Earth Syst. Dynam., 11, 139–159, 2020
https://doi.org/10.5194/esd-11-139-2020

Earth System Dynamics
Open Access

Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land

Lea Beusch, Lukas Gudmundsson, and Sonia I. Seneviratne

Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

Correspondence: Lea Beusch (lea.beusch@env.ethz.ch)

Received: 27 June 2019 – Discussion started: 2 July 2019
Revised: 8 November 2019 – Accepted: 6 January 2020 – Published: 17 February 2020

Abstract. Earth system models (ESMs) are invaluable tools to study the climate system's response to specific greenhouse gas emission pathways. Large single-model initial-condition and multi-model ensembles are used to investigate the range of possible responses and serve as input to climate impact and integrated assessment models. Thereby, climate signal uncertainty is propagated along the uncertainty chain and its effect on interactions between humans and the Earth system can be quantified. However, generating both single-model initial-condition and multi-model ensembles is computationally expensive. In this study, we assess the feasibility of geographically explicit climate model emulation, i.e., of statistically producing large ensembles of land temperature field time series that closely resemble ESM runs at a negligible computational cost. For this purpose, we develop a modular emulation framework which consists of (i) a global mean temperature module, (ii) a local temperature response module, and (iii) a local residual temperature variability module. Based on this framework, MESMER, a Modular Earth System Model Emulator with spatially Resolved output, is built. We first show that to successfully mimic single-model initial-condition ensembles of yearly temperature from 1870 to 2100 on grid-point to regional scales with MESMER, it is sufficient to train on a single ESM run, but separate emulators need to be calibrated for individual ESMs given fundamental inter-model differences. We then emulate 40 climate models of the Coupled Model Intercomparison Project Phase 5 (CMIP5) to create a "superensemble", i.e., a large ensemble which closely resembles a multi-model initial-condition ensemble. The thereby emerging ESM-specific emulator parameters provide essential insights on inter-model differences across a broad range of scales and characterize core properties of each ESM. Our results highlight that, for temperature at the spatiotemporal scales considered here, it is likely more advantageous to invest computational resources into generating multi-model ensembles rather than large single-model initial-condition ensembles. Such multi-model ensembles can be extended to superensembles with emulators like the one presented here.

1 Introduction

The range of simulated climate responses to external radiative forcing is affected by both internal variability and inter-model differences (Hawkins and Sutton, 2009; Deser et al., 2012; Taylor et al., 2012). While inter-model uncertainty is typically accounted for by considering simulations from several climate models (Meehl et al., 2007; Taylor et al., 2012; Eyring et al., 2016), uncertainty due to internal climate variability is often quantified by running the same climate model a number of times with slightly different initial conditions (Deser et al., 2012; Fischer et al., 2013; Kay et al., 2015; Leduc et al., 2019).

As climate model ensembles are inherently expensive to run, there is an interest in approximating Earth system model (ESM) output with computationally cheap emulators. In the field of climate science, the term emulator is used for a variety of statistical models which learn from existing runs of complex climate models to infer properties of runs which have not been generated yet. This makes it possible to explore the phase space at a lower computational cost. ESM emulators target different aspects of the climate system. For example, some emulators focus on the impacts of sub-grid-scale parameterizations (Rougier et al., 2009; Williamson et al., 2013). Others target the effect of greenhouse gas emission scenarios on global mean temperature (Meinshausen et al., 2011; Goodwin, 2016) or on regional mean climate fields (Santer et al., 1990; Tebaldi and Arblaster, 2014; Tebaldi and Knutti, 2018). There are also emulators for regional-scale internal climate variability (Castruccio and Genton, 2016; Alexeeff et al., 2018; Link et al., 2019). Recently, the first attempts have been made to emulate the full dynamics of simple general circulation models (Scher, 2018; Scher and Messori, 2019).

In this study, the term emulator is used to refer to a computationally cheap statistical tool which generates additional realizations of land temperature field time series for a specific greenhouse gas emission pathway at a yearly resolution. The presented emulator thus produces realizations which closely resemble initial-condition ensemble members of the considered ESMs. In the context of large multi-model ensembles, our computationally cheap emulator can be used to produce look-alikes of large initial-condition ensembles for every model within the multi-model ensemble, resulting in a "superensemble", i.e., a large ensemble which closely resembles a multi-model initial-condition ensemble.

To build this statistical temperature emulator, an overarching modular framework is proposed and put into the context of previous work in Sect. 2. The employed data and terminology are described in Sect. 3, and the specific implementation of the framework is introduced in Sect. 4. To visualize the characteristics and capabilities of the emulator, detailed results are shown for four example ESMs in Sect. 5, before applying the emulator to the large CMIP5 (Coupled Model Intercomparison Project Phase 5; Taylor et al., 2012) multi-model ensemble containing 40 climate models in Sect. 6. In Sect. 7, the results are discussed, and finally, in Sect. 8, the conclusions and an outlook are provided.

2 A framework for end-to-end climate model emulation

We propose an additive framework for temperature emulation at the yearly scale for a specific greenhouse gas emission pathway, which can be summarized as

$$T_{s,t} = f\left(T_t^{\mathrm{glob}}\right) + \eta_{s,t}, \tag{1}$$

where the local temperature $T_{s,t}$ at grid point s and time t is defined as a response to the global mean temperature T_t^{glob}, indicated by the function $f()$, and a stochastic local residual temperature variability term $\eta_{s,t}$. Contributions from physical feedbacks other than the ones captured within the global mean temperature signal are thus neglected. The assumption of an underlying additivity is in line with frequently employed approaches in uncertainty analysis in climate science (Hawkins and Sutton, 2009) and in climate change detection and attribution studies (Allen and Stott, 2003).

Our framework requires three modules: a global mean temperature module, a module for the grid-point-level temperature response to the global mean temperature, and a local residual temperature variability module. In the following, we place existing literature within these modules before discussing the connections to our emulator. As this study is primarily concerned with temperature, we focus solely on this variable in our literature review. However, several of the referenced studies also treat additional variables such as precipitation (e.g., Tebaldi and Arblaster, 2014; Seneviratne et al., 2016; Wartenburger et al., 2017) or cloud cover (e.g., Osborn et al., 2016).

2.1 Global mean temperature module

Global mean temperature is often an output of computationally efficient simple energy balance climate models (Meinshausen et al., 2011; Goodwin, 2016). While such models provide an estimate of the global mean temperature trend, they do not produce interannual global mean temperature variability. To obtain an ensemble of global mean temperature variability, statistical models which account for temporal autocorrelation can be used (Brown et al., 2015).

2.2 Local temperature response module

Pattern scaling is a frequently employed approach to relate local temperature to global mean temperature and is also used to emulate warming patterns across emission scenarios (Santer et al., 1990; Mitchell, 2003; Tebaldi and Arblaster, 2014). It was originally introduced by Santer et al. (1990), and different implementations exist (Mitchell, 2003). Most often, temperature fields are averaged over a late 21st century multi-decadal time period and the associated average global mean temperature is obtained (Tebaldi and Arblaster, 2014). This pattern is then linearly interpolated to a desired global mean temperature. An alternative is to extract the pattern from a transient simulation at the time when the simulation reaches the desired global mean temperature (Herger et al., 2015; Seneviratne et al., 2016; King et al., 2017). Other approaches include carrying out a linear regression (Lynch et al., 2017) or fitting a linear mixed-effect model (Alexeeff et al., 2018) to global mean temperature at each grid point individually. The most important assumption underlying pattern scaling is that local mean temperatures are linearly related to global mean temperature and that this relationship is consistent across forcing scenarios. For surface temperature on land this assumption is satisfactorily met (Mitchell, 2003; Tebaldi and Arblaster, 2014; Seneviratne et al., 2016; Wartenburger et al., 2017; Osborn et al., 2018). However, for strong mitigation scenarios and under strong aerosol forcing, pattern scaling is less accurate (May, 2012; Levy et al., 2013). Additionally, it is assumed that external forcing and internal variability are independent, which may not always be true (Lopez et al., 2014).

More complex local response emulation methods are rare and often directly conditioned on CO_2 concentration profiles instead of global mean temperature (Castruccio et al., 2014; Holden and Edwards, 2010). For instance, it has been proposed to employ past trajectories of atmospheric CO_2 to model regional temperatures with an infinite distributed lag model to capture nonlinear behavior in spatial patterns for regional-scale emulation (Castruccio et al., 2014) and within global space–time models (Castruccio and Stein, 2013). Other authors use singular value decomposition to emulate decadal temperature fields across scenarios while accounting for complex spatiotemporal feedbacks (Holden and Edwards, 2010; Holden et al., 2014).

While the focus is usually on emulating the pattern associated with the global mean temperature trend, patterns associated with physical modes of variability such as the El Niño–Southern Oscillation and the Pacific Decadal Oscillation can additionally be derived (McKinnon and Deser, 2018).

2.3 Local residual temperature variability module

Several approaches exist to emulate local residual temperature variability based on observations and climate model simulations (Castruccio and Stein, 2013; Osborn et al., 2016; McKinnon et al., 2017; Alexeeff et al., 2018; Link et al., 2019). Observations can be employed to avoid climate model biases but are limited by rather short observational records when deriving the local temperature variability properties (Osborn et al., 2016; McKinnon et al., 2017; McKinnon and Deser, 2018). The simplest approach is to detrend observed temperature time series and obtain additional realizations by shifting the starting date of the time series (Osborn et al., 2016). More realizations have been generated by resampling the spatial fields of detrended observed local temperature variability in blocks of 2 years (McKinnon et al., 2017). The approach was later refined to explicitly account for physical modes of variability to further reduce temporal autocorrelation in the resampled fields (McKinnon and Deser, 2018).

When employing ESMs instead, longer time series and multiple realizations are available to derive the statistical properties of the local residual temperature variability. Several authors fit autoregressive (AR) models to a set of climate model runs to account for temporal autocorrelation when emulating local residual temperature variability (Castruccio and Stein, 2013; Castruccio et al., 2014; Castruccio and Genton, 2016; Bao et al., 2016). Thereby, the spatial dependence in the innovation terms of the AR models can be considered by parameterizing their covariance structure with a Matérn covariance function (Castruccio and Stein, 2013; Castruccio and Genton, 2016; Bao et al., 2016). Alternatively, detrended ESM runs can be decomposed into their principal components, and their phases can be randomly perturbed to generate additional realizations of local residual temperature variability (Link et al., 2019).

All approaches listed so far rely on the assumption that local residual temperature variability is stationary in time, which is known not to be fulfilled everywhere. Olonscheck and Notz (2017) and references therein provide a comprehensive overview of possible changes in temperature variability in the historical time period and the business-as-usual greenhouse gas emission scenario for the large CMIP5 multi-model ensemble. They find that the strongest and most likely changes will occur over oceans but also point out land regions where variability is projected to change in the future. During the historical time period, they identify only weak changes in the variability. To account for such temporal nonstationarities, it has been proposed to resample the detrended temperature fields of large single-model initial-condition ensembles within a certain window size around a global mean temperature level (Alexeeff et al., 2018). To enlarge the number of fields to sample from, a method has additionally been developed to stochastically emulate spatially nonstationary Gaussian fields with a LatticeKrig model (Nychka et al., 2018).

2.4 This study

While most studies focus on one or two of the modules required to mimic an initial-condition ensemble, this study proposes a framework which incorporates all three components. Since only 12 out of 40 CMIP5 models provide several initial-condition members, it is essential to test to what extent an emulator trained on a single run is able to approximate both its training run and additional independent initial-condition members. We thus emulate the full CMIP5 multi-model ensemble based on single training runs and create a superensemble which accounts for inter-model uncertainty across all 40 climate models. To the best of our knowledge, this study is the first to implement an emulator which mimics an initial-condition ensemble based on a single training run and applies it to such a large multi-model ensemble.

