MESMER-M: an Earth System Model emulator for spatially resolved monthly temperature

Shruti Nath[1,2], Quentin Lejeune[1], Lea Beusch[2,*], Sonia I. Seneviratne[2], and Carl-Friedrich Schleussner[1,3]

[1]Climate Analytics, Berlin, Germany
[2]Institute of Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[3]Integrative Research Institute on Transformations of Human-Environment Systems (IRI THESys) and Geography Department, Humboldt-Universität zu Berlin, Berlin, Germany
[*]Now at: Center for Climate Systems Modeling (C2SM), ETH Zurich, Zurich, Switzerland and MeteoSwiss, via ai Monti 146, 6605 Locarno Monti, Switzerland

Correspondence: shruti.nath@climateanalytics.org

Abstract. The degree of trust placed in climate model projections is commensurate to how well their uncertainty can be quantified, particularly at timescales relevant to climate policy makers. On interannual to decadal timescales, model projection uncertainty due to [..[1]]natural variability dominates on the local scale and is imperative to [..[2]]describing near-term and seasonal climate events, but [..[3]]difficult to quantify owing to the computational constraints [..[4]]of producing large ensembles. To this extent, emulators are valuable tools for approximating climate model runs, allowing for exploration of the [..[5]]uncertainty space surrounding [..[6]]selected climate variables at a [..[7]]substantially reduced computational cost. Most emulators, however, operate at annual to seasonal timescales, leaving out monthly information that may be essential to assessing climate impacts. This study extends the framework of an existing spatially resolved, annual-scale Earth System Model (ESM) emulator (MESMER, Beusch et al. (2020)) by a monthly downscaling module (MESMER-M), thus providing local monthly temperatures from local yearly temperatures. We first linearly represent the mean response of the monthly temperature cycle to yearly temperatures using a simple harmonic model, thus maintaining month to month correlations and capturing changes in intra-annual variability. We then construct a month-specific local variability module which generates spatio-temporally correlated residuals with [..[8]]yearly temperature and month dependent skewness incorporated within. The performance of the resulting emulator is demonstrated on 38 different ESMs from the 6th phase of the Coupled Model Intercomparison Project (CMIP6). The emulator is furthermore benchmarked using a simple Gradient Boosting Regressor based, [..[9]]physically-informed model trained on biophysical information. We find that while regional-scale, biophysical feedbacks may induce non-uniformities in the yearly to monthly temperature downscaling relationship, statistical emulation of regional effects shows comparable skill to

---

[1]removed: internal variability dominates
[2]removed: understanding
[3]removed: hard
[4]removed: on
[5]removed: model
[6]removed: select
[7]removed: significantly
[8]removed: month and yearly temperature
[9]removed: physical

approaches with physical representation. Thus, MESMER-M is able to statistically generate ESM-like, large initial-condition ensembles of spatially explicit monthly temperature fields, thereby providing monthly temperature probability distributions
20 which are of critical value to impact assessments.

1 Introduction

Climate model emulators are computationally [..[10]]inexpensive devices that derive simplified statistical relationships from existing climate model runs to then approximate model runs that have not been generated yet. By reproducing runs from deterministic, process-based climate models at a [..[11]]substantially reduced computational time, climate model emulators
25 facilitate exploration of the uncertainty space surrounding model representation of climate responses to specific forcings. A wide toolset of Earth System Model (ESM) emulators exists with capabilities ranging from investigating the effects of greenhouse gas emission scenarios on global to regional mean annual climate fields [..[12]](Meinshausen et al., 2011; Tebaldi and Arblaster, 2014) to looking at regional-scale annual, seasonal and monthly [..[13]]natural climate variabilities (Link et al., 2019; McKinnon and Deser, 2018; Alexeeff et al., 2018; Castruccio et al., 2019).
30   Recently, a Modular Earth System Model Emulator with spatially Resolved output (MESMER) (Beusch et al., 2020) has been developed with the ability to represent [..[14]]Earth System Model (ESM) specific forced local responses alongside projection uncertainty arising from natural climate variability. It does so using a combination of pattern scaling and a natural climate variability module, to generate grid-point level, yearly temperature realisations that emulate the properties of ESM initial-condition ensembles. By training on [..[15]]different ESMs individually, MESMER is furthermore able to [..[16]]account
35 for inter-ESM differences in forced local responses and natural climate variability, thus approximating a multi-model initial condition ensemble, i.e. "super-ensemble"[..[17]]. The probability distributions of grid-point level, yearly temperatures generated by MESMER [..[18]]could be especially relevant when used as input data for simulation of impacts that depend on this variable. MESMER thus offers the perspective to improve our [..[19]]description of the likelihood of future impacts under multiple scenarios.
40   Considering the importance of monthly and seasonal information [..[20]]when assessing the impacts of climate change [..[21]](Schlenker and Roberts, 2009; Wramneby et al., 2010; Stéfanon et al., 2012; Zhao et al., 2017), extending the MESMER approach to grid point level monthly temperatures appears desirable. Such holds additional value in assessing the evolving

---

[10]removed: cheap
[11]removed: significantly
[12]removed: (Meinshausen et al., 2011)
[13]removed: internal
[14]removed: model uncertainty due to
[15]removed: multiple ESMs
[16]removed: create
[17]removed: which accounts for uncertainties arising from model formulation
[18]removed: are
[19]removed: understanding
[20]removed: in
[21]removed: (Zhao et al., 2017; Schlenker and Roberts, 2009; Wramneby et al., 2010; Stéfanon et al., 2012)

likelihoods of future impacts, as the temperature response at monthly timescales displays heterogeneities distributive onto seasonal to monthly variabilities and therefore uncertainties, which are otherwise unapparent at annual timescales. In particular, winter months can warm disproportionately more than summer months [..[22]](Fischer et al., 2011; Holmes et al., 2016; Loikith and Neelin, 2019), which in turn leads to non-stationarity in the amplitude of the seasonal cycle (i.e. intra-annual temperature variabilities) with evolving yearly temperatures (i.e. the intra-annual temperature response is heteroscedastic with regard to yearly temperature) (Fischer et al., 2012; Huntingford et al., 2013; Thompson et al., 2015; Osborn et al., 2016). Additionally, given that monthly temperature distributions have been observed to display non-Gaussianity, evolving yearly temperatures may cause disproportionate effects on their tail extremes, leading to changes in skewness (Wang et al., 2017; Sheridan and Lee, 2018; Tamarin-Brodsky et al., 2020).

This study focusses on extending MESMER's framework [..[23]]by a local monthly downscaling module (MESMER-M). This enables the estimation of projection uncertainty due to natural variability as propagated from annual to monthly timescales since MESMER-M builds upon MESMER, which has already been validated as yielding spatio-temporally accurate variabilities. In constructing MESMER-M, we furthermore place emphasis on representing heteroscedasticity of the intra-annual temperature response as well as changes in skewness of individual monthly distributions in a spatio-temporally accurate manner. The structure of this study is as follows: we first introduce the framework of the emulator under Section 3.1 and the approach to verification of the emulator performance under Section 3.2, we then provide the calibration results of the emulator under section 4 and verification results under section 5, after which we proceed to the conclusion and outlook under section 6.

2 Data and Terminology

In the analysis, 38 CMIP6 models (Eyring et al., 2016) are considered, using simulations for the SSP5-8.5 scenario high-emission scenario (O'Neill et al., 2016), so as to first explore the emulator's applicability to the extreme end of Greenhouse Gas (GHG) induced warming. Where an ESM's initial-condition ensemble set contains more than one member, it is split into a training set (used for emulator calibration) and a test set (used for emulator cross-validation). [..[24]]Systematic approaches in getting the best train-test split [..[25]]exist, such as that employed by (Castruccio et al., 2019) which requires presence of a large-ensemble and considers the balance between the stability in inference (represented for example by variability) of the emulator, and benefits for reduction in computational costs. Since the primary purpose of MESMER-M is to provide the best possible emulations based on available training material, such approaches are optional however and would only limit the demonstration of MESMER-M to CMIP6 ESMs with large-ensembles. Here we implement a simple 70%-30%, and for models with more than 20 ensemble members a 50%-50%, train-test split so as to maintain a good balance between training time and model performance. A summary of the CMIP6 models used, their associated modeling groups

---

[22]removed: (Holmes et al., 2016; Loikith and Neelin, 2019; Fischer et al., 2011)

[23]removed: to consist of

[24]removed: The

[25]removed: is done in a 70-30 manner

and the initial condition ensemble members present within the training and test sets are given in Table [..[26]]A1 in Appendix A. All ESM runs are obtained at a monthly resolution and are bilinearly interpolated to a spatial resolution of $2.5° \times 2.5°$(Brunner et al., 2020). The emulator is trained on yearly averaged temperature values. The term 'temperature' here refers to anomalies

75 of surface air temperature (standardly named 'tas') relative to the [..[27]]annual climatological mean over the reference period of 1870-1899.

3 Methods

3.1 MESMER

MESMER is an ESM-specific emulator built to produce spatially resolved, yearly temperature fields by considering both the

80 local mean response and the local variability surrounding the mean response. Within MESMER, local [..[28]]temperatures T for a given grid point s and year y are emulated as follows (Beusch et al., 2020):

$$T_{s,y} = g_s(T_y^{glob}) + \eta_{s,y} = \beta[..^{29}]^{forced}{}_s \cdot T[..^{30}]^{glob,forced}{}_y + \beta_s^{var} \cdot T_y^{glob,var} + \beta_s^{intercept} + \eta_{s,y} \tag{1}$$

Where g_s is the local mean response to global mean temperatures T^{glob} and consists of a [..[31]]multiple, linear regression on the smooth [..[32]]$T^{glob,forced}$ (capturing the trend in [..[33]]T^{glob}) and variability $T^{glob,var}$ components of T^{glob}[..[34]], with

85 coefficients β_s^{forced} and β_s^{var} respectively, and intercept term $\beta_s^{intercept}$. More details on the extraction and representation of $T^{glob,forced}$ and $T^{glob,var}$ can be found in Beusch et al. (2020). $\eta_{s,y}$ represents the [..[35]]residual variability around the mean response.

3.2 MESMER-M

We divide MESMER-M into a mean response module and a residual variability module, each calibrated on ESM simulation

90 output data for each grid point individually according to the procedure described in this section and summarised in Figure 1. Such division of a modelling exercise has previously been done in other climate model emulations [..[36]](Tebaldi and Arblaster, 2014; Alexeeff et al., 2018; Link et al., 2019; Beusch et al., 2020) and comes with its underlying assumptions. The primary assumption in our case is that the ESM monthly temperatures are distinctly separable into a mean response component and a residual variability component. Traditionally the mean response module is designed to be dependent on a certain forcing

---

[26]removed: A1 in Appendix A

[27]removed: mean temperature value

[28]removed: temperature anomalies

[31]removed: multivariate

[32]removed: $T^{glob,smooth}$

[33]removed: T glob

[34]removed: with coefficients β_s^{smooth}

[35]removed: local variability surrounding

[36]removed: (Link et al., 2019; Alexeeff et al., 2018; Tebaldi and Arblaster, 2014; Beusch et al., 2020)

95   (in this case local, yearly temperatures), while the variability module is space-time dependent [..[37]](i.e. varying with time and across the spatial domain). Given the expected changes in monthly skewness with evolving yearly temperature however, we furthermore propose both a space-time and temperature dependent variability module. [..[38]]other forcings, such as [..[39]]land cover, have [..[40]]a considerably smaller impact on monthly temperature response and [..[41]]therefore do not need to be explicitly included in our framework. Given the modular approach we take however, this assumption could potentially be

100   remedied in the future with the addition of separate modules which isolate the signals of [..[42]]other such forcings.

3.2.1   Mean response module

The mean response module was conceived to simply but convincingly represent the monthly [..[46]]mean response to yearly temperature, as well as the changes in the amplitude of the seasonal cycle with yearly temperatures (intra-annual variability) at a grid-point level. To that end, we employ a harmonic model consisting of a Fourier Series, [..[47]]which conserves both month-

105   to-month correlations and yearly temperatures. The amplitude terms fitted as linear combinations of yearly temperature, centred around a linear function of yearly temperature as shown in equation 2,

$$T^{meanresp.}_{m,s,y} \; f_s(T_{s,y}, m) = \beta_{0,s} + \beta_{1,s} \cdot T_{s,y} + \sum_{i=1}^{n} \left[g_{i,s}(T_{s,y}) \cdot sin(\frac{i\pi(m\%12+1)}{6}) + h_{i,s}(T_{s,y}) \cdot cos(\frac{i\pi(m\%12+1)}{6}) \right] \quad (2)$$

where T is temperature, m, s and y refer to month, space (i.e. grid-point) and year indexes, $\%$ is modulo and g_i and h_i are linear functions of [..[48]]T for the ith ordered term of the Fourier Series. [..[49]]Since the monthly cycle revolves around its

110   yearly temperature, fitting results for [..[50]]β_0 and β_1 coefficients had negligible effects (β_0=0 and β_0=1) and for simplicity's sake, we show the Fourier Series as centred around yearly temperature values within the [..[51]]Results section. In choosing the order (n) of the harmonic model we sought to have the optimal compromise between model complexity and accuracy. [..[52]]We balance between the model complexity and accuracy by calculating the Bayesian Information Criterion (BIC) of the harmonic models fitted [..[53]]for each ESM and each grid point with orders $n = 1, \ldots, 6$ and choosing the order with the

115   lowest BIC score.

---

[37]removed: .

[38]removed: A second assumption made is that other external

[39]removed: changes in

[40]removed: generally little impact on the

[41]removed: are therefore not

[42]removed: such external

[46]removed: cycle's

[47]removed: with

[48]removed: $T_{s,y}$

[49]removed: An added benefit of using a Fourier Series is that both monthto-month correlations and yearly mean temperatures are well conserved.

[50]removed: β0 and β1

[51]removed: results

[52]removed: For this, at each grid point we calculated the

[53]removed: with orders $n = 1, \ldots, 8$ and chose
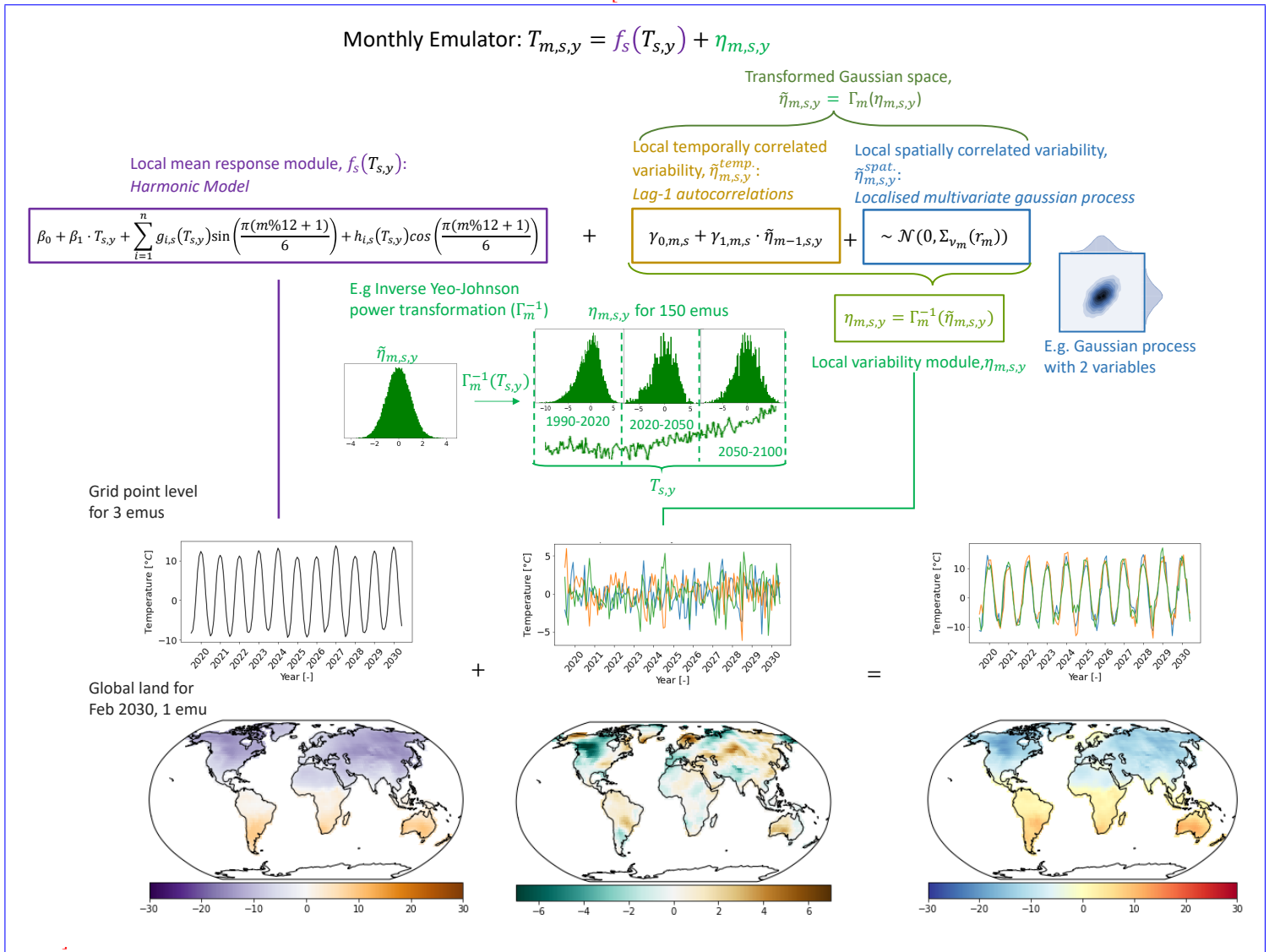
Figure 1. [..[44]]Modular framework for [..[45]]the monthly emulator illustrated for an example ESM. The local mean response module is represented using a harmonic model (Fourier Series). The local variability module starts with a Yeo-Johnson transformation before fitting the AR(1) process (distinguished into temporal-varying $\tilde{\eta}_{m,s,y}^{temp.}$ and spatial-varying $\tilde{\eta}_{m,s,y}^{spat.}$ components). An inverse Yeo-Johnson transformation is applied on emulated realizations drawn from the AR(1) process. Example data are also shown for the multivariate gaussian process used within $\tilde{\eta}_{m,s,y}^{space}$ and the inverse Yeo-Johnson transformation. 3 emulation examples are shown as time series for one grid point, and 1 emulation example is also shown as a global map.

