf{C}_d \mathbf{P}_{t|t-1} \mathbf{C}'_d. \tag{31}$$

Our a posteriori state estimate is simply our a priori estimate $\hat{\mathbf{x}}_{t|t-1}$ shrunk toward the observation $\mathbf{y}_t$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t \tilde{\mathbf{y}}_t \tag{32}$$

where the shrinkage amplitude is the optimal Kalman gain

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{C}'_d \mathbf{S}_t^{-1}. \tag{33}$$

Table 2. CMIP5 climate model expansions (Geoffroy et al. 2013a).

Model	Expansion
BCC-CSM1.1	Beijing Climate Center, Climate System Model, version 1.1
BNU-ESM	Beijing Normal University–Earth System Model
CanESM2	Canadian Earth System Model, version 2
CCSM4	Community Climate System Model, version 4
CNRM-CM5	Centre National de Recherches Météorologiques Coupled Global Climate Model, version 5
CSIRO-Mk3.6.0	Commonwealth Scientific and Industrial Research Organization Mark, version 3.6.0
FGOALS-s2	Flexible Global Ocean-Atmosphere-Land System Model gridpoint, second spectral version
GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory Earth Science Model 2M
GISS-E2-R	Goddard Institute for Space Studies Model E, coupled with Russell ocean model
HadGEM2-ES	Hadley Centre Global Environmental Model 2, Earth System
INM-CM4	Institute of Numerical Mathematics Coupled Model, version 4.0
IPSL-CM5A-LR	L'Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with NEMO, low resolution
MIROC5	Model for Interdisciplinary Research on Climate, version 5
MPI-ESM-LR	Max Planck Institute Earth System Model, low resolution
MRI-CGCM3	Meteorological Research Institute Coupled General Circulation Model, version 3
NorESM1-M	Norwegian Earth System Model, intermediate resolution

The covariance of the a posteriori estimate is

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \mathbf{C}_d) \mathbf{P}_{t|t-1} (\mathbf{I} - \mathbf{K}_t \mathbf{C}_d)' + \mathbf{K}_t \mathbf{\Sigma}_t \mathbf{K}'_t. \tag{34}$$

The measurement postfit residual is

$$\tilde{\mathbf{y}}_{t|t} = \mathbf{y}_t - \mathbf{C}_d \hat{\mathbf{x}}_{t|t}. \tag{35}$$

In the complete absence of observational noise the recursions may still be computed by setting $\mathbf{\Sigma}_t$ equal to a diagonal matrix with each diagonal element a very small number.

e. Model likelihood

Since the k-box model is a causal linear filter (i.e., system states depend on past states and past inputs but not on future states and future inputs), we can factorize the likelihood function of the temperature observations

TABLE 3. Estimated parameters for optimal k-box emulators of ESMs in CMIP5. In all cases $k = 3$. For physical units and descriptions of parameters see Table 1. MMM refers to the k-box model fitted to the average of the datasets from all 16 ESMs.

Model	γ	C_1	C_2	C_3	κ_1	κ_2	κ_3	ε	σ_η	σ_ξ	$F_{4\times CO_2}$
BCC-CSM1.1	2.9	5.3	12.3	49	1.21	1.7	0.79	1.28	0.46	0.40	7.1
BNU-ESM	2.3	4.0	9.9	85	0.94	1.6	0.71	0.98	0.60	0.66	7.4
CanESM2	2.5	4.6	11.1	66	1.01	1.8	0.81	1.24	0.53	0.52	7.9
CCSM4	2.1	4.4	13.0	70	1.28	2.3	1.05	1.44	0.49	0.49	8.0
CNRM-CM5.1	11.5	4.0	9.6	90	1.14	2.4	0.60	0.90	0.83	0.41	7.2
CSIRO-Mk3.6.0	1.7	3.6	16.0	63	0.59	2.4	1.15	1.73	0.70	0.50	6.1
FGOALS-s2	2.3	4.3	8.1	135	0.86	2.2	1.11	1.19	0.82	0.66	7.9
GFDL-ESM2M	3.3	4.8	10.2	114	1.34	2.6	1.13	1.19	0.77	0.56	6.9
GISS-E2-R	1.6	4.9	31.6	107	1.82	1.7	4.66	1.46	0.32	0.30	8.3
HadGEM2-ES	1.7	3.6	9.5	99	0.54	2.4	0.63	1.59	0.43	0.32	6.4
INM-CM4	1.6	4.3	7.9	275	1.66	2.7	0.81	0.78	0.33	0.32	6.3
IPSL-CM5A-LR	1.9	2.7	16.7	101	0.73	2.4	0.63	1.21	0.50	0.38	6.5
MIROC5	1.8	4.7	17.9	139	1.55	1.7	1.33	1.18	0.54	0.89	8.7
MPI-ESM-LR	2.5	4.4	13.7	70	1.12	2.0	0.91	1.44	0.68	0.71	8.9
MRI-CGCM3	2.6	4.5	14.5	61	1.26	2.2	0.71	1.22	0.56	0.40	6.8
NorESM1-M	2.2	5.2	13.4	105	1.08	2.6	1.29	1.50	0.52	0.47	7.0
MMM	1.9	5.1	11.2	89	1.03	2.0	0.99	1.29	0.15	0.15	7.2

$$\mathscr{L}(\mathbf{y}_1,\ldots,\mathbf{y}_n;\boldsymbol{\theta}) = \prod_{t=1}^{n} \mathscr{L}(\mathbf{y}_t|\mathbf{y}_{t-1},\ldots,\mathbf{y}_1;\boldsymbol{\theta}), \tag{36}$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters. For numerical stability it is preferable to compute the log-likelihood

$$\ell(\mathbf{y}_1,\ldots,\mathbf{y}_n;\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell(\mathbf{y}_t|\mathbf{y}_{t-1},\ldots,\mathbf{y}_1;\boldsymbol{\theta}), \tag{37}$$

which can be calculated recursively using the prefit residuals and their corresponding covariances from the Kalman filter

$$\ell(\mathbf{y}_1,\ldots,\mathbf{y}_t;\boldsymbol{\theta}) = \ell(\mathbf{y}_1,\ldots,\mathbf{y}_{t-1};\boldsymbol{\theta}) - \frac{1}{2}[2\log(2\pi) + \log(|\mathbf{S}_t|) + \tilde{\mathbf{y}}_t'\mathbf{S}_t^{-1}\tilde{\mathbf{y}}_t]. \tag{38}$$

Evaluation of (38) requires the distribution of $\mathbf{x}_0|\boldsymbol{\theta}$, upon which $\mathbf{y}_1$ depends, to be known. If we assume that at the beginning of a dataset the system is in a state of preindustrial equilibrium then $E(\mathbf{x}_0) = \mathbf{0}$ with some covariance matrix to be derived from the model parameters (see appendix C). For an abrupt 4 × CO$_2$ climate model experiment, the first element of $\mathbf{x}_0$ (corresponding to radiative forcing) has an expected value, given the model parameters, of $F_{4\times CO_2}$.

TABLE 4. Characteristic time scales τ_i, surface temperature response coefficients a_i, equilibrium climate sensitivity (ECS), and transient climate response (TCR) of optimal k-box emulators fitted to ESMs in CMIP5. Column ΔAIC shows the decrease in AIC moving from two to three boxes. For physical units and descriptions of parameters see Table 1. MMM refers to the k-box model fitted to the average of the datasets from all 16 ESMs.