3 Data and terminology

3.1 Data sources and terminology

Runs from 40 CMIP5 climate models (Taylor et al., 2012) covering the historical time period (1870–2005) and the business-as-usual greenhouse gas emission scenario RCP8.5 (2006–2100; Riahi et al., 2011) are employed. To calibrate the emulator, a single run per climate model is used. This run is referred to as the training run (Table 1). For 12 out of 40 CMIP5 climate models more than one initial-condition member is available. These additional independent initial-condition ensemble members are referred to as test runs (Table 1). A special focus is on four ESMs with differing

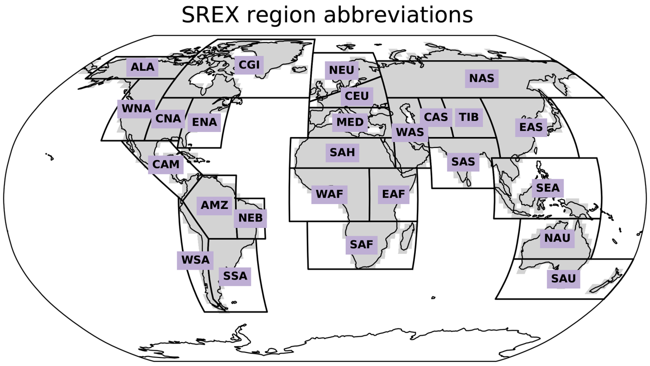

Figure 1. Map of the SREX regions and their abbreviations. The considered land grid points are shown in gray.

model genealogies (Knutti et al., 2013), namely CanESM2, CESM1(CAM5), HadGEM2-ES, and MPI-ESM-LR. All climate models, the associated modeling groups, and the number of initial-condition members employed here are listed in Table A1.

Additionally, stratospheric aerosol optical depth is used as a proxy for volcanic activity during the historical time period. This aerosol dataset was originally described by Sato et al. (1993) and later updated to cover the considered time period.

3.2 Data processing

Here, we focus on surface temperature anomaly at a yearly resolution. Temperature fields were bilinearly interpolated onto a $2.5° \times 2.5°$ grid, resulting in 3043 land grid points for each climate model. Yearly mean temperatures were computed at each grid point, and the average over the reference period of 1870–1899 in the training run at the respective grid points was subtracted. In the text, for simplicity reasons, we use the term "temperature" when referring to "yearly surface temperature anomaly". For the stratospheric aerosol optical depth, the globally averaged yearly time series is employed.

Whenever regional averages are shown, area-weighted means are referred to. The regions employed in this study are 26 land regions defined in the Special Report on Managing the Risks of Extreme Events (SREX) (Seneviratne et al., 2012) as well as global mean and global land mean (Fig. 1). While global mean refers to the average across all grid points, global land mean refers to the average across all land grid points excluding Antarctica.

4 Methods

4.1 Framework implementation

4.1.1 General approach

We follow the framework introduced in Sect. 2 to emulate temperature fields at the yearly scale for a specific greenhouse gas emission pathway. The chosen implementation is called MESMER, which stands for Modular Earth System Model Emulator with spatially Resolved output, and is shown in Fig. 2. Detailed information for each individual module is provided in the following sections. In short, the global mean temperature T_t^{glob} is split into a trend and a variability term, both of which contribute linearly to the local temperature $T_{s,t}$. The residual local temperature variability $\eta_{s,t}$ is modeled as an AR(1) process with spatially correlated innovations.

To calibrate the emulator, a single run spanning 231 years (1870–2100) per model is used. For the calibration, the global mean temperature trajectory and the associated land temperature fields are required.

4.1.2 Global mean temperature module

In the global mean temperature module, additional realizations of global mean temperature time series T_t^{glob} are generated. For this purpose, T_t^{glob} is separated into a trend $T_t^{\text{glob,trend}}$ shared by all emulations and a variability term $T_t^{\text{glob,var}}$ which varies between individual emulations:

$$T_t^{\text{glob}} = T_t^{\text{glob,trend}} + T_t^{\text{glob,var}}. \tag{2}$$

In $T_t^{\text{glob,trend}}$, smooth forcing $T_t^{\text{glob,sm}}$ and abrupt changes induced by volcanic eruptions $T_t^{\text{glob,volc}}$ are accounted for in an additive way:

$$T_t^{\text{glob,trend}} = T_t^{\text{glob,sm}} + T_t^{\text{glob,volc}}. \tag{3}$$

First, $T_t^{\text{glob,sm}}$ is derived by locally weighted scatterplot smoothing (LOWESS) of T_t^{glob}.

In a next step, $T_t^{\text{glob,volc}}$ is approximated as the linear response of the residuals of the smooth trend, i.e., $T_t^{\text{glob}} - T_t^{\text{glob,sm}}$, to stratospheric aerosol optical depth AOD_t with regression coefficients λ_0 and λ_1:

$$T_t^{\text{glob,volc}} = \lambda_0 + \lambda_1 \cdot AOD_t. \tag{4}$$

The time series of global mean temperature variability $T_t^{\text{glob,var}} = T_t^{\text{glob}} - T_t^{\text{glob,trend}}$ is modeled as an AR process of order p with coefficients $\alpha_0, \ldots, \alpha_p$ such that

$$T_t^{\text{glob,var}} = \alpha_0 + \sum_{k=1}^{k=p} \alpha_k \cdot T_{t-k}^{\text{glob,var}} + \epsilon_t$$
$$\text{with} \quad \epsilon_t \sim \mathcal{N}(0, \sigma), \tag{5}$$

whereby ϵ_t is a white noise innovation term drawn from a Gaussian distribution with mean zero and standard deviation σ.

In this study, the LOWESS smoothing window length is 50 years with weights decaying with increasing distance according to a tricube weight function. The regression coefficients for the forced response to volcanic eruptions are obtained with the ordinary least squares (OLS) algorithm. The

Table 1. Terms used to refer to different climate model runs throughout this study.

Name	Description	Application	Figures
Training run	Climate model run (1870–2100) used to calibrate the emulator parameters	– emulator calibration – emulator evaluation in terms of fitting the training run	3, 7, 8, 9, 10, 11, 12
Test run	Independent initial-condition ensemble member (1870–2100) not used to calibrate the emulator parameters	– emulator evaluation in terms of mimicking a climate model initial-condition ensemble	4, 5, 6, 8, 9, 10, 11, 12

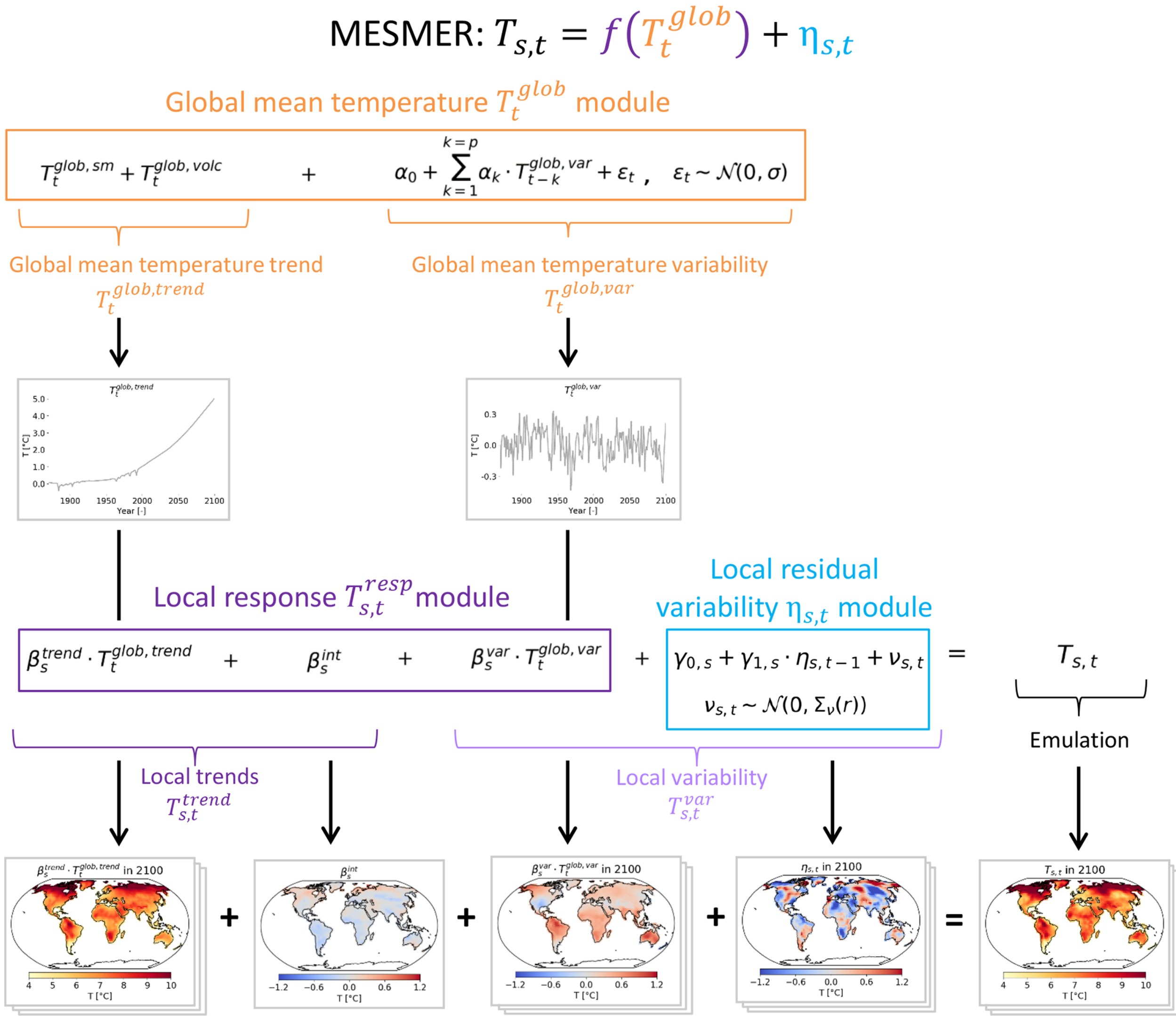

Figure 2. Illustration of the emulation framework with the MESMER implementation.

coefficients of the AR process are fit by means of maximum likelihood, and the Bayesian information criterion (BIC) is employed to select its order p with the maximum considered order being 8.

4.1.3 Local temperature response module

The local temperature response module translates the global mean temperature signal into a grid-point-level response $T_{s,t}^{\mathrm{resp}}$. Motivated by the pronounced linear scaling of regional land temperatures with global mean temperature (Seneviratne et al., 2016; Wartenburger et al., 2017), the local response is expressed as

$$T_{s,t}^{\text{resp}} = f\left(T_t^{\text{glob}}\right) = f\left(T_t^{\text{glob,trend}}, T_t^{\text{glob,var}}\right)$$
$$= \beta_s^{\text{trend}} \cdot T_t^{\text{glob,trend}} + \beta_s^{\text{int}} + \beta_s^{\text{var}} \cdot T_t^{\text{glob,var}}, \quad (6)$$

with regression coefficients β_s^{trend}, β_s^{int}, and β_s^{var}, whereby β_s^{int} represents the intercept term. Hence, the responses of the local mean temperature to $T_t^{\text{glob,trend}}$ and $T_t^{\text{glob,var}}$ are separately taken into account.

In this study, the linear regression coefficients are estimated with OLS at each grid point.

4.1.4 Local residual temperature variability module

The local residual temperature variability $\eta_{s,t}$ refers to the spatiotemporally correlated residual variability, which cannot be accounted for through a response to T_t^{glob}. This variability is assumed to be Gaussian in nature (see Fig. S1 in the Supplement for the results of a Shapiro–Wilk test for normality) and stationary in time, which makes it possible to model the time series as local AR(1) processes with spatially correlated innovations (Humphrey and Gudmundsson, 2019). Hence, additional realizations of $\eta_{s,t}$ are generated stochastically according to

$$\eta_{s,t} = \gamma_{0,s} + \gamma_{1,s} \cdot \eta_{s,t-1} + \nu_{s,t}$$
$$\text{with} \quad \nu_{s,t} \sim \mathcal{N}(0, \Sigma_\nu(r)), \quad (7)$$

whereby $\gamma_{0,s}$ and $\gamma_{1,s}$ are the coefficients of the AR model, and $\nu_{s,t}$ represents spatially correlated innovations drawn from a multivariate Gaussian with mean zero and covariance matrix $\Sigma_\nu(r)$ (Cressie and Wikle, 2011).

For an AR(1) process, $\Sigma_\nu(r)$ can be analytically derived from the covariance matrix of the residual variability $\Sigma_\eta(r)$ with

$$\Sigma_\nu(r)_{i,j} = \sqrt{1-\gamma_{1,i}} \cdot \sqrt{1-\gamma_{1,j}} \cdot \Sigma_\eta(r)_{i,j}, \quad (8)$$

whereby the indices i and j refer to spatial locations s (Cressie and Wikle, 2011).

