AGU PUBLICATIONS

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1002/2017MS000947

A Caveat Note on Tuning in the Development of Coupled Climate Models

Dietmar Dommenget[1] and **Michael Rezny**[1]

[1]School of Earth, Atmosphere and Environment, Monash University, Clayton, VIC, Australia

Key Points:
- Climate model tuning concepts are tested and compared versus flux corrections
- Climate model tuning is unlikely to improve the model significantly, but does introduce artificial errors
- Flux corrections are not perfect, but perform better than model tuning

Correspondence to:
D. Dommenget,
dietmar.dommenget@monash.edu

Citation:
Dommenget, D., & Rezny, M. (2018). A caveat note on tuning in the development of coupled climate models. *Journal of Advances in Modeling Earth Systems*, *10*, 78–97. https://doi.org/10.1002/2017MS000947

Received 16 FEB 2017
Accepted 27 OCT 2017
Accepted article online 24 NOV 2017
Published online 13 JAN 2018

Abstract State-of-the-art coupled general circulation models (CGCMs) have substantial errors in their simulations of climate. In particular, these errors can lead to large uncertainties in the simulated climate response (both globally and regionally) to a doubling of CO_2. Currently, tuning of the parameterization schemes in CGCMs is a significant part of the developed. It is not clear whether such tuning actually improves models. The tuning process is (in general) neither documented, nor reproducible. Alternative methods such as flux correcting are not used nor is it clear if such methods would perform better. In this study, ensembles of perturbed physics experiments are performed with the Globally Resolved Energy Balance (GREB) model to test the impact of tuning. The work illustrates that tuning has, in average, limited skill given the complexity of the system, the limited computing resources, and the limited observations to optimize parameters. While tuning may improve model performance (such as reproducing observed past climate), it will not get closer to the "true" physics nor will it significantly improve future climate change projections. Tuning will introduce artificial compensating error interactions between submodels that will hamper further model development. In turn, flux corrections do perform well in most, but not all aspects. A main advantage of flux correction is that it is much cheaper, simpler, more transparent, and it does not introduce artificial error interactions between submodels. These GREB model experiments should be considered as a pilot study to motivate further CGCM studies that address the issues of model tuning.

Plain Language Summary State-of-the-art climate models are highly complex models of the earth's climate. To achieve optimal simulations of the present climate these climate models are tuned. The tuning process as such is not well understood and it is unclear how good it performs. Alternative approaches, such as not tuning, but introducing correction terms, are less common. The study presented here illustrates the problems that tuning of a climate model introduce and shows that the alternative, of not tuning, but introducing corrections terms, is likely to be the better strategy.

1. Introduction

State-of-the-art coupled general circulation models (CGCMs) are highly complex models of the earth's climate system that are based on coupling a set of subsystem models (see sketch Figure 1). CGCMs can simulate many aspects of the climate system well, but they also have substantial biases relative to observations (Flato et al., 2013; Reichler & Kim, 2008). Furthermore, there is also significant spread in CGCM projections of future climate change that result from uncertainties in model physics (Collins et al., 2010; Knutti, 2008; Knutti et al., 2008; Knutti & Sedlacek, 2013; Murphy et al., 2004; Stainforth et al., 2005).

The CGCM biases often result from uncertainties (or errors) in the simulation of subscale (i.e., $<$100 km) processes of the climate system, which are parameterized (e.g., formation of clouds, small-scale convection, precipitation, etc.). A fundamental problem in the development of CGCMs is to minimize the effect of uncertainties associated with such parameterization schemes. One approach is to run multiple CGCM simulations with the physical parameters perturbed within their expected range of uncertainty (Collins et al., 2010; Murphy et al., 2004; Stainforth et al., 2005); however, this approach becomes unpractical for state of the art CGCMs due to limitations in computing power.

The most common method for developing CGCMs is to fine-tune ("tuning") the parameterization schemes (which may take numerous iterations) until the differences between the simulated and observed climate

© 2017. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

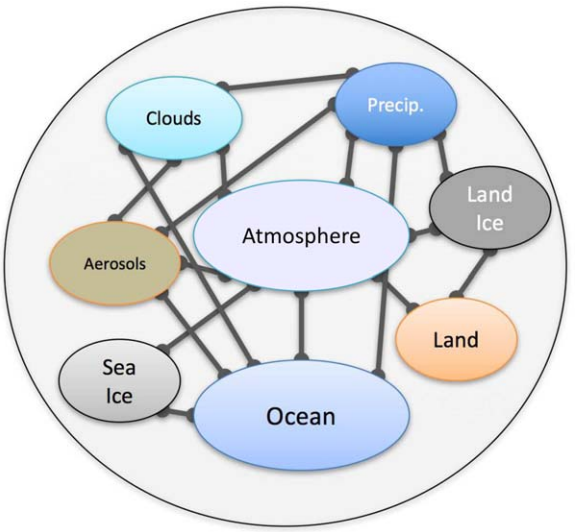

Figure 1. Sketch conceptually illustrating a climate model (the outside circle) and its submodels (named elements within the outer circle). The submodels and interactions shown are just examples and should not be considered an accurate or complete representation.

are minimized (Golaz et al., 2013; Hourdin et al., 2017; Mauritsen et al., 2012; Tang et al., 2016; Tett et al., 2013). This approach is computationally expensive and requires a substantial amount of time (i.e., months to years). It is important to note here that this tuning process is in general not documented, that is, it is unclear what information has been used for the tuning of parameters (the cost function) and it is unclear what parameters have been tuned (Mauritsen et al., 2012). Subsequently this tuning is difficult to reproduce. The complex nature of the climate system will make it very difficult to evaluate the value of tuned climate models when it is unclear what information is used to tune them (as also discussed in Ginzburg and Jensen (2004) and Lenhard and Winsberg (2010).

Another important consideration is whether tuning actually improves the model performance as such. For instance, if a model is tuned to minimize a cost function that is based on the simulated mean, observed 20th century climate, it is unclear whether the tuned model will perform equally well for other climatic states (e.g., pre-20th century or future projections). Furthermore, it is also unclear whether the tuning process actually improved the model physics. That is to say, that the tuned parameters make the model parameterization fit better to the observed physical processes. Potentially, tuning may introduce compensating errors in different submodels to artificially reduce the cost function.

An alternative or addition to model tuning is to include flux correction terms (e.g., Collins et al., 2006; Irvine et al., 2013; Manabe & Stouffer, 1988; Sausen et al., 1988; Schneider, 1996). A flux correction acts to adjust the CGCM-simulated mean state to be closer to that observed, without altering or improving model parameters. This approach has been used in a number of studies, but is not currently used in state-of-the-art CGCMs (Cubasch et al., 2001; Solomon et al. 2007). CGCM simulations without the use of flux corrections are considered to be more reliable than those using flux corrections; however, this has not been demonstrated in any CGCM experiments. Indeed, models that do not use flux correction, do employ parameter tuning to avoid climate model drifts.

This study aims to understand the effects parameter tuning in CGCM models in a realistic setup. We focus on the following two main questions:

1. Can a CGCM be improved by parameter tuning?
2. Is tuning better than flux correction?

Here we assume that "improving the CGCM" means that the simulated control climate is closer to the observed, the model physics are closer to the "true" physics and the simulation of the response to CO_2 forcing is closer to the "true" response. The flux correction approach, in general, tries to improve the CGCM without any changes to any submodel, which tuning otherwise does. An alternative way of formulating the questions in respect to alternative methods to tuning may therefore be:

1. Can we improve a CGCM *without* improving any submodel/parameterization scheme?

We approach these questions by assuming the model is already "perfect," which means that the control climate, model physics, and response to CO_2 forcing of this "perfect" reference model are the "truth." We further define a realistic optimization (tuning) scheme with limited resources and information, which mimics the real-world development of CGCMs. The reference model used here is the Globally Resolved Energy Balance (GREB) model (Dommenget & Floter, 2011; here after referred to as DF11). The GREB model has been in used in a perturbed physics study to address the role of mean state biases and flux corrections in climate change projections (Dommenget, 2016; here after referred to as D16). Parts of the GREB model have also been used for other perturbed physics optimization studies (Zhang et al., 2015, 2016). The GREB model is not a CGCM, but has a much simpler structure. It still represents aspects of the climate system realistically, but is used in this study primarily for a first-order approximation. Thus, the study presented here should be considered a pilot study that should provide motivation for further studies with CGCMs.

The paper is organized as follows: section 2 will introduce the simulation data, methods, and models used. The optimization (tuning) approach, the concept of flux corrections, and the perturbed physics ensembles are described in section 3. The main results of this study are presented in section 4, which is followed by a summary, discussion and conclusion in the final section 5.

2. Data, Model, and Methods

2.1. CGCM Simulations Data

Surface temperature (T_{surf}) data are taken from the fifth Coupled Model Intercomparison Project (CMIP5, Taylor et al., 2012) for the same 36 models used by D16 (see Table 1 in D16). The spread in T_{surf} for the control climate (1970–1999 in the Historical simulations) and the change in simulated T_{surf} (2070–2099 from the RCP 8.5 scenario) are assessed in this study.

2.2. The GREB Model

The GREB model is a three-layer (atmosphere, surface, and subsurface ocean), global climate model with a horizontal grid spacing of 3.75° longitude × 3.75° latitude. GREB simulates the thermal (long-wave) and solar (short-wave) radiation in the atmosphere, heat transport in the atmosphere by isotropic diffusion and advection with the mean winds, the hydrological cycle (evaporation, precipitation, and water vapor transport), a simple ice/snow albedo feedback, and heat uptake in the subsurface ocean. The mean winds and total cloud cover climatologies are prescribed seasonally, and flux correction is used to keep the model close to the observed climate. Thus, the GREB model is conceptually very different from the CGCM simulations in CMIP5, as atmospheric and the oceanic circulations are not simulated. An important limitation of the GREB model is that the response to external forcings or model parameter perturbations do not involve circulation or cloud feedbacks, which are relevant in CGCM simulations (Bony et al., 2006).

Furthermore, GREB does not have any internal (natural) variability, as daily weather systems are not simulated. Subsequently, the control climate or response (from now on, the use of "response" refers to the climate response to doubling CO_2 concentrations) to external forcings can be estimated from one single year.

