



MESMER-RCM: a probabilistic climate emulator for regional warming projections

Hao Pan^{1,4}, Lukas Gudmundsson¹, Mathias Hauser¹, Jonas Schwaab^{2,3}, Yann Quilcaille¹, and Sonia I. Seneviratne¹

¹Institute for Atmospheric and Climate Science, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

²WSL, Institute for Snow and Avalanche Research (SLF), Davos Dorf, Switzerland

³Institute for Environmental Planning, Leibniz Universität Hannover, Hanover, Germany

⁴Department of Geography, Institute for Climate and Carbon Neutrality, The University of Hong Kong, Hong Kong, China

Correspondence: Hao Pan (haopan8080@connect.hku.hk)

Received: 26 June 2025 – Discussion started: 2 July 2025

Revised: 30 October 2025 – Accepted: 2 December 2025 – Published: 12 February 2026

Abstract. Regional Climate Model (RCM) emulators enable rapid and computationally efficient RCM projections given Global Climate Model (GCM) inputs, complementing dynamical downscaling by approximating physical representations with statistical models. However, while existing RCM emulators perform well in deterministic emulations, they do not sample internal RCM variability and remain computationally expensive. Here, we present MESMER-RCM, a probabilistic RCM emulator designed for spatially resolved annual 2 m temperature. MESMER-RCM is a generative model that enables both data-efficient learning and interpretability. It can generate large ensembles of synthetic, yet physically plausible, RCM realizations, capturing the internal RCM variability at a fraction of the computational cost. This work offers a fast and reliable RCM emulation framework, supporting finer-scale what-if analyses of regional climate responses and informing local adaptation and mitigation strategies.

1 Introduction

Warming at regional scales is a critical aspect of climate change as humans and natural systems are affected by local temperatures and not the global mean. Warming patterns vary spatially with geographical features, with amplification in urbanized and deforested regions (IPCC, 2023), as well as in mountainous areas (Mountain Research Initiative EDW Working Group, 2015). These heterogeneous warming patterns necessitate finer-scale climate projections, often achieved using Regional Climate Models (RCMs) downscaling Global Climate Model (GCM) simulations (Soares et al., 2022; Careto et al., 2022; Jacob et al., 2020; Roberts et al., 2019). However, these high-resolution simulations are computationally expensive and the need to explore uncertainty within multiple models and scenario pathways signif-

icantly increases computational demands. In addition, representing internal climate variability requires large initial-condition ensembles (Eyring et al., 2024; Schneider et al., 2023; Giorgi, 2019; Hawkins and Sutton, 2009), further escalating the overall cost.

Recent advancements in climate emulators address these challenges by simplifying model complexity, thus enabling rapid climate projections under various scenario pathways. Among these, RCM emulators have gained attention as efficient surrogates for RCMs, providing insights into regional climate projections. While both RCM emulators and statistical downscaling methods aim to bridge the gap between coarse-resolution GCM outputs and fine-scale regional climate information, they differ in their setup and objectives. RCM emulators are trained on GCM-RCM model chain simulations to reproduce RCM outputs, whereas statistical

downscaling methods rely on empirical relationships derived from observational data to establish connections between large-scale and local-scale climate variables. Thus, RCM emulators and statistical downscaling methods can inform each other. Leveraging both deep neural networks (Doury et al., 2023, 2024; Van Der Meer et al., 2023; Baño-Medina et al., 2022; Passarella et al., 2022) and classical statistic approaches (Hobeichi et al., 2023; Boé and Mass, 2022), recent advancements in RCM emulators have achieved remarkable accuracy and efficiency. Moreover, efforts have been made to enhance interpretability of RCM emulators (González-Abad et al., 2023; Baño-Medina et al., 2022) and to evaluate transferability across emission scenarios, initial-condition ensembles, and GCMs (Hernanz et al., 2024, 2022; Bano-Medina et al., 2023).

Despite these advancements, current RCM emulators focus primarily on deterministic prediction and do not fully account for the internal RCM variability, which is essential for useful regional climate projections. Recently, deep learning-based generative models have shown promising performance in probabilistic ensemble downscaling by enabling realistic sampling of internal variability (Lopez-Gomez et al., 2024; Watt and Mansfield, 2024; Ling et al., 2024; Bischoff and Deck, 2024; Tomasi et al., 2024; Mardani et al., 2024; Wan et al., 2023). However, the inherent complexity of these models poses challenges for interpretability and requires significant computational resources during training and inference. Furthermore, these highly flexible models may fail to achieve optimal performance due to the currently limited sample size of GCM-RCM simulations that are available for training. To address these gaps, we present MESMER-RCM, a probabilistic ensemble RCM emulator designed for spatially-resolved annual 2 m temperature projections. MESMER-RCM is a generative model that achieves both interpretability and data-efficient learning. This study focuses on an RCM emulator tailored for the Europe (EURO-CORDEX, Gutowski et al., 2016), with potential applicability to other regions. MESMER-RCM can be integrated with GCM emulators such as MESMER (Beusch et al., 2020b), providing a foundation for building an robust GCM-RCM emulator chain for future applications.

2 Data

MESMER-RCM is trained and tested using simulations from the EURO-CORDEX-CMIP5 experiment (Gutowski et al., 2016; Taylor et al., 2012). EURO-CORDEX-CMIP5 provides RCM simulations forced by the output from a specific GCM, named a GCM-RCM model chain. We use RCM simulations with a spatial resolutions of 0.44° and re-grid GCM data to a common resolution of 2.5°. The dataset aligns with the one used in the Fischer et al. (2018). This study focuses on annual mean 2 m air temperature, for the period 1971 to 2099, and land grid points in the RCM domain (−25 to

45° E, 26 to 72° N). Each RCM simulation consists of 129 years of data and land 5929 grid points. The dataset contains 42 simulations (Table S1 in the Supplement), from five RCMs and nine GCMs under the Representative Concentration Pathways (RCP) 2.6, 4.5, and 8.5 scenario pathways (Meinshausen et al., 2011).

The simulation matrix for GCM-RCM combinations in EURO-CORDEX is sparse, with SMHI-RCA4 dominating the available model chains (Table S1). Additionally, not all model chains provide all scenarios and there are only few which contributed initial-condition ensembles. These limitations and their implications for the methodology are discussed in detail in Sect. 5.3. Nonetheless, the dataset provides a diverse set of GCM-RCM combinations under key scenario pathways, making it suitable to evaluate the performance of MESMER-RCM in emulating different GCM-RCM model chains and exploring various scenario pathways.

3 MESMER-RCM: A framework for regional climate model emulation

3.1 MESMER-RCM structure

MESMER-RCM explicitly models the annual mean temperature at an RCM grid point i and time step t as a linear combination of the RCM response to the driving GCM and a residual climate variability term:

$$T_{i,t}^{\text{RCM}} = f(T_{s,t}^{\text{GCM}})_{i,t} + \eta_{i,t}^{\text{RCM}} \quad (1)$$

where $f(T_{s,t}^{\text{GCM}})_{i,t}$ denotes the deterministic response of the RCM at grid point i (for all i) to the GCM temperature at grid point s , and $\eta_{i,t}^{\text{RCM}}$ represents the residual variability. The deterministic response term captures the systematic relationship between GCM and RCM simulations, while the residual variability term approximates the additional internal variability implied by the RCM as a stochastic process. The schematic overview of MESMER-RCM is shown in Fig. 1.