3.2.2 Residual variability module

The local residual variability (otherwise simply called 'residuals'), i.e. the difference between actual local monthly temperature and its mean monthly response to variations in local annual temperature (given by the harmonic model[..[54]]), is assumed to be [..[55]]solely a manifestation of [..[56]]intra-annual variability processes. It thus can be thought of as short-term spatio-temporally correlated patterns. [..[57]]Similar to previous MESMER developments[..[58]]

[..[59]]

[..[60]][..[61]], we represent the residual variability [..[62]]

[..[63]]

[..[64]]

[..[65]]using an AR(1) process. Since the residual variability distributions display month-dependent [..[66]]skewness (see example Shapiro-Wilkes test for January and July in appendix [..[67]]B), which is non-stationary with regard to yearly temperatures, we first transform them into a stationary Gaussian space before fitting the AR(1) process. This strategy has already been pursued in problems concerning skewness in residual precipitation variability (Frei and Isotta, 2019). Precisely, we use the monotonic Yeo-Johnson transformation ($\Gamma_{m,s}$), which accounts for both positive and negative data values, to locally normalise monthly residuals (Yeo and Johnson, 2000).

$$\Gamma_{\mathsf{m,s}}(\eta_{\mathsf{m,s,y}}, \lambda) = \begin{cases} \frac{(\eta_{m,s,y}+1)^{\lambda}-1}{\lambda} & \eta_{m,s,y} \geq 0, \quad \lambda \neq 0 \\ \log(\eta_{m,s,y}+1) & \eta_{m,s,y} \geq 0, \quad \lambda = 0 \\ -\frac{(-\eta_{m,s,y}+1)^{2-\lambda}-1}{2-\lambda} & \eta_{m,s,y} < 0, \quad \lambda \neq 2 \\ -\log(-\eta_{m,s,y}+1) & \eta_{m,s,y} < 0, \quad \lambda = 2. \end{cases} \tag{3}$$

---

[54]removed: described under section 4.1.1

[55]removed: mostly

[56]removed: interannual

[57]removed: We follow an approach similar

[58]removed: and divide this module further into two parts as shown in equation 5 (Beusch et al., 2020).

[60]removed: Where $\eta_{m,s,y}$ is the total residual variability (otherwise simply called 'residuals') at month, grid-point and year m, s and y and $\eta_{m,s,y}^{time}$ and $\eta_{m,s,y}^{space}$ are its respective time-dependent and space-dependent components.

[61]removed: After a time lag of 1

[62]removed: is expected to have rapidly decaying covariation such that long term patterns (if any) covary with yearly temperature. $\eta_{m,s,y}^{time}$ is thus modelled using lag-1 autocorrelations as shown in equation 6.

[64]removed: Where $\gamma_{0,m,s}$ and $\gamma_{1,m,s}$ are coefficients fitted per month.

[65]removed: Following this, the rest of the residual variability , represented as $\eta_{m,s,y}^{space}$, requires to not only render spatial correlation structures but also the

[66]removed: , non-stationary skewness within residuals (see

[67]removed: C). Since operating in a Gaussian data space is a prerequisite to infer correlation structures in stochastic modelling, we use a trans-Gaussian process relying on a power transformation to normalise data before sampling from it

$\Gamma_{m,s}$ relies on a λ parameter to deduce the shape of a distribution and normalise accordingly. Non-stationarity in [..[68]]month-specific residual skewness with regard to yearly temperature is taken into account by defining the λ parameter as a logistic function of yearly temperature, as shown in equation 4[..[69]],

$$\lambda_{m,s,y} = \frac{2}{1 + \xi_{0,m,s} \cdot e^{\xi_{1,m,s} \cdot T_{s,y}}} \tag{4}$$

[..[70]]where $\xi_{0,m,s}$ and $\xi_{1,m,s}$ are coefficients fitted using maximum likelihood. The performance improvement of having a $\lambda_{m,s,y}$ parameter that is dependent on yearly temperature compared to a case where it is invariant ($\lambda_{m,s}$) is demonstrated by additional tests shown in the Appendix [..[71]]B (shown for January and July). Since majority of ESMs profit from yearly temperature dependence in λ, we for simplicity and consistency reasons use $\lambda_{m,s,y}$ in all ESMs [..[72]]

The AR(1) process is applied on the locally power transformed residual variability. Following the approach of (Beusch et al., 2020), temporal correlations are accounted for in the deterministic component of the AR(1) process whereas spatial correlations are accounted for in the stochastic noise part, and hence $\widetilde{\eta}_{m,s,y}$ can be expressed as shown in equation 5,

$$\widetilde{\eta}_{m,s,y} = \widetilde{\eta}^{\text{temp.}}_{m,s,y} + \widetilde{\eta}^{\text{spat.}}_{m,s,y} \tag{5}$$

where $\widetilde{\eta}_{m,s,y}$ is the locally normalised residual variability at month, grid-point and year m, s and y and $\widetilde{\eta}^{temp.}_{m,s,y}$ and $\widetilde{\eta}^{spat.}_{m,s,y}$ are its respective temporally-varying and spatially-varying components.

The AR(1) process accounts for autocorrelations up to a time lag of 1, and is suited in representing the residual variability which is assumed to have rapidly decaying covariation such that longer term patterns (if any) covary with yearly temperature. $\widetilde{\eta}^{temp.}_{m,s,y}$ is shown in equation 6,

$$\widetilde{\eta}^{\text{temp.}}_{m,s,y} = \gamma_{0,m,s} + \gamma_{1,m,s} \cdot \widetilde{\eta}_{m-1,s,y} \tag{6}$$

where $\gamma_{0,m,s}$ and $\gamma_{1,m,s}$ are coefficients fitted per month. $\gamma_{1,m,s}$ is constrained to be between and including -1 and 1.

Following this, $\widetilde{\eta}^{spat.}_{m,s,y}$ (i.e. the stochastic noise part of the AR(1) process) needs to account for the spatial cross-correlations between grid-points. It is modelled through a localised monthly multivariate [..[73]]gaussian process, and thus [..[74]]dampens spatial covariations with increasing distance, as shown in equation 7[..[75]],

$$[..[76]]\widetilde{\eta}^{\text{spat.}}{}_{m,s,y} \sim [..[77]]\mathcal{N}(0, \Sigma_{\nu_m}(r_m))[..[78]] \text{ with } 1500\text{km} \leq r_m \leq 8000\text{km} \tag{7}$$

---

[68]removed: monthly

[69]removed: .

[70]removed: Where

[71]removed: D (only

[72]removed: $\eta^{space}_{m,s,y}$, is modelled using the transformed residuals

[73]removed: trans-gaussian

[74]removed: preserves spatial correlation structures up to a certain

[75]removed: .

[..[79]]where $\mathcal{N}(0, \Sigma_{\nu_m}(r_m))$ is a multivariate [..[80]]gaussian process with means 0 and covariance matrix, Σ_{ν_m}[..[81]]. As the number of land grid-points are much more than the number of temperature field samples, Σ_{ν_m} is rank deficient and is thus localized by point-wise multiplication with the smooth Gaspari Cohn correlation function (Gaspari and Cohn, 1999) which has exponentially vanishing correlations with distance r_m and was used in previous MESMER fittings (Beusch et al., 2020). r_m is chosen per month in a similar cross-validation with a leave-one-out approach as previous MESMER fittings (Beusch et al., 2020) using distances of 1500 km to 8000 km at 250 km intervals. The localized empirical covariance matrix, Σ_{ν_m}, is derived [..[82]]analytically based on the mathematical expectations [..[83]]for the covariance of innovations of an AR(1) process (Matalas, 1967; Richardson, 1981), as shown in Equation 8[..[84]],

$$\Sigma_{\nu_m}(r_m) = \sqrt{1 - \gamma_{1,m,i}^2} \cdot \sqrt{1 - \gamma_{1,m,j}^2} \cdot \Sigma_{\widetilde{\eta}_m}(r_m) \tag{8}$$

[..[85]]where $\Sigma_{\widetilde{\eta}_m}$ is the empirical covariance matrix constructed across all grid points for a given month from the locally normalised empirical residuals. When generating emulated realizations from the AR(1) process, we apply the inverse Yeo-Johnson transformation to obtain the final residual variability terms.

$$\eta_{m,s,y} = \Gamma_{m,s}^{-1}(\widetilde{\eta}_{m,s,y}) \tag{9}$$

3.3 Evaluating emulator performance

3.3.1 Mean response verification

To evaluate how well the [..[86]]seasonal cycle's mean response, $f_s(T_{s,y})$, is captured we calculate [..[87]]Pearson Correlation coefficients between the emulated values obtained from the harmonic model and their training run values across the whole globe for each month, with each grid point weighted equally. This not only gives an idea of how well the magnitude of mean response changes correspond, but also how in phase the emulated monthly cycles are with training run monthly cycles. Where test runs are available, their correlations to the harmonic model results are also calculated to assess how well the harmonic model can represent data it [..[88]]has not been trained on. Ideally, the test run correlations should be more or less equal to those of the training runs, with anything [..[89]]substantially lower indicating overfitting and anything [..[90]]substantially higher

---

[79]removed: Where $\Gamma_{m,s}^{-1}$ is a local inverse Yeo-Johnson transformation and

[80]removed: trans-gaussian

[81]removed: , localised using a Gaspari-Cohn function (Gaspari and Cohn, 1999) with radius

[82]removed: based off

[83]removed: of the residual variabilities' standard deviations once lag-1 autocorrelations are removed, so as to decouple the spatial component from the temporal component

[84]removed: .

[85]removed: Where

[86]removed: monthly

[87]removed: the grid-point level

[88]removed: hasn't

[89]removed: much

[90]removed: much

indicating an [..[91]]non-representative training set (i.e. further modifications in the train to test splitting would have to be considered).

3.3.2   Residual variability verification

In order to evaluate how well the emulator reproduces the deviations from the harmonic model simulated by the ESMs, 50 emulations are generated per training run. First, we check that short-term temporal features are sufficiently captured for each individual grid point: each residual variability sequence is decomposed into its continuous power spectra, from which we verify, by computing Pearson Correlation coefficients, that the [..[92]]top 50 highest [..[93]]frequency bands within the training run residual variabilities appear with similar power spectra in the corresponding emulated residual variabilities. Second, we verify [..[94]]to which degree the spatial covariance structure is [..[95]]preserved by calculating monthly spatial cross correlations across the residuals produced by each individual emulation and, by using Pearson Correlations, comparing them to those of their respective training runs. Where test runs are available, a similar verification between them and training runs is done, thus yielding an approximation of how actual ESM initial-condition ensemble members [..[96]]would relate to each other.

3.3.3   Regional-scale ensemble reliability verification

The [..[97]]full emulator, consisting of both the mean response (i.e. harmonic model) and residual variabilty modules, is evaluated for its representation of regionally area-weighted, averaged monthly temperatures of all 26 SREX regions [..[98]](Seneviratne et al. (2012), see Appendix D for details on SREX regions) [..[99]]at each individual month. Global land results always constitute area weighted averages across all land grid points excluding Antarctica. We assess how well the emulator can reproduce the 5th, 50th and 95th quantiles of the respective ESM initial-condition ensemble over the periods of 1870-2000 and 2000-2100, by means of quantile deviations [..[100]]as previously done by Beusch et al. (2020). A step-by-step process for calculating monthly quantile deviations of the ESM from the emulator is as follows:

1. Calculate the regional average of a given SREX region for each ESM run and each of its respective emulations at a given month.

2. Extract the qth emulated quantile at each time step from the [..[101]]set of regionally averaged emulations.

---

[91]removed: under representative

[92]removed: frequency bands with the

[93]removed: power spectra

[94]removed: that

[95]removed: maintained

[96]removed: should

[97]removed: emulated monthly temperature values are evaluated across

[98]removed: (Seneviratne et al., 2012) for each individual month (see Appendix B

[99]removed: .

[100]removed: following the same approach as Beusch et al. (2020). To do so, we first extract the emulated quantile values for

[101]removed: distribution spanning the full set of emulations at that time step, thus creating smooth time series of the emulated quantiles.For each of the above named time periods, we then compute the ESM's quantile respective to the emulated quantile as the proportion of time steps that the ESM training runs-and where available test runs -are below the emulated quantile values. The

3. Within a defined time period (e.g. 1870-2000) calculate the proportion of time-steps the rergionally averaged ESM value is less than the respective emulated q^{th} quantile value. The resulting number is q_{ESM}

4. The quantile deviation of the ESM [..[102]]from the [..[103]]

   emulator is then given as $q_{ESM} - q$

205 By drawing comparisons between the quantile deviations obtained at the 2 time periods considered, we can evaluate whether inter-annual variations in monthly temperature distributions are sufficiently captured. Since the magnitude of global warming varies between both time periods, such comparison will additionally help pinpoint whether the emulator sufficiently captures the expected changes in temperature skewness under a changing climate (Wang et al., 2017; Sheridan and Lee, 2018; Tamarin-Brodsky et al., 2020).

3.4 Benchmarking MESMER-M using a simple physical approach

210 Any variability in monthly temperatures that cannot be explained by variability in yearly temperature alone, is stochastically accounted for in MESMER-M[..[104]]'s local residual variability module (see Sect. 3.2.2), following existing downscaling theory (Berner et al., 2017; Arnold, 2001). Hence, month and season-dependent variability linked to physical drivers such as atmospheric processes (Neale et al., 2008; Deser et al., 2012), e.g.[..[105]], the El Niño-Southern Oscillation 215 (ENSO), and biophysical feedbacks (Potopová et al., 2016; Xu and Dirmeyer, 2011; Jaeger and Seneviratne, 2011; King, 2019; Tamarin-Brodsky et al., 2020), e.g., snow-albedo feedbacks, is not explicitly modelled but instead represented by a stochastic process. Nevertheless, first order changes in the characteristics of these variabilities across warming levels can be approximated within MESMER-M since the skewness of MESMER-M's [..[106]]residual variability emulations depends on the yearly temperature. In this section, we delineate a framework to verify that this statistical approach, based on a single 220 input variable of yearly temperature, can sufficiently imitate properties of the monthly temperature response which otherwise result from intra-annual variability processes. We primarily verify for representation of secondary biophysical feedbacks [..[107]]as biophysical variables are obtainable as direct output from the ESMs, whereas accounting for modes of climate variability and atmospheric processes would require further data processing and analysis. Furthermore, some effects of

---

[102]removed: computed quantile from the actual emulated quantile is then assessed, where positive deviations mean that the emulated quantile is warmer than that of

[103]removed: ESM run and vice versa.

[104]removed: is designed to provide a purely statistical representation of monthly temperatures in a spatially resolved manner, taking only local yearly temperatures as input. Intra-annual processes which are otherwise unapparent on yearly timescales may however emerge on monthly timescales, introducing additional variability and -due to possible secondary biophysical feedbacks- month and season specific non-uniformities (

[105]removed: non-linearities, non stationarities, changes in distributional properties)(Potopová et al., 2016; Xu and Dirmeyer, 2011; Jaeger and Seneviratne, 2011; King, 2019; Tamarin-Brodsky et al., 2020) into the monthly temperature response.As described in Section 3.1.2 and following existing downscaling theory (Berner et al., 2017; Arnold, 2001), such variability is stochastically accounted for

[106]removed: local residual variability module by sampling month-specific, spatially correlated variability terms from a transformed Gaussian space, where non-uniformities are somewhat also inferred by allowing the skewness to covary with local yearly temperatures

[107]removed: .

atmospheric processes can follow from or manifest in biophysical variables, e.g. as seen by Allen and Zender (2011), and
225 hence are indirectly accounted for by using biophysical variables.

To isolate the contribution of secondary biophysical feedbacks to the monthly temperature response, we consider them as inducing the residual differences between the ESM and harmonic model realisations. This follows from the harmonic model representing the expected direct mean response to evolving yearly temperatures, with any systematic departure from it being driven by secondary forcers. To rudimentarily represent these contributions, a simple, [..[108]]physically-informed model
230 consisting of a suite of Gradient Boosting Regressors (GBRs) (Hastie et al., 2009) is built for each ESM. Each GBR within the suite represents one grid point and is trained to predict the local residual differences using local biophysical variable values (see Table [..[109]]1) as predictors. Predictors are chosen so as to best represent the intra-annual variation of radiative and thermal fluxes alongside their evolution under changing yearly temperatures. The list of predictors is complemented by local yearly temperature values and month values in their harmonic form (hence [..[110]]$\frac{\pi(n\%12+1)}{6}$, $n = 1$ for January, etc.) to
235 account for month dependencies in [..[111]]the residual differences and yearly temperature influences (if any) left behind within the residuals.