To estimate $\Sigma_\eta(r)$, the empirical covariance matrix $\tilde{\Sigma}_\eta$ is computed. However, $\tilde{\Sigma}_\eta$ is rank deficient because substantially fewer temperature field samples are available than there are land grid points. Thus, $\tilde{\Sigma}_\eta$ needs to be regularized to obtain a robust estimate of the covariations between the grid points. For this purpose, we employ localization, an approach which is well established in the field of data assimilation (Carrassi et al., 2018). Localization retains anisotropy on regional scales, which is an important asset when stochastically modeling temperature variability since anisotropy is a prevalent feature due to physical factors such as the prevailing wind direction and geometry of mountainous terrain. To localize $\tilde{\Sigma}_\eta$, it is point-wise multiplied with a smooth correlation function $G(r)$ with exponentially vanishing correlations with distance:

$$\Sigma_\eta(r) = \tilde{\Sigma}_\eta \circ G(r), \quad (9)$$

whereby $\circ$ denotes the Hadamard product. Here, $G(r)$ is the numerically efficient Gaspari–Cohn function (Gaspari and Cohn, 1999), which vanishes beyond 2 times the localization radius L:

$$G(r) = \begin{cases} 1 - \frac{5}{3}\cdot r^2 + \frac{5}{8}\cdot r^3 + \frac{1}{2}\cdot r^4 - \frac{1}{4}\cdot r^5, \\ \quad \text{if } 0 \le r < 1, \\ 4 - 5\cdot r + \frac{5}{3}\cdot r^2 + \frac{5}{8}\cdot r^3 - \frac{1}{2}\cdot r^4 + \frac{1}{12}\cdot r^5 \\ \quad -\frac{2}{3}\cdot r^{-1}, \\ \quad \text{if } 1 \le r < 2, \\ 0, \\ \quad \text{if } r \ge 2, \end{cases} \quad (10)$$

with $r = \frac{d}{L}$ and d the geographical distance between two grid points.

In this study, the AR(1) coefficients are fit at each grid point by means of maximum likelihood. In our framework implementation, the obtained intercept terms $\gamma_{0,s}$ are effectively zero, as the local response module already contains an intercept term (Eq. 6). The localization radius to regularize $\tilde{\Sigma}_\eta$ is determined by cross-validation with a leave-one-out approach. Localization radii between 1000 and 4750 km every 250 km are tested. Thereby, the empirical covariance matrix is estimated based on 230 years, and the likelihood to draw the field of the left-out year from the regularized matrix is computed. This process is repeated until every year has been left out once for every localization radius. The respective log-likelihood values for each localization radius are summed up across the left-out years, and the radius which is associated with the maximum likelihood is chosen.

4.2 Evaluating the emulator

The emulator's performance is evaluated on the training run and – where available – on test runs. While the evaluation on the training run indicates how successfully this framework implementation captures the training run, the evaluation on the test runs serves as a proxy for the emulator's capability in mimicking true ESM initial-condition ensembles. For the evaluation, 1000 emulations are generated for each climate model.

4.2.1 Local trend verification

The local trends $T_{s,t}^{\text{trend}}$ are shared by all emulations and serve as an estimate of the externally forced response with

$$T_{s,t}^{\text{trend}} = \beta_s^{\text{trend}} \cdot T_t^{\text{glob,trend}} + \beta_s^{\text{int}}. \quad (11)$$

To evaluate how well the emulated local trends capture true climate model runs, the Pearson correlation of $T_{s,t}^{\text{trend}}$ with $T_{s,t}$ of the corresponding training run is computed. For climate models with test runs, the correlation coefficient is additionally computed between $T_{s,t}^{\text{trend}}$ and each test run.

4.2.2 Local variability verification

The local variability $T_{s,t}^{\text{var}}$ is different in each emulation and corresponds to the internally generated natural variability:

$$T_{s,t}^{\text{var}} = \beta_s^{\text{var}} \cdot T_t^{\text{glob,var}} + \eta_{s,t}. \tag{12}$$

To compare the emulated $T_{s,t}^{\text{var}}$ to true climate model runs, an estimate for the local variability within the climate models needs to be obtained. For this purpose, the emulated local trends $T_{s,t}^{\text{trend}}$ (Eq. 11) are subtracted from the climate models' $T_{s,t}$.

To evaluate $T_{s,t}^{\text{var}}$, on the grid-point level, lag-1 temporal autocorrelations and standard deviations are considered. Additionally, spatial cross-correlations between grid points are verified. These quantities are computed for each individual emulation as well as for all climate model runs. For each quantity, the Pearson correlation coefficient between each individual emulation and the training run is calculated. Additionally, the correlation between each individual test run and the respective training run is computed where test runs are available. These correlations between the climate model runs serve as benchmark values for the correlations between the emulations and the training run.

4.2.3 Regional-scale ensemble reliability verification

On regional scales, the emulated temperatures $T_{s,t}$ (Eq. 1) are evaluated visually and quantitatively in terms of ensemble reliability, i.e., the ability to capture the distribution of ESM runs with an ensemble of emulations (Weigel, 2012). For the visual verification, regionally averaged emulated time series are compared to climate model runs for global land, Central Europe (CEU), and Southeastern South America (SSA). In the quantitative verification, the emulator's ability to reliably reproduce a set of ESM quantiles (5 %, 50 %, 95 %) is evaluated in all 27 land regions. Smooth time series of the emulated quantiles are obtained based on the 1000 emulations, and the percentage of time slots during which a climate model run is below these emulated quantiles is counted. This is done for the training run and – where available – also for the test runs. Additionally, the counting is carried out for each individual emulation. The resulting deviations of the individual emulations from the emulated quantiles can be compared to the deviation the climate model runs exhibit from the emulated quantiles. If the climate model run deviation lies within the 95 % interval spanned by the individual emulation deviations, the climate model run is considered indistinguishable from individual emulations at this quantile.

5 Exploring emulator properties for four example ESMs

5.1 Calibration results

The parameters obtained from training the emulator on four example ESMs reveal distinct inter-ESM differences in every emulator module (Fig. 3). The global mean temperature trends diverge by 0.9 °C by the end of the 21st century. For each ESM, $T_t^{\text{glob,var}}$ is described by oscillating AR coefficients with the first lag being positive, but the AR process order and the standard deviations of the innovations vary.

In the local response module (Eq. 6), the strongest warming rates, i.e., the largest β_s^{trend} terms, are found in the northern high latitudes, but there are substantial differences in the β_s^{trend} patterns between emulators trained on different ESMs (Fig. 3). For example, the CESM1(CAM5) emulator exhibits less warming in the tropics than the others do. In all emulators, the intercept term β_s^{int} is generally small in magnitude and smooth in space. The β_s^{var} fields indicate that Alaska, the Amazon, and Australia frequently covary with global mean temperature variability. Only for HadGEM2-ES does central Asia emerge as a region of large β_s^{var} values.

The local residual variability (Eq. 7) exhibits generally less memory in the northern high latitudes than in the tropics as indicated by the lag-1 autocorrelation coefficients (Fig. 3). The innovations are largest in magnitude in high-latitude continental climates, such as northern Asia, and smallest in the tropics. However, for these quantities the patterns also differ between emulators calibrated on different ESMs. The localization radii chosen to regularize the empirical spatial covariance matrix $\tilde{\Sigma}_\eta$ range from 1750 to 2750 km.

5.2 Example realizations

Emulated temperature fields are visually indistinguishable from ESM test runs that were not used during training (Fig. 4). All fields exhibit the strongest warming and variability in the northern high latitudes. In terms of variability, CESM1(CAM5), HadGEM2-ES, and their emulations show more patchy behavior, i.e., locally more confined variability, than CanESM2 and MPI-ESM-LR.

Time series of emulations and ESM test runs averaged over global land, CEU, and SSA highlight the emulators' capability to reproduce the regionally characteristic behavior of the climate system (Fig. 5). These regions differ in terms of the warming trend and variability around this trend. The variability is smallest on the global scale since local anomalies tend to average out globally. In CEU, the warming rate and the variability are larger than in SSA.

5.3 Emulator transferability between ESMs

Figure 6 explicitly shows what the results of Sect. 5.1 and 5.2 have already hinted at, namely that an ensemble of emulations generated by an emulator calibrated on a specific ESM is capturing unique properties of that ESM, which in turn are not transferable to other ESMs. For example, the warming rate of the ensemble generated by the CESM1(CAM5) emulator is inconsistent with all three other ESMs on the global land scale. As expected, differences are also found in the variability around the trend, which is, e.g., visibly smaller

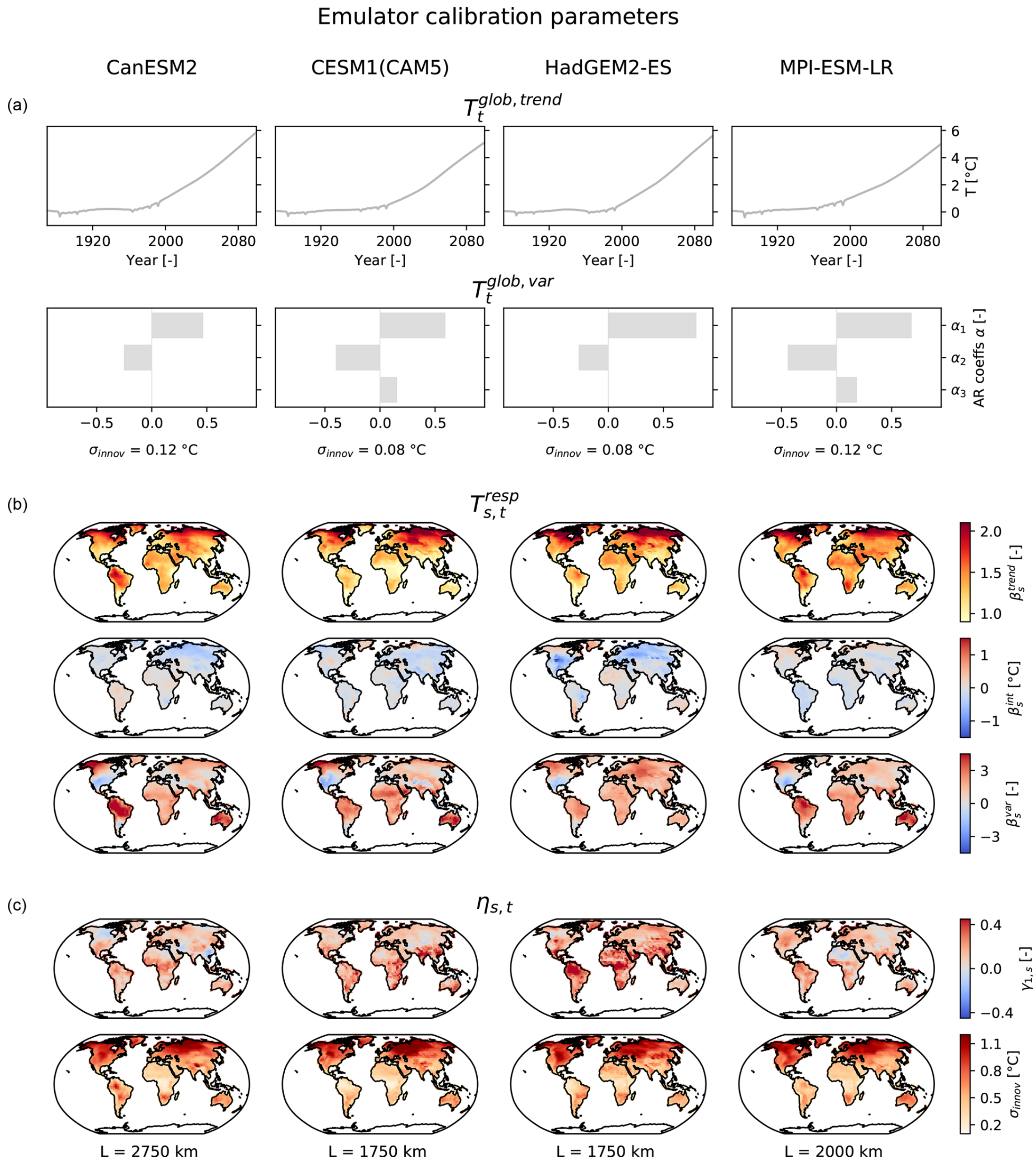

Figure 3. Emulator calibration parameters (rows) for four example ESMs (columns). **(a)** For the global mean temperature module, $T_t^{\mathrm{glob,trend}}$ and the AR coefficients plus the standard deviation of the innovations of $T_t^{\mathrm{glob,var}}$ are depicted. **(b)** For the local temperature response module, the regression coefficients are shown. **(c)** For the local residual temperature variability module, the lag-1 AR coefficients, the standard deviations of the innovations, and the localization radii are displayed.

in SSA in the CESM1(CAM5) emulations than in the runs of the other ESMs. The implications of these results are further discussed in Sect. 7.3.