Model	τ_1	τ_2	τ_3	a_1	a_2	ECS	TCR	ΔAIC
BCC-CSM1.1	1.54	7.8	162	0.28	0.33	2.9	1.9	21.0
BNU-ESM	1.32	8.8	272	0.25	0.38	3.9	2.5	17.1
CanESM2	1.34	7.6	220	0.23	0.34	3.9	2.3	21.0
CCSM4	1.05	6.1	201	0.25	0.30	3.1	1.9	29.0
CNRM-CM5.1	0.91	8.6	259	0.21	0.49	3.2	2.1	40.2
CSIRO-Mk3.6.0	1.03	6.8	315	0.14	0.18	5.2	1.9	32.0
FGOALS-s2	1.03	5.5	393	0.14	0.36	4.6	2.3	8.9
GFDL-ESM2M	0.96	5.6	262	0.20	0.38	2.6	1.5	11.2
GISS-E2-R	1.34	3.7	235	0.46	0.10	2.3	1.4	21.3
HadGEM2-ES	0.95	8.2	532	0.10	0.31	5.9	2.4	43.1
INM-CM4	0.78	5.9	551	0.23	0.52	1.9	1.4	33.0
IPSL-CM5A-LR	0.78	13.2	394	0.19	0.33	4.4	2.2	75.7
MIROC5	1.31	7.8	321	0.39	0.24	2.8	1.8	5.5
MPI-ESM-LR	1.23	7.4	231	0.26	0.29	4.0	2.3	16.4
MRI-CGCM3	1.12	9.4	190	0.27	0.36	2.7	1.7	38.5
NorESM1-M	1.12	5.9	302	0.17	0.29	3.2	1.6	13.9
MMM	1.35	6.9	273	0.20	0.34	3.5	2.0	68.8

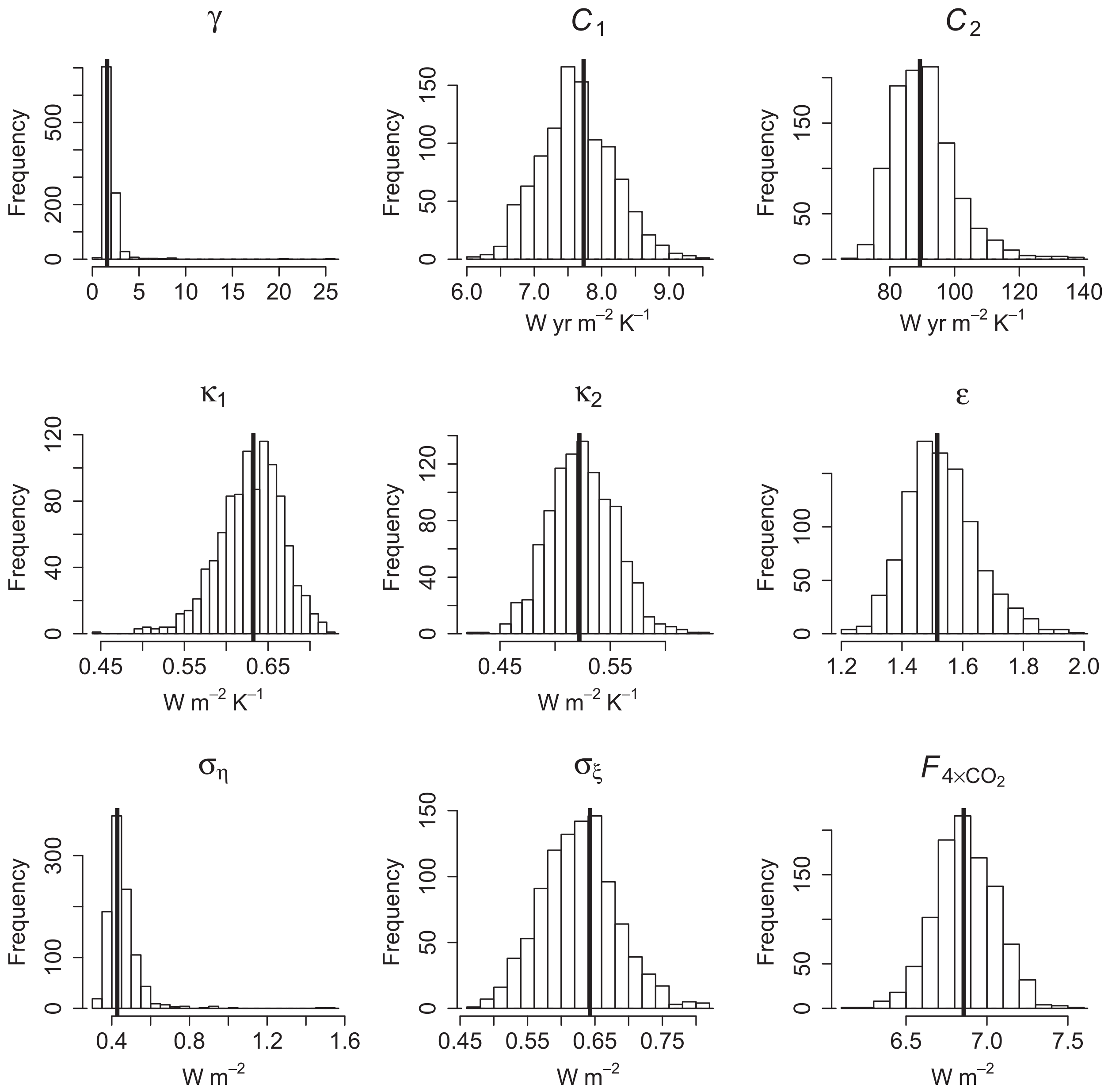

FIG. 4. Histograms showing approximate sampling distribution of the maximum likelihood estimator. The thick vertical lines indicate the true value of each parameter. Simulated datasets are from a two-box model with parameters: $\gamma = 1.58$; C_1, $C_2 = 7.73$, 89.3 W yr m^{-2} K^{-1}; κ_1, $\kappa_2 = 0.632, 0.522$ W m^{-2} K^{-1}; $\varepsilon = 1.52$; σ_η, σ_ξ, $F_{4\times \mathrm{CO}_2} = 0.428,\ 0.643,\ 6.86$ W m^{-2}.

The Kalman filter log-likelihood is essentially a weighted least squares objective function which penalizes squared one-step-ahead prediction errors (prefit residuals). The weighting applied to each prediction error is determined by its corresponding uncertainty (covariance).

f. Maximum likelihood estimation

The maximum likelihood estimator (MLE) of the model parameters $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}} \min[-\ell(\mathbf{y};\boldsymbol{\theta})], \tag{39}$$

where $\ell(\mathbf{y};\boldsymbol{\theta})$ denotes the k-box model log-likelihood function. We minimize the negative log-likelihood numerically: a modern derivative-free algorithm such as BOBYQA (Powell 2009) is well suited to this task. Standard errors and confidence intervals can be obtained using asymptotic properties of the MLE. In the limit as sample size tends to infinity the MLE $\hat{\boldsymbol{\theta}}$ is