To optimise the selection of the biophysical variables used as predictors, we first compare the performance of different [..[112]]physically-informed models trained using different sets of biophysical variables for each ESM. [..[113]]The best globally performing model is selected as a benchmark to assess how well the residual variability module, described in Section
240 3.2.2, statistically represents properties within the monthly response arising from secondary biophysical feedbacks. Pearson Correlations, between ESM test runs and harmonic model test results augmented by biophysical variable[..[114]], T_{yr} and month based physically-informed model predictions are calculated. As a measure of performance the aforementioned correlation values are given relative to those obtained when augmenting using only T_{yr} and month based [..[115]]physically-informed model predictions. This additionally allows determination of whether improvement in residual representation comes
245 from the added biophysical variable information and if so where and how. [..[116]]As we are most interested in the representation of monthly temperature distributions and the influences of biophysical feedbacks therein, we [..[117]]consider the energy distances [..[118]]of the benchmark, "[..[119]]physically-informed" emulations [..[120]]– constituting the mean response with GBR

---

[108]removed: physical

[109]removed: 1

[110]removed: $\frac{\pi(1\%12+1)}{6}$

[111]removed: residual variabilities

[112]removed: physical

[113]removed: Pearson Correlations calculated over all months

[114]removed: based physical model predictions

[115]removed: physical model predictionsare used as a measure of performance

[116]removed: The best globally performing model is selected as a benchmark to assess how well the residual variability module, described in Section 3.1.2, statistically represents properties within the monthly response arising from secondary biophysical feedbacks.

[117]removed: compare

[118]removed: from the actual ESM runs

[119]removed: physical

[120]removed: -constituting

predicted residuals added [..[121]]ontop – from the actual ESM runs and compare them to the energy distances of the statistical emulations [..[122]]– constituting the mean response with residuals from the residual variability module added ontop (as described in Section [..[123]]3.2) – from the actual ESM runs. The energy distance is a non-parametric estimate of the distance between two cumulative distribution functions (cdfs), x and y, by considering all their independent pairs of variables, X_i, X_j (i.e. pairs of physically-informed/statistical emulated values) and Y_k, Y_l [..[124]](i.e. pairs of actual ESM values) respectively.

$$D(x,y) = (2E||X_i - Y_k|| - E||X_i - X_j|| - E||Y_k - Y_l||)^{\frac{1}{2}} \tag{10}$$

Time series of the biophysical variables are obtained from CMIP6 runs. For this analysis, we only focus on ESMs which provided data for all 5 biophysical variables under consideration, for both the test and training runs used during emulator calibration.

Table 1. List of Biophysical Variables used in training the Gradient Boosting Regressor

	Variable	Abbreviation
Biophysical	Albedo	A
	Snow Cover Fraction (%)	S
	Cloud Cover Fraction (%)	C
	Sensible Heat Flux (Wm^{-2})	Hs
	Latent Heat Flux (Wm^{-2})	Hl
Other	Yearly Temperature	T_{yr}
	Month ($\frac{\pi(1\%12+1)}{6}$ ... $\frac{\pi(12\%12+1)}{6}$)	month

4 Illustration of emulator attributes

4.1 Calibration results

When calibrating the harmonic model constituting the mean response module, highest orders of the Fourier series were found in tropical to sub-tropical regions where the seasonal cycle contains a relatively small shift in temperature values (first row, Figure 2). The Arctic also displays relatively high orders chosen within the Fourier series, possibly due to higher variabilities in the response of the seasonal cycle shape with increasing yearly temperatures. In contrast, temperate regions which possess distinctly sinusoidal seasonal cycles with marked snow-driven summer to winter transitions display relatively lower orders. CanESM5 and MIROC6 show the overall highest orders, this can be traced back to the availability of significantly more

---

[121]removed: ontop- to those

[122]removed: -constituting

[123]removed: 3.1)

[124]removed: respectively

training runs, hence more information on which to train the emulator, for these two ESMs allowing for more model complexity without compromising on accuracy (refer to Table A1).

The residual variability module calibration results are shown in Figure 2 for January and July[..[125]]

. The average Yeo-johnson lambda parameter ([..[126]]$\widetilde{\lambda}_{m,s}$) displays a shift of values greater than 1 to values close to 1 in the Northern Hemisphere (30°-50°) between January and July. In general, [..[127]]$\widetilde{\lambda}_{m,s}$ values greater (less) than 1 indicate a concave (convex) transformation function owing to negative (positive) skewness, while values equal to 1 suggest minimal skewness in the input distribution (Yeo and Johnson, 2000). This explains the seasonality in [..[128]]$\widetilde{\lambda}_{m,s}$ as we expect a more negatively skewed residual distribution in the winter when the snow-albedo feedback leads to a non-linear winter-time warming (Cohen and Rind, 1991; Hall, 2004; Colman, 2013; Thackeray et al., 2019) causing the harmonic model to overestimate the mean temperature response. July displays significantly high [..[129]]$\widetilde{\lambda}_{m,s}$ values for polar latitudes (>80°) explainable by the sudden increase in warming rates during ice-free summers (Blackport and Kushner, 2016). Around the equator (-5° to 5°) we see [..[130]]$\widetilde{\lambda}_{m,s}$ values consistently higher than 1 especially in the month of July, with INM-CM5-8 and INM-CM5-0 displaying significantly high values. The source of [..[131]]the aforementioned varies model to model but mainly originates from the North-West South America and Sahel regions, alluding to the presence of some non-linear warming response in these regions.

The lag-1 auto-correlation coefficients ($\gamma_{1,m,s}$) mostly exhibit positive values across all ESMs for January, with at least 70% of grid points having values between 0 and 0.3, suggesting minimal month-to-month memory additional to the seasonal cycle. Apart from some ESMs which are more normally distributed in $\gamma_{1,m,s}$, July shows similar behaviour albeit with a larger spread in values. ACCESS-CM2 and HadGEM3-GC31-LL present themselves as outliers here with the bulk of their $\gamma_{1,m,s}$ coefficients centred around 0 for both January and July, indicating negligible autocorrelations.

Localisation radii vary from model to model and are generally higher in January than July reflecting seasonal differences in residual behaviour possibly due to boreal winter snow cover yielding larger spatial patterns. CanESM5 and MIROC6 display notably higher localisation radii, which can again be tracked back to them having more training runs: more [..[132]]time-steps are available during the leave-one-out [..[133]]cross-validation, thus making it generally possible to robustly estimate spatial correlations up to higher distances, which in turn leads to selecting larger localistaion radii. It should however be stressed

---

[125]removed: (other months are available upon request). The lag-1 auto-correlation coefficients ($\gamma_{1,m,s}$) mostly exhibit positive values across all ESMs for January, with at least 70% of grid points having values between 0 and 0.3, suggesting minimal month-to-month memory ontop of the seasonal cycle. Apart from some ESMs which are more normally distributed in $\gamma_{1,m,s}$, July shows similar behaviour albeit with a larger spread in values. HadGeM3-GC31-LL and ACCESS-CM2 present themselves as outliers here with the bulk of their $\gamma_{1,m,s}$ coefficients centred around 0 for both January and July, indicating negligible autocorrelations.

[126]removed: $\widetilde{\eta}_{m,s}$

[127]removed: $\widetilde{\eta}_{m,s}$

[128]removed: $\widetilde{\eta}_{m,s}$

[129]removed: $\widetilde{\eta}_{m,s}$

[130]removed: $\widetilde{\eta}_{m,s}$ values close to but slightly

[131]removed: this

[132]removed: information is available in the covariance matrix constructed, thus higher localisation radii perform well in fitting using the

[133]removed: approach.

that the the ESM itself is the main driver behind the calibration results (e.g. even with only one ensemble member MCM-UA-1-0 has a high localisation radius).
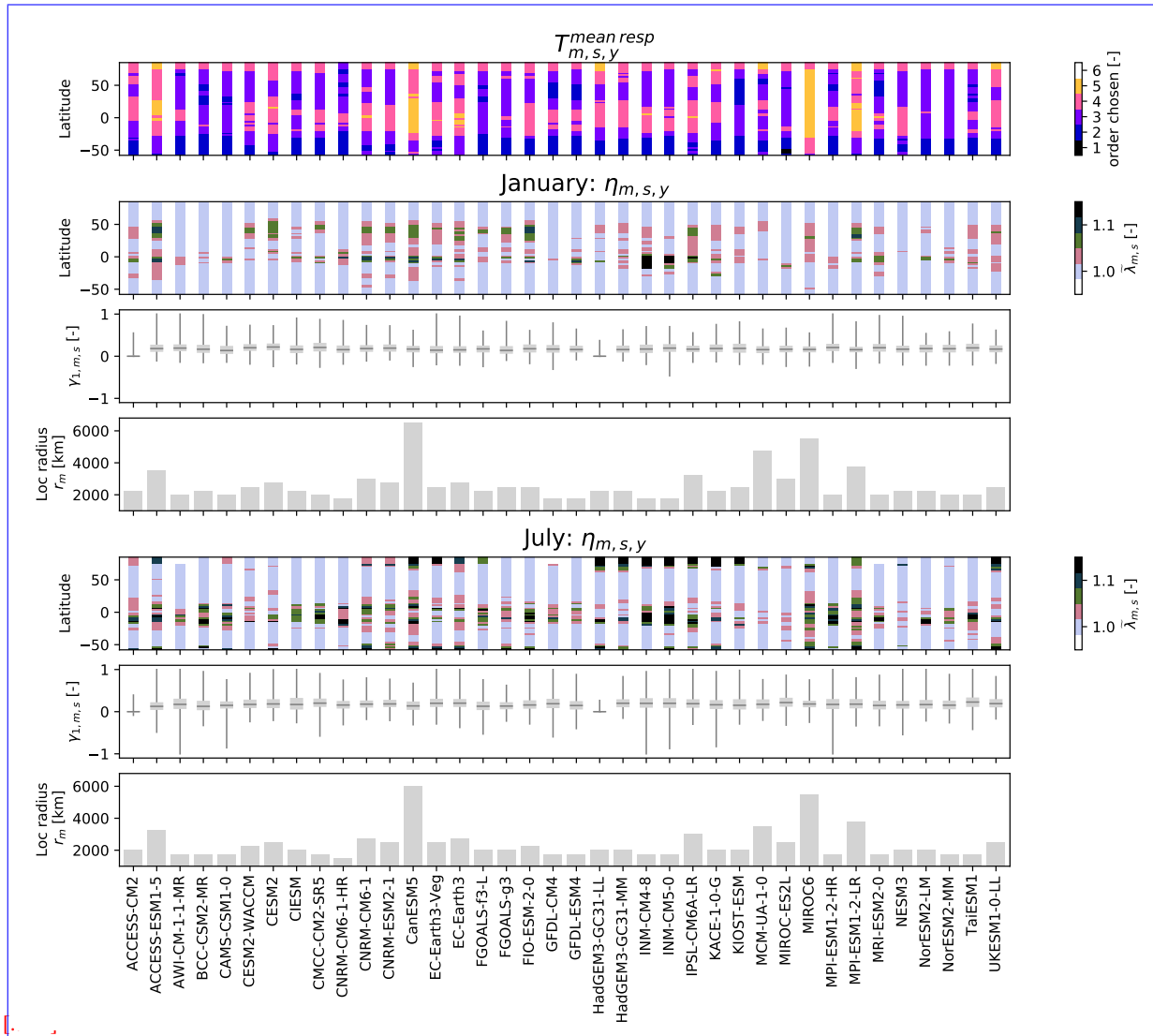
Figure 3. Calibration parameters obtained from emulator fittings for all CMIP6 models. For the mean response module, the latitudinally averaged order of the Fourier Series considered in the harmonic model is plotted against latitude (row 1). For the monthly residual variability module, parameters are displayed for January (row 2-4) and July (row 5-7). $\lambda_{m,s,y}$ coefficients averaged over latitude and all years for the local yeo-johnson transformation are plotted against latitude (row 2 and 4). The local lag-1 autocorrelation coefficients are plotted as boxplots (row 3 and 6) with whiskers covering the 0 to 1 quantile range, and the localisation radii are given as bar charts (row 4 and 7).

4.2 Regional behaviour for [..[135]]four selected ESMs

To illustrate the regional behaviour of the calibrated emulator, we focus on for four select ESMs and individually consider the emulations created by the mean response module the residual variability module. Figure 3 visually demonstrates the harmonic model constituting the mean response module (bold black lines) capturing the mean monthly temperature response for both January and July, at global and regional scales (here we show the SREX regions [..[136]]West North America, WNA, and West Africa, WAF), across all [..[137]]four ESMs. The remaining natural variability surrounding the mean response displays a month dependency across [..[138]]the four ESMs, such that January variabilities are up to double that of July both globally and in the displayed regions. These month dependencies in variabilities are well accounted for within the full emulations [..[139]]comprising both the [..[140]]mean response and the residual variability module, [..[141]]highlighting the necessity of a month specific residual variability module.

Figure 5 shows the trends in the variance of each year's monthly temperatures around the yearly mean (i.e. variance in intra-annual [..[142]]temperatures) with evolving yearly temperatures.The harmonic model is able to capture the general trends displayed by the ESMs, albeit not being able to fully account for non-linearities within them[..[143]]. For example, MPI-ESM1-2-LR [..[144]]displays an increase in [..[145]]intra-annual variance followed by stabilisation [..[146]]in WAF, which is represented by the harmonic model as a linear increase with no stabilisation[..[147]]. Construction of the [..[148]]physically-informed model outlined in Section 3.4 elucidated albedo as the main covariant to monthly temperature variability in WAF for MPI-ESM1-2-LR (see Section 5.4), indicating changes in land surface properties (possibly due to the reduction in tree cover in this region) as driving intra-annual variance changes. Such demonstrates the limitation of solely relying on yearly temperatures as input towards predicting monthly temperatures when other forcings (in this case changes in land surfaces) dominate. Other forcings rarely dominate over the yearly temperature mean response however, and usually on regional levels over short time-scales, hence the full emulations is nevertheless able to capture the overall spread.

---

[135] removed: a select 4

[136] removed: WNA and

[137] removed: 4

[138] removed: all

[139] removed: (grey lines)

[140] removed: harmonic model

[141] removed: justifying the prescription

[142] removed: variability

[143] removed: (for examplefor

[144] removed: in WAF, where

[145] removed: variability

[146] removed: is observed, and

[147] removed:). This suggests the presence of secondary, non-linear processes which cannot be fully accounted for by yearly temperature only. Where the harmonic model falls short in representing non-linearities in

[148] removed: response of the seasonal cycle's amplitude, the addition of residual variability to generate

Figure 4. Regionally averaged temperature time series (rows) of January and July for four example ESMs (columns). Three ESM ensemble runs (coloured), their respective harmonic model results (black) and 50 full emulations for each of the three patterns (grey colourscale) are plotted. Temperature values are given as anomalies with respect to the annual climatological mean [..149]over the reference period of 1870-1899. The regions are from top to bottom: global land without Antarctica, West North America (WNA), and West Africa (WAF).[..150]
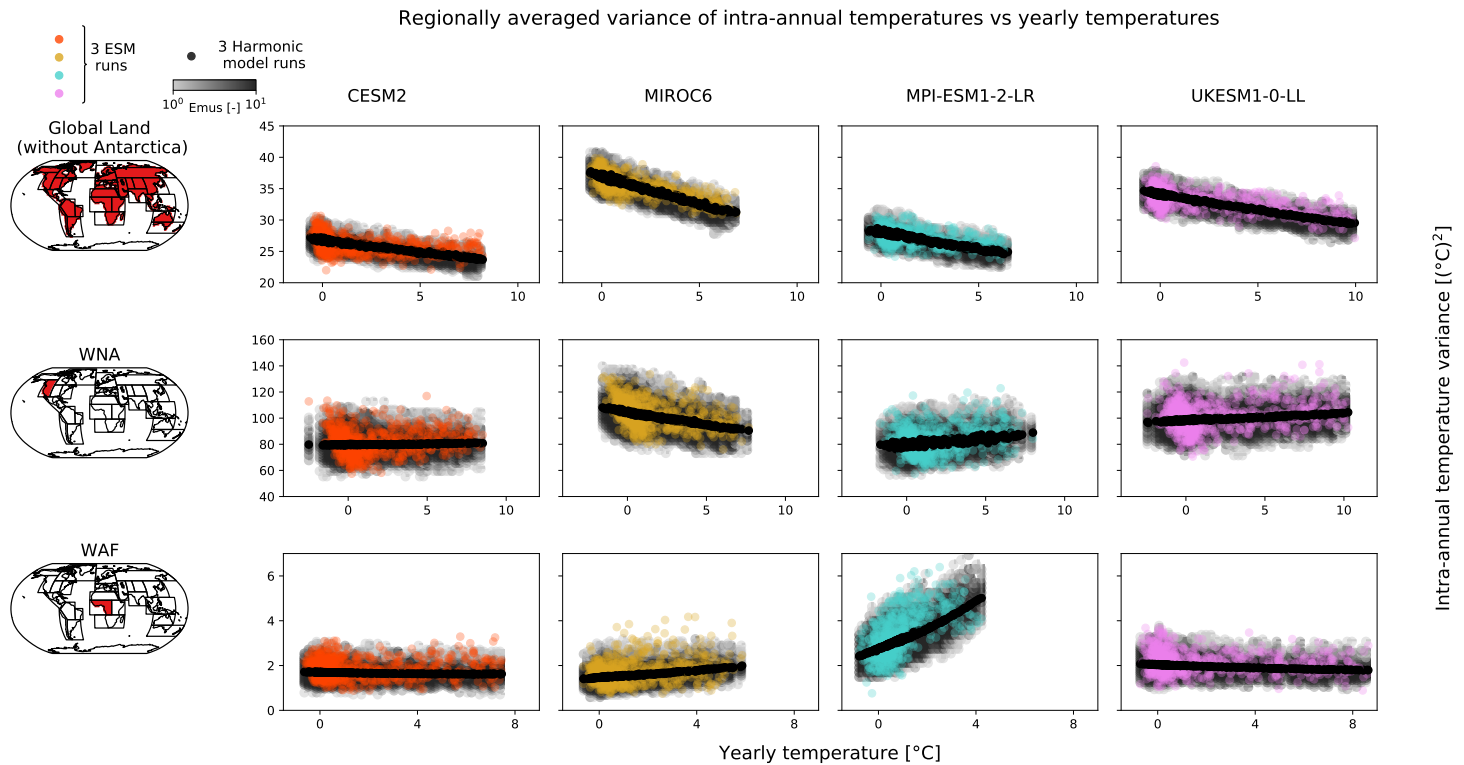
Figure 5. Regionally averaged variance of intra-annual [..¹⁵¹]temperatures (i.e. [..¹⁵²]variance of each year's monthly temperatures around the yearly mean) scatter plotted against yearly temperature (rows) for four example ESMs (columns).Three ESM ensemble runs (coloured), their respective harmonic model results (black) and 50 full emulations for each of the three patterns (grey colourscale) are plotted. Temperature values are taken as anomalies with respect to [..¹⁵³]annual climatological mean [..¹⁵⁴]over the reference period of 1870-1899. Each dot represents the temperature variance calculated from the monthly values for one individual year. The regions are from top to bottom: global land without Antarctica, West North America (WNA) and West Africa (WAF).[..¹⁵⁵]

5 Evaluating emulator performance

5.1 Mean response verification

We evaluate the ability of the harmonic model, constituting the mean response module, in capturing the seasonal cycle's response to evolving yearly temperatures. Pearson Correlations between the harmonic model and ESM training runs range from 0.7 up to almost 1 (Figure 6). Summer months exhibit the highest correlations while transition months of spring and autumn (March, April) have the lowest correlations. Such low correlations could result from the [..¹⁵⁶]inter-annual spread in the timing of snow cover decrease and increase, such that the mean response extracted does not [..¹⁵⁷]always match individual

---

¹⁵⁶removed: ESM initial-condition ensemble spread in representing

¹⁵⁷removed: fully match any individual ensemble member alone

years. Winter month correlations are generally higher than those of transition months but lower than those of summer months. This is possibly due to snow-albedo feedbacks, which induce non-linearities into the winter period mean response (Cohen and Rind, 1991; Hall, 2004; Colman, 2013; Thackeray et al., 2019) [..¹⁵⁸]leading to lower correlations than those of summer months where the response is more linear. Overall the training run correlations correspond well to test run correlations (where available) confirming good data representation within the training set and minimal (if any) model overfitting.

