Objective calibration of regional climate models

O. Bellprat,[1] S. Kotlarski,[1] D. Lüthi,[1] and C. Schär[1]

Received 8 June 2012; revised 28 September 2012; accepted 23 October 2012; published 13 December 2012.

[1] Climate models are subject to high parametric uncertainty induced by poorly confined model parameters of parameterized physical processes. Uncertain model parameters are typically calibrated in order to increase the agreement of the model with available observations. The common practice is to adjust uncertain model parameters manually, often referred to as expert tuning, which lacks objectivity and transparency in the use of observations. These shortcomings often haze model inter-comparisons and hinder the implementation of new model parameterizations. Methods which would allow to systematically calibrate model parameters are unfortunately often not applicable to state-of-the-art climate models, due to computational constraints facing the high dimensionality and non-linearity of the problem. Here we present an approach to objectively calibrate a regional climate model, using reanalysis driven simulations and building upon a quadratic metamodel presented by Neelin et al. (2010) that serves as a computationally cheap surrogate of the model. Five model parameters originating from different parameterizations are selected for the optimization according to their influence on the model performance. The metamodel accurately estimates spatial averages of 2 m temperature, precipitation and total cloud cover, with an uncertainty of similar magnitude as the internal variability of the regional climate model. The non-linearities of the parameter perturbations are well captured, such that only a limited number of 20–50 simulations are needed to estimate optimal parameter settings. Parameter interactions are small, which allows to further reduce the number of simulations. In comparison to an ensemble of the same model which has undergone expert tuning, the calibration yields similar optimal model configurations, but leading to an additional reduction of the model error. The performance range captured is much wider than sampled with the expert-tuned ensemble and the presented methodology is effective and objective. It is argued that objective calibration is an attractive tool and could become standard procedure after introducing new model implementations, or after a spatial transfer of a regional climate model. Objective calibration of parameterizations with regional models could also serve as a strategy toward improving parameterization packages of global climate models.

Citation: Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2012), Objective calibration of regional climate models, *J. Geophys. Res.*, *117*, D23115, doi:10.1029/2012JD018262.

1. Introduction

[2] Climate models are validated against observational data sets in order to judge the capability of the model and to determine model deficiencies which originate from different modeling assumptions with their related uncertainties. A major source of this uncertainty stems from the large number of parameterized physical processes within the climate model and the associated unconfined model parameters. Several studies have demonstrated the importance of this "parameter uncertainty" for the simulation of present and future climates by perturbing single and multiple model parameters within plausible parameter ranges determined by expert judgment [*Knutti et al.*, 2002; *Murphy et al.*, 2004; *Stainforth et al.*, 2005; *Klocke et al.*, 2011; *Bellprat et al.*, 2012]. Since uncertain model parameters are responsible for a large part of modeling errors, the parameter uncertainty is typically constrained by calibration or tuning methods to improve the agreement of the climate model and the available observations.

[3] Model calibration and tuning is a subject of constant debate and strongly diverging opinions [*Oreskes et al.*, 1994; *Randall and Wielicki*, 1997; *Beven*, 2002]. Because of the high risk that calibrated model processes may compensate for model errors which may not originate from the respective parameterization [*Murphy et al.*, 2007], model tuning is sometimes referred to as *bad empiricism.* It has also been suggested that parameterizations should therefore be calibrated in terms of climate processes rather than against

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

Corresponding author: O. Bellprat, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstr. 16, CH-8092 Zürich, Switzerland. (omar.bellprat@env.ethz.ch)

0148-0227/12/2012JD018262

variables of interest such as surface temperature and precipitation [*Randall and Wielicki*, 1997]. However, related observations are often not available or associated with high observational uncertainties.

[4] Despite the risk of compensating errors, the prerequisite that climate models should be able to reproduce past observations in order to project a future climate remains a common consensus. In practice, model calibration or tuning against available observations is therefore routinely performed, mainly by expert tuning. Expert tuning is a loose term referring to the subjective adjustment of model parameters, mostly neglecting parameter interactions and following no objective procedure. This practice typically lacks of transparency on the use of observations [*Knutti and Hegerl*, 2008]. Often the associated tuning does not follow a well-defined strategy, lacks proper documentation, and often happens implicitly at the workbench of scientists that aim at improving existing or implementing new parameterization schemes. Much tuning happens in a grey zone between the development and implementation of parameterization schemes, as there is little guidance where the one ends and the other begins. The lack of accepted calibration and tuning methodologies carries considerable risks. For instance, sometimes one may wonder whether some "demonstrated model improvement" indeed reflects an improved model structure, or whether it derives merely from skillful tuning efforts. Likewise, the lack of approved standards implicitly leads to increasing model complexity, as the number of tunable parameters– at least in principle – raises the prospects of model tuning (but also increases the risk of over-parameterization).

[5] The development and application of objective calibration methods has therefore recently gained much attention in climate science. In particular climate models of intermediate complexity have been subject to a wide range of objective calibration methods, as shown by, e.g., *Price et al.* [2009] using genetic algorithms, or by *Beltran et al.* [2006] using an oracle-based optimization. Also physical surrogates of general circulation models with reduced complexity or resolution have been optimized with very different approaches ranging from ensemble Kalman filters, Latin hypercubes, and Markov chain Monte Carlo integrations [*Jackson et al.*, 2004; *Jones et al.*, 2005; *Annan et al.*, 2005; *Medvigy et al.*, 2010; *Jarvinen et al.*, 2010; *Gregoire et al.*, 2011], with an overview presented in *Annan and Hargreaves* [2007]. Most of these methods are not directly applicable to computationally costly general circulation models, since typically hundreds of simulations have to be performed in order to calibrate a small set of model parameters.

[6] Here we argue that calibration should also be addressed, even with costly and complex atmospheric models, and that it should follow some well-defined standards. In particular, calibration should (1) be transparent and reproducible, (2) target a small list of key tunable parameters, (3) optimize a pre-defined performance score that accounts for uncertainties associated with observations and predictability, (4) employ an objective optimization methodology, and (5) allow for a clear separation between calibration and validation/verification periods. Similar standards have been discussed and implemented in other science fields for some time. For instance, for hydrological models, where the number of tunable parameters is overwhelming, some general principles have emerged [e.g., *Beven*, 1989; *Refsgaard and Henriksen*, 2004].

[7] A promising way to overcome the problem of computational costs is to construct a statistical surrogate model, also termed model emulator [*O'Hagan*, 2006] or metamodel [*Neelin et al.*, 2010], which is a computationally cheap representation of the climate model's sensitivity to parameter perturbations. A model emulator allows estimating the simulated climate variables of interest for a specific input of model parameters, without conducting full model simulations. Using an emulator, large ensembles of the climate model can be computed as done, e.g., in *Huber* [2011], which allows the application of comprehensive calibration methods. Different kinds of surrogate models have been considered to emulate climate models, ranging from artificial neural networks [*Knutti et al.*, 2003], parametric regression models [*Neelin et al.*, 2010] and non-parametric Gaussian process models [*Rougier et al.*, 2009].

[8] Here we present an application of a calibration framework using a second order polynomial metamodel proposed by *Neelin et al.* [2010] to a regional climate model (RCM). Other studies have already successfully used quadratic regressions to estimate parameter perturbations [*Jones et al.*, 2005], yet in contrast to the present study neglecting parameter interactions.

[9] The calibration of RCMs is particularly interesting because RCM experiments driven by re-analysis data at the lateral boundaries allow isolating the effects of regional-scale processes on error characteristics [*Suklitsch et al.*, 2010; *Bellprat et al.*, 2012, hereinafter B11]. However, to our knowledge, so far no objective calibration framework for RCMs has been presented.

[10] The present paper is structured as follows: Section 1 describes the modeling approach, sections 2 and 3 show how the model performance is assessed and used to determine important model parameters, section 4 explores the application of a metamodel and section 5 discusses the model calibration and the implications for expert tuning.

2. Development of Methodology

2.1. Regional Climate Model Approach

[11] The model used for this calibration study is the non-hydrostatic regional climate model COSMO-CLM (hereafter CCLM) version 4.8. The CCLM model is a versatile limited-area atmospheric modeling system including a whole suite of model parameterizations [*Steppeler et al.*, 2003; *Förstner and Doms*, 2004]. It is based on the non-hydrostatic compressible atmospheric equations, uses the split-explicit time stepping scheme [*Klemp and Wilhelmson*, 1978; *Wicker and Skamarock*, 2002], and is suited for applications with horizontal grid-spacings from about 100 m to 100 km. The CCLM has heavily been used for regional climate studies using real-case [e.g., *Kotlarski et al.*, 2012] and idealized configurations [e.g., *Schlemmer et al.*, 2011].