3.2 Deterministic response module

We use Lasso multiple linear regression (Lasso MLR) for mapping the deterministic RCM temperature response of GCM in Eq. (1):

$$f(T_{s,t}^{\text{GCM}})_{i,t} = a_i^{\text{int}} + \sum_{s \in N_k(i)} a_{i,s} T_{s,t}^{\text{GCM}} \quad (2)$$

The predictors are the 2 m temperature at the k GCM grid points closest to the RCM grid point i (i.e., $s \in N_k(i)$). Here, $a_{i,s}$ represents the scaling coefficients that determine the influence of GCM grid points on RCM grid point i , while a_i^{int} captures the local intercept. The scaling coefficients $a_{i,s}$ for corresponding GCM predictors are optimized by minimizing a lasso loss that combines the least squares term and an L1 normalization term, $\lambda \cdot \sum_{s=1}^k |a_{i,s}|$ (James et al., 2013). The

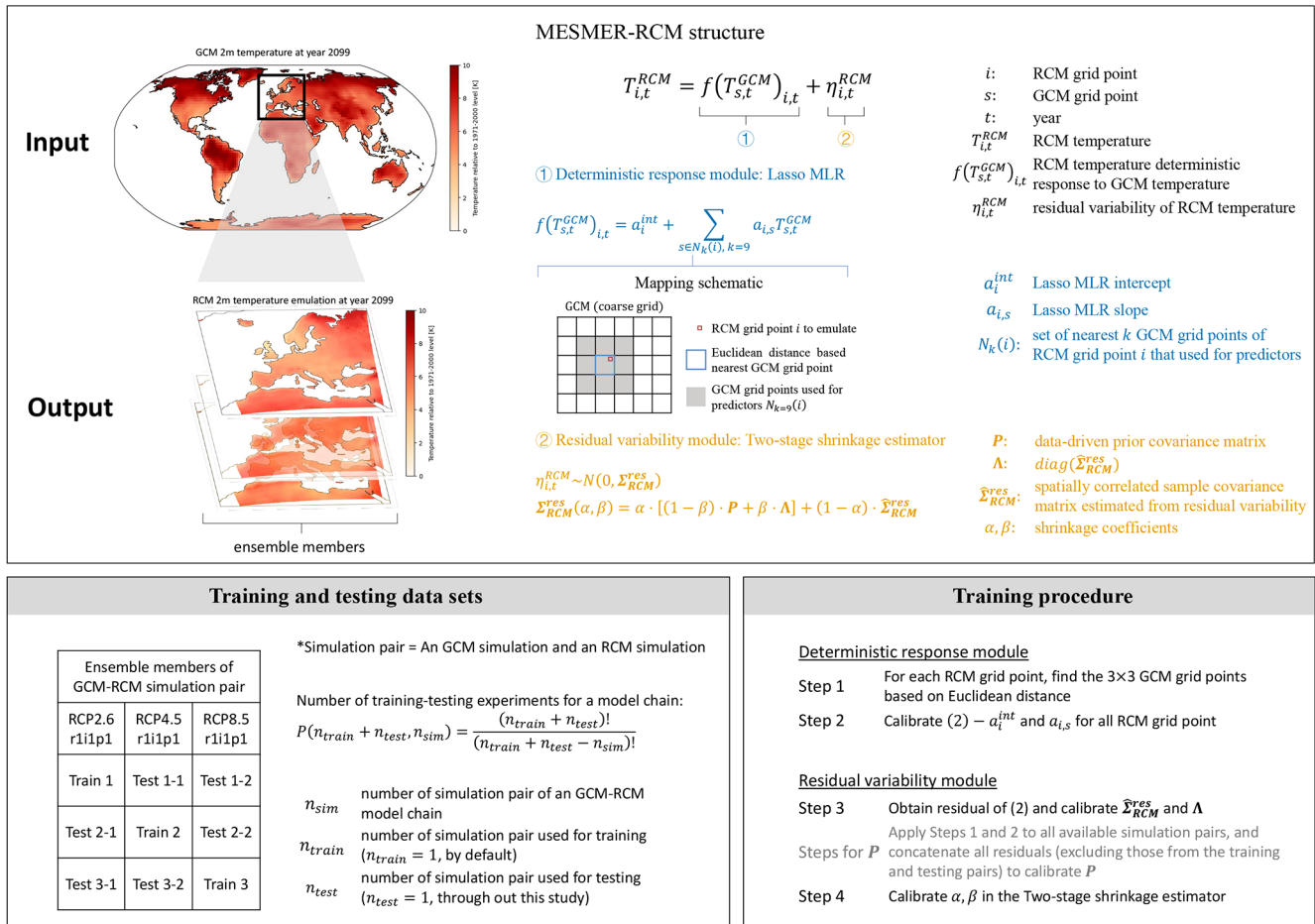


Figure 1. Schematic overview of the MESMER-RCM methodology, which statistically emulates regional climate model (RCM) outputs from global climate model (GCM) inputs. The upper box illustrates the mathematical formulation and structure of MESMER-RCM, showing how RCM temperatures are represented as statistical functions of GCM temperatures through a deterministic component and a stochastic residual. The lower-left box explains how the training and testing data sets are split. An example for a model chain with three simulation pairs is shown in the table: for each simulation pair i used for training (Train i), the remaining two simulation pairs are used for testing (Test $i - 1$ and Test $i - 2$), resulting in six training–testing experiments in total. The total number of training–testing experiments depends on the available simulation pairs, as indicated by the permutation formula in the box. The lower-right box outlines the training procedure of MESMER-RCM, which follows a modular training strategy. The steps related to the prior P are shown in gray text, as they need to be performed before Step 4 but not necessarily after Steps 1–3.

L1 regularization adaptively selects the most relevant GCM grid points for predicting RCM temperature while shrinking the scaling coefficients of less relevant contributions towards zero. In this study, we use the nine nearest GCM grid points surrounding each RCM grid point as predictors ($k = 9$). Specifically, we first identify the GCM grid point closest to the RCM grid point based on Euclidean distance, and then select the surrounding eight neighbors to form a 3×3 local grid that captures the GCM information around the RCM location.