Table 1
List of Perturbed Parameters

Name	r^2 (response) (%)	Comments
pe_1	7	CO_2 effect on emissivity; equation (5) in DF11.
pe_2	9	H_2O effect on emissivity; equation (5) in DF11.
pe_3	0.5	Residual emissivity; equation (5) in DF11.
pe_4	2	Strength of overlap band; equation (5) in DF11.
pe_5	1	Strength of CO_2 band; equation (5) in DF11.
pe_6	4	Strength of H_2O band; equation (5) in DF11.
pe_7	1	Emissivity zero off set; equation (5) in DF11.
pe_8	2	Influence of cloud cover on emissivity; equation (5) in DF11.
pe_9	5	Influence of cloud cover on emissivity; equation (5) in DF11.
pe_{10}	4	Influence of cloud cover on emissivity; equation (5) in DF11.
r_{precip}	6	Precipitation ratio; equation (11) in DF11.
r_{qviwv}	5	Regression between surface humidity and total air column water vapor; equation (8) in DF11.
c_{atmos}	5	Atmos. Coupling to surface; equation (13) in DF11.
α_{clouds}	4	Cloud albedo; equation (2) in DF11.
$\delta\alpha_{ice}$	2	Ice/snow albedo; From Figure 3a in DF11.
κ	1	Isotropic diffusion coefficient; equation (12) in DF11.
C_w	0.5	Transfer coefficient for latent cooling by evaporation; equation (7) in DF11.
$T_{sea\text{-}ice1}$, $T_{sea\text{-}ice2}$	0.3	T_{surf} range for the ice albedo feedback over oceans; From Figure 3a in DF11.
$T_{land\text{-}ice1}$, $T_{land\text{-}ice2}$	0.2	T_{surf} range for the ice albedo feedback over land; From Figure 3a in DF11.

Note. Parameter names are taken from DF11. The relative contribution of each parameter to the climate sensitivity variations, are estimated by the explained variance (r^2) based on the mean (over all regions) linear correlation of the parameter variations with the annual mean response variations in the INI-PP ensemble. Note that since the parameters interact in a non-linear way, the r^2 values do not add up to 100%. Table adapted from D16.

The near equilibrium global mean response to a doubling of CO_2 concentrations in the GREB model is +2.7°C with a seasonal warming pattern similar to those of the CMIP5 models. For a complete description of the model and the model code, see Dommenget and Floter (2011). A complete data package to run the GREB model is currently available at: http://users.monash.edu.au/~dietmard/content/GREB/GREB_model.html.

2.3. Perturbed Physics

Perturbed Physics (PP) simulations are conducted with the GREB model with the same set of perturbed parameters and associated uncertainties as in D16, see Table 1. The 21 parameters of the GREB model that are perturbed are all those that are considered to be uncertain (refer to as INI-PP simulations). The standard deviations of the parameter uncertainties have been chosen in a way that the values are within the uncertainties that can be attributed to each parameter and that none of the parameters is dominating the models climate sensitivity uncertainties. These 21 parameters include parametrizations of the thermal radiation scheme, the hydrological cycle, small-scale turbulence, and albedo of clouds, snow, and ice. Most of these do not have counterparts in CGCMs, but reflect the simplifications of the GREB model.

The GREB model's main advantage for this study is that it is computationally inexpensive (one simulated year per second on a standard personal computer), which is important as more than 2 million simulated years are considered in the study presented here. At the same time, the model produces a reasonable mean climate (e.g., T_{surf}) and response to CO_2 forcing with respect to surface temperature, T_{surf} (Dommenget & Floter, 2011). In addition, the spread (errors) in the T_{surf} responses in the PP simulations are similar to spread of the CMIP5 ensemble (see Figure 2).

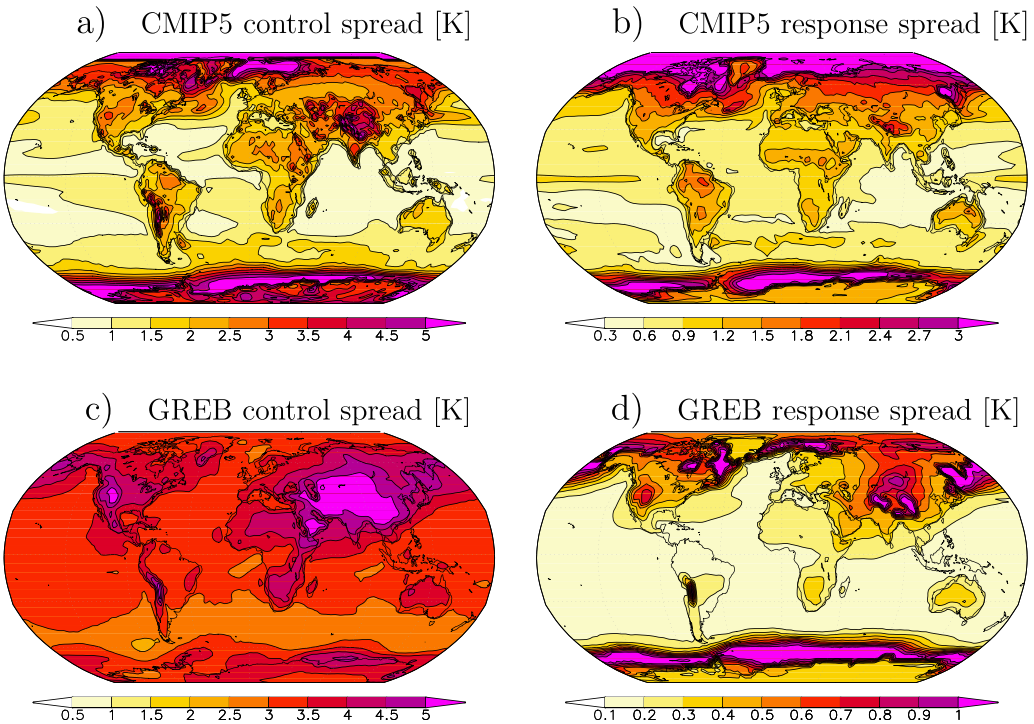

Figure 2. (a) Spread (standard deviation) of the individual CMIP5 simulations' monthly mean T_{surf} (1970–1999 climatology) relative to the CMIP5 ensemble mean T_{surf} climatology for the same period. (b) Spread of the 36 CMIP5 simulations monthly mean T_{surf} response in the RCP8.5 scenario (mean 2070–2099 minus mean 1970–2099) relative to the CMIP5 ensemble monthly mean T_{surf} response. (c) Spread of the GREB INI-PP ensemble in the control climate relative to the ensemble mean (original GREB). (d) Spread (standard deviation) of the GREB INI-PP ensemble in the control climate relative to the ensemble mean (original GREB) T_{surf} response to $2xCO_2$ forcing relative to the original GREB model response. Note, that the plots have different color scales. Figure adapted from D16.

A useful feature of the GREB model is that the mean-state climate can be controlled by flux corrections to T_{surf}, the total atmospheric water vapor content and the subsurface ocean temperatures. Since the GREB model also has no internal variability, any variations in the climate of the control or doubled CO_2 simulations must be caused by the parameter changes.

2.4. Optimization

Parameter optimization (tuning) is done using a Nelder-Mead technique (Lagarias et al., 1998; Nelder & Mead, 1965). The Nelder-Mead method is an iterative procedure that minimizes a cost function through a stepwise estimation of the gradients of the cost function and moving toward the minimum of the cost function. The cost function is defined as:

$$R = R_{climate} + R_{para}$$

$R_{climate}$ is the weighted root-mean squared error in the control climate:

$$R_{climate} = w_T R_T + w_q R_q$$

The root-mean squared errors (RMSEs) R_T and R_q are based on the area-averaged errors of T_{surf} and atmospheric total specific humidity, q_{air}, relative to the original GREB model on each grid point and calendar month. Each term is normalized by the scaling factors $w_T = 21.3\ K^{-1}$ and $w_q = 5.39 \cdot 10^{-3}\ \frac{kg}{kg}$, which are estimated from the RMSE of T_{surf} and q_{air} in the original GREB model relative to their global mean values. This scaling effectively weights both terms equally strongly in the cost function. The cost function covers almost the entire dynamical space of the GREB model, excluding only the subsurface ocean temperature and the atmospheric temperature, which are both of minor dynamical importance for this study.

R_{para} is zero if all parameters lie within ± 3 standard deviations of the parameter uncertainties σ_{pi}, and increases with a power of two if parameters are outside this interval. Essentially, R_{para} ensures that the optimization does not vary the physics parameters far beyond the $\pm 3\ \sigma_{pi}$ range.

2.5. Methods

Analyses of the mean control and response spread are based on monthly mean anomalies in the 12 month climatologies and spread is always defined by the standard deviation. In the CMIP5 ensemble, anomalies are defined relative to the ensemble mean values (12 month climatology). In the GREB model, anomalies are defined relative to the original GREB model.

In the GREB simulations any deviations from the original model can be interpreted as an error and any spread in the simulations is a spread resulting from errors. Thus, uncertainties and errors are the same in the GREB simulations. This is not the case in the CMIP ensemble in which spread or deviations from observations results from combinations of internal variability, uncertainty in observational estimates, and model errors.

The uncertainty in the local response amplitude (or pattern uncertainty) of a simulation, σ_i, is estimated (as in D16) from the normalized response pattern spread of each simulation relative to the normalized original GREB model response pattern, $\pi_i(x,y)$:

$$\pi_i(x,y) = \sqrt{\sum_{m=1}^{12} \left(\frac{T_i(m,x,y)}{\widehat{T_i}} - \frac{T_{org}(m,x,y)}{\widehat{T_{org}}} \right)^2 / 12}$$

$$\sigma_i = \sum_{x,y} w(x,y) \cdot \pi_i(x,y)$$

with the T_{surf} response for the climatological month, m, of the individual simulation, $T_i(m)$, and that of the original GREB simulation, $T_{org}(m)$, and their respective global means, $\widehat{T}_i(m)$ and $\widehat{T_{org}}\ (m)$, and the area size weight, $w(x,y)$. The normalized response pattern spread of each simulation, σ_i, gives a measure of the relative uncertainty of the local response amplitudes that is independent of the global mean response.

The response difference in the global mean of each simulation, g_i, is estimated relative to the original GREB simulation:

$$g_i = \frac{|\widehat{T_i} - \widehat{T_{org}}|}{\widehat{T_{org}}}$$

3. Concepts of Optimization (Tuning), Flux Corrections, and the Perturbed Physics Ensemble Simulations

The development of state-of-the-art CGCM simulations involves a number of steps that includes both tuning and (sometimes) flux corrections. These elements are described briefly to explain how such methods are applied to the ensemble of perturbed physics experiments here.

In the first steps of the model development, different research groups develop the subsystem models (such as those in Figure 1) from both observed processes and theoretically derived physical considerations. In this first stage, the submodel parameterizations are optimized using observationally derived thresholds. The resulting submodels (and their parameters) are then integrated into the CGCM.

In the second stage of CGCM development, simulations of historical climate are run and evaluated relative to observations. The CGCM is then optimized (relative to the observations) iteratively by developing the submodel structure further or by tuning the parameters. This in general involves running many (e.g., ~100) simulations with the CGCM until the performance of the CGCM is considered acceptable (Mauritsen et al., 2012). An important aspect of this second stage of CGCM development is that the submodel parameters are changed, despite those parameters being considered as the best fit in the initial setup. The main argument to change these parameters further is that CGCM performance is better with the tuned parameters. It is this second stage of tuning the parameters that is the focus of this study.