Figure 6. Local mean monthly response verification for [..¹⁵⁹]all CMIP6 models by means of Pearson Correlation between the harmonic model and training runs (indicated by the colour of the lower triangle), over all global land grid points (without Antarctica) for each month. The correlations between the harmonic model and test runs are given in colour in the upper triangles to see how well the harmonic model performs for data it has not seen yet (a grey upper triangle means that no test run is available for this model).[..¹⁶⁰]

325

---

¹⁵⁸removed: . Hence, even though the extracted mean response corresponds well to the individual ensemble members, correlations are still lower

5.2 Residual variability verification

To establish [..[161]]if temporal patterns within the ESM residual variabilities are successfully emulated, the correspondence of their respective power spectra [..[162]]at a grid point level is considered. Results shown in Figure 7 display the emulator's median correlations [..[163]]with the ESMs' training run power spectra lying between 0.9 and 1. This corresponds well [..[164]]to the correlations across the ESM test runs (crosses)[..[165]]. Correlations between the ESM training runs emulations for a given ESM display very little spread, which is in agreement with the near identical correlations seen amongst ESM test runs. In the example 2D histogram plot (given for CESM2), we see that the emulator [..[166]]is most successful in capturing lower power spectra [..[167]]to frequency ratios. This may be a consequence of the emulator design, as we restrict ourselves to considering only lag-1 autocorrelations such that [..[168]]lower frequencies with higher power spectra are not accounted for.[..[169]]

For verification of the residual variability's spatial component, we consider the spatial cross-correlations within [..[175]]four geographical bands centred around the grid cell for which temperature is being emulated (Figure 8) for [..[176]]the example months January and July[..[177]]. As the spatial covariance matrix within the emulator is localised (see Section [..[178]]3.2.2), its spatial cross- correlations are by design expected to diminish with increasing distances. Hence, we see the emulator performing best at distances below 1500 km, with median correlations of 0.91-0.99 which are in line with [..[179]]correlations between ESM test and training runs (crosses). Beyond 1500 km, the emulator performs progressively worse with correlations dropping below 0.1 for distances between 3000 km and 6000 km and staying there for distances larger than 6000 km, while those of test runs remain around 0.5-0.8. CanESM5 and MIROC6 are the two exceptions at distances between 3000 km and 6000 km, with correlations of 0.33-0.5 which then again drop to below 0.1 at distances larger than 6000 km. This is due to their notably larger localisation radii (see Figure 2), which leads to a slower decline of spatial cross-correlations with increasing distances as compared to other ESMs.

---

[161]removed: that

[162]removed: are

[163]removed: lying between 0.8 and 0.93.

[164]removed: - although is slightly lower than -

[165]removed: which are between 0.88 and 0.95. Correlations across

[166]removed: trained on that ESM is able to capture

[167]removed: , however, underestimates higher power spectra. This is

[168]removed: higher order temporal patterns

[169]removed: Nevertheless, the occurrence of such higher power spectra is low for the considered ESM and the overall bulk of the power spectra correlate well,

[175]removed: 4

[176]removed: each month (here we show only

[177]removed:)

[178]removed: 3.1

[179]removed: those of the ESM test
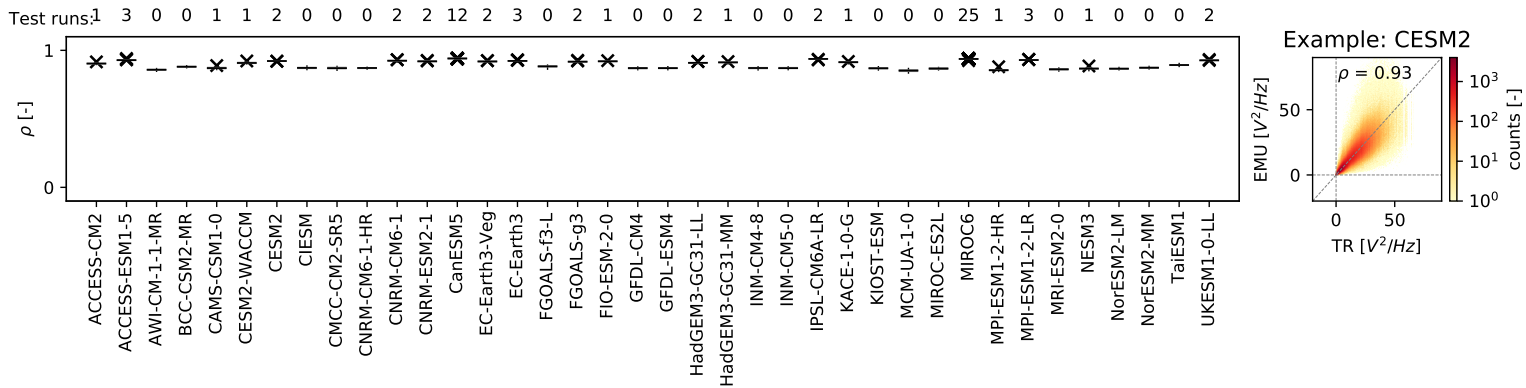
Figure 7. Verification of the time component of the local variability module by means of Pearson Correlations between the power spectra of the [..170]50 highest frequency bands present within the [..171]training runs (i.e. considering all months together) and the power spectra at which [..172]the same 50 frequency bands appear within the [..173]respective emulations (box plots, whiskers indicating 0.05 and 0.95 quantiles) calculated per grid point. 50 emulations are evaluated per training run. Where test runs are available, [..174]their correlations with training runs are also given (black crosses). The example 2D histogram shows the power spectra to frequency ratio for CESM2 training runs versus the corresponding power spectra to frequency ratio within its emulations.

5.3   Regional-scale ensemble reliability verification

Regionally aggregated [..182]5%, 50% and 95% quantile deviations of the ESM training (and where available test) runs from the [..183]emulated quantiles (derived from the full emulator consisting of both the mean response and local variability module) are plotted [..184]over the periods of 1870-2000 and 2000-2100 for example months January (Figures 9 and 11) and July (Figures 10 and 12)[..185].The 50% quantile deviations over the period of 1870-2000 in January and July (Figures 9 and 10, respectively) generally show low [..186]magnitudes (-3% to 3%). A slight regional dependency for this period is visible, where tropical/sub-tropical regions of AMZ, NEB, SSA, WAF, EAF, SAF have generally warmer (colder) emulated 50th quantiles as compared to the ESM runs, while those of the remaining regions are colder (warmer) for January (July). While January [..187]50% quantile deviations over the period of 2000-2100 remain low with less (if any) distinguishable regional dependency, July [..188]50% quantile deviations for this period increase (-10% to 10%) with an opposite pattern in regional dependency to that of 1870-2000. The increase for July in deviations could be a combined result of non-linear warming and

---

182 removed: 5th, 50th and 95th

183 removed: full emulations (

184 removed: for

185 removed: over the periods of 1870-2000 and 2000-2100 (other months are available upon request).The 50th

186 removed: values

187 removed: 50th

188 removed: 50th
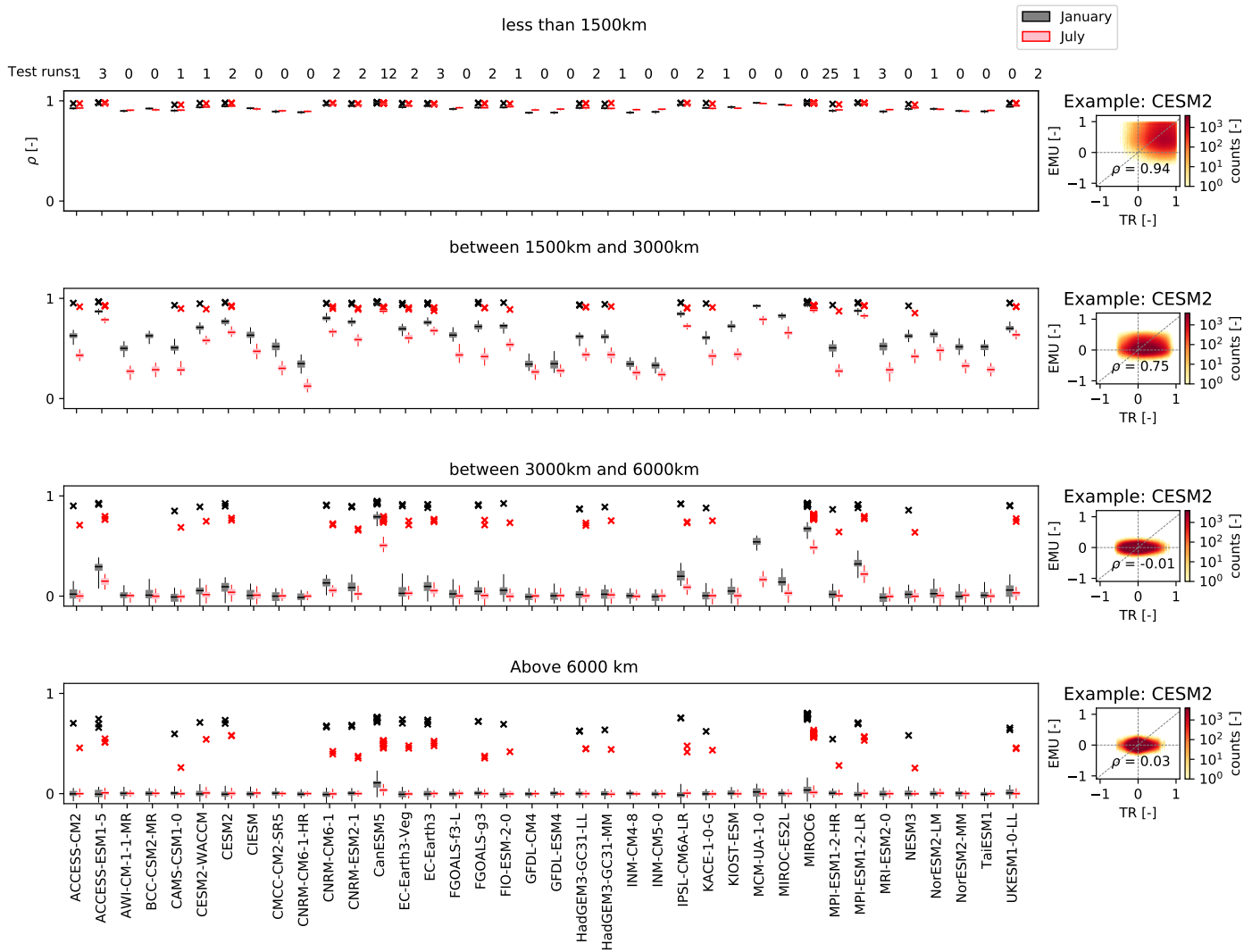
Figure 8. Verification of the spatial representation within the local variability module. Pearson correlations between ESM training run and emulated spatial cross-correlations are considered for [..[180]]four geographical bands centred around the grid cell for which temperature is being emulated (rows) at individual months of January (black boxplots) and July (red boxplots). Box plot whiskers indicate 5th and 95th quantiles. 50 emulations are evaluated per training run. Where ESM test runs are available, [..[181]]their correlations with training runs are also given for January (black crosses) and July (red crosses). Example 2D histograms of the January spatial cross-correlations for CESM2 training runs versus those of its emulations are given for each geographical band.

relatively lower variability in July temperature values as compared to those of January in the ESM simulations. [..[189]]This

---

[189]removed: Hence, the linear emulated mean response diverges from the ESM median warming, and the addition of residual variability cannot fully resolve this due to the overall higher signal to noise ratio. This shows

would indicate a limitation in the emulator's design, where delegating the representation of [..[190]]non-uniformities in the monthly temperature response to the residual variability module does not fully work in the presence of lower variabilities.

Generally, emulated [..[191]]5% (95%) quantiles are warmer (colder) than those of the ESM training and test runs. Such underdispersivity for regional averages is linked to the localisation of the spatial covariance matrix within the residual variability module, such that spatial correlations drop faster within the emulator than they do in the actual ESM. [..[192]]For January over the time period of 1870-2000[..[193]], lowest magnitudes in 5% and 95% quantile deviations is observed for Southern Hemispheric regions [..[194]](e.g. AMZ, NEB, WSA, SSA[..[195]]) , along with slight overdispersivity (see the blue 5% quantile and red 95% quantile values in their respective panels of Figure 8). Over the period of 2000-2100, this behaviour [..[196]]for January switches to Northern Hemispheric regions [..[197]](e.g. CEU, ALA, ENA, WNA, TIB) and is mostly apparent for the 95th quantile, possibly due to a decrease (increase) in January variability in the Northern (Southern) Hemisphere with increasing yearly temperatures (Holmes et al., 2016). In contrast, over both the periods of 1870-2000 [..[198]]and 2000-2100[..[199]], July consistently displays lowest magnitudes of 5% and 95% quantile deviations (with even slight overdispersivity) in Northern Hemispheric regions (e.g. WNA, ENA, NAS, WAS). The observed small regional overdispersivities hint at additional processes being at play in these regions, which are not accounted for by the emulator and that counteract the expected regional-scale underdispersivity which is inherent to the emulator's residual variability module design.

---

[190]removed: secondary, non-linear responses

[191]removed: 5th (95th

[192]removed: Underdispersivity may also be a consequence of the emulator solely relying on yearly temperatures as an input variable. To be exact, even though changes in skewness in local monthly temperature residuals with evolving local yearly temperatures is incorporated, such cannot completely represent delayed or compound effects, otherwise brought about by secondary/external feedbacks (Shinoda, 2001; Potopová et al., 2016), which more largely affect extreme quantiles (Huybers et al., 2014; Guirguis et al., 2018; Loikith and Neelin, 2019). January 5th and 95th quantile deviations

[193]removed: show lowest values in the

[194]removed: of

[195]removed: and SAF, with emulations sometimes even being slightly overdispersive. For July of the same period

[196]removed: switches to most

[197]removed: but only for select ESMs.For the same select ESMs, similar features as seen in January and July for

[198]removed: are seen again for both months in the period of

[199]removed: . An inspection of the localisation radii used within these select ESMs points at generally high values, explaining the higher dispersivity as spatial cross correlations are conserved up to larger distances.

Figure 9. January 5% (left), 50% (middle), and 95% (right) quantile deviations (colour[..²⁰⁰](of the [..²⁰¹]climate model runs from [..²⁰²]the emulated quantiles of [..²⁰³]the ESM training (top block) and test (bottom block) [..²⁰⁴]run values from their monthly emulated quantiles, over the period 1870-2000 for Global land (without Antarctica) and SREX regions (columns) across all CMIP6 models (rows). The monthly emulated quantile is computed based on 50 emulations per ESM run and quantile deviations are given as averages across the respective number of ESM training/test runs. The number of test runs averaged across is indicated in brackets next to the model names in the bottom block. Red means that the emulated quantile is warmer than the quantile of the ESM run, and vice versa for blue.