[12] The setup of CCLM for the current study is identical as in B11 with some modifications of the model physics: Other than using the cumulus convection scheme of the ECMWF Integrated Forecast System (IFS) model (P. Brockhaus et al., The ECMWF IFS convection scheme applied to the COSMO-CLM limited-area model, submitted to *Quarterly Journal of the Royal Meteorological Society*, 2011) we use

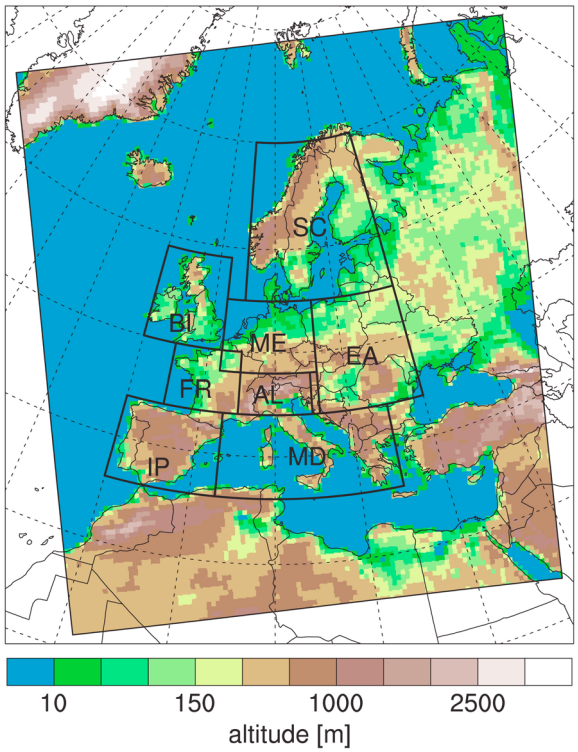

Figure 1. Model domain and PRUDENCE analysis regions: BI = British Isles, FR = France, IP = Iberian Peninsula, MD = Mediterranean, AL = Alps, ME = Mid-Europe, EA = Eastern Europe. The domain has a rotated pole and a resolution of 0.44° (∼50 km).

the default Tiedtke convection scheme [*Tiedtke*, 1989]. In addition the aerosol climatology has been changed from an default climatology of *Tanré et al.* [1984] to the higher resolution AEROCOM climatology [*Kinne et al.*, 2006], which provides more realistic estimates of aerosol loadings over Europe [*Zubler et al.*, 2011]. Further changes include a satellite derived soil albedo field from the MODIS sensor and a plant albedo field [*Houldcroft et al.*, 2009].

[13] The domain of the RCM covers a greater European region at a resolution of 0.44° as shown in Figure 1 with contours representing the model topography. The black boxes show climatic regions commonly termed PRUDENCE regions on which the analysis for this study is based. In order to be consistent with B11 we focus on the same time period from 1990 to 2000. For this period a reference simulation (REF) has been performed with model settings derived from an expert tuning process for the Coordinated Regional climate Downscaling Experiment over Europe (CORDEX, www.euro-cordex.net). Furthermore an initial condition ensemble with of five simulations from 1990 to 2000 with 6 hourly shifts of the initialization time was conducted to determine the model's internal variability. Due to computational constraints, the number of simulations in this ensemble is kept at a lower limit but is consistent with other studies assessing the internal variability of RCMs [e.g., *Roesch et al.*, 2008]. The simulations used to find optimal parameter configurations and to determine the accuracy of the metamodel are restricted to the 5-year period from 1994 to 1998 which is a sufficient integration length to reach convergence of the adopted skill metrics (see B11). These experiments were initialized with the equilibrium state as obtained from the reference simulation.

2.2. Validation Framework

[14] The parameter optimization of requires a framework to objectively assess model performance against observations. There are many ways how to measure the performance of a climate model, with choices regarding the metrics, model variables and data sets. Although there is some guidance for the validation climate models different approaches often lead to controversial outcomes [*Gleckler et al.*, 2008]. The performance of models is typically assessed with some distance measure between the model and observations [*Perkins et al.*, 2007; *Christensen et al.*, 2010]. Since one variable might be improved at the expense of some other [*Jones et al.*, 2005; *Vidale et al.*, 2003], several studies use a multivariate framework including several variables which represent dominant climate processes, as e.g. top of the atmosphere radiation, surface radiation balance, mean sea level pressure and total cloud cover [*Gleckler et al.*, 2008].

[15] In this study we use the validation framework presented in B11. Model performance is expressed as a function of 2 m temperature (T2M), precipitation (PR) and total cloud cover (CLCT). This allows to validate the variables that are often of primary interest (T2M, PR) and an additional process variable (CLCT) which plays an important role in the interaction of the three variables [*Jaeger et al.*, 2008] and which is one of the major sources of uncertainty in climate change projections [*Intergovernmental Panel on Climate Change* (*IPCC*), 2007]. The variables are validated using interannual time series of monthly means, averaged spatially for eight distinct climatic regions in Europe (PRUDENCE regions, see Figure 1). The error of these time series is measured using a performance index (*PI*),

$$PI = \left\langle \frac{\sqrt{(m-o)^2}}{(\sigma_o + \sigma_{iv} + \sigma_\epsilon)} \right\rangle, \tag{1}$$

which is a least squares estimation scaled by the interannual variability and including two sources of uncertainty, namely the internal variability and the observational uncertainty. The brackets in (1) denote the mean of monthly time series from 1994 to 1998 ($T = 60$ monthly averages), averaged over each PRUDENCE region ($R = 8$ regions), and for the three model variables (T2M, PR, CLCT, $V = 3$). *PI* is therefore the mean of $R \cdot T \cdot V = 1440$ least squares errors between the model (m) and the observations (o), scaled by the interannual variability (σ_o) expressed as the standard deviations of the observations (1990–2000), the observational uncertainty ($\sigma_{o\epsilon}$) derived from different reference data sets, and the internal variability (σ_{iv}) of the regional model derived from the initial condition ensemble. The uncertainty terms in the denominator of *PI* have the same dimensions as the spatiotemporal means. Further details about *PI*, including information about the choice of the observational data sets, can be found in B11.

[16] The error *PI* is consequently transformed into a positive defined performance score (*PS*), which is an approximation of the Gaussian likelihood:

$$PS = \exp\left(-0.5PI^2\right). \tag{2}$$

Table 1. Model Parameters Used to Calibrate the RCM[a]

Acronym	Parameter/Property	Value
rlam_heat	Scalar resistance for the latent and sensible heat fluxes in the laminar surface layer	[0.1, **1**, 10]
entr_sh	Entrainment rate for shallow convection	[$3 \cdot 10^{-5}$, $\mathbf{3 \cdot 10^{-4}}$, $3 \cdot 10^{-3}$]
qi0	Threshold for conversion of cloud ice to snow	[**0**, 10^{-4}]
uc1	Parameter controlling the vertical variation of critical relative humidity for sub-grid cloud formation	[0,**0.8**,1.6]
root_dp	Uniform factor for the root depth field	[0.5,**1**,1.5]

[a]The Value column shows the default values (in boldface) and a minimum and maximum bound for each parameter estimated in B11.

A comparison of *PS* to alternative widely used scores is discussed in B11.

2.3. Identifying Calibration Parameters

[17] In B11 a broad range of uncertain model parameters have been tested in order to estimate the associated parameter uncertainty of CCLM. Evaluation of parameters has included numerous discussions with scientists responsible for the parameterization schemes used, and a large number of model simulations. Using the definition of *PS*, the parameters which affect the CCLM performance strongest, within the range of plausible parameter bounds, were identified. One-yearlong simulations have proven sufficiently long to screen tunable model parameters regarding their relative importance in perturbing CCLM. Based on this knowledge, five model parameters have been selected for calibration, namely a factor for the laminar resistance of surface heat fluxes (*rlam_heat*), the entrainment rate for shallow convection (*entr_sc*), the autoconversion threshold for cloud ice (*qi*0), a parameter controlling the sub-grid cloud formation (*uc*1) and a uniform factor for the root depth field (*root_dp*). The parameters are summarized in Table 1 with their respective default values and related uncertainty ranges determined by expert elicitation. The sensitivity of *PS* with respect to the variation of the selected parameters is shown in Figure 2, where for each parameter (except for *qi*0) a minimum and a maximum value (black dots) have been tested and compared to the default value (red dot). The contributions to the performance score *PS* from each of the three variables (T2M, PR, CLCT) have been interpolated with a quadratic regression in order to judge on the sensitivity for each variable individually.

[18] A more detailed picture on the sensitivity of the model with respect to the selected parameters is provided in Figure 3. The figure shows the deviations of CCLM from the reference simulation when single parameters or pairs of model parameters are perturbed. Results are presented for three model variables in JJA and DJF, spatially averaged for the 8 PRUDENCE regions (horizontal axes). Compared to the bias of the reference simulation, shown in the first row of each panel, the deviations of the parameter experiments are spatially uniform in most cases. Seasonal variability of the deviations is achieved by some of the experiments as shown for the perturbation of *rlam_heat* and the response of the 2 m temperature in JJA compared to DJF. The precipitation of many experiments shows little sensitivity, indicating limited potential to overcome the current biases of precipitation. For all three model variables considered in the figure, the biases of the reference simulation show higher spatial and seasonal variability than the patters of individual parameter experiments. This suggests that several of the selected parameters need to be varied to achieve an overall improvement of the model.