An alternative to changing the model parameters is to include a flux correction term. A flux correction term is a constant (or seasonally changing) flux added to the tendency equation of any CGCM variable (e.g., surface temperature or ocean salinity) to correct the mean state toward the observed. This can also be done in combination with tuning model parameters.

In the following experiments and analysis, a "perfect" model setup is assumed to investigate the effects of tuning in the second stage of CGCM developments. The mean climate, the response to doubling of the CO_2 concentration, and the physics of the original GREB model are assumed to be the truth. A set of model parameters are then perturbed and a series of model simulations are undertaken (see Table 2). It should be noted here that CGCM uncertainties do not solely result from submodel parameter uncertainties, but also result from structural errors (e.g., choice of numerical discretization, choice of dynamical variables, equation types, etc.). The focus here on parameter uncertainties is solely to mimic a common approach in CGCM development and is set by what is possible within the framework of this study.

For each ensemble member in following optimization ensembles, we conduct a 15 year control and 2xCO_2 (560 ppm) concentration simulation with the initial conditions for both taken from the original GREB control

Table 2
Simulations for Each Member of the Optimizations Scheme and the Flux Corrections

Description	Length (years)	# iterations
Initial perturbed physics		
The PP GREB control simulation	15	1
The PP GREB 2xCO_2 simulation	15	1
Optimization scheme		
Iterations for minimizing cost function	Max. 15 Avg. ~10	Max. 300 Avg. ~300
GREB control simulation with optimized perturbed parameters	15	1
GREB 2xCO_2 simulation with optimized perturbed parameters	15	1
Flux corrections		
The PP GREB control simulation with flux correction	3	1
The PP GREB 2xCO_2 simulation with flux correction	15	1

climate. For both runs, the last year is used to estimate the climate states of the control and $2xCO_2$ simulations. The climate response to $2xCO_2$ is defined as the difference between the last year of the $2xCO_2$ experiment and the last year of the control experiment. We further do an optimization of perturbed parameters with a maximum of 300 iterations with each simulation being 15 years long for the maximum (see section 2 for details of the optimization). Another pair of control and $2xCO_2$ simulations are run for the resulting optimized parameters.

The set of simulations is completed by applying a flux correction to the perturbed physics GREB model to constrain the climatological control T_{surf} and atmospheric water vapor to be the same as in the original GREB model simulation. The flux corrections are computed for each calendar month and grid point. Another pair of control and response runs is then undertaken. By construction, the resulting control climates in these runs are identical (within numerical precision) to the original GREB model simulation. See Table 2 for a complete overview.

The optimization scheme that we have defined above does include a number of deliberate limitations that are analogous to those performed in realistic CGCM developments:

1. Estimating the control climate in the GREB model beyond a certain precision is unnecessary as the mean state of the observed climate is not perfectly known. Therefore, the optimization runs abort within a few years (approximately 10 years) if the differences in the control climate between the previous and current year are below a threshold value.
2. We restrict the optimizations to 300 iterations, as limitations in time and computing power would not allow such a large number of iterations in CGCM model development.
3. The optimization cost function only considers the mean state in the control. The cost function does not account for the correct representation of the physics or the climate response to a doubling of CO_2 concentration.

It should be noted that it is not assumed that other modeling groups use the Nelder-Mead method. The Nelder-Mead technique is simply used here as an effective and reproducible method for optimizing parameters in a similar way to other CGCM development groups.

An example of the optimization process (in which 10 parameters were perturbed) is presented in Figures 3 and 4 to illustrate the procedure. The initial set of parameter perturbations are spread around the original values (Figure 3a). The final, optimized parameters (after 300 iterations) are different relative to their initial values. Furthermore, the optimized values are also different from the original GREB model parameters (Figure 3a). The optimization method leads to an overall decrease in the cost function with successive iterations (Figure 3b). Nonetheless, the estimation of gradients in high-dimensional parameter space, during the optimization process, may lead to the cost function increasing and decreasing at successive iterations. (the "zigzag" track in Figures 3b–3d). Overall, the final optimized parameter values are still the ones with the smallest cost function values.

A smoothed track, illustrating the optimization iteration process within the parameter space, is shown in Figure 4. The track is projected onto a 2-dimenional plane, which includes the initial perturbed, the final optimized, and the original GREB parameter values. Interestingly, the final optimized point is actually further away (distance in normalized parameter space) from the original GREB parameter values than the initial parameter perturbations (Figures 3c and 4).

4. Optimized Perturbed Physics

In the following, we will discuss the results of a number of different optimization ensembles to address different characteristics that such optimizations have. First, we will explore how the increase in complexity of the model uncertainties limits the optimization. We will then focus on a number of realistic optimization problems and evaluate how these would improve or degrade the model performance on the control climate, the response to doubling of CO_2, and the model physics. The optimization approaches will further be compared against a flux corrections approach.

4.1. Limitations of Optimizations

We start our discussion of idealized optimizations by evaluating how the optimization performance for different numbers of perturbed (and subsequently optimized) parameters. Eight different experiments (the

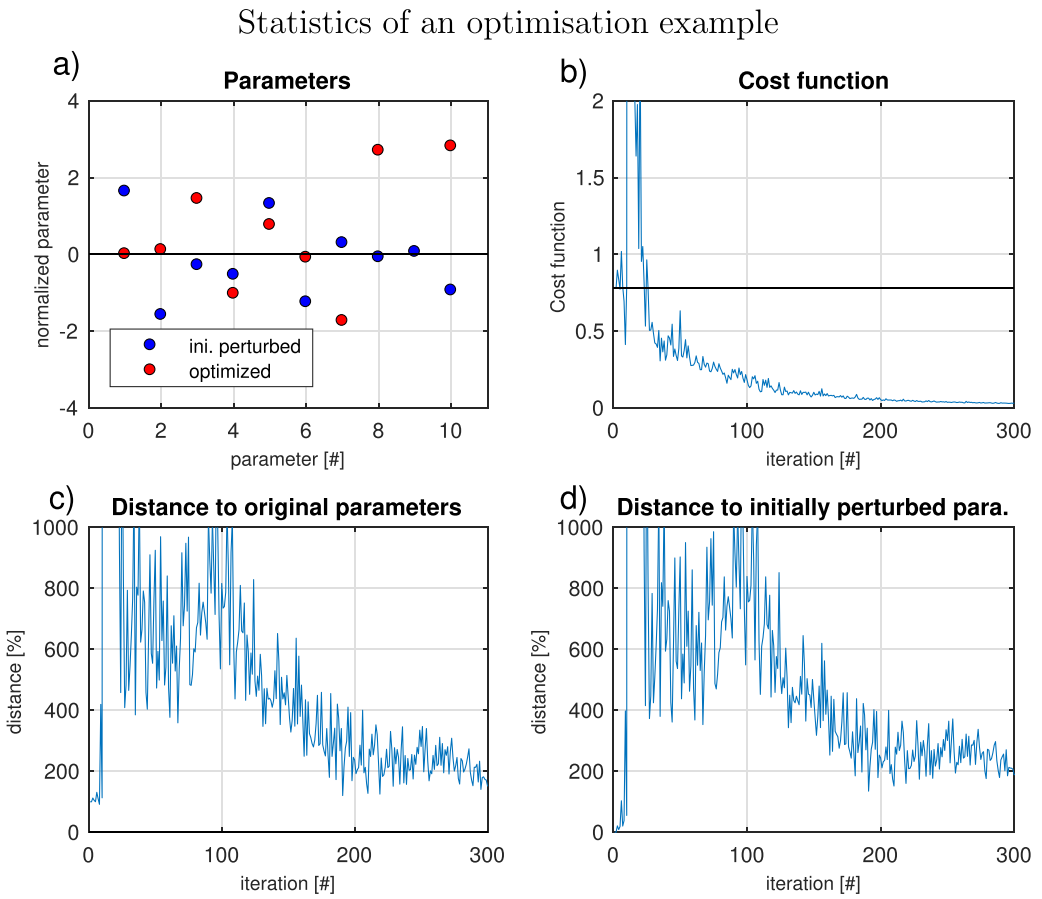

Figure 3. Some statistics of an example member from the OPT-10P ensemble: (a) Normalized perturbed parameters before the optimization (red) and after the optimization (blue). (b) Cost function values for each iteration step. The solid black line marks the initial value. (c) Distance to the original parameter values in percentage of initial distance for each iteration step. (d) Distance to the initial perturbed parameters (INI-PP) in percentage for each iteration step. The percentage is in respect to the distance of the parameters in the INI-PP to the original parameters.

first eight rows, Table 3) are generated where between 2 and 21 (i.e., all) parameters are perturbed and then optimized. Each experiment consists of an ensemble of at least 20 simulations.

The distribution across simulations of the optimized cost function values for each experiment are plotted in Figure 5. When two or three parameters are randomly perturbed and optimized, the optimization process always reduces the cost function to zero (Figure 5a) giving the original "true" model physics (distance is zero; see Figure 5b). Thus, the optimization always succeeds. When the number of perturbed parameters is increased (i.e., more complex), the optimization fails to find the original model physics and the cost function does not reduce to zero. Importantly, as the number of perturbed and optimized parameters increases, the difference between the optimized and original model physics is almost doubled (Figure 5b). Thus, by reducing the cost function, the resulting model setup is further from the "true" physics, which is undesirable and illustrated in Figures 3 and 4.

For example, if ten parameters are perturbed and optimized then the optimization space is very large (10-dimensional space). A 2-dimensional projection of this 10-dimensional space is plotted in Figure 4. The optimization process is used to minimize the cost function, which should correspond with the original GREB model parameters (the origin in Figure 4). Nonetheless, the final and original GREB model parameters are both within a region where the gradient in the cost function is very small (Figure 4). In order for the optimization to get closer to the original model parameters, the optimization has to determine these very small gradients with high precision. Such accuracy also requires precise knowledge of both the observed and simulated climate states. Neither of these above requirements can be achieved in any realistic CGCM

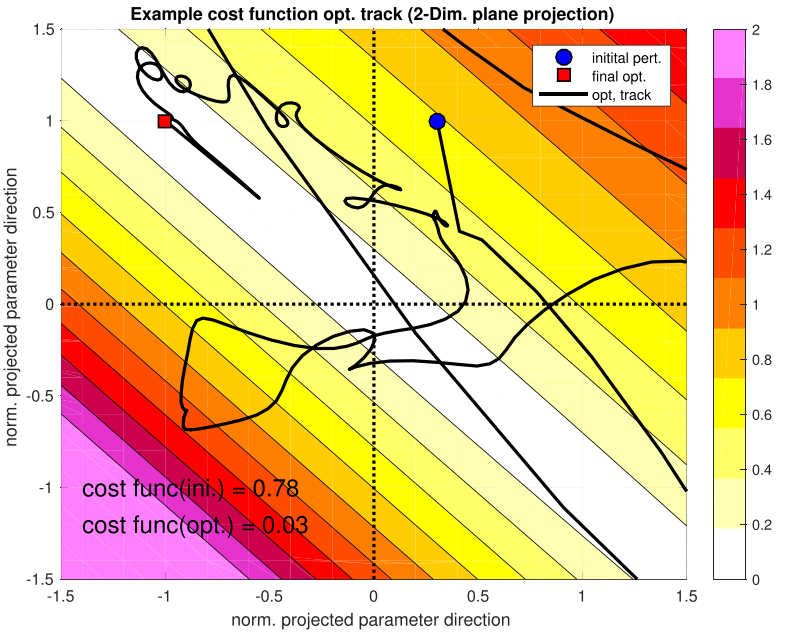

Figure 4. Track of the cost function from the initial perturbed parameters (blue circle) to the final optimized parameters (red square) projected onto a 2-dimensional plane (manifold) of the 10-dimensional parameter space for the same example member of OPT-10P as in Figure 3. The manifold was chosen by the plane that does include the initial, final optimized, and original parameter points. The track was smoothed for a better presentation. The contour shadings mark the cost function value. The axes are normalized directions in the 10-dimensional parameter space. It is normalized to have the optimized value at (–1,1), the *y* coordinate of the initial point at 1 and the original parameters at the origin.