3.3 Residual variability module

The residual variability term, $\eta_{i,t}^{RCM, res}$, is assumed to follow a multivariate Gaussian distribution (see Fig. S1 in the Supplement for the results of Shapiro–Wilk test for normality), $\eta_{i,t}^{RCM, res} \sim \mathcal{N}(0, \Sigma_{RCM}^{res})$, where Σ_{RCM}^{res} denotes the spatial covariance matrix across all considered RCM grid points, that is estimated from the residuals after subtracting the deterministic response. However, since each RCM simulation includes far more grid points than sample years, the sample covariance matrix of the residuals $\hat{\Sigma}_{RCM}^{res}$ becomes rank-deficient and needs to be regularized to improve the accuracy and stability of the estimate (Stein, 1956; Ledoit and Wolf, 2003; Pourahmadi, 2011). This is often achieved by shrinking the sample covariance matrix $\hat{\mathbf{S}}$ to a target covariance ma-

trix $\mathbf{T}$ that encapsulates prior knowledge on the dependence structure of the underlying process such that

$$\mathbf{S} = \alpha \mathbf{T} + (1 - \alpha) \hat{\mathbf{S}} \quad (3)$$

where α is the shrinkage parameter. Typical shrinkage targets are diagonal matrices such as the unit matrix since they are full rank. However, these assume independence for non-diagonal entries of the covariance matrix, which is physically not consistent with climate variability that is strongly correlated in space.

To address this limitation, we propose to leverage the full ensemble of RCM simulations to construct a data-driven prior, $\mathbf{P}$, for $\mathbf{T}$. The prior $\mathbf{P}$ is estimated by first concatenating the residuals of all RCM simulations, except those used for training and testing, and subsequent computations of an sample covariance matrix. Due to the increased sample size, $\mathbf{P}$ is nearly full rank and incorporates knowledge on the covariance structure derived from the available RCM simulations. However, $\mathbf{P}$ remains rank deficient and hence needs and is shrunk to $\mathbf{\Lambda} = \text{diag}(\hat{\mathbf{\Sigma}}_{\text{RCM}}^{\text{res}})$ which incorporates prior knowledge that sample variances typically have lower estimation errors than the covariances. The resulting shrinkage estimator can thus be written as

$$\mathbf{\Sigma}_{\text{RCM}}^{\text{res}}(\alpha, \beta) = \alpha \cdot [(1 - \beta) \cdot \mathbf{P} + \beta \cdot \mathbf{\Lambda}] + (1 - \alpha) \cdot \hat{\mathbf{\Sigma}}_{\text{RCM}}^{\text{res}} \quad (4)$$

4 Applying MESMER-RCM

4.1 Training and testing data sets

To train and test MESMER-RCM we use RCM simulations provided by the EURO-CORDEX ensemble and its driving GCM simulation from the CMIP5 ensemble (hereafter termed a simulation pair). We generally use one simulation pair for training and testing, respectively, except for some sensitivity test where we use more than one pair for training (Sect. 4.4). We train and test MESMER-RCM for each GCM-RCM model chain individually. The difference between training and testing sample thus comes from different forcing scenarios and/or initial conditions. We design the training–testing experiments by exhaustively considering all possible training–testing permutations of the available simulation pairs (Fig. 1). For instance, if there are two available simulation pairs in a model chain that differ in their scenarios, RCP4.5 and RCP8.5, we train MESMER-RCM on RCP4.5 and test on RCP8.5, and vice versa. The number of training–testing experiments for individual model chain can be calculated using the number of permutations, $P(n_{\text{train}} + n_{\text{test}}, n_{\text{sim}}) = \frac{(n_{\text{train}} + n_{\text{test}})!}{(n_{\text{train}} + n_{\text{test}} - n_{\text{sim}})!}$, where n_{sim} is the available number of simulation pairs, n_{train} and n_{test} are the number of simulation pairs used for training and testing, respectively.

In our framework, the prior $\mathbf{P}$ is decoupled from the training–testing design. For each training–testing experi-

ment, $\mathbf{P}$ is constructed from the residuals of all RCM simulations other than those selected for training and testing, and is then kept fixed and identical across both phases within the experiment.

4.2 Modular training and emulation

MESMER-RCM follows a modular training approach, where the deterministic response and the residual variability module are learned separately. Firstly, we calibrate a_i^{int} and $a_{i,s}$ by minimizing the lasso loss of Eq. (2) using coordinate descent. The lasso coefficients λ_s are determined through a cross-validation exercise based on minimum mean absolute error (James et al., 2013). The search range for λ_s spans from 10^{-4} to 10^1 with half-order-of-magnitude steps. In this study, we keep λ_s the same across all grid points as the the optimal lasso coefficients exhibit spatial smoothness.

Secondly, we derive the residual $\eta_{i,t}^{\text{RCM}}$ as the difference between the RCM temperature $T_{s,t}^{\text{RCM}}$ and the deterministic response $f(T_{s,t}^{\text{GCM}})$, the latter projected from the calibrated coefficients a_i^{int} and $a_{i,s}$ using the GCM reference temperature $T_{s,t}^{\text{GCM}}$. We then jointly calibrate the shrinkage coefficients α and β in Eq. (4) by maximizing the Gaussian log-likelihood through a grid search with 5-fold cross-validation. The search range for α is 0.1 to 1.0 in increments of 0.1, and for β it is 10^{-4} to 10^1 with half-order-of-magnitude steps. The prior $\mathbf{P}$ was constructed ahead and is kept fixed during the cross-validation experiments, while $\mathbf{\Lambda}$ and $\hat{\mathbf{\Sigma}}_{\text{RCM}}^{\text{res}}$ are re-estimated in each training fold. The log-likelihood is evaluated on the testing folds by measuring the distributional fit between the held-out data and $\mathcal{N}(0, \mathbf{\Sigma}_{\text{RCM}}^{\text{res}}(\alpha, \beta))$.

To generate an ensemble of M RCM emulations, we first apply the deterministic response module (Eq. 2), replicate its output M times, and then add spatially correlated residuals drawn from the residual variability module (Eq. 4) to each copy.

4.3 Evaluation of MESMER-RCM performance

We first evaluate the reliability of MESMER-RCM as a ensemble prediction system using a rank histogram (Wilks, 2006, 2019; Weigel et al., 2008; Jolliffe, 2012). Reliability here refers to the calibration of MESMER-RCM, and whether the RCM emulation spread well captures the RCM reference, and is hence a suitable metric to evaluate our probabilistic emulator. The idea of a rank histogram is as follow: for each ensemble prediction (or here emulation) consisting of M ensemble members, the verifying reference is combined with the ensemble members to form a vector of $M + 1$ values. The reference is ranked within this vector, where rank 1 is assigned to the smallest, and rank $M + 1$ to the largest value within the ensemble. If the reference is statistically indistinguishable from the ensemble predictions, then each rank is equally likely to occur. In such cases, the rank histogram will be flat, apart from small fluctuations

caused by sampling variability. Deviations from flatness in the rank histogram indicate MESMER-RCM miscalibration, arising either from bias in the mean due to the deterministic response module or from bias in the variance due to the residual variability module. We test the rank histogram flatness at the grid-point level using the χ^2 statistic (Wilks, 2019), and compute the χ^2 test pass rate defined as the percentage of grid points for which the null hypothesis of flatness is not rejected, for each training–testing experiment.