1870-2000 July regional-scale verification: deviation of climate model runs from emulated quantiles
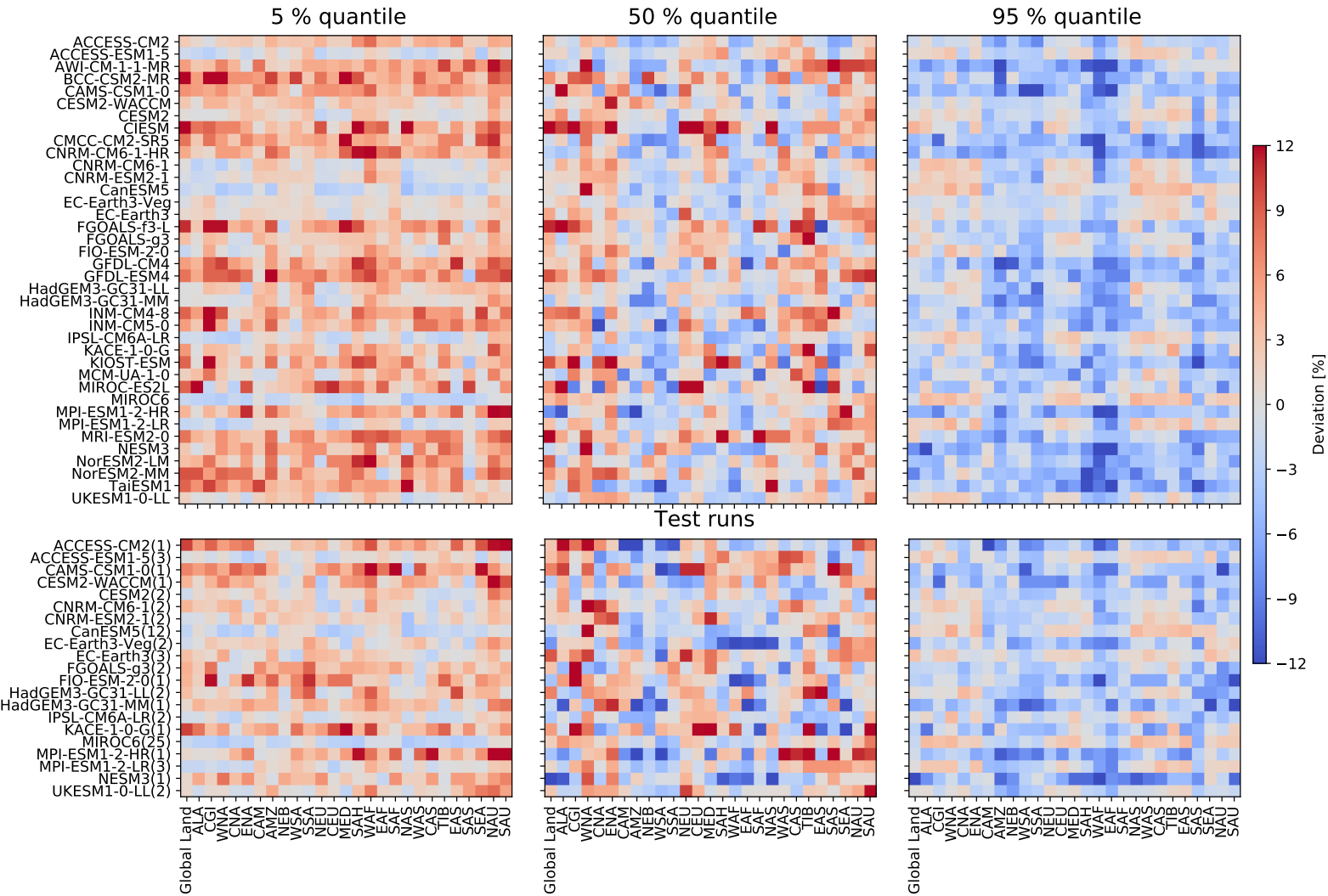
Figure 10. Same as Figure 9, but for July over the period 1870-2000
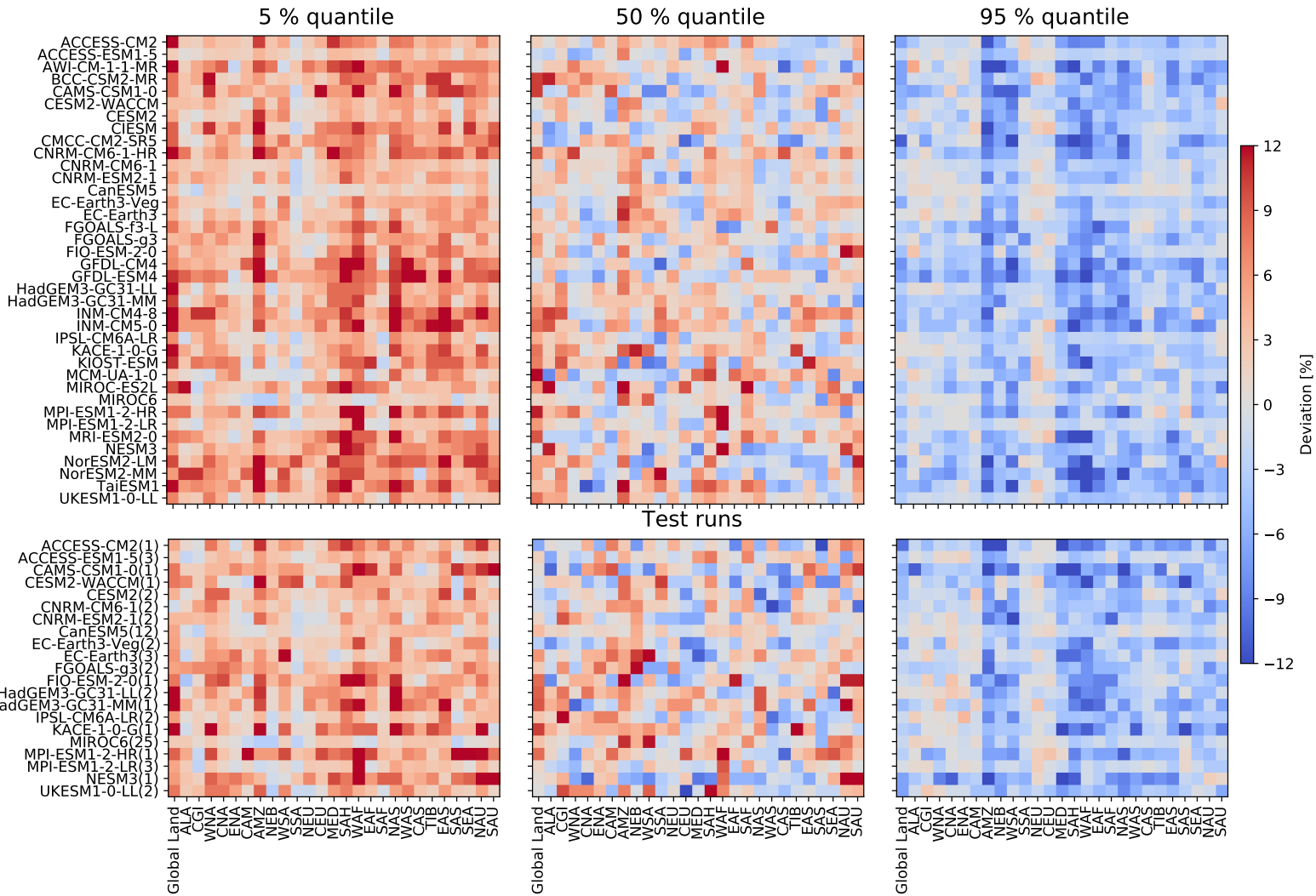
Figure 11. Same as Figure 9, but over the period 2000-2100

2000-2100 July regional-scale verification: deviation of climate model runs from emulated quantiles
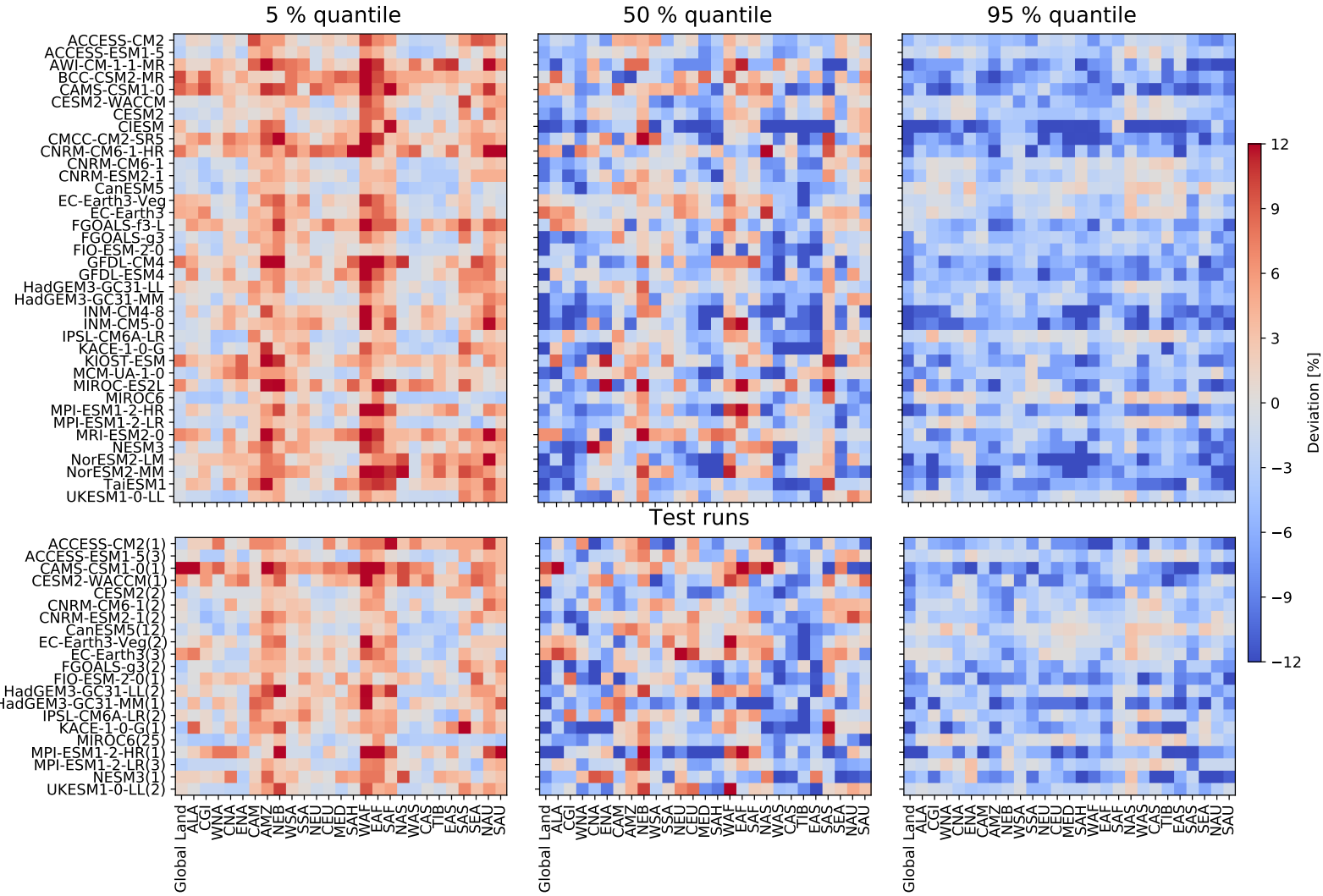
Figure 12. Same as Figure 9, but for July over the period 2000-2100

5.4 Benchmarking MESMER-M using a simple physical approach

375 In-depth analysis of the benchmarking approach outlined in Section 3.3 is conducted for [..²⁰⁵]four selected ESMs which exhibit diverse genealogies (Knutti et al., 2013; Brunner et al., 2020) (see Figure E1 in Appendix [..²⁰⁶]E for summarised results of all other ESMs). From Figure 13, [..²⁰⁷]it is evident that adding even only one biophysical variable explains part of the residual difference behaviour, with correlations between the physically-informed emulations and ESM runs (given relative to correlations between T_{yr} and $month$ informed emulations and ESM runs) over global land always being positive. Across

380 all [..²⁰⁸]four ESMs the main improvements are in Northern Hemispheric regions which possess distinct seasonal variations in snow cover namely, ALA, CGI, WNA, CNA, ENA, NEU, CEU, NAS and CAS. MIROC6 and [..²⁰⁹]MPI-ESM1-2-LR exhibit substantial improvements in other regions, notably WAF, EAF, SAS and NAU regions. It is worth noting that for most ESMs, the biophysical predictor configuration of albedo, cloud cover and snow cover (ACS) performs consistently worse than any configuration containing sensible and latent heat fluxes (H)[..²¹⁰], suggesting the presence of [..²¹¹]processes only

385 explainable using sensible and latent heat fluxes. MIROC6 is an exception to this, with both the biophysical configurations of latent heat flux (Hl) and H yielding 0 or lower relative correlations in WNA, CNA, ENA, CEU and EAS while HC displays no improvements for these regions. This could be due to colinearities between cloud cover and latent and sensible heat fluxes alongside overfitting of the [..²¹²]physically-informed model to latent and sensible heat fluxes due to confounding variabilities.

As HACS performs the best globally (appears as 1 in global land) across all [..²¹³]four ESMs we choose it as the benchmark

390 [..²¹⁴]physically-informed model to compare the residual variability module to. Figure 13 shows the energy distances of the [..²¹⁵]physically-informed (harmonic model+HACS) and statistical (full emulator = harmonic model+residual variability module) emulated cdfs to the ESM cdfs for January and July, where 0 indicates identical, and thus 'perfect' emulated cdfs. Energy distances in July for both the [..²¹⁶]physically-informed and statistical models are close to 0 indicating near perfect cdfs, with only MIROC6 and MPI-ESM1-LR showing larger distances for the full emulator in the Indo-Gangetic region,

395 South America and Central-West Africa. In contrast, January shows higher distances for both the [..²¹⁷]physically-informed and statistical model cdfs, particularly in Northern Hemispheric regions with seasonal snowfall and most notably in the full emulator of MIROC6. Overall, the statistical model performs better than the [..²¹⁸]physically-informed model for CESM2 and

---

²⁰⁵removed: 4 select

²⁰⁶removed: E

²⁰⁷removed: 2,

²⁰⁸removed: 4

²⁰⁹removed: MPIESM1-2-LR exhibit additional

²¹⁰removed: . This suggests

²¹¹removed: colinearities between albedo, snow cover and cloud cover, which limits the physical model's representation of certain residual variability behaviour, otherwise

²¹²removed: physical

²¹³removed: 4

²¹⁴removed: physical

²¹⁵removed: physical

²¹⁶removed: physical

²¹⁷removed: physical

²¹⁸removed: physical

UKESM1-0-LL and worse for MIROC6 and MPI-ESM1-2-LR. An explanation behind this could [..219]the a combination of biophysical feedbacks being more pronounced in January's Northern Hemispheric variability and that MIROC6 and MPI-ESM1-2-LR have at least [..220]four more training runs than CESM2 and UKESM1-0-LL, providing the GBR model with more training material to extract such biophysical information from. This suggests a limit to when the statistical approach performs better than the physical approach, depending on how present biophysical feedbacks are within the overall variability and how much information is available to train on. Nevertheless, without the prerequisite of having more training runs [..221]– which can be seen as an added [..222]advantage – the statistical approach taken by the full emulator generally shows better performances across most ESMs for January and July than the physical approach (Figure E2 in Appendix E). Thus, the distributional properties of local monthly temperatures as seen within ESM initial-condition ensembles can be sufficiently represented using the statistical approach outlined in this paper, which takes only local yearly temperatures as input.