Table 3
List of Perturbed Physics Experiments Discussed in This Study

Optimization experiments		
Exp. name	Members	Comments
OPT-2P	20	Global optimization with 2 perturbed and optimized parameters
OPT-3P	20	As OPT2, but with 3 parameters
OPT-4P	20	As OPT2, but with 4 parameters
OPT-5P	20	As OPT2, but with 5 parameters
OPT-7P	20	As OPT2, but with 7 parameters
OPT-10P	20	As OPT2, but with 10 parameters
OPT-15P	20	As OPT2, but with 15 parameters
OPT-21P	50	As OPT2, but with 21 parameters
INI-PP	50	As OPT21P, but without any parameters optimized. The perturbed parameters are the same as in OPT-21P for all members.
QFLX-PP	50	As INI-PP, but with flux corrections.
OPT-5/21	50	As INI-PP, but 5 out of the 21 perturbed Parameters optimized.
OPT-10/21	50	As INI-PP, but 10 out of the 21 perturbed parameters optimized.
EU-OPT-PP	50	As OPT-5/21, but with a cost function only considering the region around Europe.
AUS-OPT-PP	50	As OPT-5/21, but with a cost function only considering the region around Australia.
TPAC-OPT-PP	50	As OPT-5/21, but with a cost function only considering the region around tropical Pacific.
INI-RPL	50	As INI-PP, but 5 out of the 21 perturbed parameters are replaced with the original parameters.
OPT-RPL	50	As OPT-5/21, but 5 out of the 21 perturbed parameters are replaced with the original parameters. This can include some of the optimized parameters
OPT-RPL-OPT	50	As OPT-RPL, but the 5 optimized parameters of the OPT-5/21 ensemble are optimized again.
OPT-RPL-QFLX	50	As OPT-RPL, but with flux corrections.
OPT-COM	50	Based on OPT-5/21 with each new member having half of the parameters from on member of the OPT-5/21ensemble and the other half from another member.
OPT-COM-OPT	50	As OPT-COM, but 5 parameters are optimized again.
OPT-COM-QFLX	50	As OPT-COM, but with flux corrections.

Note. Each member of an ensemble has a different parameter set.

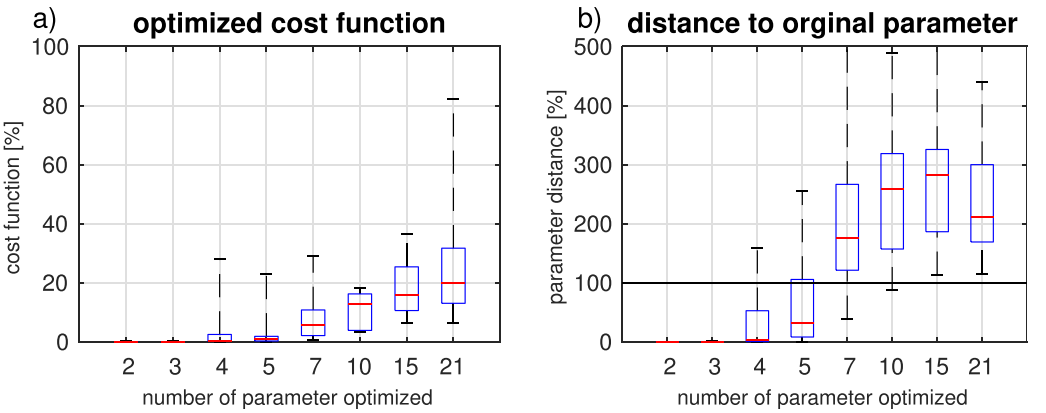

Figure 5. (a) Distributions of the final optimized cost function values for the ensembles with 2–21 parameters perturbed and optimized. Values are relative to the initial perturbed cost function value. (b) Distance of the final optimized parameters to the original parameters in values relative to the initial perturbed parameter distance. The box plot values present the median (red lines), 25% and 75% quantiles (lower and upper boundary of the blue boxes) and the +/–4 standard deviations (lower and upper whiskers).

development. In our optimization approach, we therefore also limited the optimization scheme, as discussed in the previous section, and therefore the optimization is stopped in a region with very small gradients, but still far away from the original GREB model.

In the GREB model simulations, increased precision could be achieved by significantly increasing both the length of the control runs and the number of iterations ($\gg$300). Nonetheless, such increases would result in more than 100,000 years of simulations for undertaking one optimization procedure. Such an undertaking is not realistic for CGCM developments given the computational cost. In practice, the optimization is ended without converging on the true model physics. The cost function is minimized in high-dimensional parameter space to a point that is almost indistinguishable from, but still far away from, the original ("true") physics (see Figure 4).

4.2. Resulting Optimized Control, Physics, and Response Errors

In a realistic optimization problem, all model parameters are uncertain, but not all parameters will be optimized. Model developers usually focus on a subset of parameters that they consider to be relevant for improving model performance. In order to mimic a realistic situation, a scenario in which all uncertain parameters are perturbed is considered; however, only a random subset of parameters is optimized (Table 3).

Figure 6 shows the distributions of ensemble members for different optimized error estimates. Overall, the experiments where only a subset of parameters is optimized behave similarly to those in which all parameters are optimized; however, there are some notable differences. First of all, the cost function is reduced more so when only 5 out of 21 parameters are optimized (OPT-5/21 experiment; Figure 6a). The reduction is primarily caused by making the optimization process less complex (i.e., fewer parameters). Nevertheless, the cost function is also reduced by limiting both the optimization process to 300 iterations and, the control simulation to a maximum of 15 years. The combination of both of these factors (reduced complexity and minimizing iterations/simulation time) causes the distance of the new parameters relative to the "truth" (i.e., original physics) to be smaller when fewer parameters are optimized (Figure 6c). In most cases however, the distance from the "truth" is still larger than in INI-PP ensemble.

From now on, the main focus is on the ensemble where 5 out of 21 perturbed parameters are subsequently optimized. This represents the most realistic scenario for CGCM development as only a small fraction of all uncertain parameters will be optimized in a CGCM; however, the following results are applicable to the cases with 10 or more perturbed parameters are optimized.

Differences in the control T_{surf} spread for the OPT-5/21 ensemble relative to the INI-PP ensemble are plotted in Figure 7a and are representative of a regional map of the cost function (Figure 6a). The reduction in the control T_{surf} spread is mostly uniform and illustrates that the optimization is capable of reducing the control

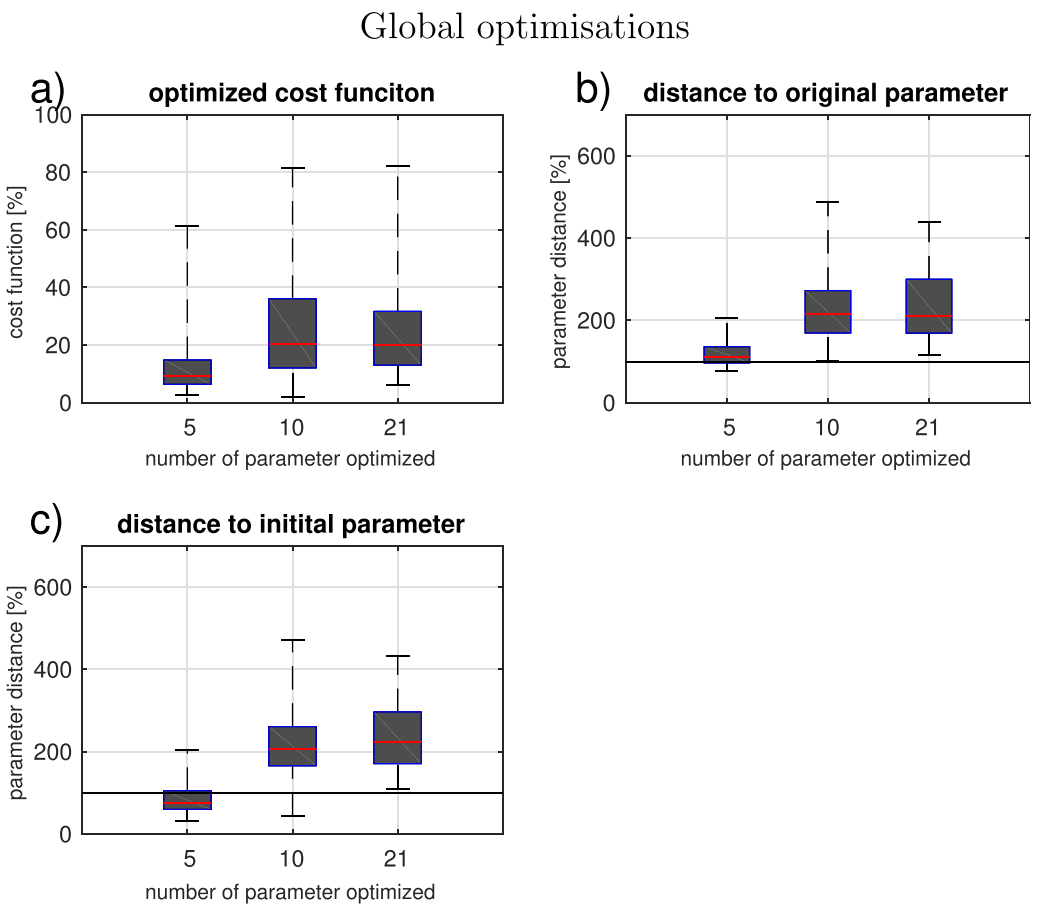

Figure 6. (a) Distributions of the final optimized cost function values, (b) distance to the original and (c) initial parameters for the ensembles OPT-5/21, OPT-10/21, and OPT-21P. Values are relative to the initial perturbed values. Box plot values as in Figure 5.

climate errors both regionally and globally. Nevertheless, the error reduction is slightly smaller over central northern Asia, northern North America, equatorial Africa, and the Southern Ocean.