Noting that the rank histogram mostly focus on the correct specification of variance (Hamill, 2001), we complement it with an evaluation of the Gaussian likelihood function as it can take into account the full covariance structure. The likelihood function measures the probability of the RCM reference for each year given the emulated distribution $\mathcal{N}(f(T_{s,t}^{\text{GCM}}), \Sigma_{\text{RCM}}^{\text{res}})$, an analogy of emulator’s ability to emulate the distribution of the RCM reference. We then compare MESMER-RCM with benchmarks using the likelihood function. We propose benchmark methods using the similar structure as MESMER-RCM. Our benchmark for the deterministic response module is a simple linear regression, where the predictor is only the nearest GCM grid point temperature ($k = 1$ in Eq. 2), while we use the Ledoit–Wolf estimator (Ledoit and Wolf, 2004) as benchmark for the residual variability module. The Ledoit–Wolf estimator is commonly used to estimate the covariance matrix in high-dimensional settings, and has wide application in finance and machine learning. Additionally, we propose a naive benchmark for residual variability module using the sample variance matrix $\mathbf{\Lambda}$ that assumes spatial independence as reference. Since $\mathbf{\Lambda}$ is full rank, innovations can be directly drawn from a multivariate Gaussian $\mathcal{N}(0, \mathbf{\Lambda})$ without regularization.

4.4 Sensitivity to Training Configurations

Given the sparsity of GCM–RCM simulations in the EURO-CORDEX ensemble (Table S1), it is crucial to evaluate how model performance varies with different training sample size. In this study, we conduct three sensitivity experiments that use different samples to train the deterministic response module, the prior $\mathbf{P}$ in the residual variability module, and both, respectively.

1. To assess the robustness of the deterministic response module (Eq. 2), we train it using concatenated GCM–RCM simulation pairs from all available ensemble members, except the pair currently used for testing (i.e., $n_{\text{train}} = n_{\text{sim}} - n_{\text{test}}$, with $n_{\text{test}} = 1$).
2. The sample–consistency trade-off in the residual variability module (Eq. 4) is evaluated by constructing the prior $\mathbf{P}$ from a reduced set of concatenated RCM residuals, selected for their higher physical consistency with the training set.

3. Combined approach: Applying both (1) and (2) simultaneously. Performance differences are evaluated using the χ^2 test pass rate.

5 Results

5.1 Calibration of MESMER-RCM

The deterministic response module is calibrated for each grid-cell individually but working with the same Lasso coefficient ($\lambda = 0.005$) across all grid-cells and model chains. The common lasso coefficient is chosen based on a cross-validation exercise which showed that the optimized Lasso coefficients exhibit spatial smoothness, with the European land mean stabilizing around 0.005 across all simulations (Fig. S2 in the Supplement). A summary of the coefficients of the deterministic response module alongside some insights to their spatial variability is provided in the Supplement (Figs. S3, S4). The sum of scaling coefficients ($\sum_{s \in N_k(i)} a_{i,s}$) shows higher values in mountainous regions and high-latitude areas, highlighting the capacity of deterministic response module to capture elevation-dependent warming and polar amplification, respectively. The scaling coefficients indicate that RCM grid points in the Swiss Alps warm faster than its northwest and southeast GCM counterparts. This pattern may be related to the north/south foehn effect over the Alpine region (Chow et al., 2013).

The shrinkage coefficients (α, β) in the residual variability module is calibrated for each model chain. The resulting log-likelihood field associated with α and β are consistent across all model chains, where the optimum β is around 10^{-2} for stabilizing $\hat{\Sigma}_{\text{RCM}}^{\text{res}}$ and α is around 10^{-1} for effectively borrowing prior information from $\mathbf{P}$. An example of such a log-likelihood field is provided in the Supplement (Fig. S5).

5.2 Example of ensemble RCM emulations

Figure 2 presents an example of RCM emulations for an GCM–RCM model chain (RCM: SMHI-RCA4, GCM: CanESM2). The emulated temperature fields closely resemble the RCM reference, providing finer-scale details, such as variations induced by topography. The time series further illustrates MESMER-RCM’s ability to adjust the GCM trajectory to align with the RCM at the grid-point level, regardless of whether the driving GCM shows a smaller or larger trend. Additionally, the ensemble range effectively captures the variability of the RCM reference, demonstrating MESMER-RCM’s capability for sampling internal RCM variability.

5.3 Evaluation of the probabilistic RCM emulator

The validation of the example GCM–RCM model chain (RCM: SMHI-RCA4, GCM: CanESM2) is shown in Fig. 3. The rank histogram (Fig. 3a) shows spatially-aggregated results, where the flatness of the rank histogram indicates over-

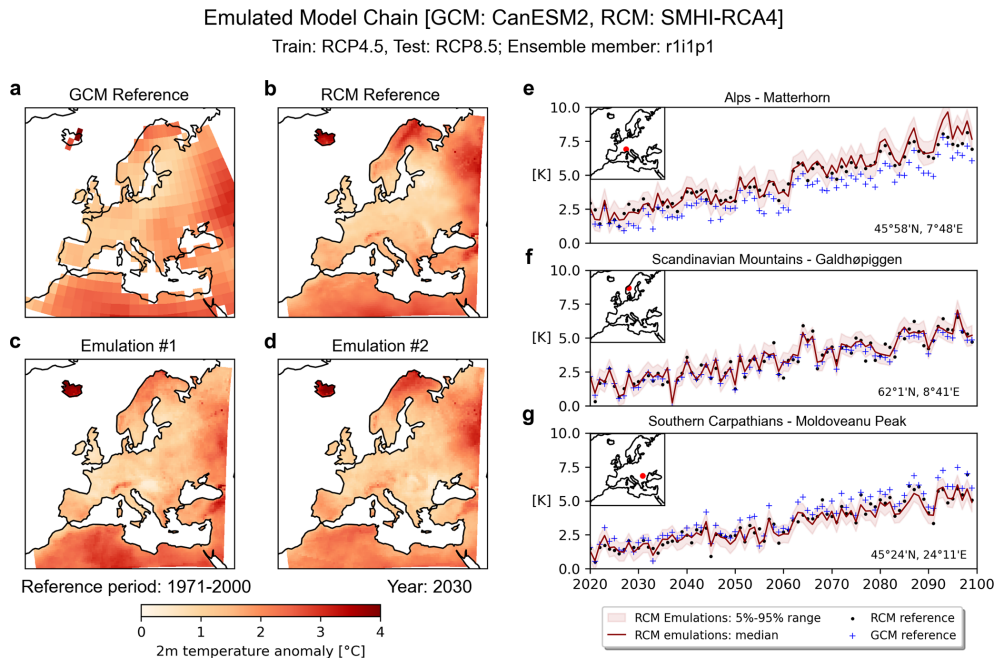


Figure 2. Examples of RCM emulations. Panels (a) to (d) show the annual mean temperature anomaly field for 2030 over land, including the GCM reference (a), RCM reference (b), and two example emulations (c–d). Panels (e) to (g) present time series of ensemble emulations at three mountainous RCM grid points, with GCM and RCM reference and the location shown in the inset. The range is computed using 1000 realizations of the emulator.