219removed: be

220removed: 4

221removed: -which

222removed: advantage-
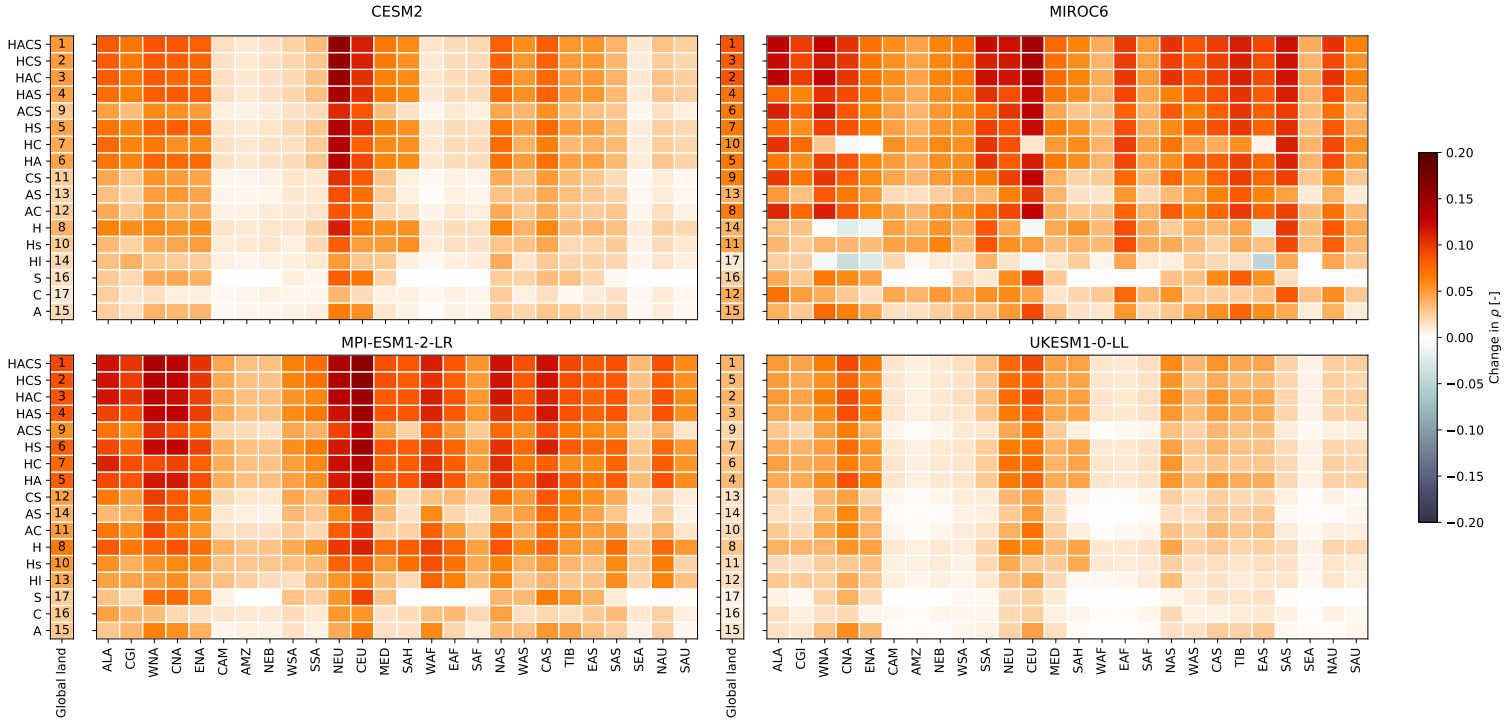
Figure 13. Global land (without Antarctica) and regional performances (columns) of the [..²²³]physically-informed model trained using different predictor sets (rows) shown for [..²²⁴]four selected CMIP6 models, for each SREX region. Acronyms for the predictor sets (y-axis tick labels) can be referred back to in Table 1. Pearson Correlations calculated over all months between test runs and harmonic model test results augmented by the [..²²⁵]physically-informed model's predictions of residual variability are considered. Here we show changes in correlations relative to those obtained when augmenting using only T_{yr} and month values as predictors. Numbers in the global land column indicate the ranking of each predictor set, where 1 is the best performing.
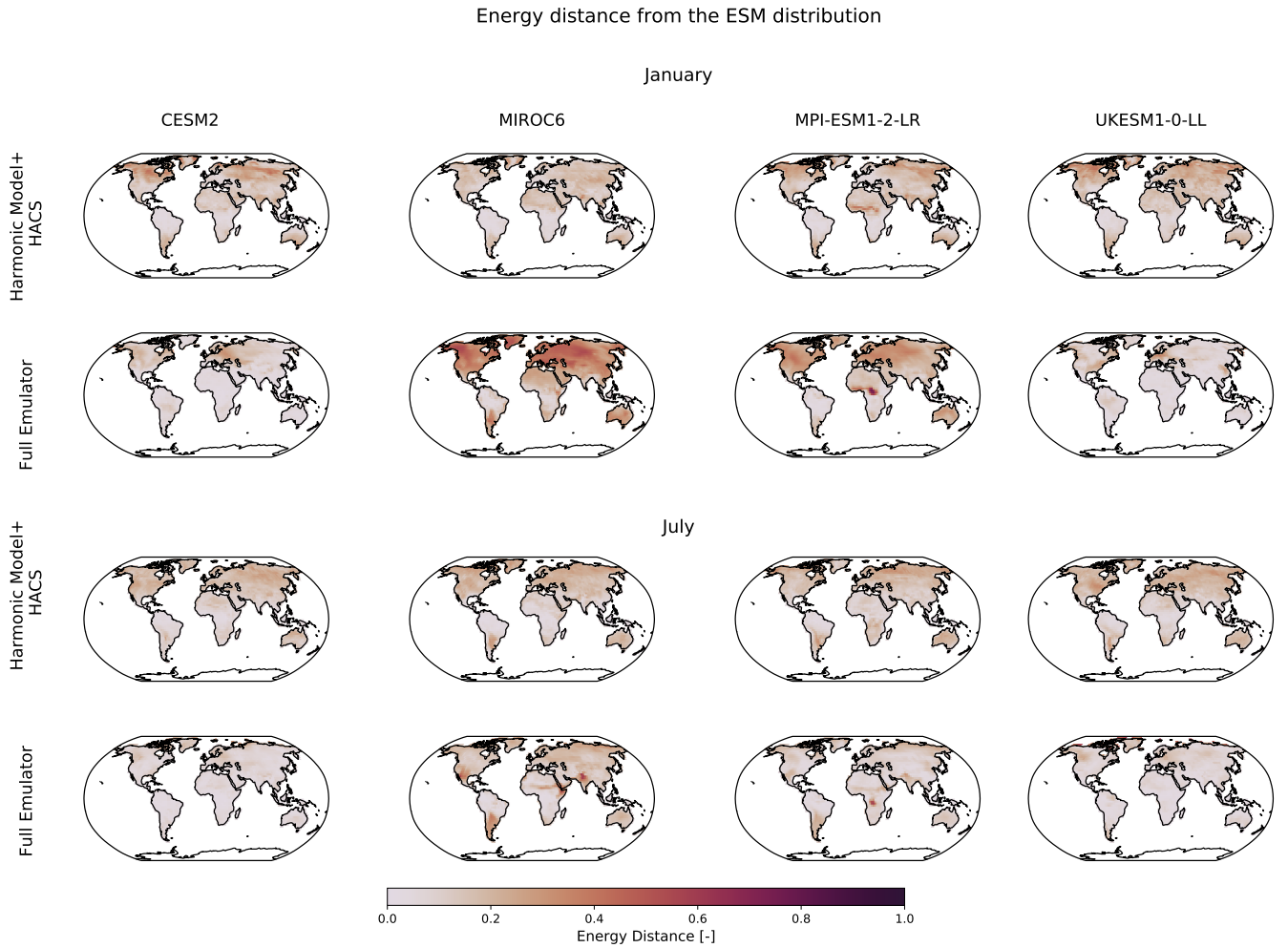
Figure 14. Comparison of the performance of the harmonic model + [..226]physically-informed HACS model to that of the full emulator for January and July of [..227]four selected CMIP6 models. The energy distance from the actual model test runs is considered, where 0 indicates the best performance.

6 Conclusion and Outlook

We extend MESMER's framework to include [..[228]]the monthly downscaling module, MESMER-M, trained for each ESM at each grid-point individually, thus providing realistic, spatially explicit monthly temperature fields from yearly temperature fields in a matter of seconds. We assume a linear response of the seasonal temperature cycle to its yearly mean values and represent it using a harmonic model. Any remaining response patterns are expected to arise from [..[229]]regional-scale, physical/intra-annual [..[230]]processes, such as changes in ENSO or snow-albedo feedbacks, and have asymmetric, non-uniform (e.g. non-linear, non-stationary, affecting variance and skewness) effects across months. To capture them, we build a month-specific residual variability module which samples spatio-temporally correlated terms, [..[231]]conserving lag-1 autocorrelations and spatial cross-correlations whilst accounting for specificities in the residual variability structure across months. By letting the skewness of the residual sampling space additionally covary with yearly temperatures, non-uniformities in secondary feedbacks are furthermore inferred through their manifestations within the monthly temperature distributions.

Verification results across all ESMs show the emulator altogether reproducing the mean monthly temperature response, as well as conserving temporal and spatial correlation patterns and regional-scale temperature distributions up to a degree sensible to its simplicity. To further assess how well the emulator is able to represent [..[232]]non-uniformities in the monthly temperature response arising from secondary biophysical feedbacks, we compare its performance to that of a simple [..[233]]physically-informed model built on biophysical information. The emulator overall reproduces the cdfs of the actual ESM just as well as, and in most cases even better than the [..[234]]physically-informed model, evidencing the validity of such a statistical approach in inferring temperature distributions and insofar the uncertainty due to natural variability within temperature realisations. Given that the uncertainty due to natural variability is a property intrinsic to climate models and largely irreducible (Deser et al., 2020), the emulator thus proves itself as a pragmatic alternative to otherwise having to generate large single-model, initial-condition ensembles.

6.1 Further emulator developments

In this study, MESMER-M was only trained on SSP5-8.5 climate scenario runs so as to demonstrate its performance over the extreme spectrum of climate response types. A further step would be to investigate the inter-scenario applicability of MESMER-M and this has already been done for MESMER with satisfactory results (Beusch et al., 2021). While we would expect the overall mean response of monthly to yearly temperatures to remain relatively stable between climate scenarios, non-uniformities arising in the local variability may be more scenario specific (e.g. due to slowing down of the snow-albedo feedback under an equilibrated climate for low-emission scenarios). Bearing this in mind, we recommend training

---

[228]removed: a

[229]removed: secondary biophysical feedbackswhich act on regional levels and

[230]removed: timescales and

[231]removed: thus

[232]removed: heterogeneities

[233]removed: physical

[234]removed: physical

MESMER-M on all available climate scenarios before using it for inter-scenario exploration. Such would provide the local Yeo-Johnson transformation with enough information on the relationship between yearly temperatures and skewness of monthly temperatures. Additional adjustments of looking at rate of yearly temperature change as a covariate to monthly temperature skewness could also be investigated.

This study demonstrates the advantage of constructing [..[235]]modular emulators such that the emulator framework can be extended according to the area of application. Additional module developments which increase the impact relevance of the emulations and improve the fidelity in global and regional representation under different climate scenarios should be given priority. A module that comes to mind would be one representing changes in land cover, such as de/afforestation, which have been historically assessed to have biophysical impacts of a similar magnitude on regional climate as the concomitant increase of [..[236]]GHGs (De Noblet-Ducoudré et al., 2012) and for which very distinct imprints on the seasonal cycle of temperatures as well as the tails of the temperature distributions have been identified (Pitman et al., 2012; Lejeune et al., 2017). Such a module would furthermore increase the emulator's relevance towards impact assessments, in light of the important land-cover changes expected to happen in the 21st century (Popp et al., 2017; O'Neill et al., 2016) and the relevance of accounting for their regional climate impacts especially in high-mitigation scenarios such as those compatible with the 1.5°C long-term temperature goal of the Paris Agreement (Seneviratne et al., 2018; Roe et al., 2019; Arneth et al., 2019)). One technical advantage of adding a land cover module would be that the effect of land cover changes can be expected to be sufficiently decoupled from the overall GHG induced temperature response. Hence, the direct local effect of such a module would not interfere with the mean temperature response as extracted within the rest of the emulator.

[..[237]]Another modular development could include an explicit representation of the main modes of climate variability, so as to strengthen MESMER-M's inter-annual variability representation. Since the coupling of such modes to the overall GHG induced temperature response is uncertain however, such inclusion would be more complicated. One possible approach could be to introduce soil moisture as an additional variable term and investigate its lag correlations to monthly temperature variabilities. Alternatively we could explore building upon existing approaches such as the one of McKinnon and Deser (2018). Bearing in mind that one key advantage of MESMER-M is that it only requires yearly temperatures as input, the added value of such a module should be critically assessed against the need for additional predictors. Another possibility could be to instead decompose the covariance matrix used in $\widetilde{\eta}_{m,s,y}^{spat.}$ (see Figure 1) so as to account for spatial cross-correlations affected by major modes of variability, again the added model complexity should be weighed against gained skill in emulation.

6.2   Potential further applications of the GBR-based physical approach

---

[235]removed: a modular emulator

[236]removed: greenhouse gases (GHGs)

[237]removed: Beusch et al. (2020) suggest

465 Beusch et al. (2020) pointed out that the ESM-specific emulator calibration results represent distinct "model IDs", containing scale-dependent information of the model structure. As a follow-up from the [..[238]]physically-informed model based benchmarking done within this study, we further propose that the residuals from the mean response module also contain ESM-specific, scale-dependent information, constituting the distinct representations and parameterisations of biophysical feedbacks within each ESM. For example, models with strong snow-albedo feedbacks and a large snow cover reduction with increasing

470 global mean temperature will show stronger warming of cold months (Fischer et al., 2011) and thus more negatively skewed residuals for those months. A step towards disentangling such process representation within the ESMs has already been made in this study, through the representation of biophysical contributions within residual variabilities using the GBR based [..[239]]physically-informed model. Further analysing the strength of the co-variations of different biophysical variables with the residuals, as identified by the [..[240]]physically-informed model, could then help isolate the exact contributions of these variables.

475 While the key physical variables contributing to temperature variability within ESMs have already been studied (Schwingshackl et al., 2018), such an analysis would further provide information on the amount by which a [..[241]]selected number of biophysical variables contribute to residual variability within each ESM. Performing a similar analysis on observational datasets and comparing the results to those of the ESMs, could then serve as a means to evaluate model representation of biophysical interactions under a changing climate.

Appendix A

Appendix B

[..[242]][..[243]]

---

[238] removed: physical

[239] removed: physical

[240] removed: physical

[241] removed: choice

[242] removed: Where regional results are shown, we consider area weighted averages for the 26 SREX regions (Seneviratne et al., 2012) as shown in Figure B1. Global land results always constitute area weighted averages across all land grid points excluding Antarctica.

[243] removed: Map of the SREX regions and their abbreviations. The considered land grid points are shown in grey.

Table A1. List of the 38 employed CMIP6 models, the modeling groups providing them, and the number of initial-condition ensemble members used in the training and test sets.

Model	Modeling Center (or Group)	Training Runs	Test Runs
ACCESS-CM2	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	2	1
ACCESS-ESM1-5	Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	7	3
AWI-CM-1-1MR	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research	1	0
BCC-CSM2-MR	Beijing Climate Center, China Meteorological Administration	1	0
CAMS-CSM1-0	Chinese Academy of Meteorological Science	1	1
CanESM5	Canadian Centre for Climate Modeling and Analysis	12	12
CESM2-WACCM	National Center for Atmospheric Research	2	1
CESM2	Community Earth System Model Contributors	3	2
CIESM	Community Earth System Model Contributors	1	0
CMCC-CM2-SR5	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1	0
CNRM-CM6-1-HR	Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	1	0
CNRM-CM6-1	Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	4	2
CNRM-ESM2-1	Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	3	2
EC-Earth3-Veg	EC-EARTH consortium	2	1
EC-Earth3	EC-EARTH consortium	7	3
FGOALS-f3-L	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS,Tsinghua University	1	0
FGOALS-g3	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS,Tsinghua University	3	1
FIO-ESM-2-0	The First Institute of Oceanography, SOA, China	2	1
GFDL-CM4	NOAA Geophysical Fluid Dynamics Laboratory	1	0
GFDL-ESM4	NOAA Geophysical Fluid Dynamics Laboratory	1	0
HadGEM3-GC31-LL	Met Office Hadley Centre	3	1
HadGEM3-GC31-MM	Met Office Hadley Centre	2	1
INM-CM4-8	Institute for Numerical Mathematics	1	0
INM-CM5-0	Institute for Numerical Mathematics	1	0
IPSL-CM6A-LR	Institut Pierre-Simon Laplace	4	1
KACE-1-0-G	National Institute of Meteorological Sciences/Korea Meteorological Administration	2	1
KIOST-ESM	Korea Institute of Ocean Science and Technology	1	0
MCM-UA-1-0	Department of Geosciences, University of Arizona	1	0
MIROC-ES2L	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	1	0
MIROC6	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	25	25
MPI-ESM1-2-HR	Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	1	1
MPI-ESM1-2-LR	Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	7	3
MRI-ESM2-0	Meteorological Research Institute 1 NESM3 Nanjing University of Information Science and Technology	1	1
NorESM2-LM	Norwegian Climate Centre	1	0
NorESM2-MM	Norwegian Climate Centre	1	0
TaiESM1	Research Center for Environmental Changes, Academia Sinica	1	0
UKESM1-0-LL	Met Office Hadley Centre	3	2

Appendix B
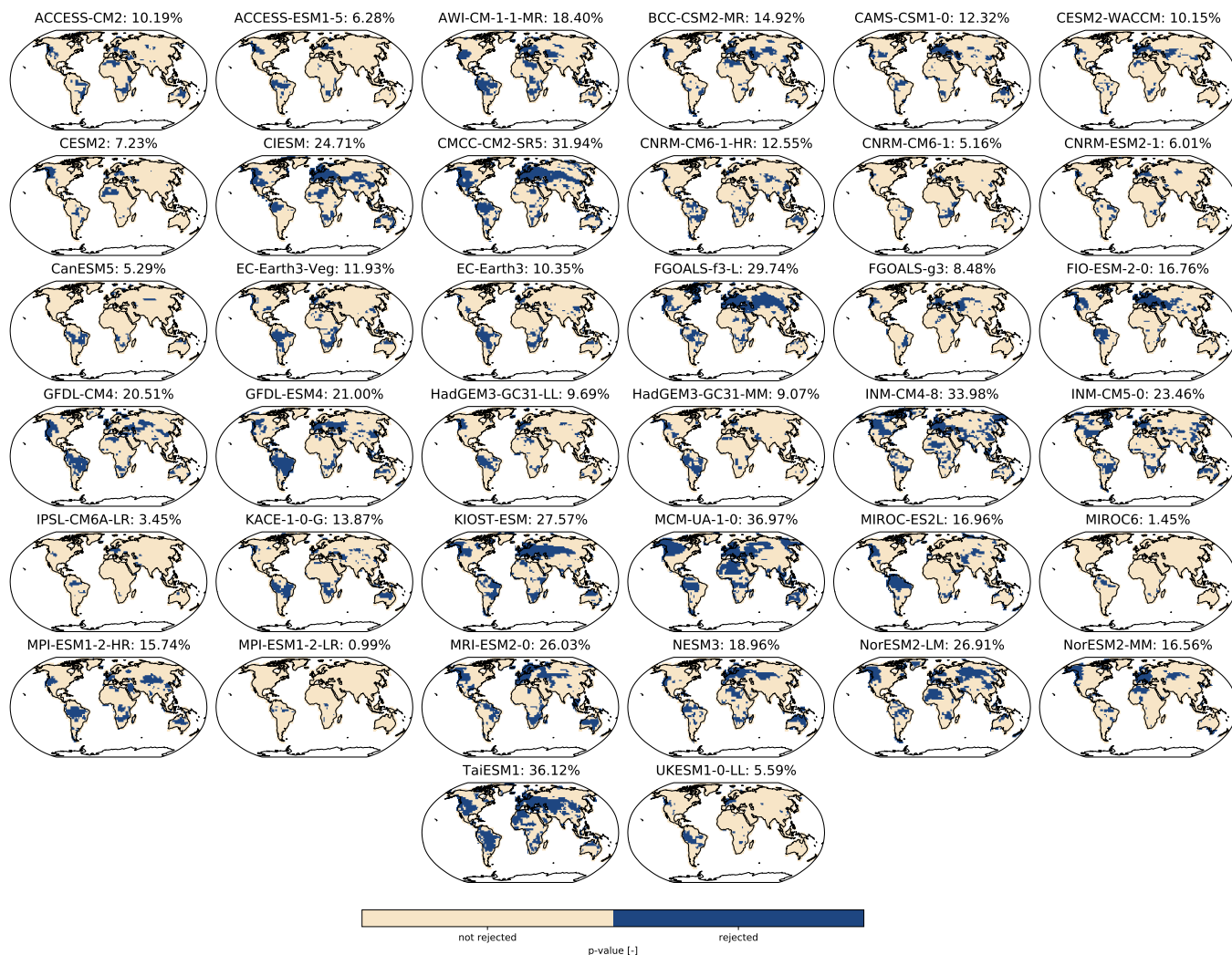
Figure B1. [..[244]]Shapiro-Wilks test for normality of January temperature residuals. The null hypothesis is that the residuals are normally distributed. A Benjamini/Hochberg multiple test correction (Benjamini and Hochberg, 1995) is applied to the p-values before plotting them.