The reduction in the control climate error (cost function) is unsurprising given that reducing error is the aim of an optimization scheme (by construction). A more important test is to see whether the optimization improves model variables that are not optimized directly (i.e., not used for the cost function). An evaluation of the $2xCO_2$ experiments provides an independent validation of the optimization method. Furthermore, as such experiments ($2xCO_2$) are also routinely run with CGCMs, such an evaluation with GREB is important.

In Figure 7c, the ensemble mean changes in the normalized T_{surf} response pattern spread (π_i; see methods) for the OPT-5/21 ensemble relative to the INI-PP ensemble, are shown. For most regions, the normalized T_{surf} response spread is reduced by 30–60% relative to the INI-PP ensemble; however, over the tropical oceans, it is higher by about 20%. The overall distribution in the T_{surf} response spread for the global mean and pattern are plotted in Figure 8. The OPT-5/21 ensemble has a slightly reduced global mean response spread relative to INI-PP ensemble and, a smaller ensemble mean spread in the response pattern (σ_i; see methods). Nevertheless, the reduction in the response spread is much smaller than in the control climate spread (compare Figure 7a versus Figure 7c). Therefore, the optimizations improve the control climate simulations considerably, but do not improve the simulated response to doubling CO_2 as much.

4.3. Optimization Versus Flux Corrections

From the INI-PP ensemble, a new ensemble with flux correction for the control climate is computed (QFLX-PP). The flux correction will force the QFLX-PP ensemble to be exactly the same as in the original GREB model and corresponds to a zero cost function value and 100% reduction in the control climate T_{surf} spread

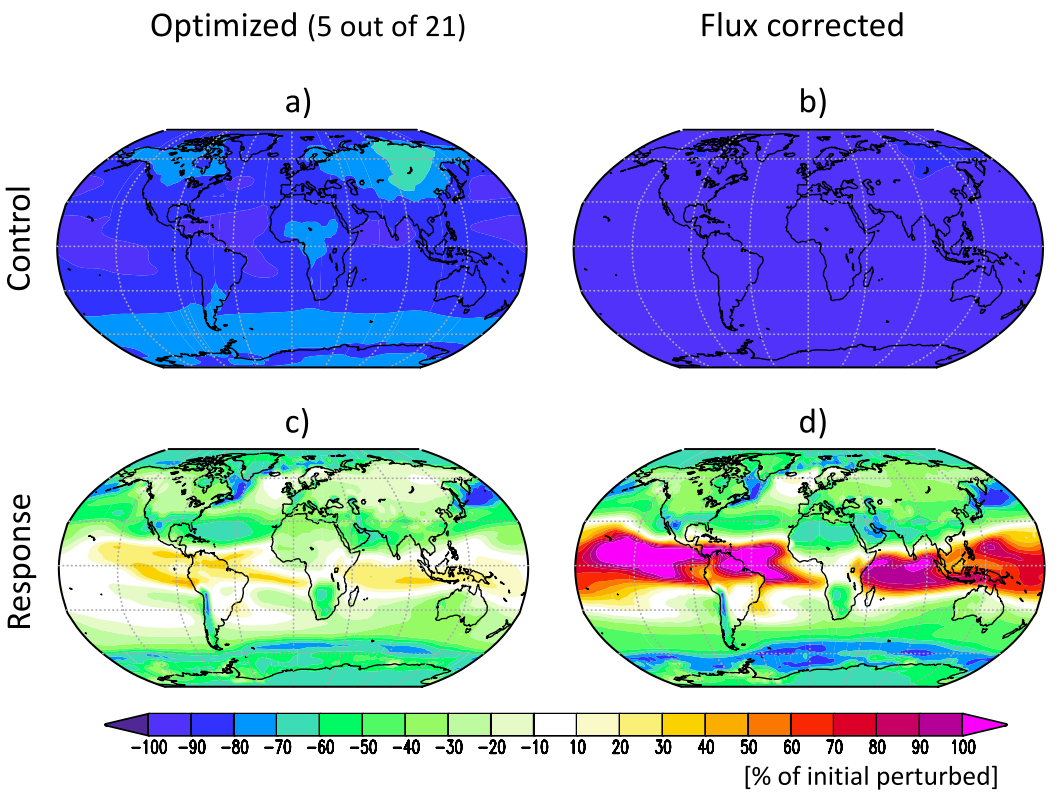

Figure 7. Changes in the spread in the (top row) T_{surf} control and (bottom row) response for the (left column) OPT-5/21 and (right column) QFLX-PP ensemble relative to the INI-PP ensemble spread. Values are in percentage of the INI-PP ensemble spread. Negative values indicate a reduction in spread relative to the INI-PP ensemble.

(see Figures 5a, 6a, and 7b). The perturbed physics are not changed with flux correction and therefore, the distance to the original (or initially perturbed) model parameters will not change (e.g., Figures 5b, 6b, and 7c).

The QFLX-PP ensemble (by definition above) simulates the exact same control climate as the original GREB simulation, but the physics (parameters) are different (perturbed). In turn QFLX-PP uses the exact same physics as the INI-PP ensemble, but the control climate is different from that of INI-PP.

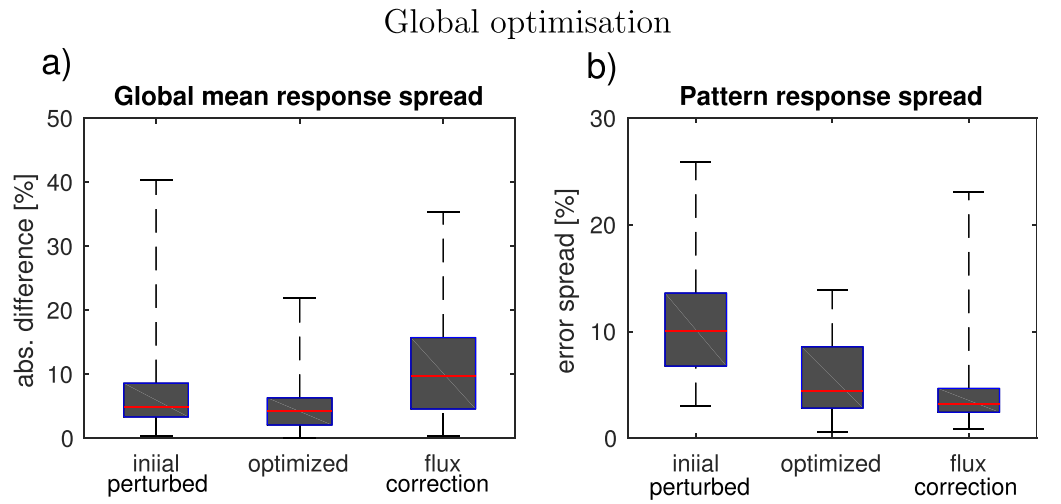

Figure 8. Mean absolute values of the relative differences in (a) the global mean T_{surf} response and (b) the T_{surf} response pattern spread for the INI-PP (initial perturbed), OPT-5/21 (optimized), and QFLX-PP (flux corrected) ensembles. Values in Figure 8a are relative to the original GREB model. For the definition of the response pattern spread, see section 2. Box plot values as in Figure 5.

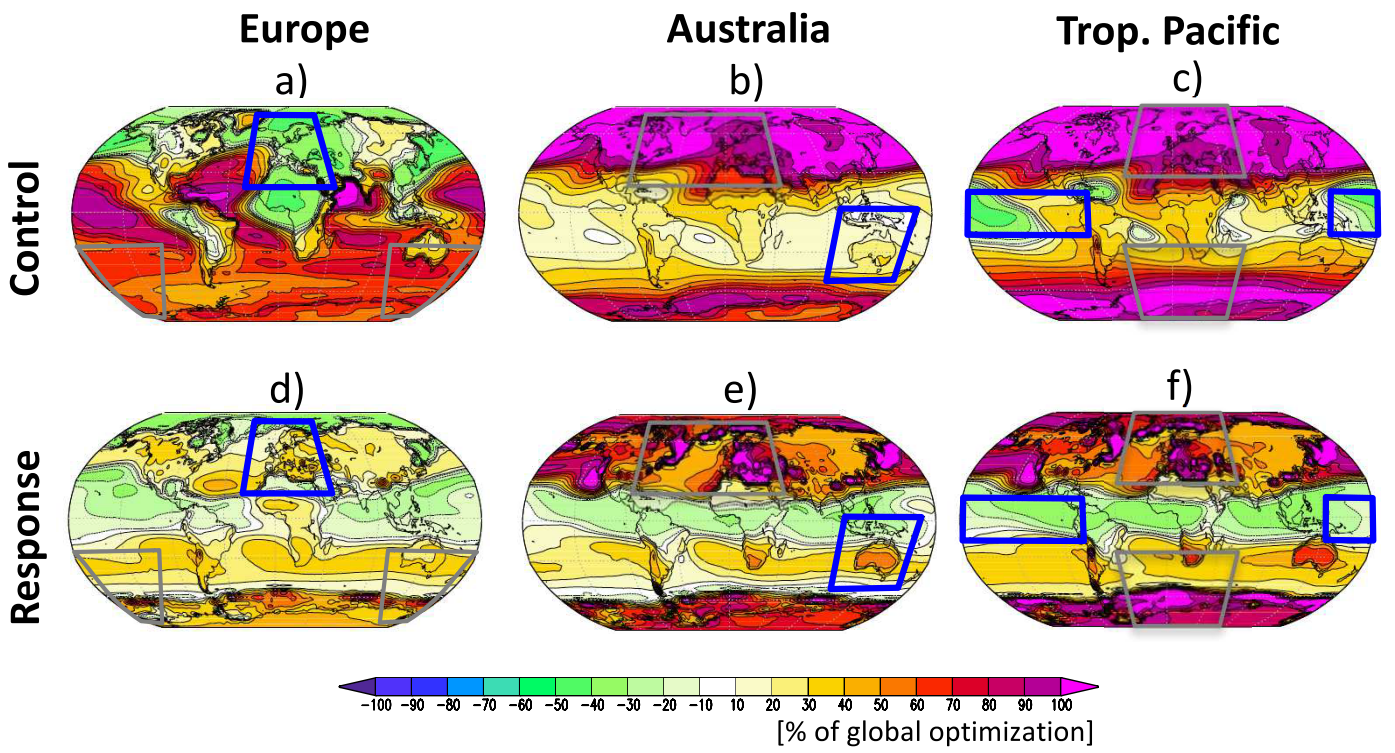

Figure 9. Changes in the spread in the (top row) T_{surf} control and (bottom row) response for the (left column) EU-OPT-PP, (middle column) AUS-OPT-PP and (left column) TPAC-OPT-PP ensemble relative to the OPT-5/21 ensemble spread. The blue boxes mark the optimization regions and the grey boxes mark regions roughly at the opposite side of the globe. Values are in percentage of the OPT-5/21 ensemble spread.