2.4. Constructing a Metamodel

[19] The fact that RCMs are computationally very expensive inhibits the application of calibration algorithms. One way to overcome this problem is to construct a computationally cheap surrogate model, hereafter termed metamodel (MM), which estimates a model quantity of interest for a given parameter input vector μ of the climate model. A computationally cheap metamodel allows thereafter to apply any optimization algorithm to estimate an optimal parameter setting.

2.4.1. Quadratic Metamodel

[20] In this study we choose to emulate the output of the regional climate model using a parametric regression model presented by *Neelin et al.* [2010]. The metamodel bases on the assumption that changes of a climate model quantity due to a parameter perturbation are smooth and thus can be approximated by a 2nd order polynomial regression. As a quadratic fit is determined by only three points, this assumption allows to fit the MM with a low number of simulations, which is crucial for computationally expensive climate models. The MM is simple and transparent in contrast to more complex emulators as e.g. neural networks [*Knutti et al.*, 2003; *Hauser et al.*, 2012] where the underlying emulator structure often remains obscure. The

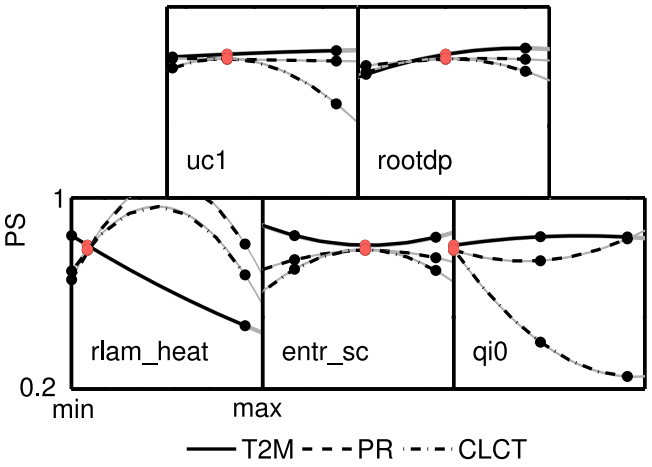

Figure 2. Sensitivity of model performance *PS* with respect to five model parameters. *PS* is computed in this case for each model variable individually in order to avoid compensating effects of the perturbations. The three lines show quadratic regressions for the three model variables (temperature T2M, precipitation PR, cloud cover CLCT) based on three simulations. The red dot shows the reference simulation (REF) with standard parameter settings, while black dots denote additional simulations with perturbed parameters (except for qi0, these are conducted at the upper and lower bound of each parameter range).

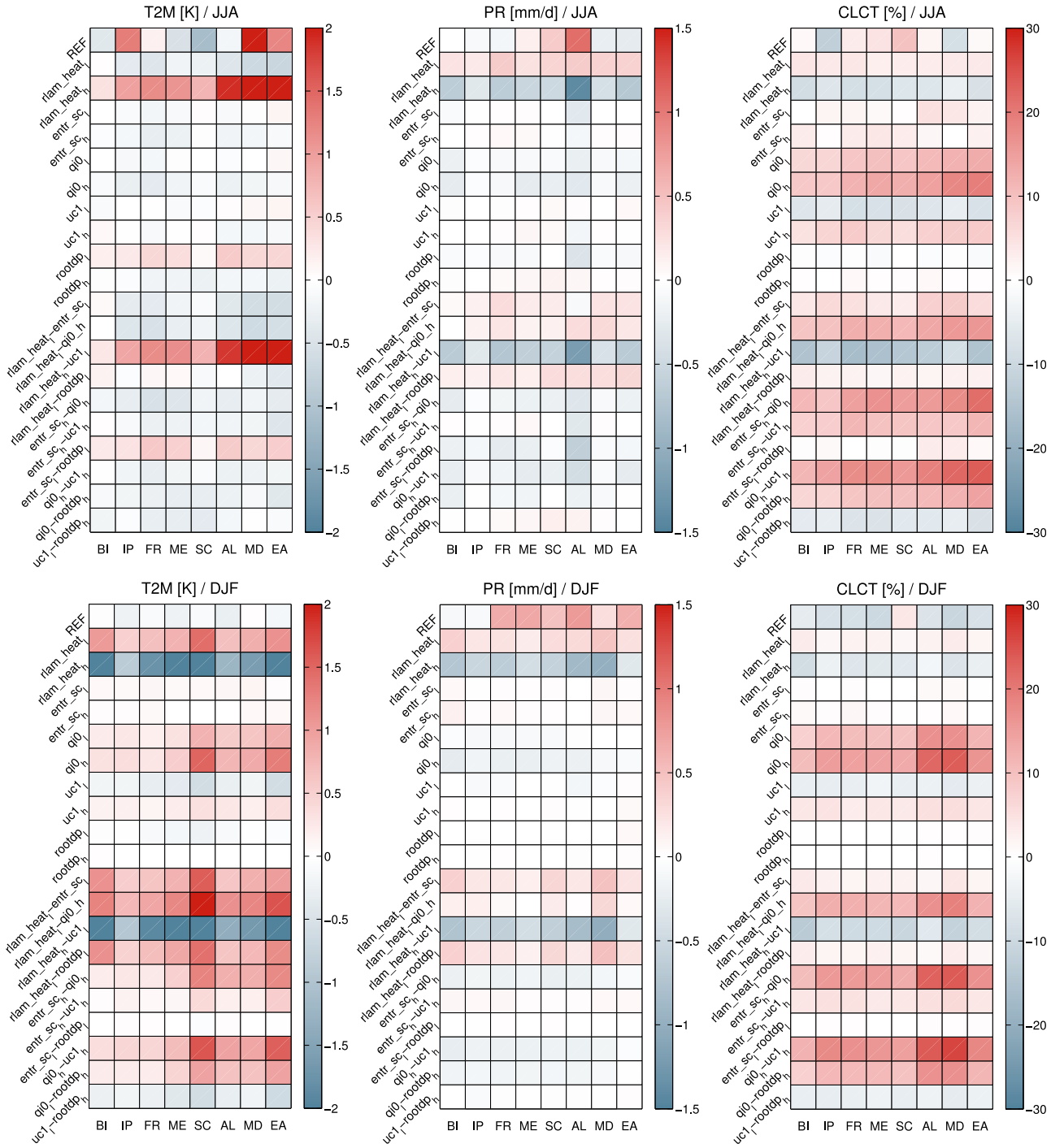

Figure 3. Sensitivity of reference simulation REF with respect to parameter perturbations for (top) JJA and (bottom) DJF. The vertical axes of each panel list the parameter perturbations. The effect of the perturbations is shown by color shading, for seasonal and regional means of T2M, PR, and CLCT averaged over the PRUDENCE regions on the horizontal axes. In the first row of each panel, the biases of the reference simulation are shown. The subsequent rows show the perturbations when using either a minimum or maximum value for single parameters, or when perturbing two parameters simultaneously. The subscripts of the experiment labels in the vertical axis denote whether the lower ("l") or higher ("h") bound of parameter has been chosen.

assumption of smoothness reduces the risk of overfitting and allows to derive analytical solutions for optimal model parameters when using simple cost functions. Interactions of parameter perturbations are approximated by a non-linear term for each parameter pair. Perturbations of more than two parameters are therefore approximated with non-linear terms of all possible pairs.

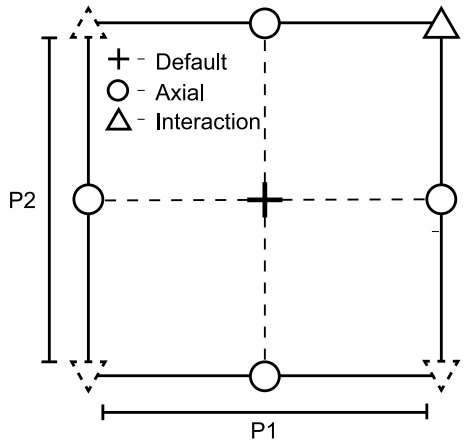

Figure 4. Schematic figure for one pair-wise parameter plane.

[21] As in *Neelin et al.* [2010] we use the following definition of a relative parameter input vector μ_* of the five parameters selected (Table 1),

$$\mu_* = \mu_p - \mu_{def}\ , \tag{3}$$

where μ_p contains the perturbed and μ_{def} the default parameter values. MM can consequently be expressed in the vector notation as,

$$\Phi^* = \Phi_{ref}\ + \mu_*^T a + \mu_*^T B \mu_*. \tag{4}$$

[22] Here the vector a contains the linear coefficients for each parameter, and the matrix B contains the quadratic and interaction terms in the diagonal and off-diagonal elements, respectively. It is further assumed without loss of generality that $B_{i,j} = B_{j,i}$.