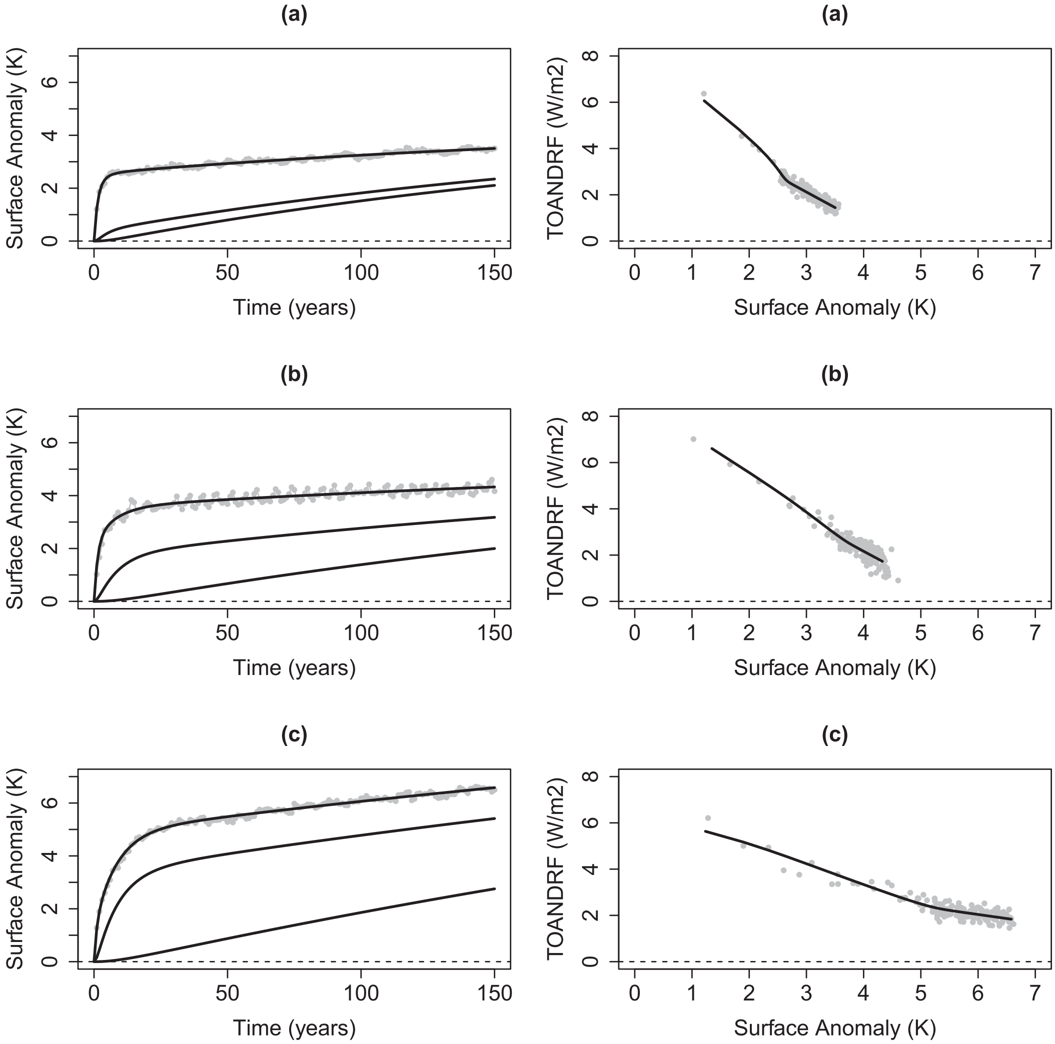

FIG. 5. Observed and fitted three-box step responses of three ESMs from CMIP5. (left) Temperature trajectories for each box. (right) TOA net downward radiative fluxes against surface temperature. Gray dots are observations while the black curves are expected box-model responses. Models are (a) GISS-E2-R, (b) MIROC5, and (c) HadGEM2-ES.

normally distributed with mean vector $\boldsymbol{\theta}$ and covariance matrix $\mathbf{I}^{-1}$ where $\mathbf{I}$ denotes the Fisher information matrix

$$\mathbf{I}_{jk} = -E\left[\frac{\partial^2 \ell(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k}\right]. \tag{40}$$

Although the Fisher information $\mathbf{I}$ depends on the values of the true parameters $\boldsymbol{\theta}$, we can obtain a consistent estimator $\hat{\mathbf{I}}$ by plugging the MLE $\hat{\boldsymbol{\theta}}$ into Eq. (40). We calculate $\hat{\mathbf{I}}$ using a numerical estimate of the Hessian of the negative log-likelihood at the MLE $\hat{\boldsymbol{\theta}}$. The estimated asymptotic sampling distribution of the MLE is then used to calculate standard errors and confidence intervals.

g. *Optimal model complexity*

The number of boxes k offers a natural parameterization of model complexity. When emulating an AOGCM with an EBM it is desirable to fit the most parsimonious model that does not significantly underperform compared to more complex models. Models with different numbers of boxes k can be compared (e.g., Caldeira and Myhrvold 2013) using Akaike's information criterion (AIC). The AIC score for a fitted model m is defined as

$$\mathrm{AIC}(m) = -2\ell(m) + 2p(m), \tag{41}$$

where ℓ is the log-likelihood and p is the number of parameters (Akaike 1974). The k-box model m_k has

$p(m_k) = 2k + 5$ since we have k heat capacities C_i, k heat transfer coefficients κ_i, a radiative forcing $F_{4\times\mathrm{CO}_2}$, two standard deviations σ_η and σ_ξ, and two dimensionless parameters γ and ε. We have

$$\mathrm{AIC}(m_k) = -2\ell(m_k) + 4k + 10. \qquad (42)$$

Competing models can be compared using the decision rule whereby for a given AOGCM we choose the number of boxes k that minimizes $\mathrm{AIC}(m_k)$.

5. Software implementation

This study has developed a package for the R software environment (R Core Team 2019) for simulation, fitting, filtering, and predicting with k-box EBMs. The package estimates parameters of k-box models from time series of GMST and TOA net downward radiative flux by numerically maximizing the likelihood function. To evaluate the model likelihood we use modern implementations of the matrix exponential (Goulet et al. 2019) and the Kalman filter (Luethi et al. 2018). The likelihood is maximized using an implementation (Johnson 2014; Ypma 2020) of the BOBYQA optimization algorithm (Powell 2009). Confidence intervals for parameter estimates are obtained using the Fisher information, as described in section 4f, where the Hessian of the likelihood function is evaluated numerically using an implementation of Richardson's extrapolation (Gilbert and Varadhan 2016). The R package, which includes the datasets used in this paper, is available for download at https://github.com/donaldcummins/EBM.

6. Simulation study

a. Methods

A simulation study was performed to investigate the feasibility of fitting k-box models to AOGCM step response data via the proposed maximum likelihood method. The step response of HadGEM2-ES from CMIP5 was used to fit a two-box model and a three-box model (optimal under AIC). HadGEM2-ES was chosen as this model has been used extensively for climate change studies. Data from HadGEM2-ES consisted of annually averaged values (see section 7 for details of CMIP5 data used). Estimated two-box model parameters were $\gamma = 1.58$; C_1, C_2 = 7.73, 89.3 W yr m^{-2} K^{-1}; κ_1, κ_2 = 0.632, 0.522 W m^{-2} K^{-1}; $\varepsilon = 1.52$; σ_η, σ_ξ, and $F_{4\times\mathrm{CO}_2}$ = 0.428, 0.643, 6.86 W m^{-2}. Estimated parameters for the three-box model were: $\gamma = 1.73$; C_1, C_2, C_3 = 3.62, 9.47, 98.7 W yr m^{-2} K^{-1}; κ_1, κ_2, κ_3 = 0.536, 2.39, 0.634 W m^{-2} K^{-1}; $\varepsilon = 1.59$; σ_η, σ_ξ $F_{4\times\mathrm{CO}_2}$ = 0.434, 0.323, 6.35 W m^{-2}. Each of the two fitted models was used to generate 1000 simulated step responses (see Fig. 2). Parameters were then estimated for each of the simulated datasets using the same maximum likelihood methodology. The resulting sets of parameter estimates form a Monte Carlo approximation to the estimator sampling distributions (see Figs. 3 and 4).

TABLE 5. Example approximate 95% confidence intervals for parameters of k-box models fitted to HadGEM2-ES. For physical units and descriptions of parameters see Table 1.

	$k = 2$			$k = 3$		
Parameter	MLE	2.5%	97.5%	MLE	2.5%	97.5%
γ	1.58	1.04	2.41	1.73	1.15	2.60
C_1	7.73	6.64	9.01	3.62	2.98	4.39
C_2	89.29	73.02	109.18	9.47	7.61	11.80
C_3				98.66	84.10	115.74
κ_1	0.63	0.56	0.71	0.54	0.46	0.63
κ_2	0.52	0.46	0.59	2.39	1.82	3.12
κ_3				0.63	0.57	0.71
ε	1.52	1.30	1.77	1.59	1.38	1.83
σ_η	0.43	0.35	0.52	0.43	0.35	0.53
σ_ξ	0.64	0.53	0.77	0.32	0.27	0.39
$F_{4\times\mathrm{CO}_2}$	6.86	6.46	7.28	6.35	6.03	6.70

b. Results

Estimator sampling distributions for the two-box and three-box models were examined for excessive bias, variance, and pairwise correlations. Results for the two-box model simulations are discussed below. Analysis of results for the three-box model leads to analogous conclusions.