6 Creating a CMIP5 superensemble

6.1 Calibration results

Figure 7 shows summary statistics of the calibrated parameters for each CMIP5 climate model, highlighting inter-model differences in each emulator module. In the Supplement, plots analogous to Fig. 3 are additionally provided for each climate model for readers interested in the geographical patterns of the emulator parameters (Figs. S2–S10).

In the global mean temperature module (Eq. 2), $T_t^{\mathrm{glob,trend}}$ ranges between 3.4 and 6.3 °C at the end of the 21st century (Fig. 7). For 45 % of the climate models, $T_t^{\mathrm{glob,var}}$ can be modeled as an AR(1) process. In the remaining ones either an AR(2) or AR(3) process is chosen. All emulators con-

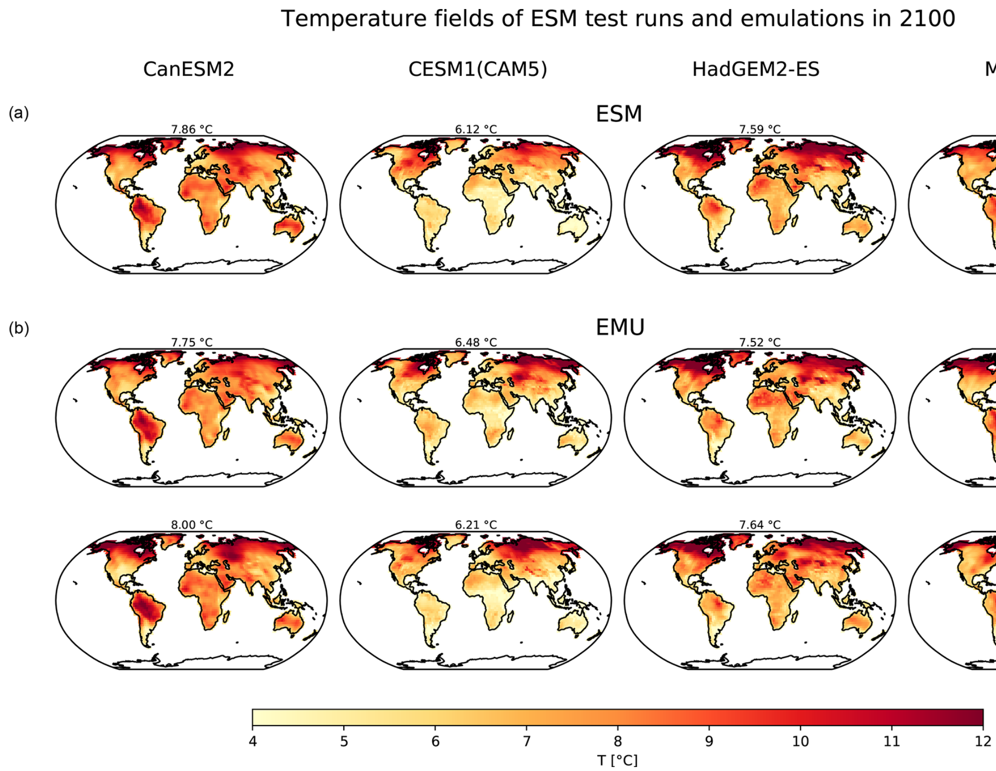

Figure 4. Temporal snapshots depicting temperature field realizations in 2100 (rows) for four example ESMs (columns). **(a)** One ESM field from a test run and **(b)** two emulations (EMUs) are shown. The temperature on top of each map refers to the global land mean.

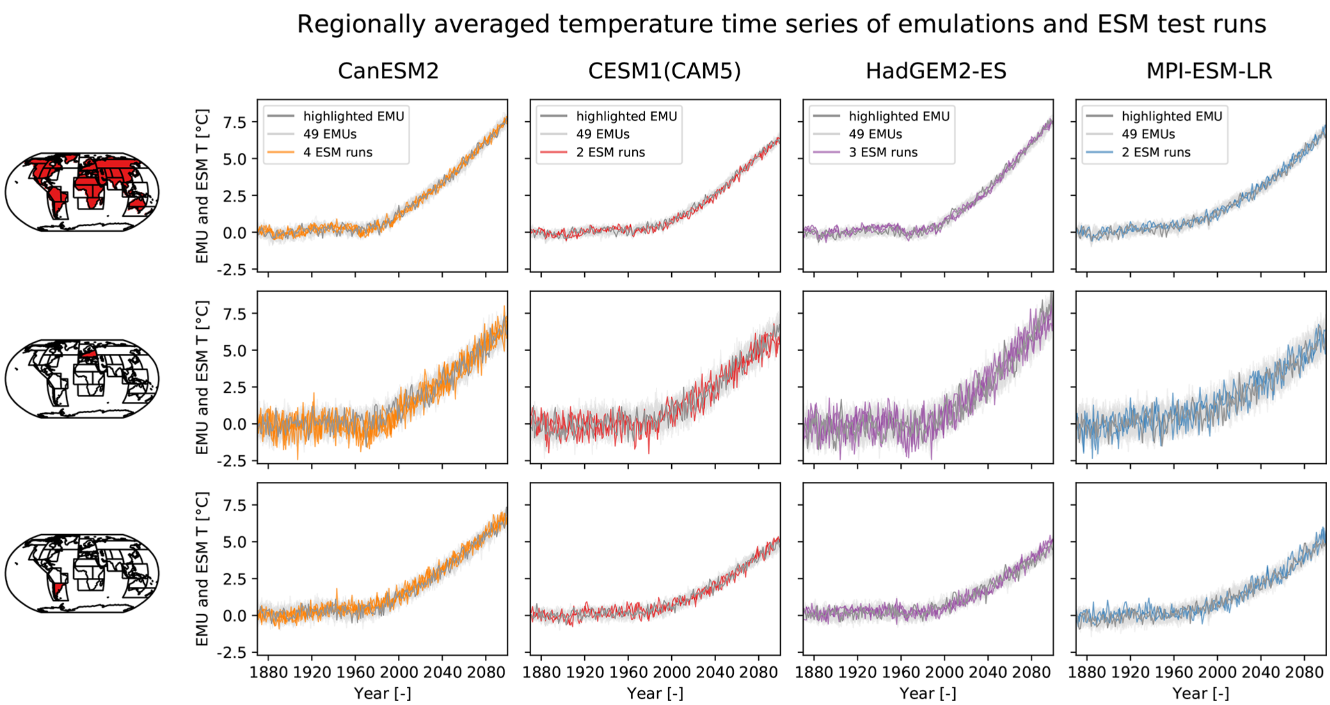

Figure 5. Regionally averaged temperature time series (rows) for four example ESMs (columns). The regions are (from top to bottom) global land, Central Europe (CEU), and Southeastern South America (SSA). In each panel, one emulation (EMU) is highlighted in dark gray and 49 other emulations are shown in light gray. Additionally, all available ESM test runs are plotted in color.

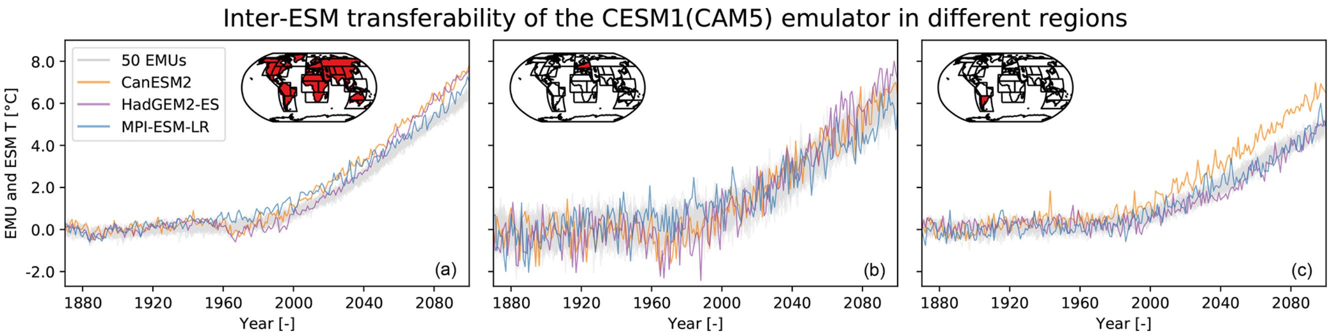

Figure 6. Time series of 50 emulations (EMUs) from the CESM1(CAM5) emulator (light gray) overlaid with runs from the three other example ESMs for three regions: global land **(a)**, CEU **(b)**, and SSA **(c)**.

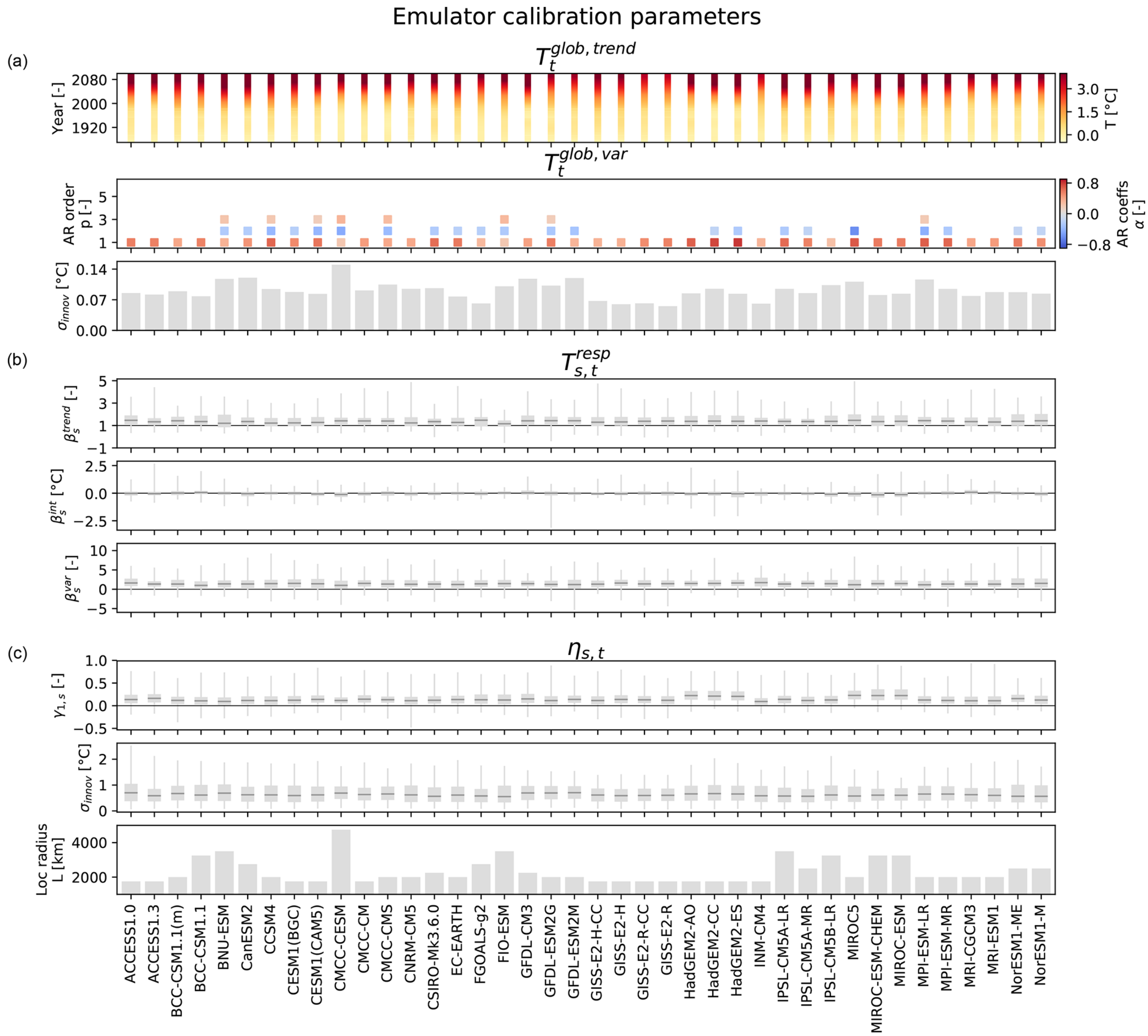

Figure 7. Emulator properties (rows) of the 40 CMIP5 climate models (columns). **(a)** For the global mean temperature module, $T_t^{\mathrm{glob,trend}}$ and the AR coefficients plus the standard deviation of the innovations of $T_t^{\mathrm{glob,var}}$ are depicted. **(b)** For the local temperature response module, the regression coefficients are shown. **(c)** For the local residual variability module, the lag-1 AR coefficients, the standard deviations of the innovations, and the localization radii are displayed. Box plots indicate the median (dark gray line), the interquartile range (gray box), and the full range of values (gray whiskers).

tain oscillating positive and negative AR coefficients with the first coefficient being positive, but they differ in the magnitude of the respective AR coefficients. The associated innovations vary in their standard deviations by a factor of almost 3 (0.06–0.15 °C).