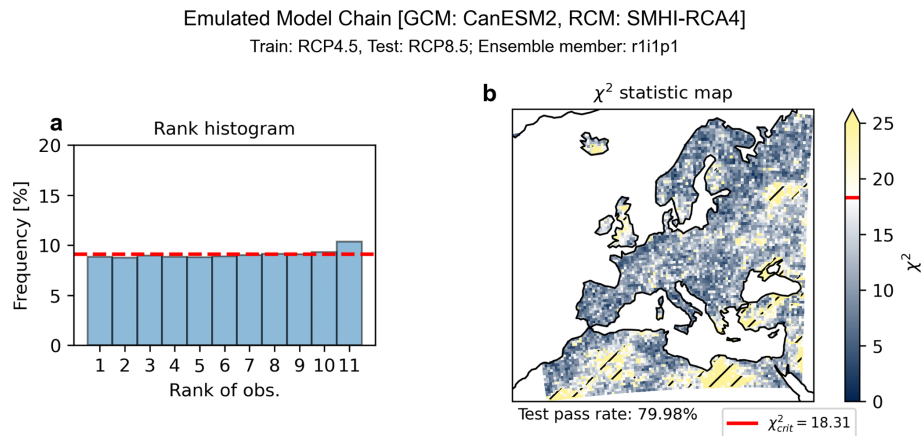


Figure 3. Evaluation of example model chain (RCM: SMHI-RCA4, GCM: CanESM2). The rank histogram based on all RCM grid points, where the red dash line indicates the ideal even distribution (a) and χ^2 test for rank histogram flatness at the RCM grid-point level (b). The hatched area in the χ^2 statistic map represents regions that fail the χ^2 test (i.e., $\chi^2 > \chi^2_{crit}$).

all reliable performance of the RCM emulator. The χ^2 test, applied at the grid-point level, shows that 80 % of RCM grid points exhibit significant rank uniformity, indicating generally reliable performance at the RCM grid-point level as well (Fig. 3b).

Extending the validation to all GCM-RCM model chains (Fig. 4), we observe fluctuations in the χ^2 test pass rate across different model chains, with a relatively large spread within individual chains. The example model chain SMHI-

RCA4_CanESM2 is one of the better performing ones, while e.g. KNMI-RACMO22E_EC-EARTH only has a pass rate of 40 %. This fluctuation can be largely understood by our conducted sensitivity experiments. Comparing to default setup, training the deterministic response module in multiple instead of a single simulation pairs (Fig. 4b) improves the robustness of the scaling coefficients, which reduces bias in temperature trend projections. The resulting spread of χ^2 pass rates across validation experiments for each individual

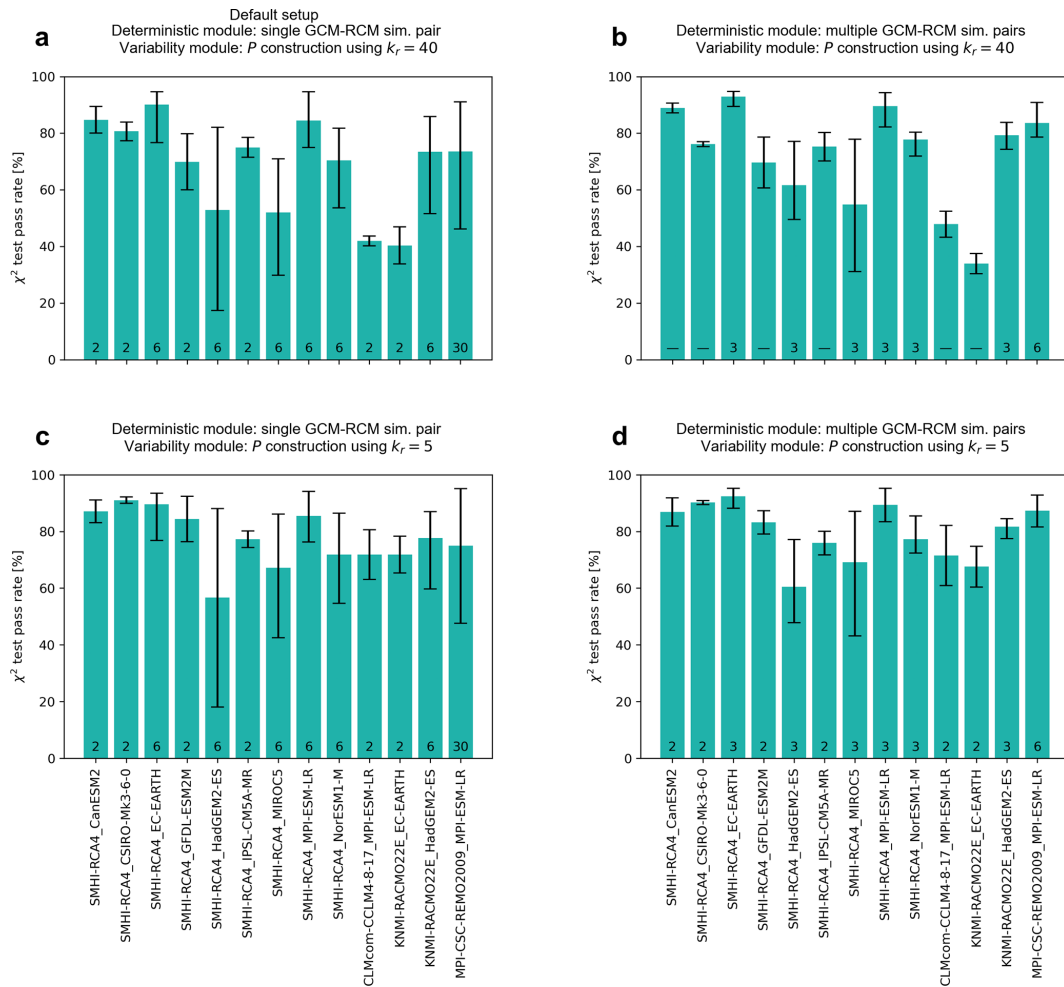


Figure 4. Comparison of training setups: χ^2 test pass rates across different GCM-RCM model chains. The x -axis lists the model chains, formatted as RCM-name_GCM-name. The number above each chain indicating the total validation experiments conducted. A dash (–) denotes cases where the sensitivity experiment could not be applied to a model chain due to data availability. In such cases, the experiment has no effect on the model chain, and the performance shown corresponds to the default setup. The prior covariance matrix $\mathbf{P}$ is constructed using residuals from the k_r nearest RCM simulations, determined by the Euclidean distance between their $\text{diag}(\hat{\Sigma}_{\text{RCM}}^{\text{res}})$ and that of the training set. Here, $k_r = 40$ corresponds to the total number of available residuals of RCM simulations.

model chain is considerably lower. For example, in MPI-CSC-REMO2009_MPI-GCM-LR, the average χ^2 test pass rate increases by 10 % compared to default setup, while the spread decreases from ± 20 % to ± 5 %. The next sensitivity experiment revisits the residual variability module, using a physically consistent subset of concatenated RCM residuals to construct the prior $\mathbf{P}$ (Fig. 4c). This improves the statistical representation of residual variability, leading to higher average χ^2 test pass rates. For instance, the pass rate increases by 30 % for KNMI-RACMO22E_HadGEM2-ES and CLMcom-CCLM4-8-17_MPI-ESM-LR. Due to limited data availability, the construction of prior $\mathbf{P}$ faces a trade-off between physical consistency and numerical stability. In cases where stability is compromised, the emulator may suffer from distortion due to ill-conditioned sampling. Nevertheless, this sensitivity experiment highlights the potential to

alleviate this trade-off by identifying and incorporating diverse, physically consistent model chains into the construction of $\mathbf{P}$, thereby improving the sampling of the residual variability. Finally, the combined approach (Fig. 4d) shows a clear additive improvement in RCM emulation, highlighting the complementary strengths of training strategies indicated by our sensitivity experiments (Fig. 4b–c). These results reveals that the observed inter-model performance variability primarily stems from unevenly distributed GCM–RCM training data, and the emulator performance is generally robust when trained with well-prepared inputs.