Pairwise parameter correlations are visible in the estimated sampling distribution of the two-box model estimator (see Fig. 3). The strongest correlation (positive) is between the parameters controlling the stochastic forcing, γ and σ_η; that is, the whiter the noise, the greater the disturbance needed at each time step to obtain the same overall level of variability. The second strongest correlation (negative) is between the climate feedback parameter κ_1 and deep ocean heat capacity C_2. The third strongest correlation (positive) is between C_1 and σ_ξ. A natural explanation for this is that when the heat capacity of the first box C_1 is increased the corresponding temperature T_1 has more inertia and hence requires a stronger stochastic disturbance amplitude σ_ξ to maintain the same level of variability. The fourth strongest correlation (negative) is between κ_2 and C_2. This correlation is related to the time taken for relaxation of the system on the longer time scale τ_2. A longer relaxation time can be achieved either by increasing the heat capacity of the second box C_2 or by reducing the heat transfer coefficient κ_2 between boxes one and two.

Model parameters can be divided, by correlation, into two disjoint sets: set (i), stochastic forcing parameters

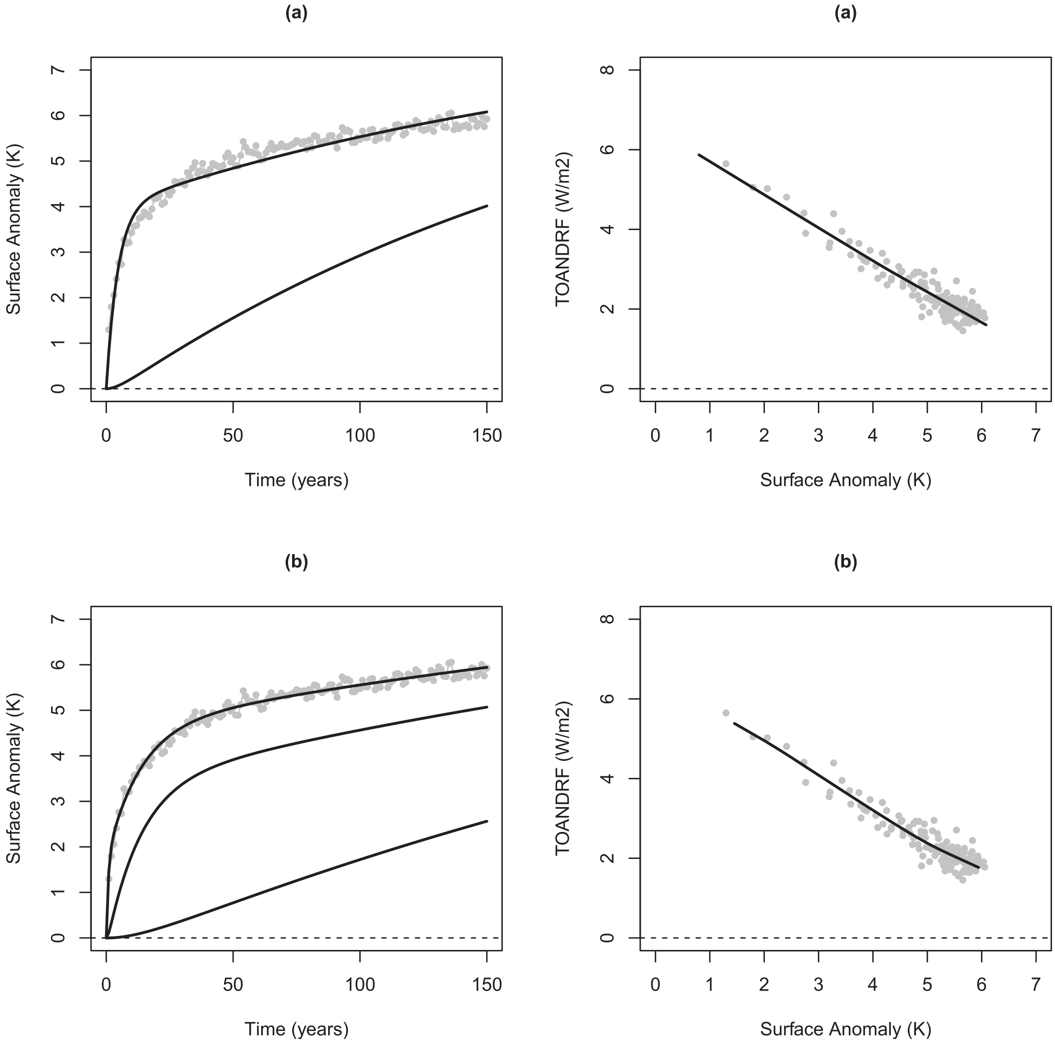

FIG. 6. (a) Two-box and (b) three-box fits to the step response of IPSL-CM5A-LR. (left) Temperature trajectories for each box. (right) TOA net downward radiative fluxes against surface temperature. Gray dots are observations while the black curves are expected box-model responses.

γ and σ_η; and set (ii), all remaining parameters. Neither γ nor σ_η is correlated with any parameter in set (ii), nor does either parameter appear well constrained by the simulated datasets, the consequence being a mutual inflation of uncertainty. The correlations between parameters in set (ii) appear to act favorably: individual parameter uncertainty in set (ii) is uniformly low with coefficients of variation mostly less than 10%. If at least one parameter in set (ii) is well constrained by observations, as appears to be the case, then uncertainty in the other parameters decreases as a result.

Estimated marginal distributions of the two-box model parameters resemble unimodal bell curves (see Fig. 4), with the notable exception of γ and σ_η. The parameter γ appears poorly bounded from above (hard to rule out very white stochastic forcing) and this uncertainty propagates into σ_η. The maximum likelihood estimator is asymptotically unbiased but in general has a finite sample bias. Estimates of all two-box model parameters display some bias. Parameters γ and σ_η have positive relative biases of 21% and 6% respectively, which is unsurprising given the skewness of their marginal distributions. For parameters in set (ii), and for both the two-box and three-box models, the magnitude of the bias is in all cases less than 5% of the parameter's true value.

c. Conclusions

The simulation study demonstrates that the proposed maximum likelihood method reliably estimates parameters of two-box and three-box box models from the step response of a typical AOGCM from CMIP5. Pairwise correlation and estimator bias were found to influence estimates

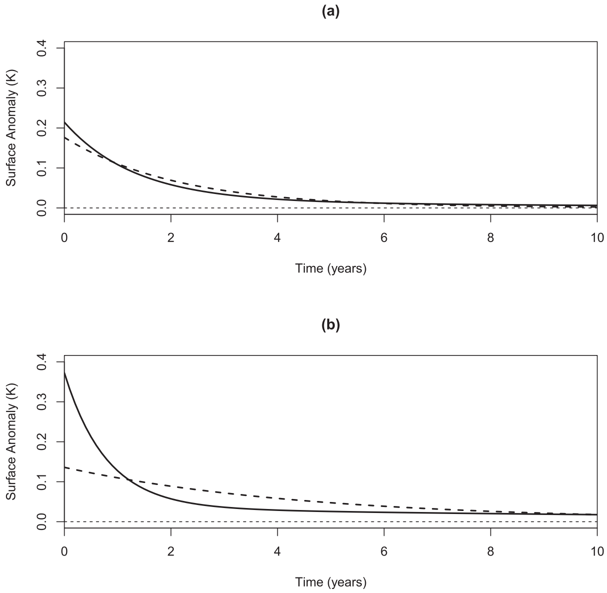

FIG. 7. Impulse responses of fitted k-box models. The curves are the expected temperature trajectories of the first box of the fitted models in response to a unit-impulse forcing. The solid and dashed curves correspond to three-box and two-box fits respectively, fitted using maximum likelihood. Models are (a) MIROC5 and (b) IPSL-CM5A-LR.

of stochastic forcing parameters γ and σ_η; however, other model parameters were not adversely affected.