In the local response module (Eq. 6), more than 80 % of the land grid points warm more quickly than the global mean, i.e., $\beta_s^{\mathrm{trend}} > 1$, in 25 out of 40 emulators (Fig. 7). Overall, the spread in the β_s^{trend} terms differs substantially between emulators trained on different climate models. The intercept terms β_s^{int} cluster closely around zero in each emulator. The fraction of outlier grid points deviating > 1 °C from zero, and hinting at suboptimal local fits, exceeds 1 % in only one of the emulators. The vast majority of land grid points are positively correlated with $T_t^{\mathrm{glob,var}}$, i.e., $\beta_s^{\mathrm{var}} > 0$, with the minimum fraction of positive correlations amounting to 82 % of the land grid points.

In the local residual variability module (Eq. 7), the year-to-year memory contribution $\gamma_{1,s}$ is overall generally small, with the 75 % quantile lying below 0.25 for 34 out of 40 emulators (Fig. 7). Only the six models of the HadGEM and the MIROC family tend to have systematically larger $\gamma_{1,s}$. While the median of the standard deviations of the innovations is similar in all calibrated emulators, the full ranges differ substantially, with the maximum between 1.3 and 2.5 °C. The selected localization radii vary between 1750 and 4750 km. Thereby, 4750 km is a strong outlier, with the second-highest localization radii amounting to 3500 km. Generally, climate models with a coarser native resolution are associated with larger localization radii (not shown).

6.2 Example realizations

Figure 8 demonstrates that the emulations nicely capture regional-scale trends and variability in the training and the test runs of the CMIP5 ensemble. The histograms also highlight that the larger sample size of the emulations by a factor of 1000 makes it possible to sample the temperature phase space better. The CMIP5 projections, and thus also the emulations, diverge substantially towards the end of the 21st century in global land and SSA but agree rather well in CEU. At the end of the 21st century, an inter-model spread of roughly 4 °C is observed in global land, with models spread out evenly across this space. In SSA, on the other hand, the bulk of the models clusters within a space of 2 °C and a few outlier models cause the overall CMIP5 spread of almost 6 °C.

6.3 Quantitative verification

6.3.1 Local trend verification

Correlation between the emulated local trends and the true climate model runs is very high in both training and test runs in all CMIP5 models, indicating that the forced trends are successfully extracted from each training run (Fig. 9). For the climate models with test runs, these correlations are nearly identical for each individual test and the training runs. The smallest correlation coefficient is 0.90, and the highest one is 0.97.

6.3.2 Local variability verification

To evaluate the local variability at the grid-point level, lag-1 temporal autocorrelations and standard deviations are considered (Fig. 10). The lag-1 temporal autocorrelation is a rather noisy parameter to estimate, and the median correlations between emulations and the training run lie between 0.67 and 0.92. Generally, the correlation of the lag-1 autocorrelations between test and training runs is smaller than the one between emulations and training runs, implying a tendency to overfit this parameter. The correlation between the standard deviations of the emulations and the training run is never below 0.98. The correlation between test and training runs is almost identical to the one between emulations and training runs. Thus, at the grid-point level the emulations reliably reproduce the stochastic variability of climate model runs.

To evaluate the spatial cross-correlations between grid points, three geographical bands are considered (Fig. 11). At all spatial scales, cross-correlations between test and training runs are higher than correlations between emulations and training runs. This is a direct consequence of the regularization, which dampens covariances between grid points as a function of distance and is thus inherent to the emulator's design. In a radius of up to 2000 km, the emulators perform best and covariations between grid points are generally well reproduced. The medians of the correlations between the emulations and the training runs span from 0.85 to 0.98. Plotting an individual example emulation against its associated training run clearly shows the dampening of the cross-correlations in the regularized emulations. Emulations of climate models with larger localization radii (Fig. 7) have by design a larger correlation with their respective training runs (Fig. 11). In a radius between 2000 and 15 000 km, the emulators perform the least well because cross-correlations for this radius are strongly dampened in the emulations, with the medians of the correlations between emulations and training runs ranging from 0.17 to 0.82. For long-range distances beyond 15 000 km, medians lie between 0.20 and 0.93. For all distances beyond 2000 km, there are large inter-model differences in the ability of the emulators to reproduce cross-correlations between grid points. Also, correlations between the spatial cross-correlations of test and training runs are generally lower and exhibit more inter-model differences at distances beyond 2000 km, highlighting that it is more difficult to estimate far-reaching spatial cross-correlation based on single ESM runs. Generally, the emulations perform better and are more comparable to test runs at distances beyond 15 000 km than between 2000 and 15 000 km, which is likely

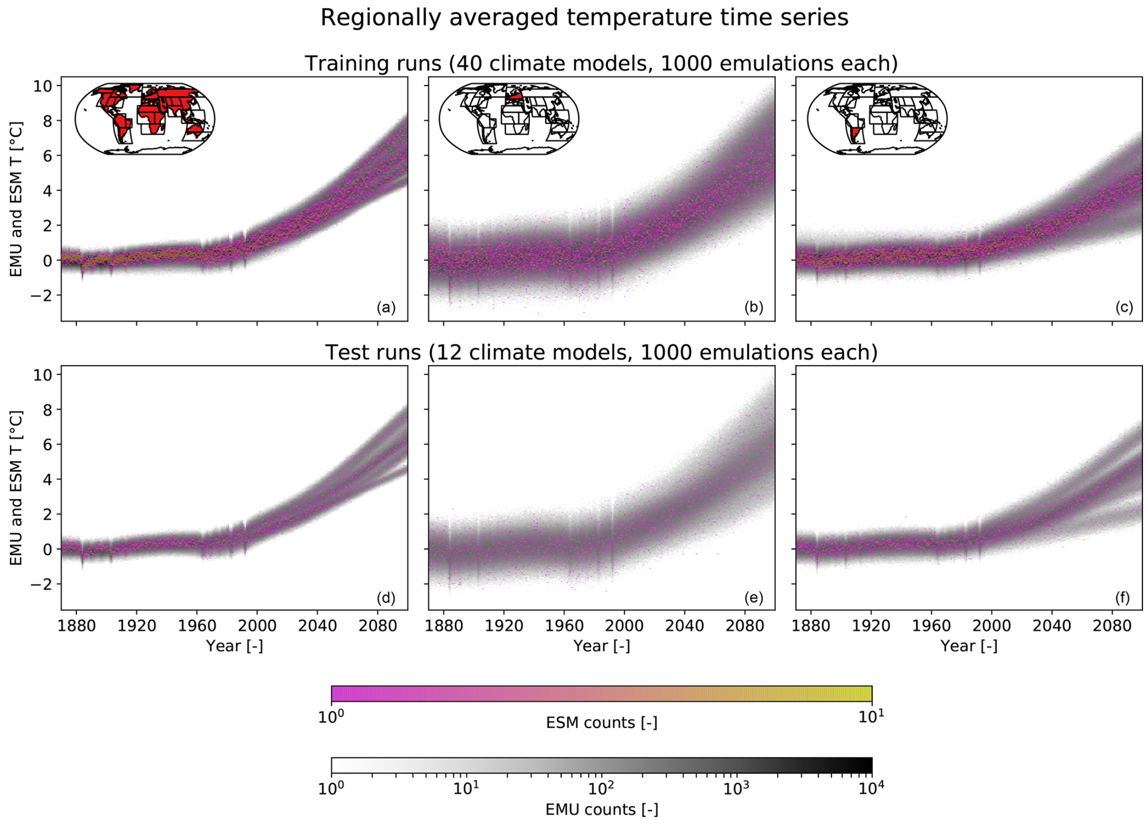

Figure 8. Regionally averaged time series as 2-D histograms for 40 CMIP5 model training runs and 1000 emulations per model **(a–c)** as well as for 12 CMIP5 models with one test run and 1000 emulations per model **(d–f)**. For the CMIP5 model runs a color map from pink to yellow is employed, and for the emulations a grayscale is used. The regions are global land **(a, d)**, CEU **(b, e)**, and SSA **(c, f)**.

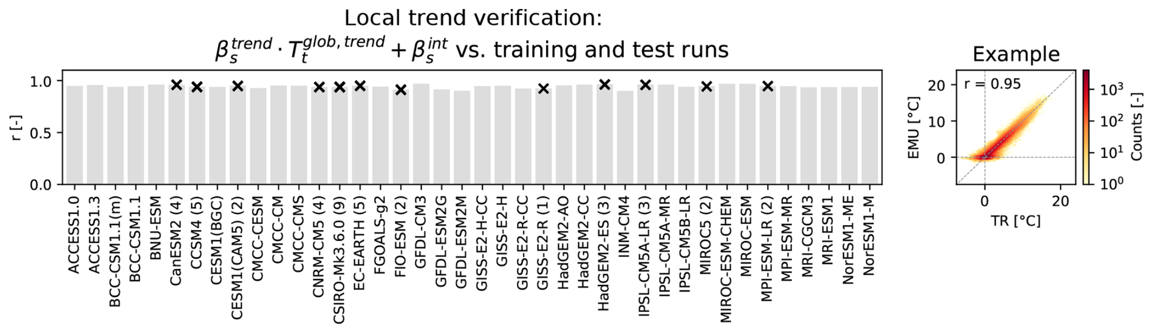

Figure 9. Local trend verification for the CMIP5 models by means of Pearson correlation between the emulated local trends and the training runs (gray bars). The example shows the associated 2-D histogram for CESM1(CAM5). For the CMIP5 models with test runs, the correlation between the emulated local trends and each individual test run is indicated by a black cross. Since these correlations are nearly identical for each test run of a specific climate model, the individual black crosses cannot be visually distinguished from one another. For all climate models with test runs, the number of available test runs is given in brackets after the model name.

due to that fact that the global correlation pattern induced by the global mean temperature variability serves as a more important driver for the longest-range correlations.

6.3.3 Regional-scale ensemble reliability verification

When considering full emulations, i.e., the local trends plus the local variability, the median is successfully emulated but the emulations are a bit underdispersive compared to the training run for the vast majority of CMIP5 models and SREX regions (Fig 12). The emulations tend to be more reli-

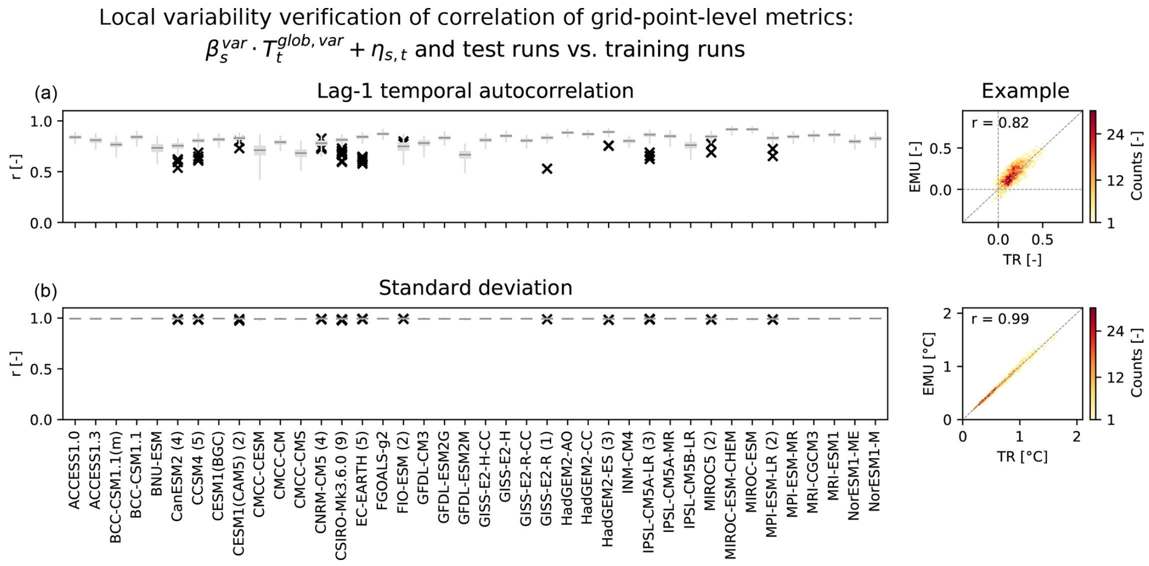

Figure 10. Local variability verification for the CMIP5 models (columns) by means of the Pearson correlation of grid-point-level lag-1 temporal autocorrelations **(a)** and standard deviations **(b)** between the 1000 individual emulations and the training runs (box plots). The examples show the associated 2-D histograms for a single emulation and the training run of CESM1(CAM5). For the CMIP5 models with test runs, the correlation between the quantity in the training run and in each individual test run is indicated by a black cross. For all climate models with test runs available, the number of test runs is given in brackets after the model name.