The improved mean state of the QFLX-PP ensemble leads to a reduction in the normalized response spread for most regions, but increases the spread over the tropical oceans (Figure 7d). The response pattern change is similar to the OPT-5/21 ensemble, but the changes are stronger in both, increases (tropics) and decreases (extratropics). The largest improvements are visible over the land and at higher latitudes (Figure 7d). Over tropical oceans, the QFLX-PP ensemble has increased relative errors. As the response spread (absolute) in the INI-PP ensemble is much larger over land and higher latitudes than over the tropical oceans (see Figure 2d), the overall spread in the QFLX-PP ensemble is smaller than in the INI-PP and in the OPT-5/21 ensemble (see Figure 8b).

The global mean response spread is larger in the QFLX-PP ensemble than in the INI-PP and in the OPT-5/21 ensemble (see Figure 8a). This is consistent with the finding that the global mean response variations in the GREB model, due to the control mean climate and perturbed physics, are anticorrelated in the INI-PP ensemble (D16). This leads to a counteracting effect between the control mean climate variations and the perturbed physics.

4.4. Optimizations With a Local Focus

Model development groups may have a tendency to develop their models based upon the representation of the climate in their local area. Thus, placing more importance on the simulation of local climate effectively replaces the global cost function with a region-specific cost function. How does such a regionally based strategy perform relative to a global cost function? To address the effect of local versus global error reduction, the cost function is restricted to only consider a subregion of the globe. Three different example regions are chosen to illustrate this: Europe, Australia, and the tropical Pacific El Nino region (EU-OPT-PP, AUS-OPT-PP, and TPAC-OPT-PP).

In Figure 9, the control climate and T_{surf} response spread of the local optimization ensembles relative to the OPT-5/21 ensemble is shown. The domains considered for the local cost functions are within the blue boxes and diametrically (approximate) opposite regions of the globe are denoted by the grey boxes.

In European case, it is clear that the spread in the control climate (T_{surf}) is reduced within the optimization domain, but has increased relative to the global optimization simulations for most other regions (Figure 9a).

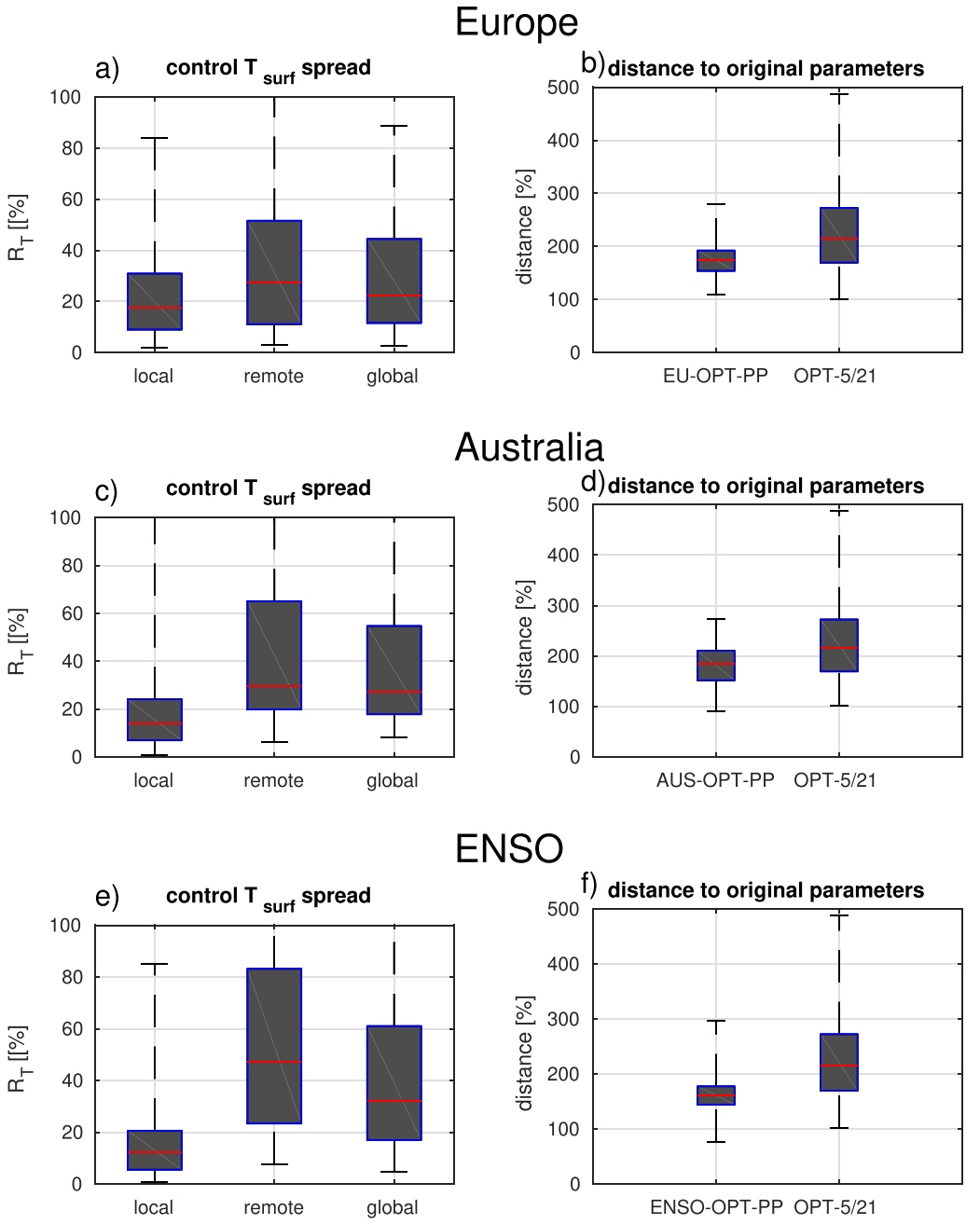

Figure 10. Distributions of the local optimizations: (left column) control T_{surf} spread and (right column) distance of the optimized parameters to the original parameters for (top row) the EU-OPT-PP, (middle row) AUS-OPT-PP, and (bottom row) TPAC-OPT-PP ensemble. The R_T values are computed for the local optimization region (local), the remote region (remote), and globally (global). Values are in percentage of the INI-PP ensemble. Box plot values as in Figure 5.

Thus, the region-specific cost function approach succeeds in reducing the local biases; however, outside the optimization region the performance actually gets worse relative to a globally optimized simulation. A similar result is seen for the El Nino region focus (Figure 9c), but not for the Australian region (Figure 9b).

In Figures 10a, 10c, and 10e, the distribution of the optimized T_{surf} control spread are shown for different regions relative to the initial perturbed T_{surf} control spread. In all three regions, the smallest relative R_T (see section 2) occurs within the optimized area whereas on the opposite side of the globe (grey boxes, Figure 9) the R_T values are clearly larger than for the global mean. It should be noted here that the global relative

R_T in all local optimization ensembles is, on average, larger than in the global OPT-5/21 ensemble (not shown), which suggests the global optimization is more efficient in reducing R_T globally.

The improved climate-mean state for the European region does not result in an improved representation of the regional response to doubling CO_2 concentrations (Figure 9d). Indeed, the local optimization over Europe results in a better representation of the regional climate response than the globally optimized simulation in regions outside Europe. This is an unexpected and unintentional result of utilizing a local optimization strategy.

The other two local optimizations also find mixed results in the normalized response pattern uncertainty. For the Australian domain, the response spread typically increases, whereas for the ENSO domain focused optimization there is a reduction in the normalized response pattern uncertainty.

In all local optimization experiments, the parameter distances to the original values increase (Figures 10b, 10d, and 10f), which indicates that the local optimization also does not result in a model that is closer to the true physics. Nevertheless, the distances from the original parameter values do not increase as much as in the global OPT-5/21 ensemble.

4.5. Introducing Model Improvements Into Optimized (Tuned) Models/Compensating Errors

A typical problem in model development is how to introduce an improved submodel (e.g., a new cloud parameterization scheme) into a tuned CGCM. In general, one would expect that replacing an inferior submodel with a superior one would improve the overall performance of the CGCM—but this is often not the case. Compensating errors are the most likely culprit for the poorer performance—i.e., an error in one submodel is compensated for by errors in the others. If the errors in the submodel are unrelated (or random as in the INI-PP ensemble), error compensation only occurs by chance and may therefore not be a significant problem; however, in an optimization scheme, this is not always true. For instance, a tuning of a CGCM may adjust the ocean mixing scheme to improve the simulation of sea surface temperature (SST), but the errors in the SST may actually be caused by e.g., flaws in the cloud scheme (as illustrated in the following experiments).

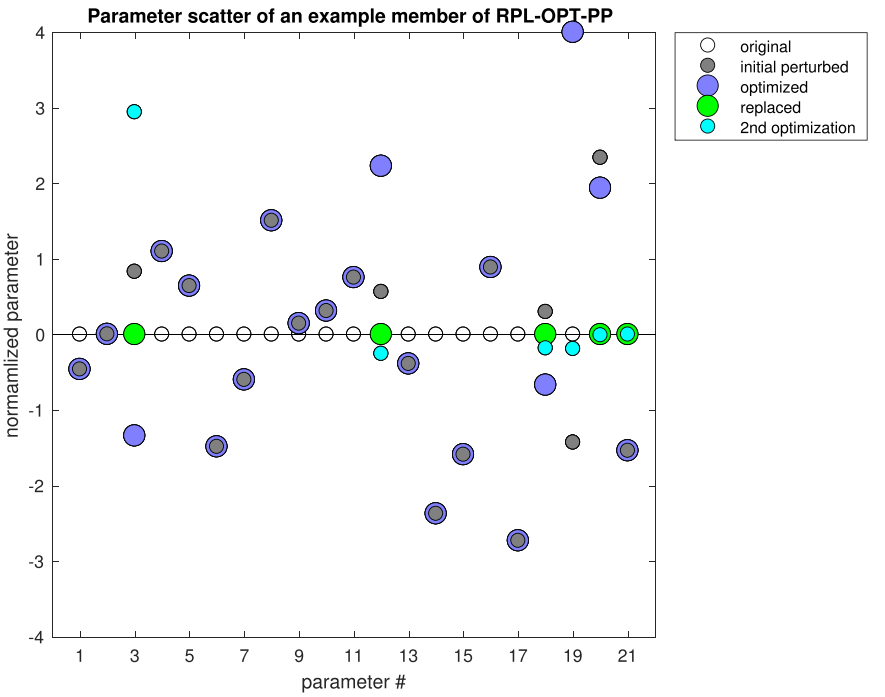

Figure 11. Normalized parameter scatter for the original model (white circles) and an example member of: the INI-PP (initial perturbed; grey), OPT-5/21 (optimized; blue), OPT-RPL (replaced; green), and the OPT-RPL-OPT (second optimization; light blue).