7. Fitting to CMIP5 climate model simulations

The R package was used to fit two-box and three-box models to the step responses of 16 ESMs from CMIP5 (see Table 2), using the same data as Geoffroy et al. (2013b). The step response data consist of values of GMST T_1 and TOA net downward radiative flux N averaged over each of 150 years in the experiment. While it is possible, in practice, to fit four-box models using the methodology described in this paper, it was decided that the upper limit in this study should be $k = 3$. It was found that fitting a fourth box typically yields an estimated characteristic time scale substantially shorter than one year, which is beyond what might reasonably be extracted from annually averaged data.

For each ESM the fitted box model with lower AIC (see section 4g) was chosen as the optimal k-box emulator. The same procedure was applied to the multimodel mean (MMM) of the 16 step-response datasets. Maximum likelihood parameter estimates are reported for these optimal fits (see Table 3), with corresponding estimates (see Table 4) of characteristic time scales τ_i, surface temperature response coefficients a_i, equilibrium climate sensitivity (ECS), and transient climate response (TCR). It should be noted that, as the shortest time scale of the three-box model is on the order of one year, estimated parameters of the first box will be affected by changes in radiative forcing due to stratospheric and tropospheric "rapid adjustments" (Chung and Soden 2015). We refer to the fits chosen using AIC as optimal k-box emulators for the remainder of this paper.

TABLE 6. Instantaneous increase in surface temperature (K) under a unit-impulse forcing scenario. Results are given for two-box and three-box maximum likelihood fits. Also given is the percentage increase moving from two to three boxes. MMM refers to the k-box model fitted to the average of the datasets from all 16 ESMs.

Model	MLE two-box	MLE three-box	Percent increase
BCC-CSM1.1	0.13	0.19	42
BNU-ESM	0.17	0.25	45
CanESM2	0.15	0.22	48
CCSM4	0.14	0.23	58
CNRM-CM5.1	0.12	0.25	106
CSIRO-Mk3.6.0	0.19	0.28	50
FGOALS-s2	0.16	0.23	42
GFDL-ESM2M	0.14	0.21	46
GISS-E2-R	0.17	0.20	17
HadGEM2-ES	0.13	0.28	114
INM-CM4	0.14	0.23	72
IPSL-CM5A-LR	0.14	0.37	174
MIROC5	0.18	0.21	22
MPI-ESM-LR	0.16	0.23	46
MRI-CGCM3	0.13	0.22	74
NorESM1-M	0.12	0.19	54
MMM	0.12	0.19	62

From Table 4 it can be seen that, for all 16 fitted ESMs, three boxes are required for optimal emulation under AIC. According to AIC the multimodel mean requires three boxes. Figure 5 shows three examples of fitted step responses for optimal k-box emulators. In all fitted models the heat capacities of the boxes increase with depth while, with the exception of GISS-E2-R, the heat transfer coefficients decrease with depth (excluding the feedback parameter κ_1). The approximate signal-to-noise ratio, calculated as $F_{4\times CO_2}/\sqrt{\sigma_\eta^2 + \sigma_\xi^2}$ for the step-forcing experiment, ranges from 7.1 to 19 with a median of 9.9. This high ratio allows us to fit models with many parameters and without excessive parameter uncertainty (see Table 5). The improved fit to the step response moving from a two-box to a three-box model is often clearly visible (e.g., Fig. 6).

The number of boxes k influences the impulse responses of the fitted box models, sometimes strongly (see Fig. 7). The mathematical definition of the impulse response is given in appendix D. For all 16 ESMs from CMIP5 the impulse response of the optimal k-box emulator runs hotter in the first few years than that of the corresponding two-box model. Moving from two to

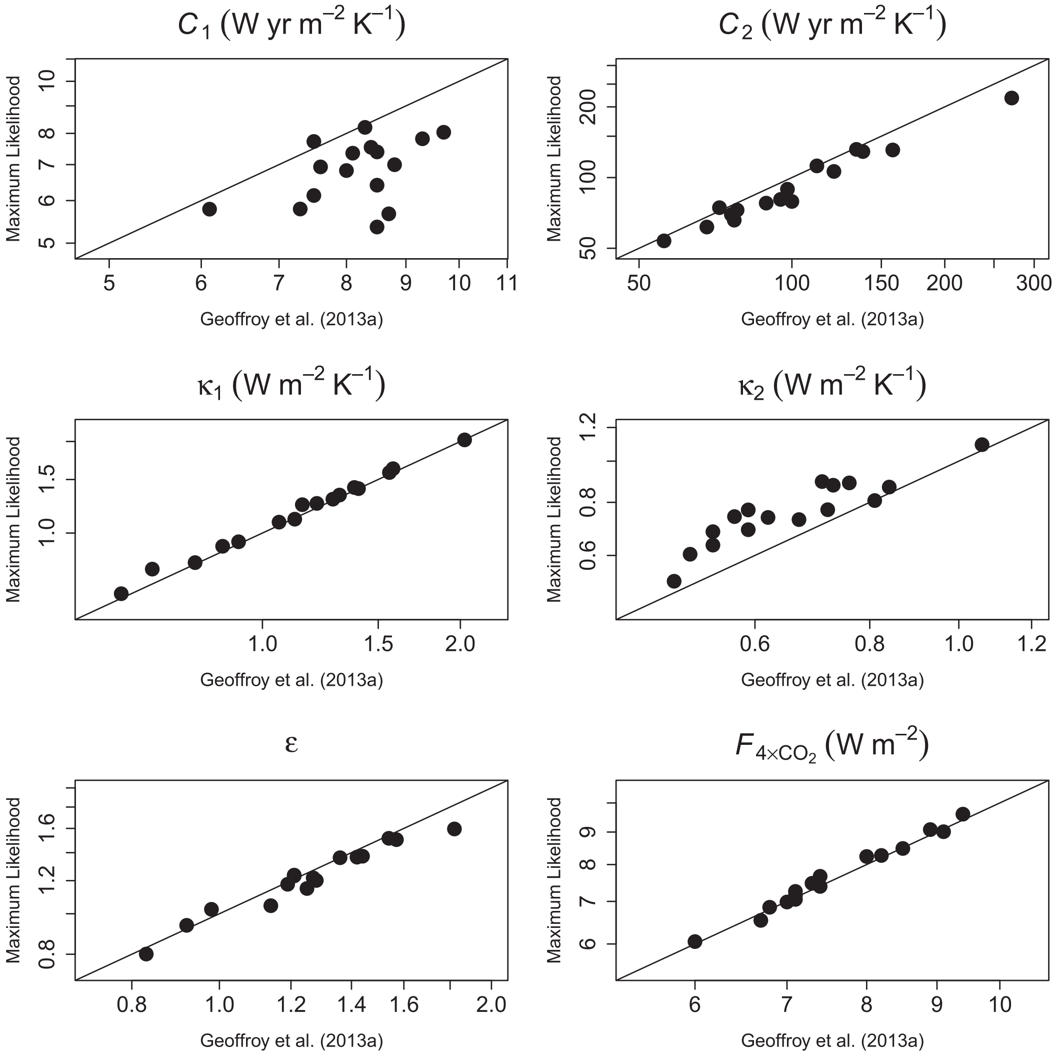

FIG. 8. Maximum likelihood parameter estimates for two-box models compared with corresponding estimates from Geoffroy et al. (2013b). Each point is one of 16 ESMs from CMIP5. The solid lines have equation $y = x$ and show where estimates are the same for both fitting methodologies. Plot axes are logarithmic.

three boxes increases the instantaneous sensitivity by between 17% and 174% (see Table 6). This suggests that when modeling the GMST response to impulse-like forcing events such as volcanic eruptions the greater flexibility of a three-box model might prove valuable.

Figure 8 compares two-box model parameter estimates obtained using maximum likelihood with those obtained by Geoffroy et al. (2013b). Maximum likelihood typically yields lower estimates of the heat capacities C_1 and C_2 but higher estimates of the heat transfer coefficient κ_2. This results in shorter estimated characteristic time scales τ_1 and τ_2 when using maximum likelihood. Estimates of the radiative parameters κ_1, ε, and $F_{4\times\mathrm{CO}_2}$ appear insensitive to the choice of fitting methodology in the case $k = 2$.