Shapiro-Wilks Test: July
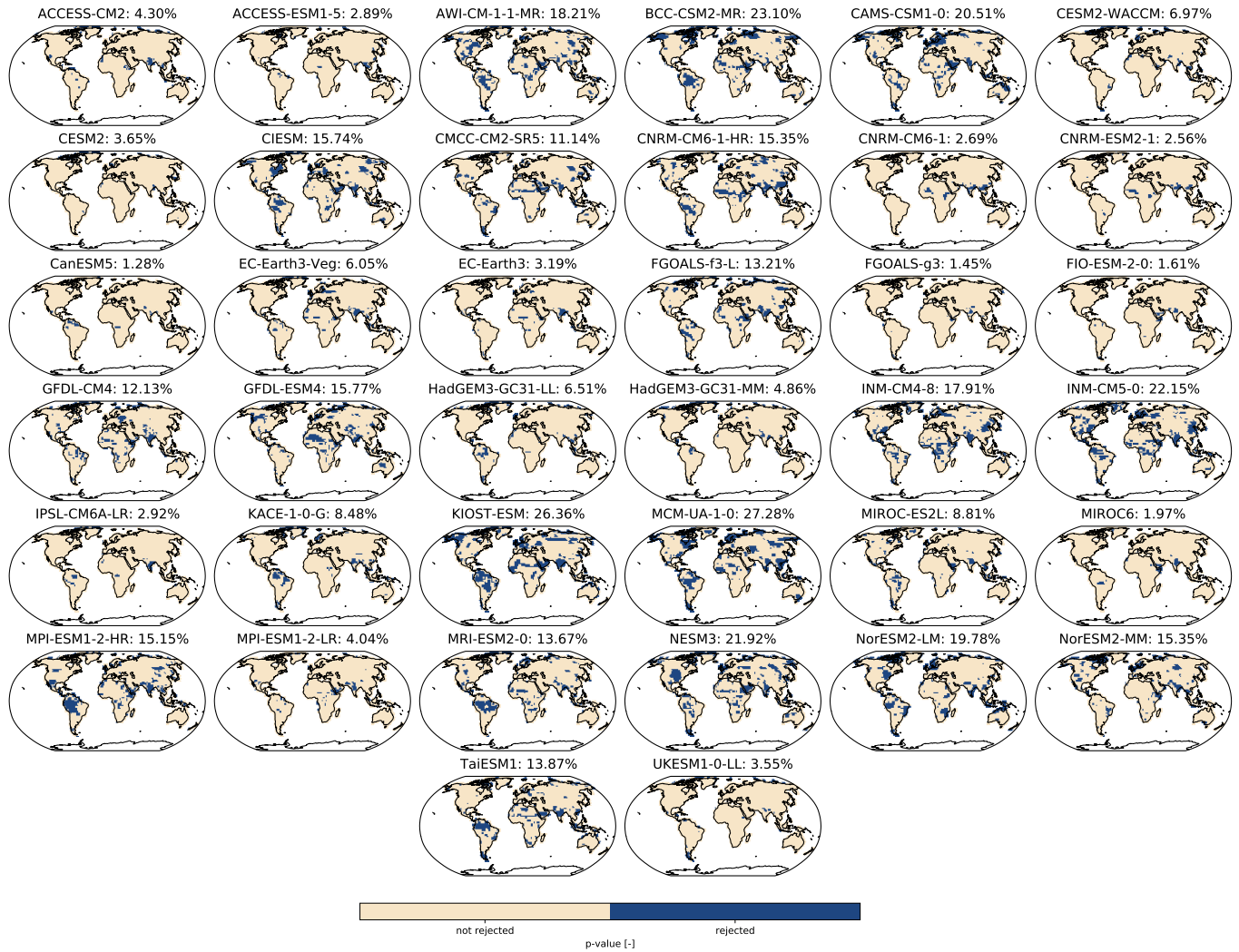
Figure B2. [..²⁴⁵]Same as Figure C1, except for [..²⁴⁶]July[..²⁴⁷].[..²⁴⁸]

Appendix C: [..[249]]

Figure C1. Likelihood ratio test comparing the performance of January's Yeo-Johnson transformations when using just one single lambda parameter ($\lambda_{m,s}$) vs when using a yearly temperature dependent lambda parameter ($\lambda_{y,m,s}$). The null hypothesis is that the $\lambda_{m,s}$ based transformation performs better than the $\lambda_{y,m,s}$ based transformation. A Benjamini/Hochberg multiple test correction (Benjamini and Hochberg, 1995) is applied to the p-values before plotting them.

---

[249]removed: Likelihood-ratio test for Yeo-Johnson transformation

Likelihood Ratio test for Yeo-Johnson power transformation: July
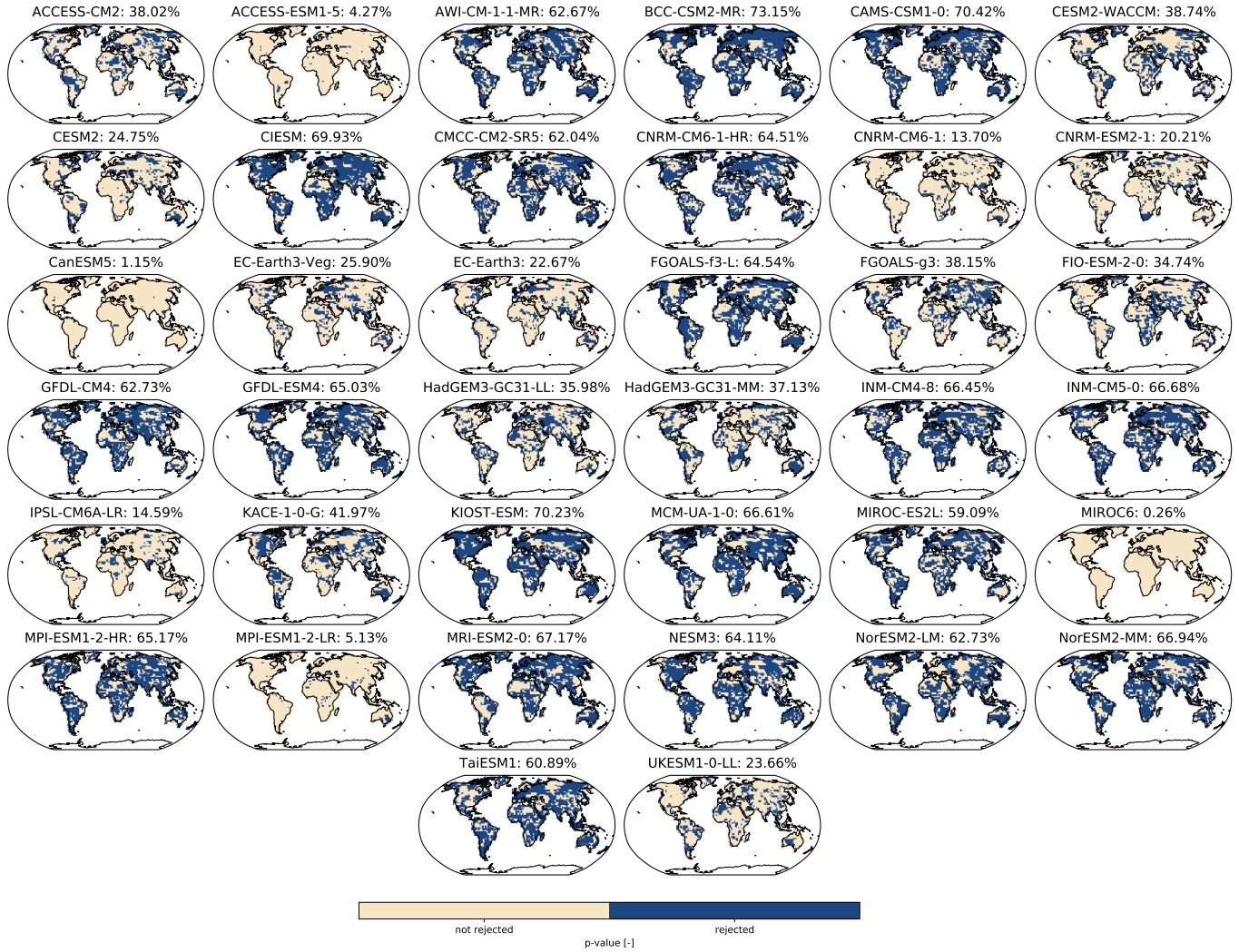
Figure C2. [..²⁵⁰]Same as Figure D1, except for July[..²⁵¹].[..²⁵²]

485 # Appendix D: [..²⁵³]
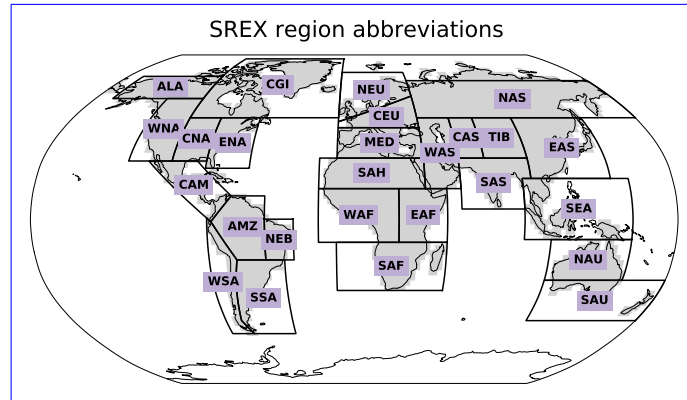
Figure D1. Map of the SREX regions and their abbreviations. The considered land grid points are shown in grey.

---

²⁵³removed: Exploration of processes explaining the deviations from the harmonic model

Appendix E
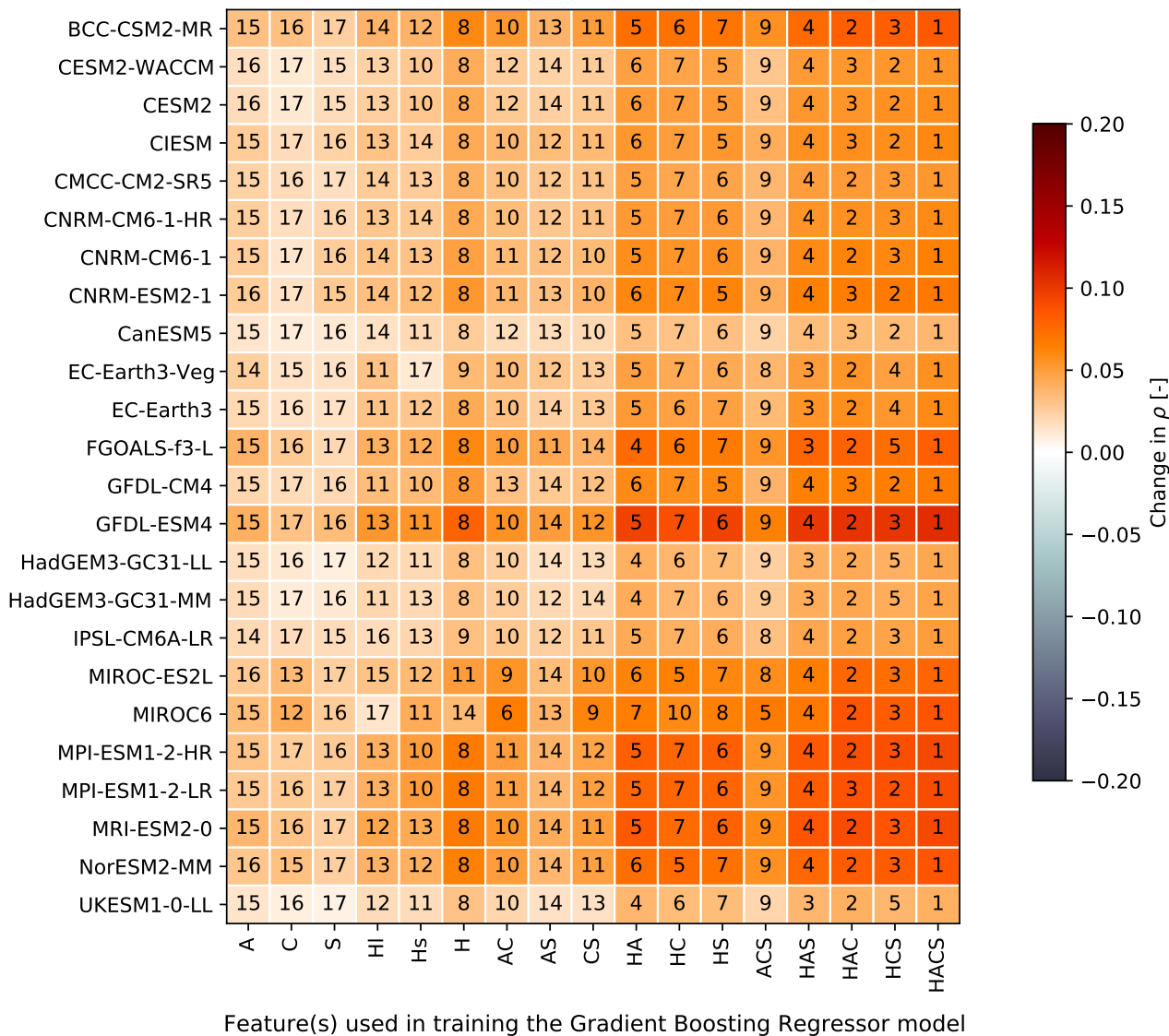
Figure E1. [..²⁵⁴]Same as Figure 12, except that all CMIP6 models [..²⁵⁵]are shown for the [..²⁵⁶]global land ([..²⁵⁷]without Antarctica)[..²⁵⁸].[..²⁵⁹]
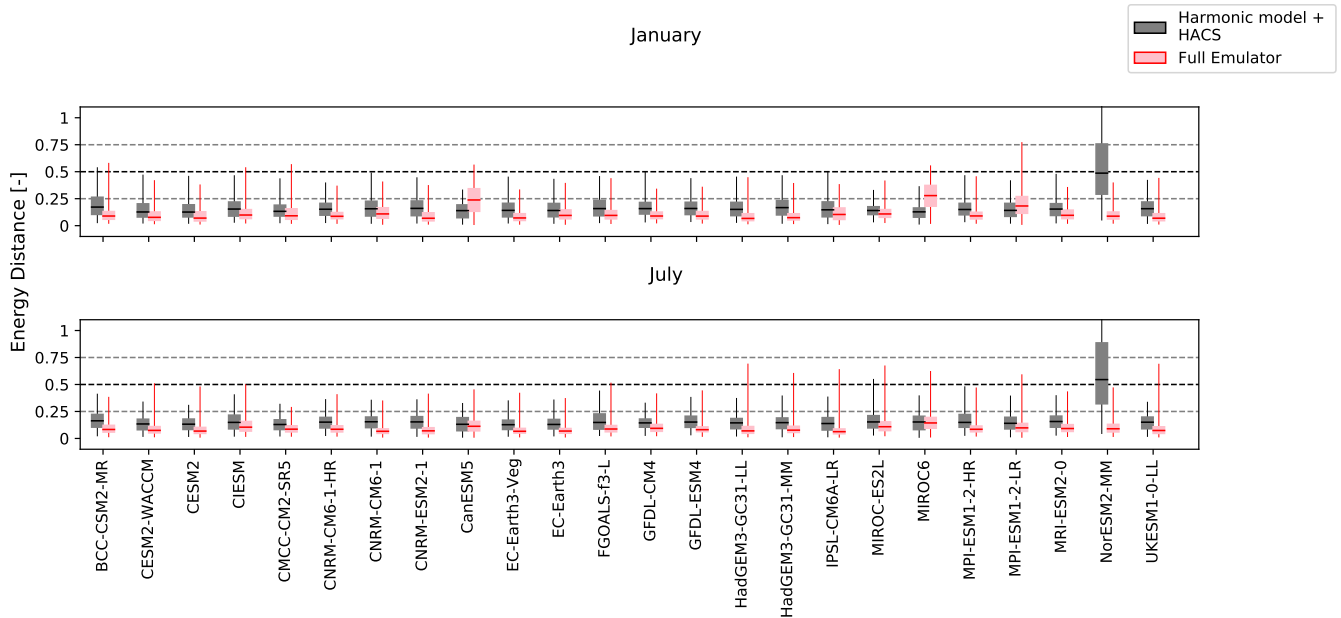
Figure E2. Comparison of the performance of the harmonic model + [..²⁶⁰]physically-informed HACS model to that of the full emulator for January and July of all CMIP6 models. The energy distance from the actual model test runs is considered, where 0 indicates the best performance. Boxplot whiskers indicate 5th and 95th quantiles.