[23] Φ^* and Φ_{ref} denote a model field of the perturbation and the reference simulation that is being used to derive the model performance, for instance Φ_{ref} can be the monthly 2 m temperature averaged over the Mediterranean in the year 1994 using default parameter settings. For each these model fields (1440 in total, see section 2.2) an independent metamodel is regressed, which consequently allow to compute the PS value for a given parameter input vector μ_*.

[24] Performing the vector operations yields an alternative notation of MM which illustrates the contribution of linear, quadratic and interaction terms, here shown for two parameters ($\mu_{*,1}$, $\mu_{*,2}$),

$$\Phi^* = \Phi_{ref}\ + \overbrace{\mu_{*,1}a_1 + \mu_{*,2}a_2}^{linear} + \overbrace{\mu_{*,1}^2 B_{1,1} + \mu_{*,2}^2 B_{2,2}}^{quadratic} + \overbrace{2\mu_{*,1}\mu_{*,2}B_{1,2}}^{interaction}. \tag{5}$$

[25] The minimum design points, also termed the saturated design, necessary to analytically estimate the parameters of MM (a and B) are illustrated for one pair-wise plane in Figure 4. In each pairwise plane the center point is given by the reference simulation (REF) with default parameter settings, complemented with four axial points (circular symbols in Figure 4) which correspond to simulations with minimum and maximum values of each parameter, keeping all other parameters at their default values. For each parameter μ_i the two axial points can be used to estimate the linear term a_i and the quadratic term $B_{i,j}$. In order to solve the interaction term $B_{i,j}$ at least one corner point (triangular symbols in Figure 4) of each pair-wise plane has to be simulated, where two parameters are varied at the same time. Depending on the number of parameters (N) the number of simulations needed for the saturated design is given by $2N + \frac{(N-1)N}{2}$, which corresponds to 20 simulations for five parameters.

[26] To estimate the interaction terms more accurately the remaining three corner points have also been simulated for this study as done in *Neelin et al.* [2010], which leads to additional 30 simulations. Since the linear system is consequently overdetermined, the interaction parameters are estimated using least squares with the corresponding four corner point simulations. Furthermore, one additional simulation has been performed to cover the relatively wide distance of the default value of *rlam_heat* to its maximum value compared to the distance to the minimum value. The additional simulation is used to constrain the linear and quadratic term of *rlam_heat*, again using least squares.

[27] The approach to predict the model quantities rather than estimating the model performance directly, which would require only one regression, is advantageous for several reasons and has therefore also been applied in *Neelin et al.* [2010]. First, as most performance metrics including *PS* are a measure of a squared model error, the quadratic transformation of the performance metric exerts higher non-linearities in the performance space which are therefore more difficult to regress. Second, predicting the model quantities allows a more versatile application as different performance metrics can be considered for a calibration and the understanding of parameter perturbations and their interactions is improved when remaining in a physical space.

[28] Once the metamodel MM has been fitted to the design points, the CCLM skill of any parameter combination can be estimated. In Figure 5 all pair-wise planes for the five parameters with contours of the performance score *PS* of the planes are shown. Strong non-linearities emerge as e.g. shown in the top right panel of *rlam_heat* and *qi*0. The colored circles show the performance of the simulated design points used to fit MM. The good agreement between all the circle colors and the contours show that MM is able to model the non-linearities of the performance hyper-plane, at least at the design points. The panels also show that parameter values with higher *PS* values than the reference, shown as black dashed contour lines, occur within the bounds of the parameter ranges.

2.4.2. Accuracy of Metamodel

[29] In order to quantify the precision of MM to emulate the output of the CCLM model, an additional independent ensemble of simulations has been performed. The ensemble consists of ten simulations with a space-filling Latin hypercube design [*McKay et al.*, 2000], in which the parameter ranges are sampled with the same number of intervals as the number of simulations performed (ten and of equal length for this study). The parameter combinations are thereafter randomly selected such that each combination uses a

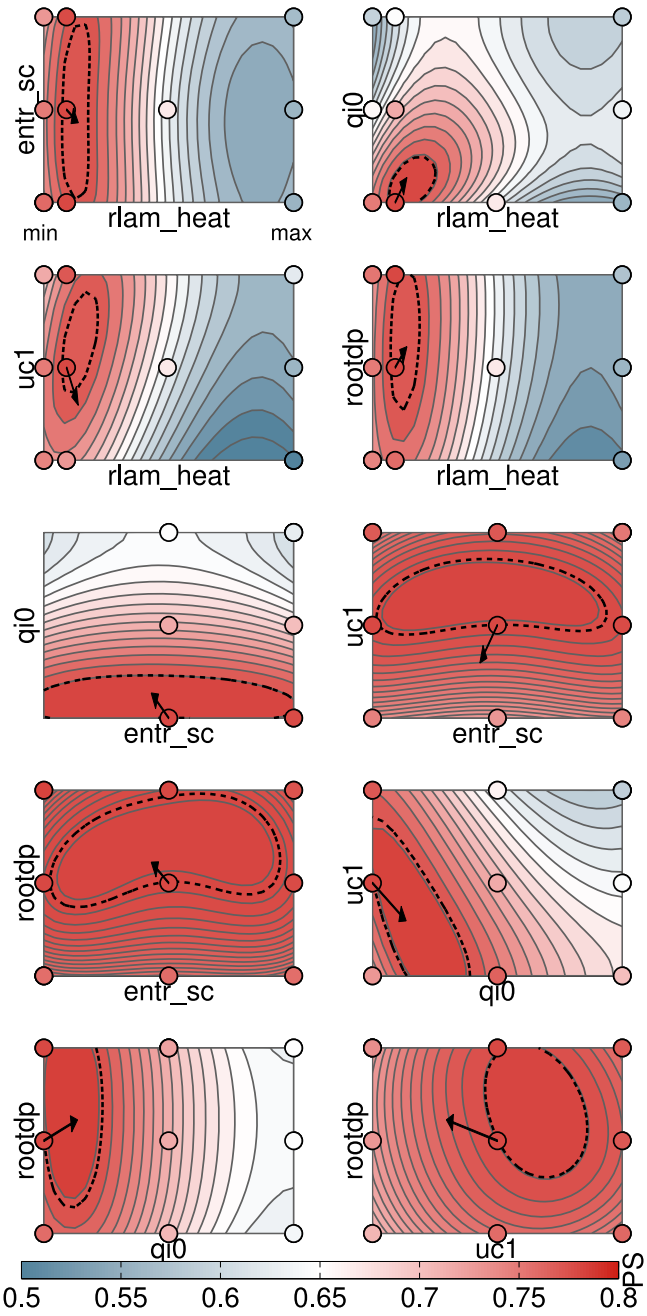

Figure 5. Pairwise planes of a five-dimensional parameter space. The contour lines show the model performance on the pair-wise planes estimated with the quadratic metamodel when keeping all other parameters constant. The colored circles show the simulated performance of the design points used to fit the metamodel. The good agreement of the contours and the circles highlight the capability of the metamodel to capture non-linearities induced by parameter perturbations. The black dashed contours indicates regions of improved model performance compared to the reference simulation. The arrows point from the default parameter values toward the optimal parameter values identified when all five parameters are varied.

different value for each parameter. Latin hypercubes prove to have optimal space-filling properties, which has in addition been optimized by maximizing the Euclidean distance of the parameter combinations for this study.

[30] In Figure 6 the precision of MM is illustrated by comparing the simulated time series of spatial averages of all independent ten simulations (4800 values for each model variable) with the predicted values from the metamodel. The dots lying on the black straight line show spatial averages which are accurately reproduced by the metamodel whereas the cloud of deviations from the line indicates the error of MM, with a gray band showing the 95% percentile range of the deviations. The fact that most of the predicted points lie in a very narrow band for all model variables illustrates the high precision of MM, which also supported by the R^2 statistic. The individual panels of each model variable in addition show that the simulated distributions of the three model variables are almost identically reproduced by the metamodel.

[31] To compare the imprecision of the metamodel with the uncertainties considered in the denominator of performance index (*PI*), we compute the standard deviations σ_{MM} of the errors when emulating the independent ensemble. It is important to consider at this point that the metamodel MM is fitted and compared to noisy simulations with a standard deviation of σ_{iv}, which leads to an overestimation of the imprecision when computing differences between these noisy signals [*Separovic et al.*, 2012]. Taking this overestimation into account (see Appendix A) σ_{MM} is on the median level 0.14 K for T2M, 0.07 mm/d for PR, and 0.72% for CLCT. In Figure 7 this additional uncertainty is compared to the individual terms of *PI*. The figure shows that the imprecision of MM is much smaller than the observational uncertainty and of about the same magnitude as the internal variability for PR and CLCT. In order to quantify how the imprecision of MM affects the estimation of *PI* and hence the performance score (*PS*), we perturb the time series of the reference simulation with the imprecision (σ_{MM}) using Gaussian white noise. The perturbed time series allow thereafter estimating a resulting standard deviation of *PS*, which amounts to about $0.005 \cdot PS$.