Under the proposed observation model (see section 4c), a fitted k-box model can be combined with temperature and forcing data to filter the (possibly noisy) observations and estimate the temperatures of the unobserved boxes (see Fig. 9). In this way we can see the attenuation of natural variability in temperature with increasing depth. The thermal inertia of the deep ocean boxes with their large heat capacity means that in the CO_2 quadrupling experiment the noise in these boxes is dwarfed by the signal. Filtering and hidden state estimation with k-box models is not restricted to step responses or AOGCM experiments, but rather is applicable to any combination of global temperature and radiative forcing data, including the observational record.

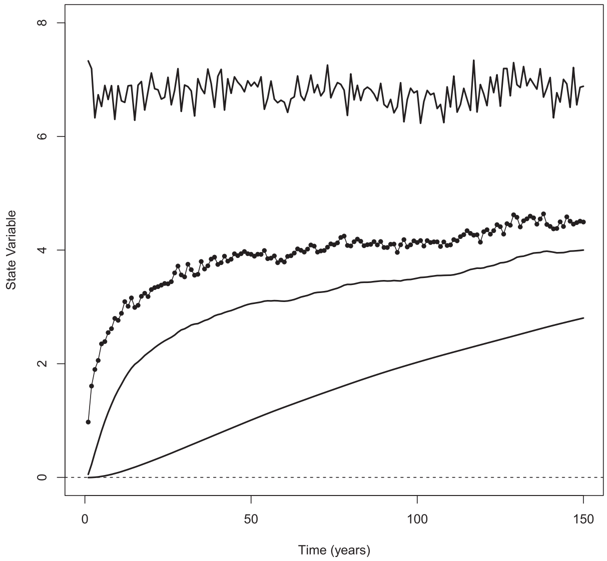

FIG. 9. Reconstructed three-box model state variables in the MRI-CGCM3 step-forcing experiment. The dots are observed surface temperatures $T_1(t)$ while the solid curves are reconstructed time series of the latent variables in their respective units. Latent variables are, from top to bottom, radiative forcing $F(t)$ (in W m^{-2}) and deep ocean box temperatures $T_2(t)$ and $T_3(t)$ (in K).

8. Summary

The k-box energy balance model in this paper offers a simple but flexible representation of the response of global mean surface temperature to radiative forcing, both deterministic and stochastic, over a range of time scales. Parameter estimation for this class of model is nontrivial: since we can typically observe the temperature of only the first box, we have a situation where for $k \geq 2$ at least half of the model state variables are latent. We have shown how, by finding a state-space representation of the linear dynamic system and evaluating the likelihood recursively via the Kalman filter, maximum likelihood estimates of all model parameters may be obtained.

The k-box model is a linear time-invariant system and thus characterized by its response to a step forcing, a forcing scenario that has been simulated in AOGCM experiments. A simulation study has been carried out to investigate the feasibility, reliability, and performance of the proposed method when applied to step-response data. The proposed method has been found to reliably estimate the k-box model parameters.

An important advantage of maximum likelihood estimation is that optimal model complexity can be chosen using information criteria. To demonstrate this, two-box and three-box models were fitted to each of a set of 16 Earth system models from CMIP5 with the optimal number of boxes chosen by Akaike's information criterion. It was found that for all 16 AOGCMs three boxes are required for optimal k-box emulation. Results obtained via maximum likelihood estimation were compared with equivalent results from the method of Geoffroy et al. (2013b). It was found that estimates of some model parameters differ systematically depending on the choice of fitting method. The number of boxes, k, was found to influence the impulse responses of the fitted models, sometimes strongly. These results suggest that, under impulse-like forcing scenarios, AOGCM responses might be better emulated using three-box models.

Finally, an example has been presented showing how a fitted k-box model can be combined with temperature and forcing data to reconstruct the temperatures of unobserved boxes corresponding to the deep ocean. Noise filtering and hidden state estimation using k-box AOGCM emulators are possible wherever we have a combination of global temperature and radiative forcing data, including the observational record.

Acknowledgments. We thank Chris E. Forest and two anonymous reviewers for their thoughtful and constructive comments, which helped improve the manuscript. We thank Olivier Geoffroy for kindly providing the climate model datasets used in this study. We thank Stefan Siegert for helpful discussions regarding the Kalman filter. We acknowledge the World Climate Research Program'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 A

Proof that Eigenvalues of Matrix A are Real and Non-Positive when ε = 1

Consider the $k \times k$ matrices $\mathbf{A}^{\dagger}$ and $\mathbf{D}$ where

$$A^{\dagger}_{i,j} = \begin{cases} -(\kappa_i + \kappa_{i+1})/C_i, & \text{if} \quad i = j, \\ \kappa_j \Big/ \sqrt{C_i C_j}, & \text{if} \quad j = i+1, \\ \kappa_i \Big/ \sqrt{C_j C_i}, & \text{if} \quad j = i-1, \\ 0, & \text{otherwise}, \end{cases} \tag{A1}$$

and $\mathbf{D}$ is a diagonal matrix with leading diagonal $(1, \sqrt{C_2/C_1}, \sqrt{C_2 C_3/C_1 C_2}, \ldots)'$. If $\varepsilon = 1$ the matrix $\mathbf{A}$ is similar to $\mathbf{A}^{\dagger}$ by the similarity transform $\mathbf{A} = \mathbf{D}^{-1}\mathbf{A}^{\dagger}\mathbf{D}$.

Since matrix $\mathbf{A}^{\dagger}$ is real and symmetric it must have all real eigenvalues. By similarity, all eigenvalues of $\mathbf{A}$ are therefore real. Applying the Geršgorin circle theorem (Geršgorin 1931) to $\mathbf{A}$ it follows that the eigenvalues of $\mathbf{A}$ are also non-positive.

APPENDIX B

Discretization of the Full *k*-Box Model

Equation (12) can be rearranged as follows:

$$\dot{\mathbf{x}}(t) - \mathbf{A}^{+}\mathbf{x}(t) = \mathbf{b}u(t) + \mathbf{w}(t). \tag{B1}$$

Multiplying by the integrating factor $e^{-\mathbf{A}^{+}t}$ we have

$$\frac{d}{dt}[e^{-\mathbf{A}^{+}t}\mathbf{x}(t)] = e^{-\mathbf{A}^{+}t}\mathbf{b}u(t) + e^{-\mathbf{A}^{+}t}\mathbf{w}(t). \tag{B2}$$

Integrating with respect to time,

$$e^{-\mathbf{A}^{+}t}\mathbf{x}(t) = \int_{s=-\infty}^{t} e^{-\mathbf{A}^{+}s}\mathbf{b}u(s) + e^{-\mathbf{A}^{+}s}\mathbf{w}(s)\,ds, \tag{B3}$$

so that, multiplying by $e^{\mathbf{A}^{+}t}$, we obtain

$$\mathbf{x}(t) = \int_{s=-\infty}^{t} e^{\mathbf{A}^{+}(t-s)}\mathbf{b}u(s) + e^{\mathbf{A}^{+}(t-s)}\mathbf{w}(s)\,ds. \tag{B4}$$