Author contributions. QL, CFS and SIS identified the need to extend MESMER's framework by a monthly downscaling module. SN designed the monthly downscaling module with support and guidance from QL and LB. SN led the analysis and drafted the text with help from QL in developing the storyline. All authors contributed to interpreting results and streamlining the text.

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

Acknowledgements. We acknowledge that this study was conducted as part of the LAMACLIMA project, receiving funding from the German Federal Ministry of Education and Research (BMBF) and the German Aerospace Center (DLR) as part of AXIS, an ERANET initiated by JPI Climate (grant no. 01LS1905A), with co-funding from the European Union (grant no. 776608). SIS acknowledges partial support from the European Research Council (ERC) through the Proof-of-Concept Project MESMER-X (Project number: 964013). We furthermore thank
495 Lukas Gudmundsson, Joel Zeder and Christoph Frei for their invaluable statistical insight into the development and analysis of the emulator modules. Finally, we thank the climate modelling groups listed in Table A1 for producing and making available the CMIP6 model outputs, [..²⁶¹]as well as Urs Beyerle and Lukas Brunner for downloading the CMIP6 data and pre-processing them.

---

²⁶¹removed: without which there would be no training material for the emulator

References

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, R. J. and Zender, C. S.: Forcing of the Arctic Oscillation by Eurasian snow cover, Journal of Climate, 24, 6528–6539, https://doi.org/10.1175/2011JCLI4157.1, 2011.

Arneth, A., Barbosa, H., Benton, T., Calvin, K., Calvo, E., Connors, S., Cowie, A., Davin, E., Denton, F., van Diemen, R., Driouech, F., Elbehri, A., Evans, J., Ferrat, M., Harold, J., Haughey, E., Herrero, M., House, J., Howden, M., Hurlbert, M., Jia, G., Gabriel, T. J., Krishnaswamy, J., Kurz, W., Lennard, C., Myeong, S., Mahmoud, N., Delmotte, V. M., Mbow, C., McElwee, P., Mirzabaev, A., Morelli, A., Moufouma-Okia, W., Nedjraoui, D., Neogi, S., Nkem, J., Noblet-Ducoudré, N. D., Pathak, L. O. M., Petzold, J., Pichs-Madruga, R., Poloczanska, E., Popp, A., Pörtner, H.-O., Pereira, J. P., Pradhan, P., Reisinger, A., Roberts, D. C., Rosenzweig, C., Rounsevell, M., Shevliakova, E., Shukla, P., Skea, J., Slade, R., Smith, P., Sokona, Y., Sonwa, D. J., Soussana, J.-F., Tubiello, F., Verchot, L., Warner, K., Weyer, N., Wu, J., Yassaa, N., Zhai, P., and Zommers, Z.: IPCC SR: Climate Change and Land, An IPCC Special Report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems, p. 43, 2019.

Arnold, L.: Hasselmann's program revisited: the analysis of stochasticity in deterministic climate models, in: Stochastic Climate Models, edited by Imkeller, P. and von Storch, J.-S., pp. 141–157, Birkhäuser Basel, Basel, https://doi.org/10.1007/978-3-0348-8287-3_5, 2001.

Benjamini, Y. and Hochberg, Y.: benjamini_hochberg1995, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289–300, https://www.jstor.org/stable/2346101, 1995.

Berner, J., Achatz, U., Batté, L., Bengtsson, L., De La Cámara, A., Christensen, H. M., Colangeli, M., Coleman, D. R., Crommelin, D., Dolaptchiev, S. I., Franzke, C. L., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajaajaajan, S., Palmer, T. N., Penland, C., Sakradzijaja, M., Von Storch, J. S., Weisheimer, A., Weniger, M., Williams, P. D., and Yano, J. I.: Stochastic parameterization toward a new view of weather and climate models, Bulletin of the American Meteorological Society, 98, 565–587, https://doi.org/10.1175/BAMS-D-15-00268.1, 2017.

Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth System Model temperatures: from global mean temperature trajectories to grid-point level realizations on land, Earth System Dynamics, 11, 139–159, https://doi.org/10.5194/esd-11-139-2020, 2020.

Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., and Seneviratne, S.: From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: MAGICC (v7.5.1) – MESMER (v0.8.1) coupling, Geoscientific Model Development Discussions, pp. 1–26, https://doi.org/10.5194/gmd-2021-252, 2021.

Blackport, R. and Kushner, P. J.: The transient and equilibrium climate response to rapid summertime sea ice loss in CCSM4, Journal of Climate, 29, 401–417, https://doi.org/10.1175/JCLI-D-15-0284.1, 2016.

Brunner, L., Pendergrass, A., Lehner, F., Merrifield, A., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth System Dynamics Discussions, pp. 1–23, https://doi.org/10.5194/esd-2020-23, 2020.

Castruccio, S., Hu, Z., Sanderson, B., Karspeck, A., and Hammerling, D.: Reproducing internal variability with few ensemble runs, Journal of Climate, 32, 8511–8522, https://doi.org/10.1175/JCLI-D-19-0280.1, 2019.

Cohen, J. and Rind, D.: The Effect of Snow Cover on the Climate, Journal of Climate, 4, 689–706, https://doi.org/10.1175/1520-0442(1991)004<0689:TEOSCO>2.0.CO;2, 1991.

535 Colman, R. A.: Surface albedo feedbacks from climate variability and change, Journal of Geophysical Research Atmospheres, 118, 2827–2834, https://doi.org/10.1002/jgrd.50230, 2013.

De Noblet-Ducoudré, N., Boisier, J. P., Pitman, A., Bonan, G. B., Brovkin, V., Cruz, F., Delire, C., Gayler, V., Van Den Hurk, B. J., Lawrence, P. J., Van Der Molen, M. K., Müller, C., Reick, C. H., Strengers, B. J., and Voldoire, A.: Determining robust impacts of land-use-induced land cover changes on surface climate over North America and Eurasia: Results from the first set of LUCID experiments, Journal of

540 Climate, 25, 3261–3281, https://doi.org/10.1175/JCLI-D-11-00338.1, 2012.

Deser, C., Phillips, A. S., Tomas, R. A., Okumura, Y. M., Alexander, M. A., Capotondi, A., Scott, J. D., Kwon, Y. O., and Ohba, M.: ENSO and pacific decadal variability in the community climate system model version 4, Journal of Climate, 25, 2622–2651, https://doi.org/10.1175/JCLI-D-11-00301.1, 2012.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E.,

545 Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nature Climate Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

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, Geoscientific Model Development, 9, 1937–1958,

550 https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Fischer, E. M., Lawrence, D. M., and Sanderson, B. M.: Quantifying uncertainties in projections of extremes-a perturbed land surface parameter experiment, Climate Dynamics, 37, 1381–1398, https://doi.org/10.1007/s00382-010-0915-y, 2011.

Fischer, E. M., Rajczak, J., and Schär, C.: Changes in European summer temperature variability revisited, Geophysical Research Letters, 39, 1–8, https://doi.org/10.1029/2012GL052730, 2012.

555 Frei, C. and Isotta, F. A.: Ensemble Spatial Precipitation Analysis From Rain Gauge Data: Methodology and Application in the European Alps, Journal of Geophysical Research: Atmospheres, 124, 5757–5778, https://doi.org/10.1029/2018JD030004, 2019.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, https://doi.org/10.1256/smsqj.55416, 1999.

Guirguis, K., Gershunov, A., Cayan, D. R., and Pierce, D. W.: Heat wave probability in the changing climate of the Southwest US, Climate

560 Dynamics, 50, 3853–3864, https://doi.org/10.1007/s00382-017-3850-3, 2018.

Hall, A.: The role of surface albedo feedback in climate, Journal of Climate, 17, 1550–1568, https://doi.org/10.1175/1520-0442(2004)017<1550:TROSAF>2.0.CO;2, 2004.

Hastie, T., Tibshirani, R., and Friedman, J.: Elements of Statistical Learning Ed. 2, Springer, 2009.

Holmes, C. R., Woollings, T., Hawkins, E., and de Vries, H.: Robust future changes in temperature variability under greenhouse gas forcing

565 and the relationship with thermal advection, Journal of Climate, 29, 2221–2236, https://doi.org/10.1175/JCLI-D-14-00735.1, 2016.

Huntingford, C., Jones, P. D., Livina, V. N., Lenton, T. M., and Cox, P. M.: No increase in global temperature variability despite changing regional patterns, Nature, 500, 327–330, https://doi.org/10.1038/nature12310, 2013.

Huybers, P., Mckinnon, K. A., Rhines, A., and Tingley, M.: U.S. daily temperatures: The meaning of extremes in the context of nonnormality, Journal of Climate, 27, 7368–7384, https://doi.org/10.1175/JCLI-D-14-00216.1, 2014.

570 Jaeger, E. B. and Seneviratne, S. I.: Impact of soil moisture-atmosphere coupling on European climate extremes and trends in a regional climate model, Climate Dynamics, 36, 1919–1939, https://doi.org/10.1007/s00382-010-0780-8, 2011.

King, A. D.: The drivers of nonlinear local temperature change under global warming, Environmental Research Letters, 14, https://doi.org/10.1088/1748-9326/ab1976, 2019.

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

Lejeune, Q., Seneviratne, S. I., and Davin, E. L.: Historical land-cover change impacts on climate: Comparative assessment of LUCID and CMIP5 multimodel experiments, Journal of Climate, 30, 1439–1459, https://doi.org/10.1175/JCLI-D-16-0213.1, 2017.

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, Geoscientific Model Development, 12, 1477–1489, https://doi.org/10.5194/gmd-12-1477-2019, 2019.

Loikith, P. C. and Neelin, J. D.: Non-Gaussian cold-side temperature distribution tails and associated synoptic meteorology, Journal of Climate, 32, 8399–8414, https://doi.org/10.1175/JCLI-D-19-0344.1, 2019.

Matalas, N. C.: Mathematical assessment of synthetic hydrology, Water Resources Research, 3, 937–945, https://doi.org/10.1029/WR003i004p00937, 1967.

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

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

Neale, R. B., Richter, J. H., and Jochum, M.: The impact of convection on ENSO: From a delayed oscillator to a series of events, Journal of Climate, 21, 5904–5924, https://doi.org/10.1175/2008JCLI2244.1, 2008.

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J. F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., and Sanderson, B. M.: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6, Geoscientific Model Development, 9, 3461–3482, https://doi.org/10.5194/gmd-9-3461-2016, 2016.

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.

Pitman, A., De Noblet-Ducoudré, N., Avila, F., Alexander, L., Boisier, J.-P., Brovkin, V., Delire, C., Cruz, F., Donat, M., Gayler, V., Hurk, B. v. d., Reick, C., and Voldoire, A.: Effects of land cover change on temperature and rainfall extremes in multi-model ensemble 3 simulations, Earth System Dynamics, p. 213–231, 2012.

Popp, A., Calvin, K., Fujimori, S., Havlik, P., Humpenöder, F., Stehfest, E., Bodirsky, B. L., Dietrich, J. P., Doelmann, J. C., Gusti, M., Hasegawa, T., Kyle, P., Obersteiner, M., Tabeau, A., Takahashi, K., Valin, H., Waldhoff, S., Weindl, I., Wise, M., Kriegler, E., Lotze-Campen, H., Fricko, O., Riahi, K., and Vuuren, D. P.: Land-use futures in the shared socio-economic pathways, Global Environmental Change, 42, 331–345, https://doi.org/10.1016/j.gloenvcha.2016.10.002, 2017.

Potopová, V., Boroneanţ, C., Možný, M., and Soukup, J.: Driving role of snow cover on soil moisture and drought development during the growing season in the Czech Republic, International Journal of Climatology, 36, 3741–3758, https://doi.org/10.1002/joc.4588, 2016.

Richardson, C. W.: Stochastic simulation of daily precipitation, temperature, and solar radiation, Water Resources Research, 17, 182–190, https://doi.org/10.1029/WR017i001p00182, 1981.

Roe, S., Streck, C., Obersteiner, M., Frank, S., Griscom, B., Drouet, L., Fricko, O., Gusti, M., Harris, N., Hasegawa, T., Hausfather, Z., Havlík, P., House, J., Nabuurs, G. J., Popp, A., Sánchez, M. J. S., Sanderman, J., Smith, P., Stehfest, E., and Lawrence, D.: Contribution of the land sector to a 1.5 °C world, Nature Climate Change, 9, 817–828, https://doi.org/10.1038/s41558-019-0591-9, 2019.

Schlenker, W. and Roberts, M. J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change, Proceedings of the National Academy of Sciences, 106, 15 594–15 598, https://doi.org/10.1073/pnas.0906865106, 2009.

Schwingshackl, C., Hirschi, M., and Seneviratne, S. I.: Global Contributions of Incoming Radiation and Land Surface Conditions to Maximum Near-Surface Air Temperature Variability and Trend, Geophysical Research Letters, 45, 5034–5044, https://doi.org/10.1029/2018GL077794, 2018.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., Mc Innes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., Rusticucci, M., Semenov, V., Alexander, L. V., Allen, S., Benito, G., Cavazos, T., Clague, J., Conway, D., Della-Marta, P. M., Gerber, M., Gong, S., Goswami, B. N., Hemer, M., Huggel, C., Van den Hurk, B., Kharin, V. V., Kitoh, A., Klein Tank, A. M., Li, G., Mason, S., Mc Guire, W., Van Oldenborgh, G. J., Orlowsky, B., Smith, S., Thiaw, W., Velegrakis, A., Yiou, P., Zhang, T., Zhou, T., and Zwiers, F. W.: Changes in climate extremes and their impacts on the natural physical environment, Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, 9781107025, 109–230, https://doi.org/10.1017/CBO9781139177245.006, 2012.

Seneviratne, S. I., Wartenburger, R., Guillod, B. P., Hirsch, A. L., Vogel, M. M., Brovkin, V., Van Vuuren, D. P., Schaller, N., Boysen, L., Calvin, K. V., Doelman, J., Greve, P., Havlik, P., Humpenöder, F., Krisztin, T., Mitchell, D., Popp, A., Riahi, K., Rogelj, J., Schleussner, C. F., Sillmann, J., and Stehfest, E.: Climate extremes, land-climate feedbacks and land-use forcing at 1.5C, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376, https://doi.org/10.1098/rsta.2016.0450, 2018.

Sheridan, S. C. and Lee, C. C.: Temporal Trends in Absolute and Relative Extreme Temperature Events Across North America, Journal of Geophysical Research: Atmospheres, 123, 889–11, https://doi.org/10.1029/2018JD029150, 2018.

Shinoda, M.: Climate memory of snow mass as soil moisture over central Eurasia, Journal of Geophysical Research Atmospheres, 106, 33 393–33 403, https://doi.org/10.1029/2001JD000525, 2001.

Stéfanon, M., Drobinski, P., D'Andrea, F., and De Noblet-Ducoudré, N.: Effects of interactive vegetation phenology on the 2003 summer heat waves, Journal of Geophysical Research Atmospheres, 117, 1–15, https://doi.org/10.1029/2012JD018187, 2012.

Tamarin-Brodsky, T., Hodges, K., Hoskins, B. J., and Shepherd, T. G.: Changes in Northern Hemisphere temperature variability shaped by regional warming patterns, Nature Geoscience, 13, 414–421, https://doi.org/10.1038/s41561-020-0576-3, 2020.

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.

Thackeray, C. W., Derksen, C., Fletcher, C. G., and Hall, A.: Snow and Climate: Feedbacks, Drivers, and Indices of Change, Current Climate Change Reports, 5, 322–333, https://doi.org/10.1007/s40641-019-00143-w, 2019.

Thompson, D. W., Barnes, E. A., Deser, C., Foust, W. E., and Phillips, A. S.: Quantifying the role of internal climate variability in future climate trends, Journal of Climate, 28, 6443–6456, https://doi.org/10.1175/JCLI-D-14-00830.1, 2015.

Wang, Z., Jiang, Y., Wan, H., Yan, J., and Zhang, X.: Detection and attribution of changes in extreme temperatures at regional scale, Journal of Climate, 30, 7035–7047, https://doi.org/10.1175/JCLI-D-15-0835.1, 2017.

Wramneby, A., Smith, B., and Samuelsson, P.: Hot spots of vegetation-climate feedbacks under future greenhouse forcing in Europe, Journal of Geophysical Research Atmospheres, 115, 1–12, https://doi.org/10.1029/2010JD014307, 2010.

645 Xu, L. and Dirmeyer, P.: Snow-atmosphere coupling strength in a global atmospheric model, Geophysical Research Letters, 38, 1–5, https://doi.org/10.1029/2011GL048049, 2011.

Yeo, I.-K. and Johnson, R. A.: A New Family of Power Transformations to Improve Normality or Symmetry Author (s): In-Kwon Yeo and Richard A . Johnson Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : http://www.jstor.org/stable/2673623, Biometrika, 87, 954–959, 2000.

650 Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J. L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., and Asseng, S.: Temperature increase reduces global yields of major crops in four independent estimates, Proceedings of the National Academy of Sciences of the United States of America, 114, 9326–9331, https://doi.org/10.1073/pnas.1701762114, 2017.