2.4.3. Parameter Interactions and Non-linearities

[32] The relative importance of linear and non-linear perturbations as well as the parameter interactions can be studied with the metamodel by computing the corresponding terms in (5) embraced by the brackets. The contribution of these terms to the estimated model field is shown in Figure 8 for each parameter, with linear terms ($\mu_{*,i}a(i)$) in the transformed vector a^*, non-linear terms ($\mu^2_{*,i}B_{i,i}$) in the diagonal and interaction terms ($2\mu_{*,i}\mu_{*,j}B_{i,j}$) in the off-diagonal of the transformed matrix B^*. The terms have been normalized by the observed interannual variability σ_o in order to average between all model variables. The figure shows that the strongest sensitivities of CCLM originate from the perturbations of the parameters *rlam_heat* and *qi*0, both having large values for the linear and non-linear contributions. Generally the linear and non-linear contributions are of a similar magnitude for all parameters, demonstrating the importance to include non-linear terms in the metamodel to estimate parameter perturbations. Strong non-linearities of parameter perturbations are found by many studies and have implications when computing ensemble means of perturbed physics ensembles or even multimodel ensembles as discussed in *Neelin et al.* [2010].

[33] To assess how well the quadratic assumption captures the non-linearity of the parameter space three additional independent simulations between the design points of the parameter axis of *rlam_heat* have been performed. The

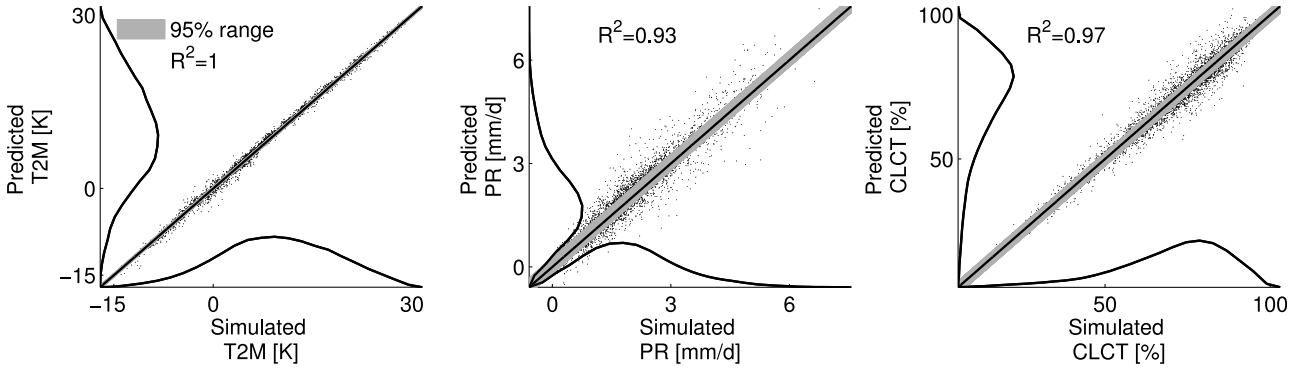

Figure 6. Performance of the metamodel to predict independent simulations. The three panels show for each model variable the spatial means of ten independent simulations over the 8 subregions, as simulated by the model and predicted by the meta model. The deviations of the black line with a slope of one show the errors of the metamodel in either predicting too low or too high values of the respective model variable. As a gray bar the 95% percentile of all points are shown, hence all points which lie outside of the gray band correspond to only 5% of all data points (4800 in total). The narrow width of the 95% level and the very high R^2 statistic for all variables illustrate the high accuracy of the metamodel. The additional two black curves show the projections of the density of the data points to the two axes.

parameter *rlam_heat* has been chosen for this purpose. since it is responsible for the strongest variability of the model and has a strong non-linear term as shown in Figure 8. The variations of 2 m temperature, precipitation and total cloud cover due to changes of *rlam_heat* are shown in Figure 9. The non-linearities induced by the parameter perturbations are reproduced well by the metamodel, supporting the assumption of a quadratic regression, in particular for precipitation and total cloud cover. The non-linearity of the 2 m temperature seems to be of a slightly higher order than modeled by the metamodel, which leads to a relative larger error compared to the other two variables. A more flexible regression might achieve higher accuracy in this case.

[34] For the remaining calibration parameters the non-linearity of the model behavior is not explicitly analyzed, yet the high accuracy of MM to reproduce independent simulations supports the assumption that the model behavior is well captured by a 2nd order polynomial. An initial screening of

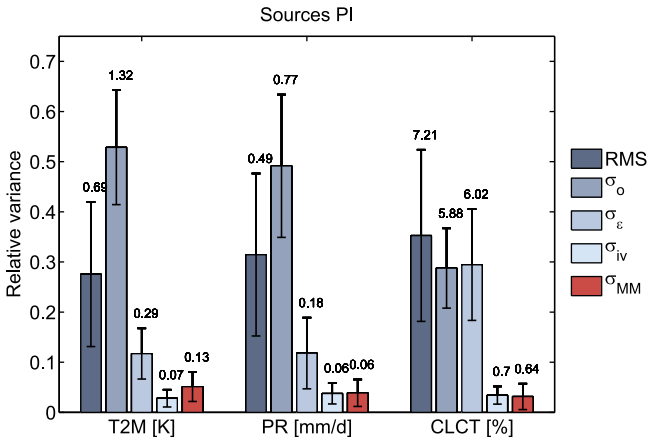

Figure 7. Comparison of the imprecision of the metamodel (σ_{MM}, red column) to all sources of the performance index (*PI*, blue columns) for T2M, PR and CLCT separately. The individual columns show the median values of all terms, whereas the error bars show the inter-quartile range derived from all spatial means considered to compute *PI*. To compare the terms for all three model variables all terms are scaled to sum up to 1 for each model variable. The original values are shown on top of each column with the dimension given by each model variable in the horizontal axis. The figure shows that the imprecision of the metamodel is small compared to the other sources of uncertainty in *PI*.

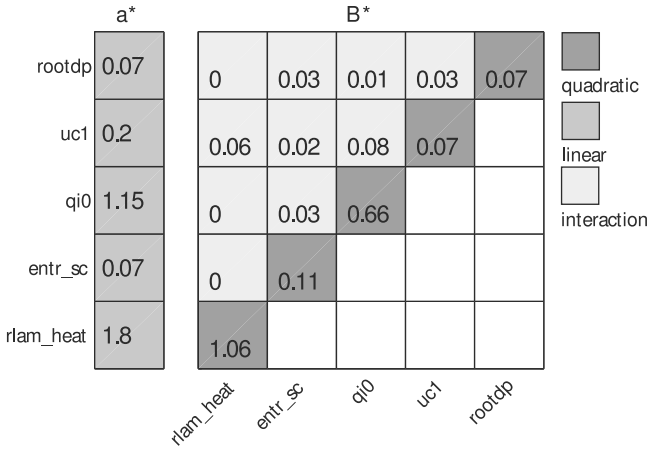

Figure 8. Linear, quadratic and interaction terms contributing to the estimation of CCLM with the metamodel when perturbing the calibration parameters as shown in equation (5). In the vector a^* (gray) the linear terms, in the diagonal of the matrix B^* (dark gray) the quadratic terms and in the off-diagonal the interaction terms (light gray) of the metamodel are shown. In order to average between the three model variables the terms are normalized by the interannual variability of each variable. To compute the contribution of each term the maximum bound of each parameter is used respectively. The dominant parameters are *qi*0 and *rlam_heat*. The interaction terms are relatively small compared to the linear and non-linear terms ranging at a similar magnitude.

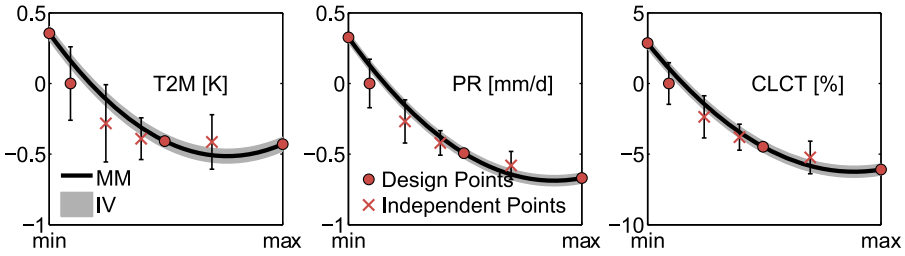

Figure 9. Response of T2M, PR and CLCT to changes in *rlam_heat* shown as black dots averaged for all spatial means. Red dots denote the design points used to fit the metamodel and red crosses show independent simulations. The black line shows the quadratic metamodel. The deviation from the simulated points is denoted with error bars expressing the average regression error. The internal variability of the model is shown as gray shade. The quadratic regression holds well for PR and CLCT but shows some deficiencies for T2M.

the smoothness of the RCM response to parameter perturbations by performing two additional simulations between the design points for each parameter axis would increase the confidence that a quadratic model captures well the induced perturbations. Such a screening would also support the selection of model parameters used to calibrate the model but would also add additional expenses to the tuning process.