As a linear function of Gaussian random variables $\mathbf{x}(t)$ is itself Gaussian and hence fully characterized by its mean and covariance. Since $E[\mathbf{w}(t)] = \mathbf{0}$ for all t,

$$E[\mathbf{x}(t)|\mathbf{x}(t-1)] = e^{\mathbf{A}^{+}}\mathbf{x}(t-1) + \int_{s=t-1}^{t} e^{\mathbf{A}^{+}(t-s)}\mathbf{b}u(s)\,ds, \tag{B5}$$

where, assuming $u(s) = u(t-1)$ for $s \in [t-1, t)$,

$$\int_{s=t-1}^{t} e^{\mathbf{A}^{+}(t-s)}\mathbf{b}u(s)\,ds = \int_{s=t-1}^{t} e^{\mathbf{A}^{+}(t-s)}\mathbf{b}u(t-1)\,ds \tag{B6}$$

$$= [-(\mathbf{A}^{+})^{-1}e^{\mathbf{A}^{+}(t-s)}\mathbf{b}u(t-1)]_{s=t-1}^{t} \tag{B7}$$

$$= -(\mathbf{A}^{+})^{-1}(\mathbf{I} - e^{\mathbf{A}^{+}})\mathbf{b}u(t-1) \tag{B8}$$

$$= (\mathbf{A}^{+})^{-1}(e^{\mathbf{A}^{+}} - \mathbf{I})\mathbf{b}u(t-1). \tag{B9}$$

For the covariance we have

$$\mathrm{cov}[\mathbf{x}(t)|\mathbf{x}(t-1)] = \mathrm{cov}\left[\int_{s=t-1}^{t} e^{\mathbf{A}^{+}(t-s)}\mathbf{w}(s)\,ds\right] \tag{B10}$$

$$= \mathrm{cov}\left[\int_{s=0}^{1} e^{\mathbf{A}^{+}(1-s)}\mathbf{w}(s)\,ds\right] \tag{B11}$$

$$= \mathrm{cov}\left[\int_{s=0}^{1} e^{\mathbf{A}^{+}s}\mathbf{w}(s)\,ds\right], \tag{B12}$$

where, since $\mathbf{w}(t)$ is white noise and hence uncorrelated in time,

$$\mathrm{cov}\left[\int_{s=0}^{1} e^{\mathbf{A}^{+}s}\mathbf{w}(s)\,ds\right] = \int_{s=0}^{1} \mathrm{cov}[e^{\mathbf{A}^{+}s}\mathbf{w}(s)]\,ds \tag{B13}$$

$$= \int_{s=0}^{1} e^{\mathbf{A}^{+}s}\mathrm{cov}[\mathbf{w}(s)]e^{\mathbf{A}^{+'}s}\,ds \tag{B14}$$

$$= \int_{s=0}^{1} e^{\mathbf{A}^{+}s}\mathbf{Q}e^{\mathbf{A}^{+'}s}\,ds. \tag{B15}$$

For additional information on this type of discretization scheme see section 4.3 of Ljung (1987).

APPENDIX C

Marginal Covariance of the Stochastic Response

The *k*-box model is a linear dynamic system. Therefore the response to a linear combination of inputs is equal to the sum of the responses to individual inputs. In this way we can separate the model responses to deterministic and stochastic forcing components. The stochastic component of the response is driven by a purely stochastic input and may be written

$$\mathbf{x}(t) = \mathbf{A}_d\mathbf{x}(t-1) + \mathbf{w}_d(t), \tag{C1}$$

which is a vector autoregressive process of order one [VAR(1)]. The matrix-valued auto-cross-covariance function $\mathbf{\Gamma}(h)$ is defined as

$$\mathbf{\Gamma}(h) = \mathbf{\Gamma}(-h) = E[\mathbf{x}(t)\mathbf{x}(t+h)'], \tag{C2}$$

where the lag h is an integer. We seek the marginal auto-cross-covariance matrix $\mathbf{\Gamma}(0)$, which is the a priori covariance of $\mathbf{x}_0$ in the Kalman filter. Define the backshift operator B such that

$$B\mathbf{x}(t) = \mathbf{x}(t-1). \tag{C3}$$

We can write

$$(\mathbf{I} - \mathbf{A}_d B)\mathbf{x}(t) = \mathbf{w}_d(t) \tag{C4}$$

and

$$\mathbf{x}(t) = (\mathbf{I} - \mathbf{A}_d B)^{-1}\mathbf{w}_d(t) \tag{C5}$$

$$= (\mathbf{I} + \mathbf{A}_d B + \mathbf{A}_d^2 B^2 + \cdots)\mathbf{w}_d(t). \tag{C6}$$

The geometric series converges when the VAR(1) process is stationary (i.e., all eigenvalues of $\mathbf{A}_d$ lie within the unit circle in the complex plane).

$$\boldsymbol{\Gamma}(0) = E[\mathbf{x}(t)\mathbf{x}(t)'] \tag{C7}$$

$$= E[(\mathbf{I} + \mathbf{A}_d B + \cdots)\mathbf{w}_d(t)\mathbf{w}_d(t)'(\mathbf{I} + \mathbf{A}_d' B + \cdots)]. \tag{C8}$$

Since

$$E[B^i\mathbf{w}_d(t)\mathbf{w}_d(t)'B^j] = \mathbf{Q}_d\delta_{ij}, \tag{C9}$$

where δ_{ij} denotes the Kronecker delta, we have

$$\boldsymbol{\Gamma}(0) = \mathbf{Q}_d + \mathbf{A}_d\mathbf{Q}_d\mathbf{A}_d' + \mathbf{A}_d^2\mathbf{Q}_d\mathbf{A}_d^{2\prime} + \cdots. \tag{C10}$$

The infinite series can be computed as follows using the vec operator and the Kronecker product (Luetkepohl 1991).

$$\mathrm{vec}[\boldsymbol{\Gamma}(0)] = \mathrm{vec}(\mathbf{Q}_d) + \mathrm{vec}(\mathbf{A}_d\mathbf{Q}_d\mathbf{A}_d') + \mathrm{vec}(\mathbf{A}_d^2\mathbf{Q}_d\mathbf{A}_d^{2\prime}) + \cdots \tag{C11}$$

$$= \mathrm{vec}(\mathbf{Q}_d) + (\mathbf{A}_d \otimes \mathbf{A}_d)\mathrm{vec}(\mathbf{Q}_d) + (\mathbf{A}_d^2 \otimes \mathbf{A}_d^2)\mathrm{vec}(\mathbf{Q}_d) + \cdots \tag{C12}$$

$$= (\mathbf{I} - \mathbf{A}_d \otimes \mathbf{A}_d)^{-1}\mathrm{vec}(\mathbf{Q}_d). \tag{C13}$$

Note $(\mathbf{I} - \mathbf{A}_d \otimes \mathbf{A}_d)$ is invertible because eigenvalues of $\mathbf{A}_d \otimes \mathbf{A}_d$ are products of eigenvalues of $\mathbf{A}_d$ and hence have modulus < 1 when the VAR(1) process is stationary.

APPENDIX D

Analytical Responses under Idealized Forcing Scenarios

a. Unit step forcing

The k-box model response under a unit step-forcing scenario

$$F_{\mathrm{step}}(t) = \begin{cases} 1, & \text{if} \quad t \ge 0, \\ 0, & \text{otherwise}; \end{cases} \tag{D1}$$

can be written

$$\mathbf{x}_{\mathrm{step}}(t) = \frac{1}{\kappa_1}(\mathbf{1} - e^{\mathbf{A}t}\mathbf{1}), \tag{D2}$$

where $\mathbf{1}$ denotes the vector of ones $(1, \ldots, 1)'$. The unit-forced equilibrium temperature is $1/\kappa_1$, which is obtained by setting Eq. (7) equal to zero and solving for $T_1 = \cdots = T_k$. As the k-box model is linear, transient relaxation to the new equilibrium temperature is exponential.

b. Unit impulse forcing

Differentiating $\mathbf{x}_{\mathrm{step}}(x)$ with respect to time we obtain the response to a unit-impulse forcing

$$F_{\mathrm{imp}}(t) = \delta(t), \tag{D3}$$

where $\delta(t)$ denotes the Dirac delta function:

$$\mathbf{x}_{\mathrm{imp}}(t) = -\frac{1}{\kappa_1}\mathbf{A}e^{\mathbf{A}t}\mathbf{1}. \tag{D4}$$

This follows from the fact that an impulse is the time derivative of a step forcing.

c. Transient climate response

Integrating $\mathbf{x}_{\mathrm{step}}(x)$ with respect to time and scaling appropriately we obtain the transient climate response (TCR)

$$\mathbf{x}_{\mathrm{TCR}}(t) = \frac{\log 1.01}{\log 4}\int_{s=0}^{t}\frac{F_{4\times\mathrm{CO}_2}}{\kappa_1}(\mathbf{1} - e^{\mathbf{A}s}\mathbf{1})\,ds \tag{D5}$$

$$= \frac{\log 1.01}{\log 4}\frac{F_{4\times\mathrm{CO}_2}}{\kappa_1}[\mathbf{t} - \mathbf{A}^{-1}(e^{\mathbf{A}t} - \mathbf{I})\mathbf{1}], \tag{D6}$$

which is the response to atmospheric CO_2 concentration increasing at a rate of 1% yr^{-1} starting at time $t = 0$. This follows from the fact that an exponentially increasing CO_2 input is equivalent to a sequence of superimposed 1.01 × CO_2 step-forcing inputs. By linearity, the k-box model response to this superposition of forcing inputs is a superposition of the corresponding temperature outputs.