We further assess MESMER-RCM’s ability to emulate distribution of RCM reference using the log-likelihood, and compare it with different benchmarks (Fig. 5). Overall, MESMER-RCM provides more accurate and stable estimation of the RCM reference distribution compared to

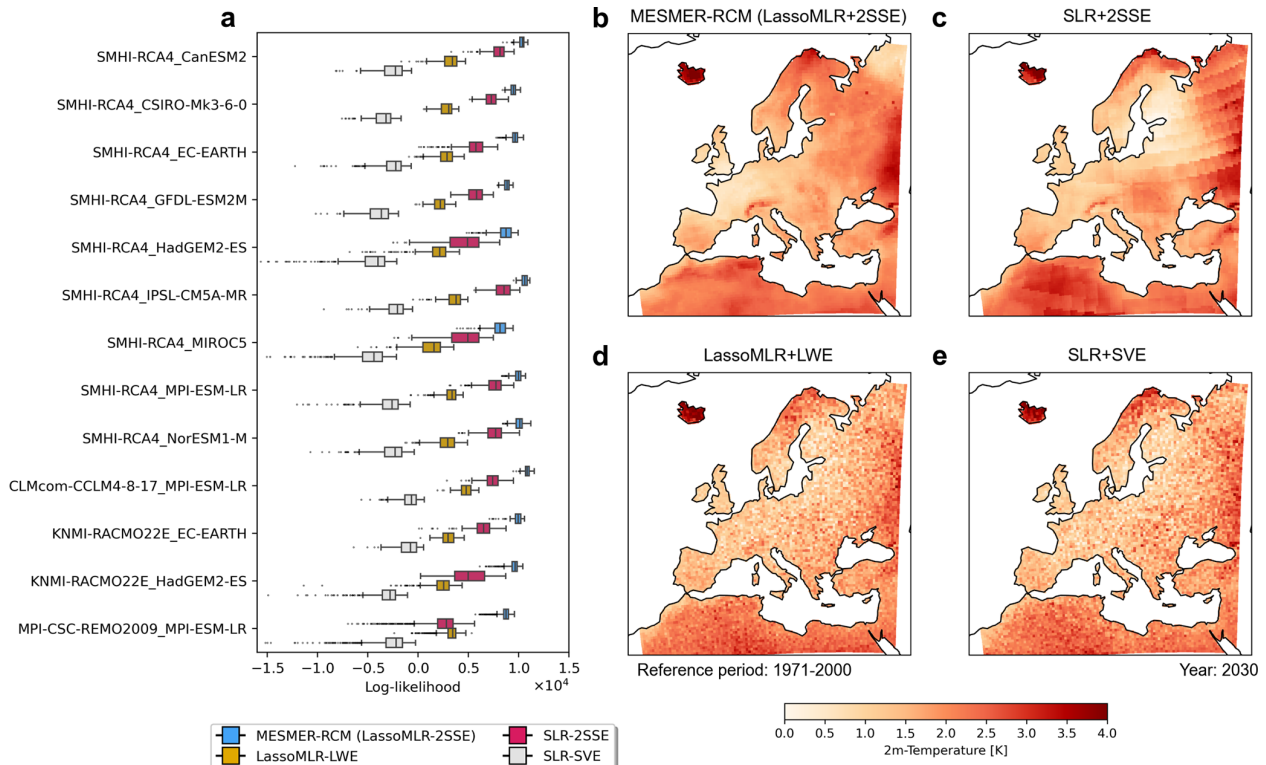


Figure 5. Comparison of MESMER-RCM with benchmark methods based on the log-likelihood of the RCM emulation, computed under a multivariate normal assumption from the spatial Gaussian distribution. **(a)** Yearly log-likelihood test performance aggregated into boxplots for each emulator across training–testing experiments of individual model chains. LassoMLR: lasso multiple linear regression; 2SSE: Two-stage shrinkage estimator; SLR: simple linear regression; LWE: Ledoit–Wolf estimator; SVE: sample variance estimator. **(b–e)** Example of emulation from MESMER-RCM and benchmark methods for the CanESM2-SMHI-RCA4 model chain, with emulators trained on RCP4.5 and tested on RCP8.5.

the benchmark methods, for all GCM-RCM model chains (Fig. 5a). The emulation generated by MESMER-RCM shows higher fidelity and physical plausibility (Fig. 5b). Comparing MESMER-RCM and SLR+2SSE, the simple linear regression as a nearest-neighbor method often leads to blocky artifacts (Allebach, 2005), which leaves behind the GCM grid texture and results in less smooth temperature fields (Fig. 5c). This distortion of temperature fields reflects a form of underfitting, which can be effectively improved by adding predictors and normalization using LassoMLR. Comparing MESMER-RCM and LassoMLR-LWE, the latter generates emulations with a noisy texture (Fig. 5d), because the sample covariance matrix is excessively shrunk to the unit matrix in LWE due to the high rank deficiency of the sample covariance matrix. MESMER-RCM has an advantage in this high-dimensional setting because the residual variability module adopts a more physically plausible prior for the shrinkage target and complement with two stage shrinkage procedure for ensuring numerical stability. This leads to a more favorable bias–variance trade-off by reducing estimation error while adding less bias.

6 Conclusions

This study develops a probabilistic RCM emulator, MESMER-RCM, which demonstrates reliable performance in emulating annual 2 m temperature projections over European land. MESMER-RCM is capable of exploring regional-scale warming under various scenario pathways and sampling internal RCM variability. Its modular structure enhances model interpretability and ensures computational efficiency, offering a simpler yet effective alternative to commonly used deep learning approaches. MESMER-RCM has the potential to be extended to other regions provided adequate simulation data are available. Our sensitivity experiments further underscore that emulator behavior depends strongly on the training data and that a sufficiently large set of GCM–RCM simulations is essential for building robust RCM emulators.

Future work on MESMER-RCM will leverage its ability to generate high-resolution ensemble projections from coarse-resolution data. Specifically, MESMER-RCM can be further integrated with GCM emulators to form a GCM-RCM emulator chain, adding value to GCM emulators by enhancing

their spatial resolution. For instance, MESMER generates spatially resolved annual mean temperature fields from arbitrary global mean temperature inputs. Coupling MESMER with MESMER-RCM enables regional climate projections across a wide range of GCM–RCM combinations and scenario pathways, thereby facilitating comprehensive what-if analysis. Such a framework holds strong potential for informing regional climate risk assessments and supporting policy-relevant decision-making.