[35] The interaction terms shown in the off-diagonal of matrix B in Figure 8 are overall relatively small for the set of parameters considered in this study. Highest interaction are obtained between the parameter affecting the sub-grid scale cloud formation *uc*1 and the threshold for ice auto-conversion *qi*0, which both strongly affect the total cloud cover. The weak interaction between the parameters may be a result of the fact that every parameter originates from a different model parametrization. Since the parameter interactions are weak one might consider to omit these terms as their estimation is relatively expensive in comparison to the estimation of the linear and quadratic terms. Omitting the interaction terms may therefore be reasonable in case of low computational resources and little indication of strong parameter interactions. In the case of the five parameters selected for this study, setting the interactions terms to zero increases the error when estimating the model fields on average by about 20% for T2M, 20% for PR and 100% for CLCT. This decrease of the accuracy of MM shows that at least part of the parameter interactions are well captured by MM and that the interaction terms are particularly important to model the cloud cover fields for the parameters considered. The most important interaction is between qi0 and uc1, which both affect cloud cover CLCT.

3. Model Calibration Results

[36] Having established a computationally cheap surrogate for the RCM, we proceed with the calibration and present the respective results in comparison to previous versions of CCLM. We choose to sample the parameter space with a Latin hypercube as done for the independent ensemble (see section 2.4.2) and as also done in *Gregoire et al.* [2011], but using a much larger number of one million parameter

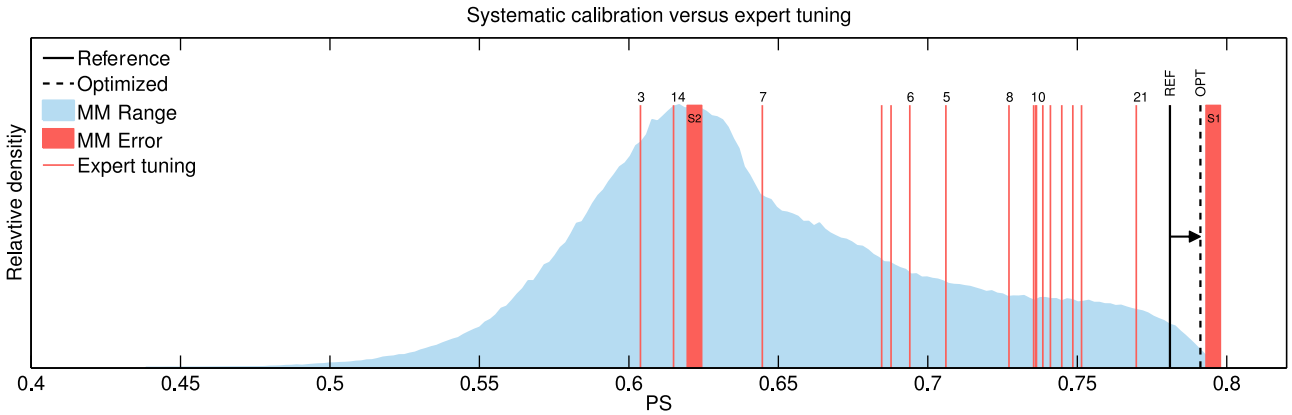

Figure 10. Calibration range estimated with the quadratic metamodel (MM) when computing one million parameter combinations from a Latin hypercube experiment. The blue area shows the empirical probability density of the performance assessed by the metamodel. The solid black line corresponds to the reference simulation (REF), which at the same time is the optimal simulation resulted from the expert tuning. The black dashed line shows the optimized simulation (OPT), where the black arrow shows the improvement achieved which corresponds to a reduction of the model error of about 7%. The two red bands show the spread of two sub-samples (S1,S2) with a range of $0.005 \cdot PS$ corresponding to the estimated uncertainty (1σ) of MM. The red lines denote the performance of the simulations of the expert-tuned ensemble LONG.

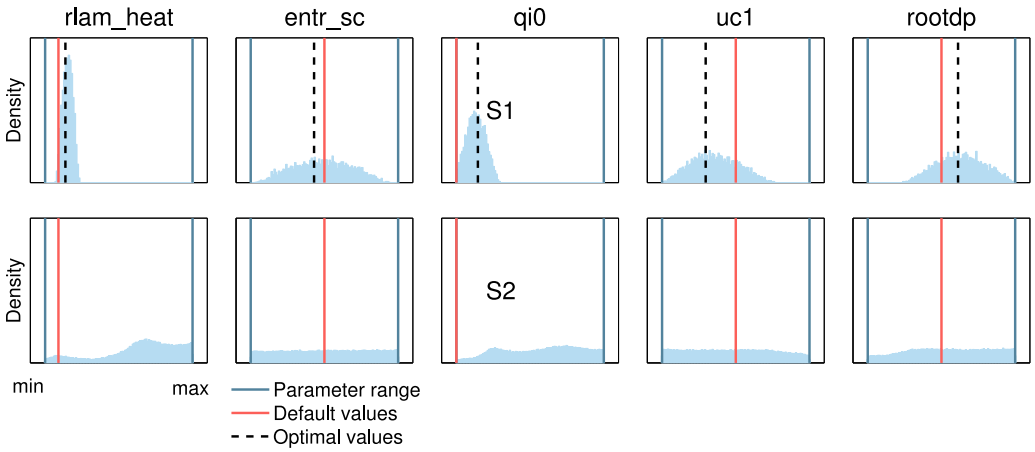

Figure 11. Empirical densities of the parameter values of the sub-samples S1 and S2 drawn from one million parameter combinations shown in Figure 10. The individual parameter combinations within each sample perform equally well given the uncertainty of the metamodel in predicting the model performance. S1 contains the best performing parameter combinations and the sub-sample S2 is drawn at level of the highest density of the one million parameter combinations. The red line in each panel shows the default parameter value and the black line shows the parameter combination of the best performing simulation (OPT).

combinations. Out of this large sample, optimal parameter configurations can be identified which improve the model performance with respect to a performance metric. The distribution of the model performance (*PS*) of the one million parameter combinations is shown in Figure 10 as a blue area spanning a wide range of *PS* values between 0.45 and 0.8. The parameter combination leading to the highest model performance is shown in Figure 4 as arrows pointing to the optimal values with respect to the default values. An experiment with the respective parameter combination has then been carried out with CCLM for the entire validation period from 1990 to 2000 model and referred to as OPT.

3.1. Constrained Model Parameters

[37] Determining a single parameter combination as a calibration result is ambiguous as many other and in principle substantially different parameter combinations may perform almost equally well. Since the metamodel is attached with an uncertainty in predicting the model performance, the determined optimal parameter setting is only a sample out of a distribution of parameter combinations with indistinguishable model performance. It is hence more meaningful to compare parameter distributions rather than single values in order to show how the observations constrain the model parameters. We determine for this purpose two sub-samples shown in Figure 10, one at the 99.9% percentile of the distribution (S1) and another sub-sample at level of the highest density of the million parameter combinations (S2). Both samples cover a range equal to the uncertainty of the metamodel (*PS*).

[38] The empirical parameter densities of the sub-sample containing the best performing parameter combinations are shown in Figure 11 (top). The figure shows that given the observations considered in $0.005 \cdot PS$ the resistance for the surface fluxes (*rlam_heat*) and the threshold for the ice autoconversion (*qi*0) are constrained toward slightly higher values compared to the default values shown as a red line. The calibration further yields a reduction of the sub-grid cloud formation parameter (*uc*1) by about 30% and an overall increase of the root depths (*rootdp*) by about 30%. The highest densities of the parameter values for the entrainment for shallow convection (*entr_sc*) lie at the default level, yet the observations impose a weak constraint on this parameter. Figure 11 (bottom) shows the sub-sample drawn at the level where most of the parameter combinations occur (cf. Figure 9) and where the parameter distributions are therefore almost uniform. The differences between these two sub-samples illustrate the effects of imposing an observational constraint on the parameter distributions, similar to a Bayesian inference applied in many calibration procedures [*Villagran et al.*, 2008; *Jarvinen et al.*, 2010; *Annan and Hargreaves*, 2007]. Consequently the observationally constrained sub-sample allows drawing substantially different parameter combinations for perturbed physics studies, which are equally appropriate and agree well with the observations.