New simulations are constructed based on the INI-PP and OPT-5/21 in order to mimic the methods of replacing inferior submodels that are used to improve CGCMs (replace the inferior submodel with a "perfect" one). For this we randomly replaced five perturbed parameters in the INI-PP and OPT-5/21 ensembles against the original (true) parameters, which is illustrated in Figure 11. Both control and response experiments are performed with the two new experiments (INI-RPL and OPT-RPL; see Table 3).

The INI-RPL ensemble essentially behaves as expected: the parameter uncertainties always reduce (Figure 12a) and the control climate spread reduces in most cases (Figure 12b). This also holds for the response spread (Figures 12c and 12d). The OPT-RPL simulations, in turn, behave quite differently. The parameter uncertainties reduce, but the relative uncertainties are distributed differently due to the different OPT-5/21 ensemble parameter spread. The global mean response spread also improves (Figure12c); however, there is a much larger spread in the control climate (Figure 12b) and, also an increased response pattern spread (Figure 12d) that is now higher relative to the OPT-5/21 ensemble. The results therefore suggest that the optimized parameters of the OPT-5/21 ensemble have a large degree of error compensation. If a few optimized parameters with compensating errors in other parameters are replaced, the errors associated with the other parameters are visible.

The OPT-RPL simulations need to be optimized again in order to improve their performance relative to the OPT-5/21 ensemble. A second optimization, based on the OPT-RPL ensemble (OPT-RPL-OPT), does indeed improve their behavior relative to the OPT-5/21 ensemble in all aspects (Figure 12). The application of a flux

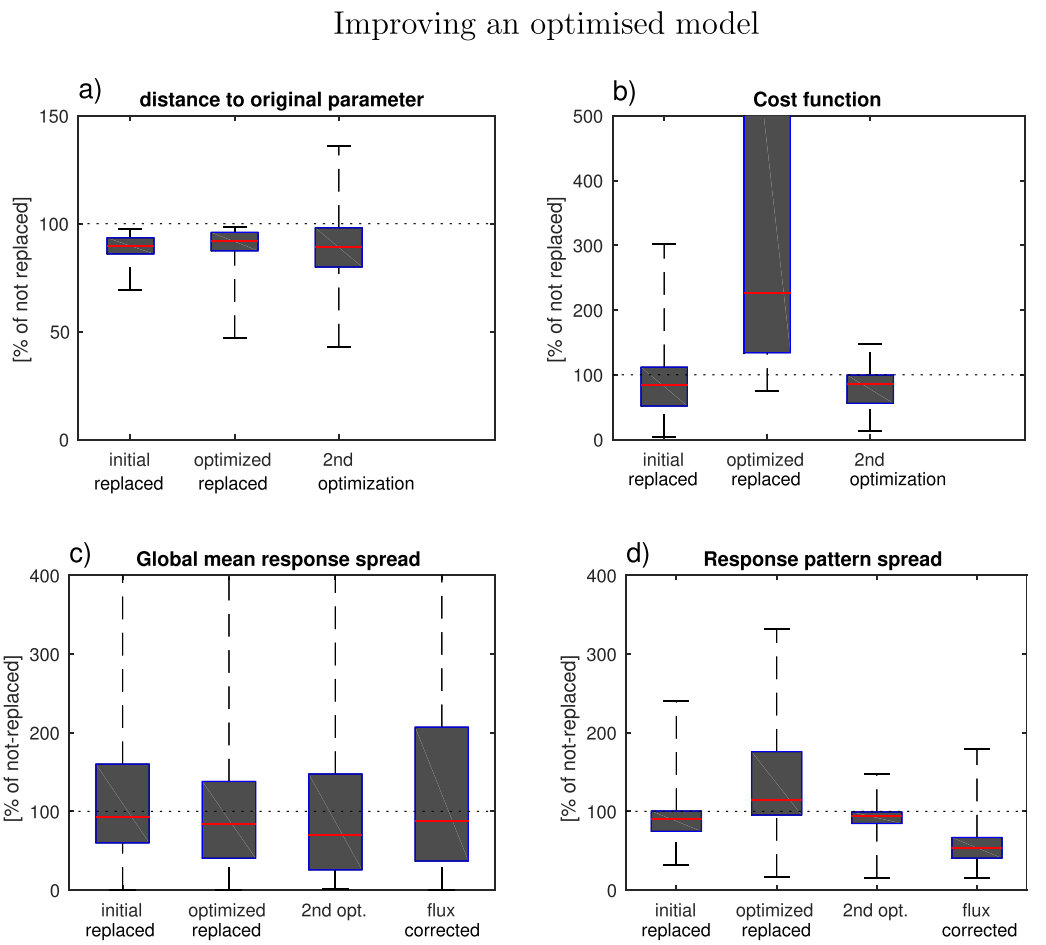

Figure 12. Statistics for the replaced parameters ensembles: Distance to (a) the original parameters, (b) the cost function, (c) global mean response spread, and (d) the normalized response pattern spread for the INI-RPL (initial replaced), the OPT-RPL (optimized replaced), the OPT-RPL-OPT (second optimization), and for the RPL-QFLX (flux corrected) ensembles. Values for the INI-RPL ensemble are in percentage of the INI-PP ensemble and all others are in percentage of the OPT-5/21 ensemble.

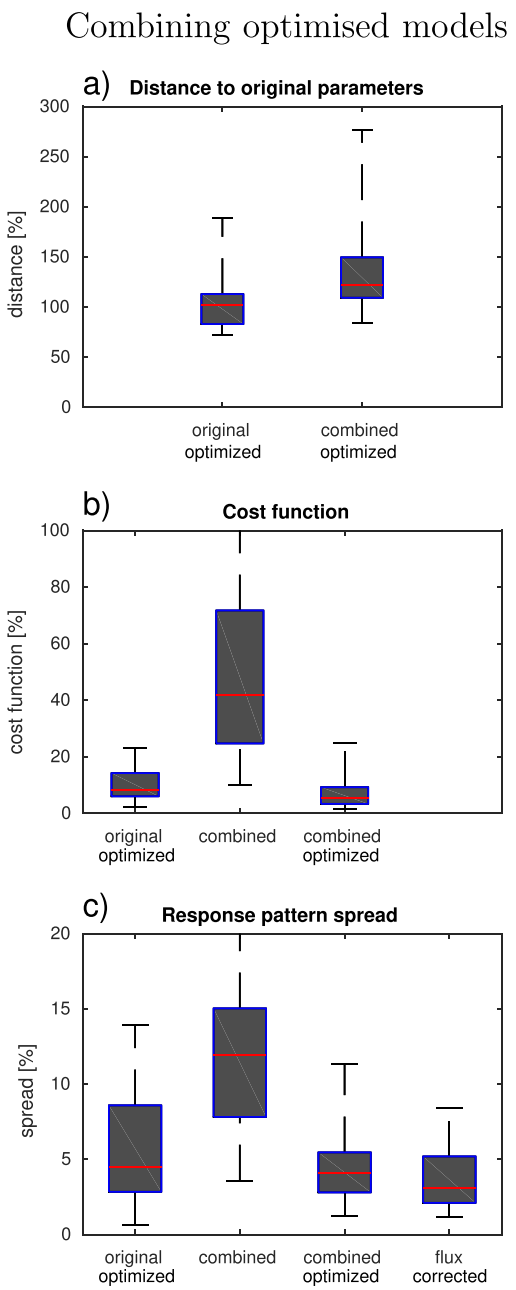

Figure 13. Statistics for the combined parameters ensembles: Distance to (a) the original parameters, (b) the cost function, and (c) the normalized response pattern spread for the original optimized (OPT-5/21), the combined (OPT-COM), the combined optimized (OPT-COM-OPT), and for the flux corrected (OPT-COM-QFLX) ensembles. Percentage values in Figure 13a are relative to the OPT-5/21 ensemble and in Figure 13b relative to the INI-PP ensemble.

correction term to the OPT-RPL ensemble (OPT-RPL-QFLX) results in a similar, but slightly larger, spread in the global mean and a better representation of response pattern uncertainty (Figure 12c,d).

It is important to note here that the parameters are optimized (i.e., "tuned") a second time such that parameters (see for instance parameter 3, 12 and 18 in Figure 11) are changed from their original "perfect" values. The optimization therefore, actually introduces artificial errors into these parameters from minimizing the cost function.

4.6. Combining Optimized Submodels

Another problem in CGCM development arises from attempting to combine two submodels from different CGCMs (e.g., the atmosphere model from one CGCM combined with the ocean model from another CGCM). It could be expected that combining any two submodels from different CGCMs should result in two equally good CGCMs. Given the above discussion however, it is more likely that error compensation that arises from the optimization process may actually lead to larger model errors, as illustrated below.

Based on the OPT-5/21 ensemble, two ensemble members are randomly combined by taking one half of the parameters from one member and the other half from the other member (OPT-COM). A 50 member ensemble is created in this way, see Table 3. The parameter set by which we split the models into halves are the same for all ensemble members and are not chosen related to any parameter uncertainties, but solely chosen by splitting the GREB model into roughly equal halves. Thus, each member of the OPT-COM ensemble is a combination of two members from the OPT-5/21 ensemble.

As expected, the uncertainties in the OPT-COM ensemble are increased for the control and response climate states to the OPT-5/21 ensemble (Figures 13b and 13c). The optimization of the two submodels in the OPT-5/21 ensemble, which now form a new model in the OPT-COM ensemble, has led to compensating errors in submodels parts that are not part of the new model any more. These missing compensating errors now lead to increased uncertainties. Again, a second optimization is required to improve model performance (OPT-COM-OPT). The OPT-COM-OPT simulations have lower uncertainties in the control and response climates relative to the OPT-5/21 ensemble (Figures 13b and 13c); however, the model physics are further away from the "true" physics than in the OPT-5/21 ensemble (Figure 13a).

Again, if a flux correction is applied to the OPT-COM ensemble, the normalized response pattern spread is smaller than that of a second optimization (Figure 13c).

5. Summary, Discussions, and Conclusion

This study addresses the problems associated with tuning of climate models and how it would relate to an alternative of flux corrections. A perfect model approach is undertaken—i.e., assume that the original GREB model simulation is the "truth" and that the PP experiments are representative of typical model errors. A Nelder-Mead optimization scheme with realistic limitations (section 3) is applied in an attempt to mimic the tuning strategies of CGCM development groups. The overall aim of this work is to elucidate the problems associated with tuning fully coupled global climate models (e.g., CGCM), but not on tuning a specific submodel.

The first important result shows that the GREB model is highly complex and that the optimization of the perturbed model parameters is very difficult under the given constraints. If more than four uncertain parameters are considered, the optimization will not converge and the cost function will remain significantly larger than zero. In realistic situations, where all parameters are uncertain and a small subset is tuned, the optimization process is able to reduce the errors in the control mean climate (cost function) substantially (80–90%). However, the optimized parameters are typically further away from the original parameter values than the initial perturbed parameters, creating numerous compensating errors that reduce the cost function.