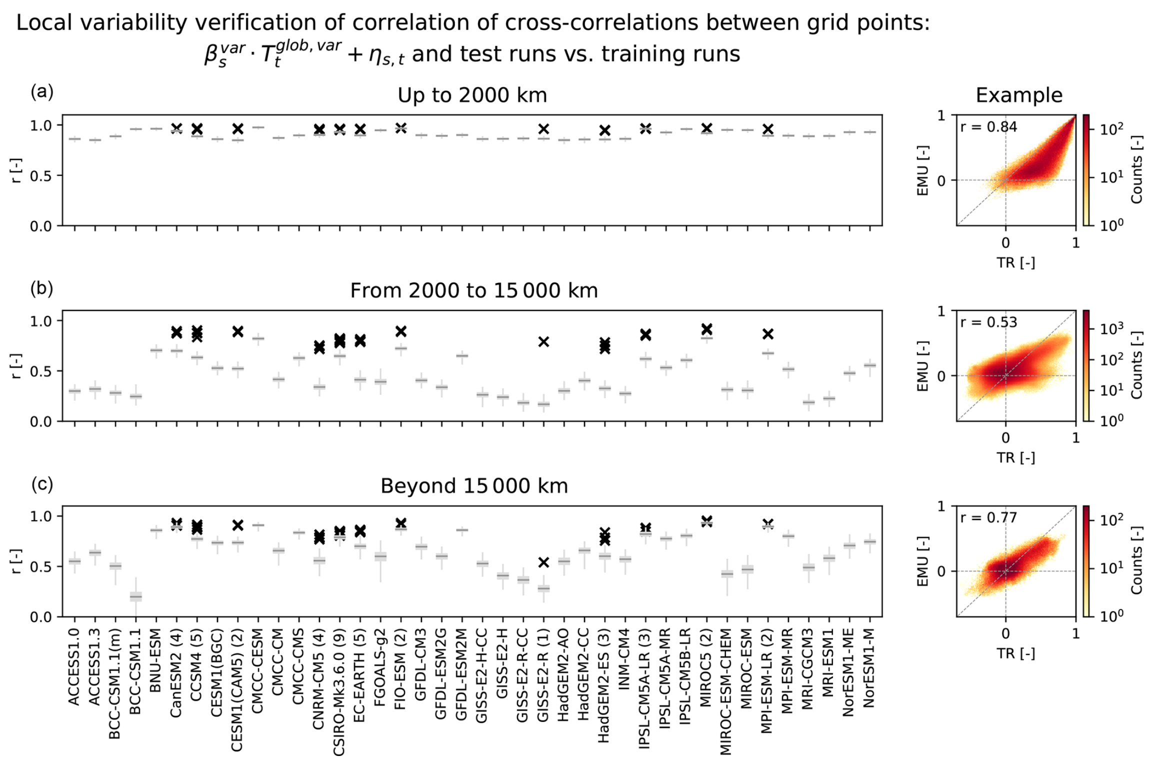

Figure 11. Local variability verification for the CMIP5 models (columns) by means of the Pearson correlation of cross-correlations between grid points in three geographical bands (rows) between the 1000 individual emulations and the training runs (box plots). The geographical bands cover distances below 2000 km **(a)**, between 2000 and 15 000 km **(b)**, and beyond 15 000 km **(c)**. The examples show the associated 2-D histograms for a single emulation and the training run of CESM1(CAM5). For the CMIP5 models with test runs, the correlation between the quantity in the training run and in each individual test run is indicated by a black cross. For all climate models with test runs available, the number of test runs is given in brackets after the model name.

able for climate models with larger localization radii (Fig. 7). In North Asia (NAS), the underdispersion is strongest for most models (Fig. 12). The only region where the emulations are fully reliable is global land. The underdispersion on the SREX regional scales is related to the regularization, which dampens covariances between grid points as a function of distance between them and is thus inherent to the emulator's design. The results are qualitatively similar for the test runs but, as expected, the deviations from the emulated quantiles tend to be larger in magnitude than for the training runs. For most climate models, the strongest deviations in the median of the test runs are observed in global land, Canada–Greenland–Iceland (CGI), and South Australia–New Zealand (SAU). Out of all climate models, the least optimal fit is obtained for MIROC5, with the emulated median being systematically warmer than the training and especially the test run medians in many regions.

7 Discussion

7.1 Emulator design choices and their implications

7.1.1 Modular framework

A modular framework is chosen for the climate model emulation because of its manifold advantages. First, the calibrated parameters of each emulator module can be used for climate model intercomparison over a wide range of scales since they can be readily visualized and easily interpreted (Sects. 5.1 and 6.1). Second, the modular framework renders it straightforward to substitute each emulator module with approaches other than the ones chosen here. For example, alternative approaches for the global mean temperature trend (e.g., Meinshausen et al., 2011), for the local response module (e.g., Tebaldi and Arblaster, 2014; Alexeeff et al., 2018), or for the local residual temperature variability (e.g., Link et al., 2019) could be employed. Third, if the modeling task were to change, additional predictors could easily be integrated. For example, precipitation emulation would likely require human-induced aerosol emissions as an additional predictor in the local response module (Frieler et al., 2012).

7.1.2 Emulating temperature trends

In this study, an estimate of $T_t^{\mathrm{glob,trend}}$ is retrieved with a simple statistical model from the training run (Sect. 4.1.2). However, it could alternatively be considered to obtain $T_t^{\mathrm{glob,trend}}$ from a simple energy balance model (Meinshausen et al., 2011). This would open avenues towards emulating initial-condition ensembles across different $T_t^{\mathrm{glob,trend}}$ trajectories and thus different emission scenario pathways.

To translate $T_t^{\mathrm{glob,trend}}$ into a local temperature in the local response module, a linear approach is chosen (Sect. 4.1.3). The thereby obtained regression coefficients β_s^{trend} represent well-known climate phenomena. The enhanced warming over land compared to the global mean (Sutton et al., 2007; Hartmann et al., 2013) at many grid points is captured by $\beta_s^{\mathrm{trend}} > 1$ (Figs. 3 and 7). The Arctic amplification (Serreze and Barry, 2011) manifests itself in the large β_s^{trend} values in northern high latitudes (Fig. 3). The overall good performance in capturing local trends is in line with the pronounced linear scaling of regional land temperatures with global mean temperature (Seneviratne et al., 2016; Wartenburger et al., 2017) and the widely used linear pattern scaling approaches (Mitchell, 2003; Tebaldi and Arblaster, 2014; Lynch et al., 2017; Osborn et al., 2018).

7.1.3 Emulating temperature variability

Spatially coherent local variability is introduced in two emulator modules, namely in the local response module as the local response to $T_t^{\mathrm{glob,var}}$ (Sect. 4.1.3) and in the local residual variability module (Sect. 4.1.4). The local variability is an essential ingredient in mimicking initial-condition ensembles as visualized by comparing the regionally averaged time series of our emulations with simple pattern scaling results which contain no local variability module (Fig. S11). In this study, and all other studies cited in the following paragraphs, the local temperature variability is assumed to be stationary in time, which is not fulfilled everywhere in the business-as-usual greenhouse gas emission scenario (see Sect. 2.3 and Olonscheck and Notz, 2017).

$T_t^{\mathrm{glob,var}}$ can be regarded as the globally aggregated signal of all physical modes of variability (Sect. 4.1.2), with the calibrated emulators accounting for memory of up to 3 years (Fig. 7). While the linear translation of $T_t^{\mathrm{glob,var}}$ to a grid-point-level temperature response is purely statistical in nature, physically meaningful patterns nevertheless emerge in the β_s^{var} patterns. For example, for many climate models, β_s^{var} tends to resemble an El Niño–Southern Oscillation pattern (Trenberth, 1997) with the Amazon, Australia, and Alaska covarying, while the southeastern USA exhibits the opposite temperature sign (Figs. 3 and S2–10). Qualitatively similar results could alternatively be obtained by stochastically generating time series of major physical modes of variability and translating those to the grid-point level (McKinnon and Deser, 2018).

Local residual variability is modeled as an AR process with spatially correlated innovations (Sect. 4.1.4). While several other authors have employed AR models with spatially correlated innovations (Castruccio and Stein, 2013; Castruccio and Genton, 2016; Bao et al., 2016), they all chose a parametric approach to model the covariance between grid points. However, in this study, a nonparametric approach is employed, which retains regional-scale anisotropy in the underlying data.

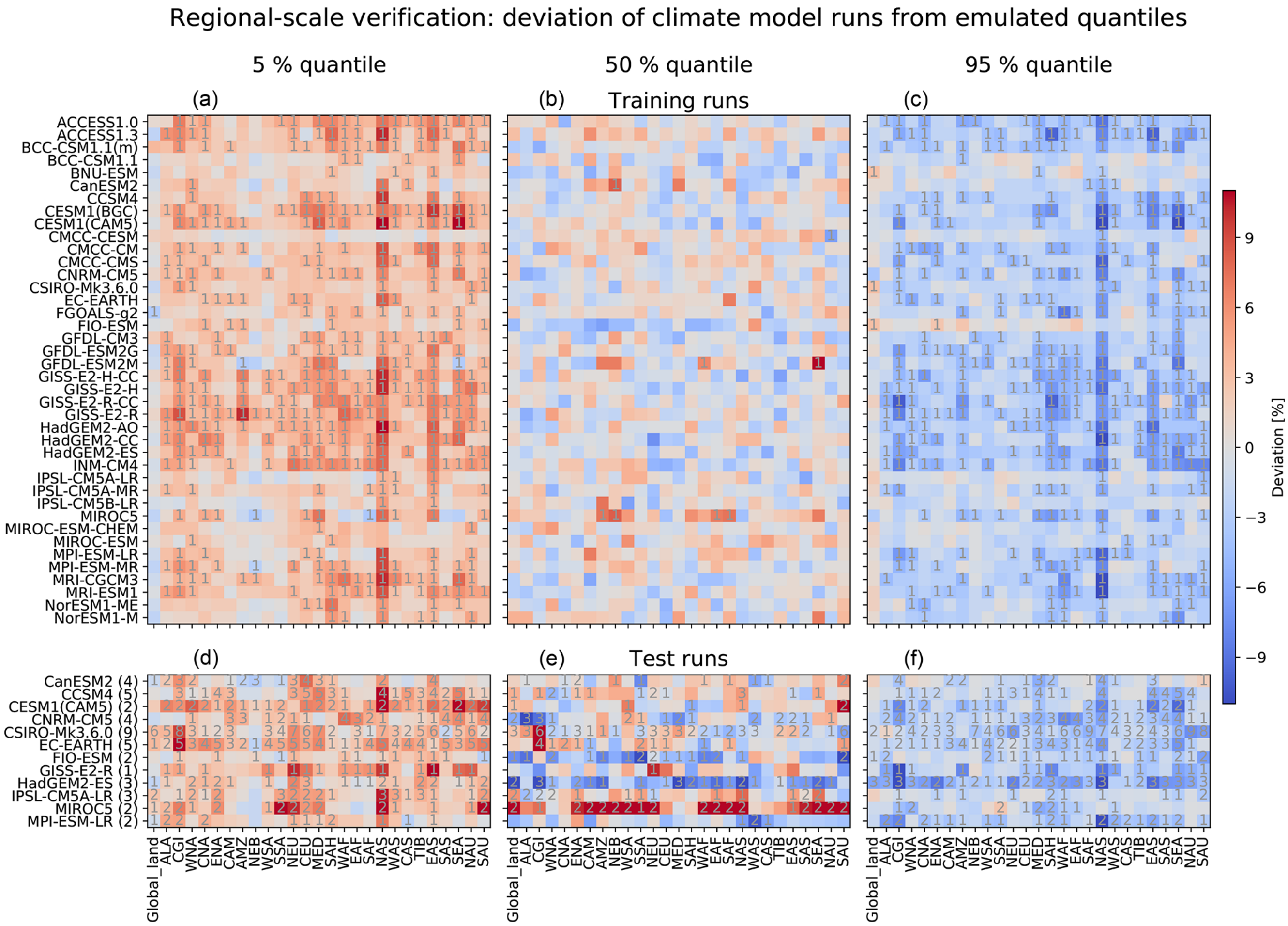

Figure 12. Deviation of climate model runs from the emulated 5 % **(a, d)**, 50 % **(b, e)**, and 95 % **(c, f)** quantile for CMIP5 models (rows) and regions (columns). The emulated quantile is computed based on 1000 emulations per climate model. The deviation of the climate model run from the emulated quantile is given in color. Red means that the emulated quantile is warmer than the quantile of the climate model run, and blue means that it is colder. The gray numbers indicate how many climate model run deviations lie outside the 95 % interval spanned by the deviations of single emulations from the emulated quantiles. If the climate model run lies outside this interval, it is no longer considered indistinguishable from the emulations. The deviation from the training run is shown in panels **(a)–(c)**, and the average deviation across all available test runs is shown in panels **(d)–(f)**. The number of test runs averaged across is indicated in brackets behind the model names.