3.2. Objective Calibration Versus Expert Tuning

[39] The calibration procedure yields a reduction of the model error of about 7% in *PI* with respect to the reference simulation. Thus, due to the immanent uncertainty of the MM (see Appendix A) the model performance (PS) of the optimal simulation is slightly lower than expected from the MM alone. The improvement of the model performance is achieved for the 5 years used for the calibration (1994–1998) and also for the remaining 5 independent years of the period 1990–2000 (omitting the year 1990 for spin-up reasons) for which both OPT and REF have been simulated with identical initial conditions. The only moderate improvement shows that the reference has previously already been expert tuned, as it is common in practice. In order to compare the expert tuning process against an objective calibration as presented for the current study, we use an additional ensemble termed LONG, which is described in detail in B11. The simulations of ensemble LONG were performed after a new model version release in order to test different physical parameter choices and to find an optimal setting for the CORDEX simulations based on subjective choices. Ensemble LONG is hence a typical *ensemble of opportunity* commonly performed during model

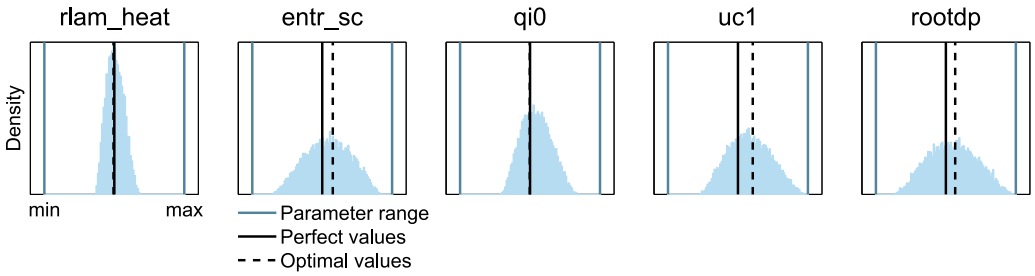

Figure 12. Calibration results of a perfect model experiment. The observations in the performance score are for this purpose substituted with a CCLM simulation with known parameter configurations. For this independent simulation parameter values at the center of the parameter ranges are chosen, shown as black solid lines for each model parameter. The histograms in blue show the empirical densities of the parameter configurations which reproduce most accurately the perfect model simulation, and the black dashed line shows the optimal parameter configuration as derived with the calibration procedure.

development and therefore subject to substantial expert tuning. The reference simulation used for the present study corresponds to the optimal setting determined by expert tuning. In Figure 10 the red lines show the individual simulations of ensemble LONG labeled with an internal numbering system. Interestingly, the reference simulation is also the best simulation of ensemble LONG in terms of the quasi-objective skill metric (*PS*), showing some coherence between subjective evaluation of model performance and the applied skill metric. Overall, the range sampled with MM is much wider than the range covered by the ensemble LONG demonstrating the higher potential of an objective calibration, which at the same time is also much more efficient and transparent. Re-calibration as it is commonly done after e.g. resolution changes, implementation of new processes and spatial transfer of the geographical domain of an RCM, would therefore be greatly facilitated using a framework as the one applied in this study. Exploring the full performance range additionally allows to intercompare different model versions more objectively, since otherwise the intercomparison will implicitly be biased by differences in expert tuning devoted to the two model versions.

3.3. Perfect Model Approach

[40] As an additional proof of the concept, a perfect model approach is performed. Here the observational fields in *PS* are substituted by a CCLM simulation with known parameter configurations. The sigma terms in the denominator of *PS* are for this purpose kept unchanged such that the same weighting of the different model variables is used. As in *Hauser et al.* [2012] we substitute the observations with an independent simulation in which the parameter values are set at the center value of all parameter uncertainties. In Figure 12 the calibration results are illustrated in a similar manner as in Figure 11 where the distributions of the parameter combinations which reproduce the perfect model simulations most accurately are shown. The optimal simulation shown as a black dashed line finds parameter values close to those chosen for the perfect model simulation, with some inaccuracies for the parameters for which the model is less sensitive (*entr_sc*, *uc*1 and *rootdp*).

4. Conclusion

[41] We have presented an application of an objective calibration for regional climate models using a quadratic metamodel presented in *Neelin et al.* [2010]. Five parameters which dominate important parametrized processes are calibrated using an objective framework presented in *Bellprat et al.* [2012] and compared to the expert tuning process common in climate modeling. The key steps of the calibration are: definition of a objective performance function, selection of important model parameters for calibration, construction of a cheap metamodel applicable to parameter variations in a control integration, and sampling the parameter space to identify optimal parameter configurations.

[42] The main conclusions we draw from this study are summarized in the following:

[43] 1. The applied framework for the objective calibration of regional climate models is feasible in terms of computational demands and effective as verified with a perfect model approach. To optimize five selected model parameters typically 20–50 simulations are needed, which corresponds to a total of 100–250 model years.

[44] 2. Quadratic regressions capture the non-linearities induced by parameter perturbations rather accurately and thus allow the formulation and application of an efficient metamodel.

[45] 3. The calibration allowed to reduce the model error of an expert tuned model by about 7%. For the model under consideration, the calibration yields slight changes in two parameters (resistance used for the surface fluxes, and threshold for ice auto-conversion) and significant changes in three parameters (relating to roots depths, sub-grid cloud formation, and entrainment in shallow convection).

[46] 4. The comparison of the objective calibration expert tuning favors the use of the objective method, due to the high efficiency, wider calibration range and transparency of the method.

[47] 5. Providing observed lateral boundary conditions allows to calibrate RCMs for short time periods (in our case 5 years), which is more efficient and potentially more robust than the calibration of GCMs (which would require a much longer integration time).

[48] 6. Intercomparison of different models as well as the implementations of new model parameterization or the spatial transfer of regional climate models for coming "global" regional climate change efforts could be facilitated by applying systematic methods for re-calibration.

[49] The calibration of physical parameterizations using RCM simulations driven by re-analysis data at the lateral boundaries has an important advantage compared to the

calibration of a parameterization package in global climate simulations: Constraining the large scale flow in the interior of the RCM domain reduces the risk of error compensations between resolved and parameterized processes. The objective calibration of parameterizations could hence be a promising strategy to test and improve physical packages of GCMs, in particular for unified model systems, where global and regional model components share the same model physics. We are aware that more flexible metamodels such as Gaussian process models would probably increase the accuracy of the metamodel as there is an indication that parameter perturbations are not fully described with a second order polynomial. However, as the accuracy of the metamodel satisfies the needs of our calibration we think it is reasonable to apply this simpler approach, which is easy to handle and to reproduce. The presented methodology therefore could potentially reach a wider application targeting the problem of parameter estimation for computationally expensive climate models.

Appendix A: Imprecision of the Metamodel

[50] The metamodel is fitted to model output with a noise exhibited by the internal variability with a standard deviation (σ_{iv}) derived from an initial condition ensemble. The uncertainty of the metamodel therefore consists of the imprecision of the metamodel (σ_{MM}) and the internal variability (σ_{iv}) of the data toward it is fitted. To estimate the imprecision of the metamodel the difference between ten independent model realizations and the predicted model fields of the metamodel is computed. Assuming zero mean Gaussian statistics the variance of the differences computed (σ^2_{MM-ind}) equals the sum of the variances of the independent realizations (σ_{iv}) and the metamodel ($\sigma_{iv} + \sigma_{MM}$), hence

$$\sigma^2_{MM-lh} = \sigma^2_{iv} + (\sigma_{iv} + \sigma_{MM})^2. \tag{A1}$$

Solving for σ_{MM} yields,

$$\sigma^2_{MM} + 2\sigma_{iv}\sigma_{MM} + 2\sigma^2_{iv} - \sigma^2_{MM-lh} = 0, \tag{A2}$$

which has the positive solution at,

$$\sigma_{MM} = -\sigma_{iv} + \sqrt{\sigma^2_{MM-lh} - \sigma^2_{iv}}. \tag{A3}$$

The Gaussian assumption is justified by the fact that spatial as well as temporal averages are considered.

[51] **Acknowledgments.** We acknowledge the E-OBS data set from the ENSEMBLES project and the data providers in the ECA&D project (http://eca.knmi.nl).The Center for Climate Systems Modeling (C2SM) at ETH Zurich is acknowledged for providing technical and scientific support. All simulations have been conducted at the Swiss Center for Scientific Computing (CSCS, Manno). Finally we would like to thank David Neelin for the help to apply the method.

References

Annan, J. D., and J. C. Hargreaves (2007), Efficient estimation and ensemble generation in climate modelling, *Philos. Trans. R. Soc. A*, *365*(1857), 2077–2088.

Annan, J. D., D. J. Lunt, J. C. Hargreaves, and P. J. Valdes (2005), Parameter estimation in an atmospheric GCM using the Ensemble Kalman Filter, *Nonlinear Proc. Geophys.*, *12*(3), 363–371.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2012), Exploring perturbed physics ensembles in a regional climate model, *J. Clim.*, *25*(13), 4582–4599.