The optimization did improve the simulated response to a doubling of CO_2 even though this is not accounted for by the cost function. The optimized improvements in the CO_2 response are, however, much smaller than those found in the control climate (e.g., compare Figure 7a versus Figure 7c).

In the flux correction approach, the normalized pattern response to CO_2 forcing is better than in the optimized simulations, but the global mean spread increases. The spread increase for the global mean is caused by the compensating effects (anticorrelation) between the uncertainties resulting from the GREB model physics and those due to the mean state (D16).

Replacing a subsystem model in an optimized model against a perfect subsystem model, thus improving the physics of the model, does not improve the simulation of the model. This is due to the compensating errors that the optimization introduces into the optimized model. In a similar way, if we combine submodels of different optimized models, the simulations will significantly degrade, which does not happen in a perturbed model (without optimizations) or in the flux corrected model.

The reasons why the optimization creates compensating errors and, why the flux correction approach is in some cases superior to the tuning approach needs some further discussion. The errors introduced by perturbing model physics in the GREB model can conceptually be split into two parts: A "direct" error from the perturbed physics and an "indirect" error from the biased mean-state climate (D16). The direct error results into incorrect tendencies of the model given the right boundary conditions due to the perturbed physics. The indirect error is caused by the biased mean state, which is a direct result of the perturbed physics. This biased mean state causes incorrect tendencies of the model even with the correct physics, because many feedbacks and processes in the climate system are mean state-dependent.

In the perturbed physics optimization of the GREB model, the optimization scheme has to correct for both kind of errors at the same time. Since, the cost function was defined by the mean-state climate and not by the "true" physics, it will tend to reduce the indirect error by the mean state and not the errors in the perturbed physics. For instance, if the model has too large humidity in some tropical regions, which do result from biases in the mean state surface temperatures, then the optimization may correct these by changing the parameters of the hydrological cycle. While this may succeed in reducing the cost function, it does so by introducing artificial errors in the physics of the hydrological cycle model to compensate for errors in the physics that are in other processes of the GREB model. It does create compensating errors.

In general, this will be similar in CGCM model development. Within the CGCM simulation, the errors in each submodel produce errors that are partly due to the error in the submodel's parameterization and partly due to the biased input they get from the other submodels. In the tuning of the CGCM, the parameters of each submodel are adjusted to compensate for both. For instance, if a CGCM simulation produces biases in cloud cover in a particular region this can result from the error in the cloud model or from the incorrect input (biased mean state in other climate variables). The development team of the CGCM may come to the conclusion to tune the parameters of the cloud model, even though it may indeed be problems in other submodels leading to the biased input into the cloud model. They come to this conclusion, because they have only limited information and the system is simply too complex to get to the right conclusions.

The flux correction approach improves the simulations in many cases, because it directly corrects the indirect error (mean-state climate) problem. By achieving a mean T_{surf} and atmospheric humidity that is as in the original model it provides a more realistic input for all submodel processes. This reduces one part of the errors and helps to get better tendencies despite having perturbed physics. This, however, is not always the case and indicates that the flux correction approach is better when the balance between direct and indirect errors is toward the indirect errors.

In the optimization simulations, the GREB model is also producing better response to CO_2 forcing than the initial perturbed physics ensembles, because it has a mean state closer to the original GREB model. Though, it performs not as good as the flux correction approach in most aspects, despite both having now very similar mean-state climate, because it has tuned physics that are now further away from the "true" physics of the GREB model than in the initial perturbed physics ensembles. Thus, the optimization approach has succeeded in reducing the indirect error, but at the same time has increased the direct errors for the response to CO_2 forcing. It has created compensating errors that only become visible if it is applied to a problem that is significantly different to what it has been tuned for.

Assuming the results from this GREB model study hold for realistic CGCM developments, we can conclude the following:

1. **Tuning of CGCMs is unlikely to improve the model physics**: The tuning process will reduce the cost function, but the actual physics are not closer to the "true" physics. Subsequently, the models appear to represent the climate better, but when a detailed analysis is undertaken on process-basis or under different climatic states (e.g., paleo or future climate change scenarios), substantial model errors will occur. Tuning will be more successful in limited complexity problems, such as a submodel system of a CGCM forced with observed boundary conditions (e.g., a sea ice model or 1–5 parameters of the GREB model). Here the tuning may find the physically better parameters.
2. **Tuning actually makes coupled model development (and improvement) more difficult**: Climate model tuning introduces compensating errors that effectively make all submodels of the CGCM dependent on each other. Thus, it is almost certain that replacing or exchanging submodels causes degradation in the mean climate state and further optimization is required. The optimization will then (again) tend to move the model physics away from the "truth."
3. **Flux correction represents a useful alternative**: Flux correction does not alter the model physics and therefore no optimization is required. Furthermore, error compensation introduced by the tuning process will also not be included. Another benefit is that flux corrections are easy to document and quantify—the correction is simply an input for the CGCM (and can be turned off). Thus, it can be tested how strongly they affect the model performance. Flux correction also makes it easier to undertake model improvements and combine different model parts as compensating errors from tuning are not included. Nonetheless, they are not without problems and in some circumstances, may cause a degradation of the model performance (e.g., the global mean response to CO_2 forcing in the GREB model).
4. **Climate models can be improved without improving any of the submodels**: Concepts like flux corrections or anomaly coupling can reduce the interactions of errors between submodels. This improves CGCM performance, because each submodel would see more realistic mean input variables and uses more realistic physical parameters.

A significant caveat of this study is that the GREB model and its uncertainties may be substantially different from those of a CGCM. The fact that the flux correction performs well in many aspects (e.g., Figures 6b and 8b) does not imply that it will always perform well. This will depend on the kind of application for which the model is used and on what the aim of the application is. We therefore, should consider this study as a pilot, which should motivate further studies with CGCMs to replicate the findings. It should motivate a rethink of the approaches for developing CGCMs.

Acknowledgements
We like to thank Duncan Ackerley and Rob Colman for helpful discussions and comments. We further like to thank the associate editor Thorsten Mauritsen and three anonymous referees for their many helpful comments. This study was supported by the ARC Centre of Excellence for Climate System Science, Australian Research Council (grant CE110001028). The data needed for reproducing the results of this study are published under doi:10.4225/41/5a0e63c045536.

References

Bony, S., Colmanb, R., Kattsovc, V. M., Alland, R. P., Brethertone, C. S., Dufresnea, J.-L., . . . Webb, M. J., (2006). How well do we understand and evaluate climate change feedback processes? *Journal of Climate*, *19*, 3445–3482.

Collins, M., Booth, B. B. B., Harris, G. R., Murphy, J. M., Sexton, D. M. H., & Webb, M. J. (2006). Towards quantifying uncertainty in transient climate change. *Climate Dynamics*, *27*, 127–147.

Collins, M., Booth, B. B. B., Harris, B., Murphy, G., Sexton, J. D., & Webb, M. (2010). Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Climate Dynamics*, *36*, 1737–1766.

Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J. Dix, M., Noda, A., . . . Zwiers, F. (2001). Projections of future climate change. In J. T. Houghton et al. (Eds.), *Climate change 2001: The scientific basis*. Cambridge, UK: Cambridge University Press.

Dommenget, D. (2016). A simple perturbed physics study of the simulated climate sensitivity uncertainty and its relation to control climate biases. *Climate Dynamics*, *46*, 427–447.

Dommenget, D., & Floter, J. (2011). Conceptual understanding of climate change with a globally resolved energy balance model. *Climate Dynamics*, *37*, 2143–2165.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W. . . . Rummukainen, M. , (2013). Evaluation of climate models. In T. F. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press.

Ginzburg, L. R., & Jensen, C. X. J. (2004). Rules of thumb for judging ecological theories. *Trends in Ecology & Evolution*, *19*, 121–126.

Golaz, J. C., Horowitz, L. W., & Levy, H. (2013). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40*, 2246–2251. https://doi.org/10.1002/grl.50232

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., . . . Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*, 589–602.

Irvine, P. J., Gregoire, L. J., Lunt, D. J., & Valdes, P. J. (2013). An efficient method to generate a perturbed parameter ensemble of a fully coupled AOGCM without flux-adjustment. *Geoscientific Model Development*, *6*, 1447–1462.

Knutti, R. (2008). Should we believe model predictions of future climate change?. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, *366*, 4647–4664.

Knutti, R., Allenc, M. R., Friedlingsteind, P., Gregoryef, J. M., Hegerlg, G. C., Meehlb, G. A., . . . Wigely, T. M. L. (2008). A review of uncertainties in global temperature projections over the twenty-first century. *Journal of Climate*, *21*, 2651–2663.

Knutti, R., & Sedlacek, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*, 369–373.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, *9*, 112–147.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B*, *41*, 253–262.

Manabe, S., & Stouffer, R. J. (1988). Two stable equilibria of a coupled ocean-atmosphere model. *Journal of Climate*, *1*, 841–866.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., . . . Tomassini, L., (2012). Tuning the climate of a global model. *Journal Advances in Modeling Earth Systems*, *4*, M00A01. https://doi.org/10.1029/2012MS000154

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

Nelder, J. A., & Mead, R. (1965). A simplex-method for function minimization. *The Computer Journal*, *7*, 308–313.

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*, 303.

Sausen, R., Barthel, K., & Hasselmann, K. (1988). Coupled ocean-atmosphere models with flux correction. *Climate Dynamics*, *2*, 145–163.

Schneider, E. K. (1996). Flux correction and the simulation of changing climate. *Annales Geophysicae*, *14*, 336–341.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., . . . Miller, H. L. (2007). Climate change 2007: The physical science basis. In S. Solomon et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of working group 1 to the fourth assesment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., & Frame, D. J., . . . Allen, M. R., (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*, 403–406.

Tang, Y. L., Li, L. J., Dong, W. J., & Wang, B. (2016). Reducing the climate shift in a new coupled model. *Science Bulletin*, *61*, 488–494.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of Cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498.

Tett, S. F. B., Mineter, M. J., Cartis, C., Rowlands, D. J., & Liu, P. (2013). Can top-of-atmosphere radiation measurements constrain climate predictions? Part I: Tuning. *Journal of Climate*, *26*, 9348–9366.

Zhang, X. F., Zhang, S. Q., Liu, Z. Y., Wu, X. R., & Han, G. J. (2015). Parameter optimization in an intermediate coupled climate model with biased physics. *Journal of Climate*, *28*, 1227–1247.

Zhang, X. F., Zhang, S. Q., Liu, Z. Y., Wu, X. R., & Han, G. J. (2016). Correction of biased climate simulated by biased physics through parameter estimation in an intermediate coupled model. *Climate Dynamics*, *47*, 1899–1912.