7.2 The pros and cons of training on single climate model runs

We demonstrated that, for yearly temperature at grid-point to regional scales, training on a single run per climate model is sufficient to learn the key properties of the climate system of this climate model. Early results furthermore indicated that larger single-model initial-condition ensembles, in that case a 21-member CESM ensemble, can also be successfully emulated when training on a single ESM run (Beusch et al., 2018). Since a single run was submitted for the majority of climate models participating in CMIP5 for the emission pathway considered here, requiring only one run to train the emulator provides the opportunity to emulate a much larger multi-model ensemble and thus to have the resulting superensemble account for more inter-model uncertainty. Nevertheless, it is not possible to reproduce the characteristics of a true ESM at all spatial and temporal scales when training on a single run. To obtain the best possible emulations to be used, e.g., for uncertainty propagation in climate impact or integrated assessment models, it is thus advisable to employ all available runs for training instead of just a single one for each climate model. When training on multiple runs, the parameters of the emulator can be estimated more robustly, which, among other things, results in a larger localization radius and thus the ability to reproduce farther-reaching spatial cross-correlations between grid points.

7.3 Large single-model initial-condition vs. large multi-model ensembles

Our results highlight fundamental differences between large single-model initial-condition ensembles (Deser et al., 2012; Fischer et al., 2013; Kay et al., 2015; Leduc et al., 2019; Maher et al., 2019) and large multi-model ensembles (Meehl et al., 2007; Taylor et al., 2012; Eyring et al., 2016). While multi-model ensembles are imperfect, with several ESMs exhibiting dependencies (Knutti, 2010; Bishop and Abramowitz, 2013; Sanderson et al., 2015; Abramowitz et al., 2019), multi-model uncertainty nevertheless clearly exceeds single-model initial-condition uncertainty at the yearly scale for temperature (Sect. 5.3). ESMs contained within CMIP5 differ substantially across a broad range of scales and thus sample different phase spaces in projections, which renders it necessary to train an emulator on each climate model to approximate the CMIP5 ensemble. A single-model initial-condition ensemble, on the other hand, can be successfully mimicked on grid-point to regional scales by training on a single ESM run (Sects. 5 and 6). While this lies beyond the scope of this study, the developed emulator could additionally serve as a novel tool to address the challenge of inter-model dependencies. Differences between climate models could be quantified in terms of their emulator parameters, and subsequently a subset of models with sufficiently divergent parameters could be selected to base projections on. Additionally, observations could be used to constrain the emulated ensemble by providing validation measures for the emulator parameters.

8 Conclusions and outlook

We introduce a modular framework for climate model emulation of yearly land temperatures and present a specific, computationally cheap implementation called MESMER, which can create plausible temperature field time series within seconds based on a single climate model training run. Our emulator consists of (i) a global mean temperature module, (ii) a local temperature response module, and (iii) a local residual temperature variability module. The global mean temperature module contains a global mean temperature trend, which is shared by all emulations, and a global mean temperature variability term, which is modeled as an AR process and varies between individual emulations. The local response module is linear in nature and consists of a separate response to the global mean temperature trend and the global mean temperature variability. The local residual variability module generates spatiotemporally correlated fields by means of locally fit AR(1) processes with spatially correlated innovations.

Since emulators approximate complex ESMs in a simplified manner, they are not able to accurately reproduce all spatiotemporal ESM characteristics. The emulator presented here, e.g., dampens covariations between grid points as a function of distance in the local residual variability module due to regularization. Thus, our emulator reliably reproduces climate model variability at the grid-point level, but the emulations are increasingly underdispersive for larger regional averages and intermediate-range spatial teleconnections cannot be accounted for. This caveat could be addressed by further improving the local residual variability module implementation with a focus on such teleconnections. Alternatively, training on several ESM runs would increase the robustness of the estimated parameters and make it possible to reproduce farther-reaching teleconnections within the current emulator setup. Nevertheless, calibrating our emulator on a single training run is sufficient to generate emulations which are visually indistinguishable from true ESM runs.

Inherent inter-ESM differences in warming trends and spatiotemporal variability make it necessary to calibrate a separate emulator for each of the 40 considered CMIP5 models. The resulting emulations successfully approximate the training run for each climate model on grid-point to regional scales. For CMIP5 models with more than one initial-condition ensemble member, it was furthermore demonstrated that the ensemble of emulations is generally able to mimic true climate model initial-condition ensembles at these scales. Hence, we argue that to sample climate signal uncertainty for yearly temperature at grid-point to regional scales, it is more advantageous to invest computational resources into generating multi-model ensembles rather than large single-model ensembles, since the latter can be readily approximated by our emulator.

Superensembles such as the one generated in this study, which contains 1000 emulations per climate model, are expected to be particularly helpful in regions with large interannual variability. There, the very sparse sampling of the temperature phase space by the CMIP5 ensemble may result in biased conclusions when solely employing the CMIP5 ensemble as an input to impact or integrated assessment models which estimate the effect of climate signal uncertainty on their quantity of interest.

The emulator is designed to be flexible enough to emulate whatever climate model run it is provided with. Hence, it is not part of the emulator's tasks to judge the realism of individual climate models. Instead, the choice of considered ESMs will depend on the scope of different applications. For example, results from emergent constraints analyses (e.g., Hall and Qu, 2006; Eyring et al., 2019) could be combined with the implementation of an emulator to derive a superensemble based on an observationally constrained set of ESMs. On the other hand, the emulator parameters could themselves be used as potential constraints that can also be derived from observations. Additionally, the emulator parameters can be regarded as an ESM-specific "model ID", which provides an interesting avenue for climate model intercomparison across a wide range of scales. Inter-model differences can be readily visualized for every emulator module, resulting in comprehensible scale-dependent insights into the underlying properties of each climate model. Future work could focus on extending the emulator to simultaneously generate multivariate output. Furthermore, it would be interesting to investigate how transferable an emulator trained on a specific greenhouse gas emission scenario is to other emission pathways and which modules would need to be modified to account for inter-scenario differences.

In conclusion, in this study we have presented a novel ESM emulator called MESMER that can be trained to represent separate ESMs based on single realizations of the respective ESMs and which has been shown to be able to emulate and expand multi-model ensembles such as CMIP5. We expect that the developed emulator can serve as a training ground for investigating the phase space of multi-model ensembles in new applications, e.g., related to the derivation of emissions scenarios or the assessment of impacts under different emissions pathways.

Appendix A

Table A1. List of the 40 employed CMIP5 models, the modeling groups providing them, and the number of initial-condition ensemble members used.

Model	Modeling center (or group)	Runs
ACCESS1.0	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	1
ACCESS1.3	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	1
BCC-CSM1.1(m)	Beijing Climate Center, China Meteorological Administration	1
BCC-CSM1.1	Beijing Climate Center, China Meteorological Administration	1
BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University	1
CanESM2	Canadian Centre for Climate Modeling and Analysis	5
CCSM4	National Center for Atmospheric Research	6
CESM1(BGC)	Community Earth System Model Contributors	1
CESM1(CAM5)	Community Earth System Model Contributors	3
CMCC-CESM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1
CMCC-CM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1
CMCC-CMS	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1
CNRM-CM5	Centre National de Recherches Météorologiques/Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	5
CSIRO-Mk3.6.0	Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	10
EC-EARTH	EC-EARTH consortium	6
FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS, Tsinghua University	1
FIO-ESM	The First Institute of Oceanography, SOA, China	3
GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory	1
GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory	1
GFDL-ESM2M	NOAA Geophysical Fluid Dynamics Laboratory	1
GISS-E2-H-CC	NASA Goddard Institute for Space Studies	1
GISS-E2-H	NASA Goddard Institute for Space Studies	1
GISS-E2-R-CC	NASA Goddard Institute for Space Studies	1
GISS-E2-R	NASA Goddard Institute for Space Studies	2
HadGEM2-AO	National Institute of Meteorological Research/Korea Meteorological Administration	1
HadGEM2-CC	Met Office Hadley Centre	1
HadGEM2-ES	Met Office Hadley Centre (additional realizations contributed by Instituto Nacional de Pesquisas Espaciais)	4
INM-CM4	Institute for Numerical Mathematics	1
IPSL-CM5A-LR	Institut Pierre Simon Laplace	4
IPSL-CM5A-MR	Institut Pierre Simon Laplace	1
IPSL-CM5B-LR	Institut Pierre Simon Laplace	1
MIROC5	Atmosphere and Ocean Research Institute (the University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine–Earth Science and Technology	3
MIROC-ESM-CHEM	Japan Agency for Marine–Earth Science and Technology, Atmosphere and Ocean Research Institute (the University of Tokyo), and National Institute for Environmental Studies	1
MIROC-ESM	Japan Agency for Marine–Earth Science and Technology, Atmosphere and Ocean Research Institute (the University of Tokyo), and National Institute for Environmental Studies	1
MPI-ESM-LR	Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	3
MPI-ESM-MR	Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	1
MRI-CGCM3	Meteorological Research Institute	1
MRI-ESM1	Meteorological Research Institute	1
NorESM1-ME	Norwegian Climate Centre	1
NorESM1-M	Norwegian Climate Centre	1

Data availability. The employed CMIP5 data are available from the public CMIP archive at: https://esgf-node.llnl.gov/projects/esgf-llnl/ (last access: 8 November 2019). The stratospheric aerosol optical depth data are provided by NASA and available at: https://data.giss.nasa.gov/modelforce/strataer/ (last access: 8 November 2019).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/esd-11-139-2020-supplement.

Author contributions. LB, LG, and SIS designed the study based on an initial idea from SIS. LB carried out the analysis and drafted the text. LG provided statistical support for the analysis. All authors contributed to interpreting the results and refining the text.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the World Climate Research Program's Working Group on Coupled Modeling, which is responsible for the Coupled Model Intercomparison Project (CMIP), and we thank the climate modeling groups (listed in Table A1 of this paper) for producing and making available their model output. We are additionally indebted to Urs Beyerle and Jan Sedláček for retrieving and preprocessing the CMIP5 data. Moreover, we would like to thank Julien Brajard, Loris Foresti, and Vincent Humphrey for their valuable input on different modules of our emulator and Erich Fischer for coming up with the term "superensemble". Lastly, we thank Robert Link and the two anonymous reviewers for their useful feedback, which helped to improve this study.

Financial support. This research has been supported by Horizon 2020 (CRESCENDO (grant no. 641816)) and FP7 Ideas: European Research Council (DROUGHT-HEAT (grant no. 617518)).

Review statement. This paper was edited by Raghavan Krishnan and reviewed by two anonymous referees.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Alexeeff, S. E., Nychka, D., Sain, S. R., and Tebaldi, C.: Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments, Climatic Change, 146, 319–333, https://doi.org/10.1007/s10584-016-1809-8, 2018.