Beltran, C., N. R. Edwards, A. B. Haurie, J. P. Vial, and D. S. Zachary (2006), Oracle-based optimization applied to climate model calibration, *Environ. Model. Assess.*, *11*(1), 31–43.

Beven, K. (1989), Changing ideas in hydrology—The case of physically-based models, *J. Hydrol.*, *105*(1–2), 157–172.

Beven, K. (2002), Towards a coherent philosophy for modelling the environment, *Proc. R. Soc. A*, *458*(2026), 2465–2484.

Christensen, J., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen (2010), Assigning relative weights to regional climate models: Exploring the concept, *Clim. Res.*, *44*, 179–194.

Förstner, J., and G. Doms (2004), Runge-Kutta time integration and high-order spatial discretization of advection—A new dynamical core for the LMK, *COSMO Newsl.*, *4*, 168–176.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Gregoire, L. J., P. J. Valdes, A. J. Payne, and R. Kahana (2011), Optimal tuning of a GCM using modern and glacial constraints, *Clim. Dyn.*, *37*(3–4), 705–719.

Hauser, T., A. Keats, and L. Tarasov (2012), Artificial neural network assisted Bayesian calibration of climate models, *Clim. Dyn.*, *39*, 137–154.

Houldcroft, C. J., W. M. F. Grey, M. Barnsley, C. M. Taylor, S. O. Los, and P. R. J. North (2009), New vegetation albedo parameters and global fields of soil background albedo derived from modis for use in a climate model, *J. Hydrometeorol.*, *10*(1), 183–198.

Huber, M. B. (2011), The Earth's energy balance and its changes: Implications for past and future temperature change, PhD thesis, ETH Zurich, Zurich, Switzerland.

Intergovernmental Panel on Climate Change (IPCC) (2007), Summary for policymakers, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 1–18, Cambridge Univ. Press, Cambridge, U. K.

Jackson, C., M. K. Sen, and P. L. Stoffa (2004), An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions, *J. Clim.*, *17*(14), 2828–2841.

Jaeger, E. B., I. Anders, D. Luethi, B. Rockel, C. Schaer, and S. I. Seneviratne (2008), Analysis of ERA40-driven CLM simulations for Europe, *Meteorol. Z.*, *17*(4), 349–367.

Jarvinen, H., P. Raisanen, M. Laine, J. Tamminen, A. Ilin, E. Oja, A. Solonen, and H. Haario (2010), Estimation of ECHAM5 climate model closure parameters with adaptive MCMC, *Atmos. Chem. Phys.*, *10*(20), 9993–10,002.

Jones, C., J. Gregory, R. Thorpe, P. Cox, J. Murphy, D. Sexton, and P. Valdes (2005), Systematic optimisation and climate simulation of FAMOUS, a fast version of HadCM3, *Clim. Dyn.*, *25*(2–3), 189–204.

Kinne, S., et al. (2006), An AeroCom initial assessment: Optical properties in aerosol component modules of global models, *Atmos. Chem. Phys.*, *6*(7), 1815–1834.

Klemp, J. B., and R. B. Wilhelmson (1978), Simulation of 3-dimensional convective storm dynamics, *J. Atmos. Sci.*, *35*(6), 1070–1096.

Klocke, D., R. Pincus, and J. Quaas (2011), On constraining estimates of climate sensitivity with present-day observations through model weighting, *J. Clim.*, *24*(23), 6092–6099.

Knutti, R., and G. C. Hegerl (2008), The equilibrium sensitivity of the Earth's temperature to radiation changes, *Nat. Geosci.*, *1*(11), 735–743.

Knutti, R., T. F. Stocker, F. Joos, and G. K. Plattner (2002), Constraints on radiative forcing and future climate change from observations and climate model ensembles, *Nature*, *416*(6882), 719–723.

Knutti, R., T. F. Stocker, F. Joos, and G. K. Plattner (2003), Probabilistic climate change projections using neural networks, *Clim. Dyn.*, *21*(3–4), 257–272.

Kotlarski, S., T. Bosshard, D. Lüthi, P. Pall, and C. Schär (2012), Elevation gradients of European climate change in the regional climate model COSMO-CLM, *Clim. Change*, *112*(2), 189–215.

McKay, M. D., R. J. Beckman, and W. J. Conover (2000), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, *42*(1), 55–61.

Medvigy, D., R. L. Walko, M. J. Otte, and R. Avissar (2010), The Ocean-Land-Atmosphere-Model: Optimization and evaluation of simulated radiative fluxes and precipitation, *Mon. Weather Rev.*, *138*(5), 1923–1939.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and M. Collins (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(7001), 768–772.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb (2007), A methodology for probabilistic predictions of

regional climate change from perturbed physics ensembles, *Philos. Trans. R. Soc. A*, *365*(1857), 1993–2028.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson (2010), Considerations for parameter optimization and sensitivity in climate models, *Proc. Natl. Acad. Sci. U. S. A.*, *107*(50), 21,349–21,354.

O'Hagan, A. (2006), Bayesian analysis of computer code outputs: A tutorial, *Reliab. Eng. Syst. Safety*, *91*(10–11), 1290–1300.

Oreskes, N., K. Shraderfrechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical-models in the Earth-sciences, *Science*, *263*(5147), 641–646.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney (2007), Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *J. Clim.*, *20*(17), 4356–4376.

Price, A. R., R. J. Myerscough, I. I. Voutchkov, R. Marsh, and S. J. Cox (2009), Multi-objective optimization of GENIE Earth system models, *Philos. Trans. R. Soc. A*, *367*(1898), 2623–2633.

Randall, D. A., and B. A. Wielicki (1997), Measurements, models, and hypotheses in the atmospheric sciences, *Bull. Am. Meteorol. Soc.*, *78*(3), 399–406.

Refsgaard, J. C., and H. J. Henriksen (2004), Modelling guidelines—Terminology and guiding principles, *Adv. Water Resour.*, *27*(1), 71–82.

Roesch, A., E. B. Jaeger, D. Lüthi, and S. I. Seneviratne (2008), Analysis of CCLM model biases in relation to intra-ensemble model variability, *Meteorol. Z.*, *17*, 369–382.

Rougier, J., D. M. H. Sexton, J. M. Murphy, and D. Stainforth (2009), Analyzing the climate sensitivity of the HadSM3 Climate Model using ensembles from different but related experiments, *J. Clim.*, *22*(13), 3540–3557.

Schlemmer, L., C. Hohenegger, J. Schmidli, C. S. Bretherton, and C. Schär (2011), An idealized cloud-resolving framework for the study of midlatitude diurnal convection over land, *J. Atmos. Sci.*, *68*(5), 1041–1057.

Separovic, L., R. de Elía, and R. Laprise (2012), Impact of spectral nudging and domain size in studies of rcm response to parameter modification, *Clim. Dyn.*, *38*(7), 1325–1343.

Stainforth, D. A., et al. (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, *433*(7024), 403–406.

Steppeler, J., G. Doms, U. Schättler, H. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric (2003), Meso-gamma scale forecasts using the nonhydrostatic model LM, *Meteorol. Atmos. Phys.*, *82*, 75–96.

Suklitsch, M., A. Gobiet, H. Truhetz, N. Awan, H. Göttel, and D. Jacob (2010), Error characteristics of high resolution regional climate models over the alpine area, *Clim. Dyn.*, *37*, 377–390.

Tanré, D., J. Geleyn, and J. Slingo (1984), First results of the introduction of an advanced aerosol-radiation interaction in ECMWF low resolution global model, in *Aerosols and Their Climatic Effects*, edited by H. Gerber and A. Deepak, pp. 133–177, A. Deepak, Hampton, Va.

Tiedtke, M. (1989), A comprehensive mass flux scheme for cumulus parameterization in large-scale models, *Mon. Weather Rev.*, *117*(8), 1779–1800.

Vidale, P. L., D. Lüthi, C. Frei, S. I. Seneviratne, and C. Schär (2003), Predictability and uncertainty in a regional climate model, *J. Geophys. Res.*, *108*(D18), 4586, doi:10.1029/2002JD002810.

Villagran, A., G. Huerta, C. S. Jackson, and M. K. Sen (2008), Computational methods for parameter estimation in climate models, *Bayesian Anal.*, *3*(4), 823–850.

Wicker, L., and W. Skamarock (2002), Time-splitting methods for elastic models using forward time schemes, *Mon. Weather Rev.*, *130*, 2088–2097.

Zubler, E. M., D. Folini, U. Lohmann, D. Lüthi, A. Muhlbauer, S. Pousse-Nottelmann, C. Schär, and M. Wild (2011), Implementation and evaluation of aerosol and cloud microphysics in a regional climate model, *J. Geophys. Res.*, *116*, D02211, doi:10.1029/2010JD014572.