REFERENCES

Akaike, H., 1974: A new look at the statistical model identification. *Selected Papers of Hirotugu Akaike*, Springer, 215–222.

Budyko, M. I., 1969: The effect of solar radiation variations on the climate of the earth. *Tellus*, **21**, 611–619, https://doi.org/10.3402/tellusa.v21i5.10109.

Caldeira, K., and N. Myhrvold, 2013: Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environ. Res. Lett.*, **8**, 034039, https://doi.org/10.1088/1748-9326/8/3/034039.

Chung, E.-S., and B. J. Soden, 2015: An assessment of methods for computing radiative forcing in climate models. *Environ. Res. Lett.*, **10**, 074004, https://doi.org/10.1088/1748-9326/10/7/074004.

Crowley, T. J., 2000: Causes of climate change over the past 1000 years. *Science*, **289**, 270–277, https://doi.org/10.1126/science.289.5477.270.

Fredriksen, H. B., and M. Rypdal, 2017: Long-range persistence in global surface temperatures explained by linear multibox energy balance models. *J. Climate*, **30**, 7157–7168, https://doi.org/10.1175/JCLI-D-16-0877.1.

Geoffroy, O., D. Saint-Martin, D. J. L. Olivie, A. Voldoire, G. Bellon, and S. Tyteca, 2013a: Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. *J. Climate*, **26**, 1841–1857, https://doi.org/10.1175/JCLI-D-12-00195.1.

——, ——, G. Bellon, A. Voldoire, D. J. L. Olivié, and S. Tytéca, 2013b: Transient climate response in a two-layer energy-balance model. Part II: Representation of the efficacy of deep-ocean heat uptake and validation for CMIP5 AOGCMs. *J. Climate*, **26**, 1859–1876, https://doi.org/10.1175/JCLI-D-12-00196.1.

Geršgorin, S., 1931: Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk SSSR Ser. Mat.*, **1**, 749–755.

Gilbert, P., and R. Varadhan, 2016: numDeriv: Accurate numerical derivatives, version 2016.8-1. R package, https://CRAN.R-project.org/package=numDeriv.

Good, P., J. M. Gregory, and J. A. Lowe, 2011: A step-response simple climate model to reconstruct and interpret AOGCM projections. *Geophys. Res. Lett.*, **38**, L01703, https://doi.org/10.1029/2010GL045208.

Goulet, V., C. Dutang, M. Maechler, D. Firth, M. Shapira, and M. Stadelmann, 2019: expm: Matrix Exponential, Log, etc, version 0.999-4. R package, https://CRAN.R-project.org/package=expm.

Gregory, J. M., 2000: Vertical heat transports in the ocean and their effect on time-dependent climate change. *Climate Dyn.*, **16**, 501–515, https://doi.org/10.1007/s003820000059.

——, and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.*, **31**, L03205, https://doi.org/10.1029/2003GL018747.

Hansen, J., M. Sato, P. Kharecha, and K. Schuckmann, 2011: Earth's energy imbalance and implications. *Atmos. Chem. Phys.*, **11**, 13 421–13 449, https://doi.org/10.5194/acp-11-13421-2011.

Hasselmann, K., 1976: Stochastic climate models, Part I. Theory. *Tellus*, **28**, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316.

Held, I. M., M. Winton, K. Takahashi, T. Delworth, F. Zeng, and G. K. Vallis, 2010: Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing. *J. Climate*, **23**, 2418–2427, https://doi.org/10.1175/2009JCLI3466.1.

Johnson, S. G., 2014: The NLopt nonlinear-optimization package. http://github.com/stevengj/nlopt.

Jonko, A., N. M. Urban, and B. Nadiga, 2018: Towards Bayesian hierarchical inference of equilibrium climate sensitivity from a combination of CMIP5 climate models and observational data. *Climatic Change*, **149**, 247–260, https://doi.org/10.1007/s10584-018-2232-0.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45, https://doi.org/10.1115/1.3662552.

Kaufmann, B., 2003: Fitting a sum of exponentials to numerical data. Accessed 21 March 2019, https://arxiv.org/abs/physics/0305019.

Ljung, L., 1987: *System Identification: Theory for the User.* Prentice-Hall, 519 pp.

Lucarini, V., F. Ragone, and F. Lunkeit, 2017: Predicting climate change using response theory: Global averages and spatial patterns. *J. Stat. Phys.*, **166**, 1036–1064, https://doi.org/10.1007/s10955-016-1506-z.

Luethi, D., P. Erb, and S. Otziger, 2018: FKF: Fast Kalman filter, version 0.1.5. R package, https://CRAN.R-project.org/package=FKF.

Luetkepohl, H., 1991: *Introduction to Multiple Time Series Analysis.* Springer-Verlag, 556 pp., https://doi.org/10.1007/978-3-662-02691-5.

Moberg, A., D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlén, 2005: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature*, **433**, 613–617, https://doi.org/10.1038/nature03265.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Powell, M. J., 2009: The BOBYQA algorithm for bound constrained optimization without derivatives. Cambridge NA Rep. NA2009/06, University of Cambridge, 39 pp., http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf.

Proistosescu, C., and P. J. Huybers, 2017: Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sci. Adv.*, **3**, e1602821, https://doi.org/10.1126/sciadv.1602821.

R Core Team, 2019: R: A language and environment for statistical computing, R Foundation for Statistical Computing, https://www.R-project.org/.

Reid, I., 2001: Estimation II [lecture notes]. 44 pp., http://www.robots.ox.ac.uk/~ian/Teaching/Estimation/LectureNotes2.pdf.

Rypdal, M., and K. Rypdal, 2014: Long-memory effects in linear response models of Earth's temperature and implications for future global warming. *J. Climate*, **27**, 5240–5258, https://doi.org/10.1175/JCLI-D-13-00296.1.

——, H.-B. Fredriksen, E. Myrvoll-Nilsen, K. Rypdal, and S. H. Sørbye, 2018: Emergent scale invariance and climate sensitivity. *Climate*, **6**, 93, https://doi.org/10.3390/CLI6040093.

Sellers, W. D., 1969: A global climatic model based on the energy balance of the Earth–atmosphere system. *J. Appl. Meteor.*, **8**, 392–400, https://doi.org/10.1175/1520-0450(1969)008<0392:AGCMBO>2.0.CO;2.

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

Tsutsui, J., 2016: Quantification of temperature response to CO2 forcing in atmosphere–ocean general circulation models. *Climatic Change*, **140**, 287–305, https://doi.org/10.1007/s10584-016-1832-9.

Tusell, F., 2011: Kalman filtering in R. *J. Stat. Software*, **39** (2), 1–27, https://doi.org/10.18637/jss.v039.i02.

Van Loan, C., 1978: Computing integrals involving the matrix exponential. *IEEE Trans. Autom. Control*, **23**, 395–404, https://doi.org/10.1109/TAC.1978.1101743.

Ypma, J., 2020: Introduction to nloptr: An R interface to NLopt. 14 pp., https://cran.r-project.org/web/packages/nloptr/vignettes/nloptr.pdf.