Data availability. CMIP5 data is accessible through <https://wcrp-cmip.org/cmip-phases/cmip5/> (last access: 14 January 2025), and CORDEX simulations through <https://cordex.org/data-access/cordex-cmip5-data/> (last access: 14 January 2025).

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/npg-33-73-2026-supplement>.

Author contributions. H.P., L.G., M.H. and S.I.S. designed the study based on an initial idea from S.I.S. H.P. developed the methods with contributions from L.G. and M.H. H.P. wrote the manuscript with contributions from all authors.

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

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modeling, the coordinating bodies behind CORDEX and CMIP5. We acknowledge and thank the climate modeling groups that have participated in the CMIP5 and EURO-CORDEX initiatives for producing and making available their model output, and the Earth System Grid Federation (ESGF) for managing the database and the infrastructure. The post-processed model data of CMIP GCMs and CORDEX RCMs are provided by the Center for Climate Systems Modeling (C2SM) and ETH Zurich (Mathias Hauser, Urs Beyerle, Jan Rajczak, Silje Sørland, Curdin Spirig, Elias Zubler).

Financial support. The authors gratefully acknowledge funding from the Joint Initiative SPEED2ZERO that received support from the ETH-Board under the Joint Initiatives scheme and from

the Horizon Europe research and innovation programme (grant no. 101081369 (SPARCCLE)).

Review statement. This paper was edited by Jie Feng and reviewed by two anonymous referees.

References

- Allebach, J. P.: Image Scanning, Sampling, and Interpolation, in: *Handbook of Image and Video Processing*, 895–XXVII, Elsevier, ISBN 978-0-12-119792-6, <https://doi.org/10.1016/B978-012119792-6/50115-7>, 2005.
- Baño-Medina, J., Manzanar, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., and Gutiérrez, J. M.: Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44, *Geosci. Model Dev.*, 15, 6747–6758, <https://doi.org/10.5194/gmd-15-6747-2022>, 2022.
- Bano-Medina, J., Iturbide, M., Fernandez, J., and Gutierrez, J. M.: Transferability and explainability of deep learning emulators for regional climate model projections: Perspectives for future applications, *arXiv*, <https://doi.org/10.48550/arXiv.2311.03378>, 2023.
- Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Cross-breeding CMIP6 Earth System Models With an Emulator for Regionally Optimized Land Temperature Projections, *Geophysical Research Letters*, 47, e2019GL086812, <https://doi.org/10.1029/2019GL086812>, 2020b.
- Bischoff, T. and Deck, K.: Unpaired Downscaling of Fluid Flows with Diffusion Bridges, Artificial Intelligence for the Earth Systems, <https://doi.org/10.1175/AIES-D-23-0039.1>, 2024.
- Boé, J. and Mass, A.: A simple hybrid statistical-dynamical downscaling method for emulating regional climate models over Western Europe. Evaluation, application, and role of added value?, *Research Square*, <https://doi.org/10.21203/rs.3.rs-1585025/v1>, 2022.
- Careto, J. A. M., Soares, P. M. M., Cardoso, R. M., Herrera, S., and Gutiérrez, J. M.: Added value of EURO-CORDEX high-resolution downscaling over the Iberian Peninsula revisited – Part 2: Max and min temperature, *Geosci. Model Dev.*, 15, 2653–2671, <https://doi.org/10.5194/gmd-15-2653-2022>, 2022.
- Chow, F. K., De Wekker, S. F., and Snyder, B. J.: Chapter 4: Understanding and Forecasting Alpine Foehn, in: *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, edited by: Chow, F. K., De Wekker, S. F., and Snyder, B. J., Springer Atmospheric Sciences, 219–260, Springer Netherlands, ISBN 978-94-007-4097-6, <https://doi.org/10.1007/978-94-007-4098-3>, 2013.
- Doury, A., Somot, S., Gadat, S., Ribes, A., and Corre, L.: Regional Climate Model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach, *Climate Dynamics*, 60, 1751–1779, <https://doi.org/10.1007/s00382-022-06343-9>, 2023.
- Doury, A., Somot, S., and Gadat, S.: On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation, 62, 8587–8613, <https://doi.org/10.1007/s00382-024-07350-8>, 2024.