Allen, M. R. and Stott, P. A.: Estimating signal amplitudes in optimal fingerprinting, part I: Theory, Clim. Dynam., 21, 477–491, https://doi.org/10.1007/s00382-003-0313-9, 2003.

Bao, J., McInerney, D. J., and Stein, M. L.: A spatial-dependent model for climate emulation, Environmetrics, 27, 396–408, https://doi.org/10.1002/env.2412, 2016.

Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth System Model Temperatures, in: Proceedings of the 8th International Workshop on Climate Informatics: CI 2018, 19–21 September 2018, Boulder, CO, USA, edited by: Chen, C., Cooley, D., Runge, J., and Szekely, E., NCAR Technical Note, Boulder, 41–44, https://doi.org/10.5065/D6BZ64XQ, 2018.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dynam., 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Brown, P. T., Li, W., Cordero, E. C., and Mauget, S. A.: Comparing the model-simulated global warming signal to observations using empirical estimates of unforced noise, Sci. Rep., 5, 1–9, https://doi.org/10.1038/srep09957, 2015.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIRES Clim. Change, 9, 1–79, https://doi.org/10.1002/wcc.535, 2018.

Castruccio, S. and Genton, M. G.: Compressing an Ensemble With Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature, Technometrics, 58, 319–328, https://doi.org/10.1080/00401706.2015.1027068, 2016.

Castruccio, S. and Stein, M. L.: Global space-time models for climate ensembles, Ann. Appl. Stat., 7, 1593–1611, https://doi.org/10.1214/13-AOAS656, 2013.

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J.: Statistical emulation of climate model projections based on precomputed GCM runs, J. Climate, 27, 1829–1844, https://doi.org/10.1175/JCLI-D-13-00099.1, 2014.

Cressie, N. and Wikle, C. K.: Statistics for spatio-temporal data, John Wiley & Sons, Hoboken, New Jersey, USA, 2011.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: The role of internal variability, Clim. Dynam., 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nat. Clim. Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Fischer, E. M., Beyerle, U., and Knutti, R.: Robust spatially aggregated projections of climate extremes, Nat. Clim. Change, 3, 1033–1038, https://doi.org/10.1038/nclimate2051, 2013.

Frieler, K., Meinshausen, M., Mengel, M., Braun, N., and Hare, W.: A scaling approach to probabilistic assessment of regional climate change, J. Climate, 25, 3117–3144, https://doi.org/10.1175/JCLI-D-11-00199.1, 2012.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Q. J. Roy. Meteor. Soc., 125, 723–757, https://doi.org/10.1256/smsqj.55416, 1999.

Goodwin, P.: How historic simulation–observation discrepancy affects future warming projections in a very large model ensemble, Clim. Dynam., 47, 2219–2233, https://doi.org/10.1007/s00382-015-2960-z, 2016.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys. Res. Lett., 33, L03502, https://doi.org/10.1029/2005GL025127, 2006.

Hartmann, D., Klein Tank, A., Rusticucci, M., Alexander, L., Brönnimann, S., Charabi, Y., Dentener, F., Dlugokencky, E., Easterling, D., Kaplan, A., Soden, B., Thorne, P., Wild, M., and Zhai, P.: Observations: Atmosphere and Surface, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgle, P., chap. 2, Cambridge, UK and New York, NY, USA, 159–254, https://doi.org/10.1017/CBO9781107415324.008, 2013.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, B. Am. Meteorol. Soc., 90, 1095–1107, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Herger, N., Sanderson, B. M., and Knutti, R.: Improved pattern scaling approaches for the use in climate impact studies, Geophys. Res. Lett., 42, 3486–3494, https://doi.org/10.1002/2015GL063569, 2015.

Holden, P. B. and Edwards, N. R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling, Geophys. Res. Lett., 37, 1–5, https://doi.org/10.1029/2010GL045137, 2010.

Holden, P. B., Edwards, N. R., Garthwaite, P. H., Fraedrich, K., Lunkeit, F., Kirk, E., Labriet, M., Kanudia, A., and Babonneau, F.: PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment, Geosci. Model Dev., 7, 433–451, https://doi.org/10.5194/gmd-7-433-2014, 2014.

Humphrey, V. and Gudmundsson, L.: GRACE-REC: a reconstruction of climate-driven water storage changes over the last century, Earth Syst. Sci. Data, 11, 1153–1170, https://doi.org/10.5194/essd-11-1153-2019, 2019.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, B. Am. Meteorol. Soc., 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1, 2015.

King, A. D., Karoly, D. J., and Henley, B. J.: Australian climate extremes at 1.5 °C and 2 °C of global warming, Nat. Clim. Change, 7, 412–416, https://doi.org/10.1038/nclimate3296, 2017.

Knutti, R.: The end of model democracy?, Climatic Change, 102, 395–404, https://doi.org/10.1007/s10584-010-9800-2, 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

Leduc, M., Mailhot, A., Frigon, A., Martel, J.-L., Ludwig, R., Brietzke, G. B., Giguére, M., Brissette, F., Turcotte, R., Braun, M., and Scinocca, J.: The ClimEx project: a 50-member ensemble of climate change projections at 12-km resolution over Europe and northeastern North America with the Canadian Regional Climate Model (CRCM5), J. Appl. Meteorol. Clim., 58, 663–693, https://doi.org/10.1175/jamc-d-18-0021.1, 2019.

Levy, H., Horowitz, L. W., Schwarzkopf, M. D., Ming, Y., Golaz, J.-C., Naik, V., and Ramaswamy, V.: The roles of aerosol direct and indirect effects in past and future climate change, J. Geophys. Res.-Atmos., 118, 4521–4532, https://doi.org/10.1002/jgrd.50192, 2013.

Link, R., Snyder, A., Lynch, C., Hartin, C., Kravitz, B., and Bond-Lamberty, B.: Fldgen v1.0: an emulator with internal variability and space–time correlation for Earth system models, Geosci. Model Dev., 12, 1477–1489, https://doi.org/10.5194/gmd-12-1477-2019, 2019.

Lopez, A., Suckling, E. B., and Smith, L. A.: Robustness of pattern scaled climate change scenarios for adaptation decision support, Climatic Change, 122, 555–566, https://doi.org/10.1007/s10584-013-1022-y, 2014.

Lynch, C., Hartin, C., Bond-Lamberty, B., and Kravitz, B.: An open-access CMIP5 pattern library for temperature and precipitation: description and methodology, Earth Syst. Sci. Data, 9, 281–292, https://doi.org/10.5194/essd-9-281-2017, 2017.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Kornblueh, L., Takano, Y., Kröger, J., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble – Enabling the Exploration of Climate System Variability, J. Adv. Model. Earth Syst., 11, 1–21, https://doi.org/10.1029/2019MS001639, 2019.

May, W.: Assessing the strength of regional changes in near-surface climate associated with a global warming of 2 °C, Climatic Change, 110, 619–644, https://doi.org/10.1007/s10584-011-0076-y, 2012.

McKinnon, K. A. and Deser, C.: Internal variability and regional climate trends in an observational large ensemble, J. Climate, 31, 6783–6802, https://doi.org/10.1175/JCLI-D-17-0901.1, 2018.

McKinnon, K. A., Poppick, A., Dunn-Sigouin, E., and Deser, C.: An "observational large ensemble" to compare observed and modeled temperature trend uncertainty due to internal variability, J. Climate, 30, 7585–7598, https://doi.org/10.1175/JCLI-D-16-0905.1, 2017.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., and Taylor, K. E.: The WCRP CMIP3 multi-model dataset: a new era in climate change research, B. Am. Meteorol. Soc., 88, 1383–1394, https://doi.org/10.1175/BAMS-88-9-1383, 2007.

Meinshausen, M., Raper, S. C. B., and Wigley, T. M. L.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration, Atmos. Chem. Phys., 11, 1417–1456, https://doi.org/10.5194/acp-11-1417-2011, 2011.

Mitchell, T. D.: Pattern Scaling. An Examination of the Accuracy of the Technique for Describing Future Climates, Climatic Change, 60, 217–242, https://doi.org/10.1023/A:1026035305597, 2003.

Nychka, D., Hammerling, D., Krock, M., and Wiens, A.: Modeling and emulation of nonstationary Gaussian fields, Spat. Stat., 28, 21–38, https://doi.org/10.1016/j.spasta.2018.08.006, 2018.

Olonscheck, D. and Notz, D.: Consistently estimating internal climate variability from climate model simulations, J. Climate, 30, 9555–9573, https://doi.org/10.1175/JCLI-D-16-0428.1, 2017.

Osborn, T. J., Wallace, C. J., Harris, I. C., and Melvin, T. M.: Pattern scaling using ClimGen: monthly-resolution future climate scenarios including changes in the variability of precipitation, Climatic Change, 134, 353–369, https://doi.org/10.1007/s10584-015-1509-9, 2016.

Osborn, T. J., Wallace, C. J., Lowea, J. A., and Bernie, D.: Performance of pattern-scaled climate projections under high-end warming. Part I: Surface air temperature over land, J. Climate, 31, 5667–5680, https://doi.org/10.1175/JCLI-D-17-0780.1, 2018.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5-A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33–57, https://doi.org/10.1007/s10584-011-0149-y, 2011.

Rougier, J., Sexton, D. M. H., Murphy, J. M., and Stainforth, D.: Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments, J. Climate, 22, 3540–3557, https://doi.org/10.1175/2008JCLI2533.1, 2009.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties, J. Climate, 28, 5150–5170, https://doi.org/10.1175/JCLI-D-14-00361.1, 2015.

Santer, B. D., Wigley, T. M. L., Schlesinger, M. E., and Mitchell, J. F. B.: Developing climate scenarios from equilibrium results. Max-Planck-Institut für Meteorologie report, Tech. Rep. 47, Hamburg, Germany, 1990.

Sato, M., Hansen, J. E., McCormick, M. P., and Pollack, J. B.: Stratospheric aerosol optical depths, 1850–1990, J. Geophys. Res., 98, 22987–22994, https://doi.org/10.1029/93JD02553, 1993.

Scher, S.: Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning, Geophys. Res. Lett., 45, 12616–12622, https://doi.org/10.1029/2018GL080704, 2018.

Scher, S. and Messori, G.: Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground, Geosci. Model Dev., 12, 2797–2809, https://doi.org/10.5194/gmd-12-2797-2019, 2019.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X.: Changes in climate extremes and their impacts on the natural physical environment, in: Managing the Risk of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), edited by: Field, C. B., Barros, V., Stocker, T. F., Qin, D., Dokken, D. J., Ebi, K. L., Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Allen, S. K., Tignor, M., and Midgley, P. M., chap. 3, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 109–230, 2012.

Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., and Wilby, R. L.: Allowable CO_2 emissions based on regional and impact-related climate targets, Nature, 529, 477–483, https://doi.org/10.1038/nature16542, 2016.

Serreze, M. C. and Barry, R. G.: Processes and impacts of Arctic amplification: A research synthesis, Global Planet. Change, 77, 85–96, https://doi.org/10.1016/j.gloplacha.2011.03.004, 2011.

Sutton, R. T., Dong, B., and Gregory, J. M.: Land/sea warming ratio in response to climate change: IPCC AR4 model results and comparison with observations, Geophys. Res. Lett., 34, L02701, https://doi.org/10.1029/2006GL028164, 2007.

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

Tebaldi, C. and Arblaster, J. M.: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations, Climatic Change, 122, 459–471, https://doi.org/10.1007/s10584-013-1032-9, 2014.

Tebaldi, C. and Knutti, R.: Evaluating the accuracy of climate change pattern emulation for low warming targets, Environ. Res. Lett., 13, 55006, https://doi.org/10.1088/1748-9326/aabef2, 2018.

Trenberth, K. E.: The Definition of El Niño, B. Am. Meteorol. Soc., 78, 2771–2778, https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2, 1997.

Wartenburger, R., Hirschi, M., Donat, M. G., Greve, P., Pitman, A. J., and Seneviratne, S. I.: Changes in regional climate extremes as a function of global mean temperature: an interactive plotting framework, Geosci. Model Dev., 10, 3609–3634, https://doi.org/10.5194/gmd-10-3609-2017, 2017.

Weigel, A. P.: Ensemble Forecasts, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T. and Stephenson, D. B., chap. 8, John Wiley & Sons, Chichester, UK, 2nd edn., 141–166, https://doi.org/10.1002/9781119960003.ch8, 2012.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dynam., 41, 1703–1729, https://doi.org/10.1007/s00382-013-1896-4, 2013.