- Eyring, V., Collins, W. D., Gentile, P., and Barnes, E. A.: Pushing the frontiers in climate modelling and analysis with machine learning, <https://doi.org/10.1038/s41558-024-02095-y>, 2024.
- Fischer, A. M., Strassmann, K. M., Kotlarski, S., Schär, C., Sørland, S., and Zubler, E. M.: CH2018 Technical Report: CH2018 – Climate Scenarios for Switzerland, Chap. 2, p. 271, National Centre for Climate Services, Zurich, ISBN 978-3-9525031-4-0, 2018.
- Giorgi, F.: Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next?, *Journal of Geophysical Research: Atmospheres*, 2018JD030094, <https://doi.org/10.1029/2018JD030094>, 2019.
- González-Abad, J., Baño-Medina, J., and Gutiérrez, J. M.: Using Explainability to Inform Statistical Downscaling Based on Deep Learning Beyond Standard Validation Approaches, *arXiv*, <https://doi.org/10.48550/arXiv.2302.01771>, 2023.
- Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F.: WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6, *Geosci. Model Dev.*, 9, 4087–4095, <https://doi.org/10.5194/gmd-9-4087-2016>, 2016.
- Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, 129, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2), 2001.
- Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, *Bulletin of the American Meteorological Society*, 90, 1095–1108, <https://doi.org/10.1175/2009BAMS2607.1>, 2009.
- Hernanz, A., García-Valero, J. A., Domínguez, M., and Rodríguez-Camino, E.: A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment, *Atmospheric Science Letters*, 23, e1087, <https://doi.org/10.1002/asl.1087>, 2022.
- Hernanz, A., Correa, C., Sánchez-Perrino, J., Prieto-Rico, I., Rodríguez-Guisado, E., Domínguez, M., and Rodríguez-Camino, E.: On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues, *Atmospheric Science Letters*, 25, e1195, <https://doi.org/10.1002/asl.1195>, 2024.
- Hobeichi, S., Nishant, N., Shao, Y., Abramowitz, G., Pitman, A., Sherwood, S., Bishop, C., and Green, S.: Using Machine Learning to Cut the Cost of Dynamical Downscaling, *Earth's Future*, 11, e2022EF003291, <https://doi.org/10.1029/2022EF003291>, 2023.
- IPCC: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1st edn., ISBN 978-1-00-915789-6, <https://doi.org/10.1017/9781009157896>, 2023.
- Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M., Benestad, R., Boberg, F., Buonomo, E., Cardoso, R. M., Casanueva, A., Christensen, O. B., Christensen, J. H., Coppola, E., De Cruz, L., Davin, E. L., Dobler, A., Domínguez, M., Fealy, R., Fernandez, J., Gaertner, M. A., García-Díez, M., Giorgi, F., Gobiet, A., Goergen, K., Gómez-Navarro, J. J., Alemán, J. J. G., Gutiérrez, C., Gutiérrez, J. M., Güttler, I., Haensler, A., Halenka, T., Jerez, S., Jiménez-Guerrero, P., Jones, R. G., Keuler, K., Kjellström, E., Knist, S., Kotlarski, S., Maun, D., Van Meijgaard, E., Mercogliano, P., Montávez, J. P., Navarra, A., Nikulin, G., De Noblet-Ducoudré, N., Panitz, H.-J., Pfeifer, S., Piazza, M., Pichelli, E., Pietikäinen, J.-P., Prein, A. F., Preuschmann, S., Rechid, D., Rockel, B., Romera, R., Sánchez, E., Sieck, K., Soares, P. M. M., Somot, S., Srnec, L., Sørland, S. L., Termonia, P., Truhetz, H., Vautard, R., Warrach-Sagi, K., and Wulfmeyer, V.: Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community, *Regional Environmental Change*, 20, 51, <https://doi.org/10.1007/s10113-020-01606-9>, 2020.
- James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning, vol. 103 of Springer Texts in Statistics, Springer New York, ISBN 978-1-4614-7137-0, <https://doi.org/10.1007/978-1-4614-7138-7>, 2013.
- Jolliffe, I. T.: Forecast verification: a practitioner's guide in atmospheric science, Wiley & Sons, 2012.
- Ledoit, O. and Wolf, M.: Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance*, 10, 603–621, [https://doi.org/10.1016/S0927-5398\(03\)00007-0](https://doi.org/10.1016/S0927-5398(03)00007-0), 2003.
- Ledoit, O. and Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, 88, 365–411, [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4), 2004.
- Ling, F., Lu, Z., Luo, J.-J., Bai, L., Behera, S. K., Jin, D., Pan, B., Jiang, H., and Yamagata, T.: Diffusion model-based probabilistic downscaling for 180-year East Asian climate reconstruction, *npj Climate and Atmospheric Science*, 7, 131, <https://doi.org/10.1038/s41612-024-00679-1>, 2024.
- Lopez-Gomez, I., Wan, Z. Y., Zepeda-Núñez, L., Schneider, T., Anderson, J., and Sha, F.: Dynamical-generative downscaling of climate model ensembles, *arXiv*, <https://doi.org/10.48550/arXiv.2410.01776>, 2024.
- Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J., and Pritchard, M.: Residual Corrective Diffusion Modeling for Km-scale Atmospheric Downscaling, *arXiv*, <https://doi.org/10.48550/arXiv.2309.15214>, 2024.
- Meinshausen, M., Raper, S. C. B., and Wigley, T. M. L.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, *MAGICC6 – Part 1: Model description and calibration*, *Atmos. Chem. Phys.*, 11, 1417–1456, <https://doi.org/10.5194/acp-11-1417-2011>, 2011.
- Mountain Research Initiative EDW Working Group: Elevation-dependent warming in mountain regions of the world, *Nature Climate Change*, 5, 424–430, <https://doi.org/10.1038/nclimate2563>, 2015.
- Passarella, L. S., Mahajan, S., Pal, A., and Norman, M. R.: Reconstructing High Resolution ESM Data Through a Novel Fast Super Resolution Convolutional Neural Network (FSR-CNN), *Geophysical Research Letters*, 49, e2021GL097571, <https://doi.org/10.1029/2021GL097571>, 2022.
- Pourahmadi, M.: Covariance Estimation: The GLM and Regularization Perspectives, *Statistical Science*, 26, <https://doi.org/10.1214/11-STS358>, 2011.
- Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., Jackson, L. C., Kuhlbrodt, T., Mathiot, P., Roberts, C. D., Schiemann, R., Seddon, J., Vannière, B., and Vidale, P. L.: Description of the resolution hierarchy of

- the global coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments, *Geosci. Model Dev.*, 12, 4999–5028, <https://doi.org/10.5194/gmd-12-4999-2019>, 2019.
- Schneider, T., Behera, S., Boccaletti, G., Deser, C., Emanuel, K., Ferrari, R., Leung, L. R., Lin, N., Müller, T., Navarra, A., Ndiaye, O., Stuart, A., Tribbia, J., and Yamagata, T.: Harnessing AI and computing to advance climate modelling and prediction, *Nature Climate Change*, 13, 887–889, <https://doi.org/10.1038/s41558-023-01769-3>, 2023.
- Soares, P. M. M., Careto, J. A. M., Cardoso, R. M., Goergen, K., Katragkou, E., Sobolowski, S., Coppola, E., Ban, N., Belušić, D., Berthou, S., Caillaud, C., Dobler, A., Hodnebrog, Å., Kartsios, S., Lenderink, G., Lorenz, T., Milovac, J., Feldmann, H., Pichelli, E., Truhetz, H., Demory, M. E., De Vries, H., Warrach-Sagi, K., Keuler, K., Raffa, M., Tölle, M., Sieck, K., and Bastin, S.: The added value of km-scale simulations to describe temperature over complex orography: the CORDEX FPS-Convection multi-model ensemble runs over the Alps, *Climate Dynamics*, 62, 4491–4514, <https://doi.org/10.1007/s00382-022-06593-7>, 2022.
- Stein, C.: Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, in: *Contribution to the Theory of Statistics*, 197–206, University of California Press, ISBN 978-0-520-31388-0, <https://doi.org/10.1525/9780520313880-018>, 1956.
- Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, *Bulletin of the American Meteorological Society*, 93, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>, 2012.
- Tomasi, E., Franch, G., and Cristoforetti, M.: Can AI be enabled to dynamical downscaling? A Latent Diffusion Model to mimic km-scale COSMO5.0_CLM9 simulations, *arXiv*, <https://doi.org/10.48550/arXiv.2406.13627>, 2024.
- Van Der Meer, M., De Roda Husman, S., and Lhermitte, S.: Deep Learning Regional Climate Model Emulators: A Comparison of Two Downscaling Training Frameworks, *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003593, <https://doi.org/10.1029/2022MS003593>, 2023.
- Wan, Z. Y., Baptista, R., Boral, A., Chen, Y.-F., Anderson, J., Sha, F., and Zepeda-Nunez, L.: Debias Coarsely, Sample Conditionally: Statistical Downscaling through Optimal Transport and Probabilistic Diffusion Models, in: *Thirty-seventh Conference on Neural Information Processing Systems*, OpenReview, <https://openreview.net/forum?id=5NxJuc0T1P> (last access: 14 January 2025), 2023.
- Watt, R. A. and Mansfield, L. A.: Generative Diffusion-based Downscaling for Climate, *arXiv*, <https://doi.org/10.48550/arXiv.2404.17752>, 2024.
- Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitar, F., and Appenzeller, C.: Probabilistic Verification of Monthly Temperature Forecasts, 136, 5162–5182, <https://doi.org/10.1175/2008MWR2551.1>, 2008.
- Wilks, D. S.: Statistical methods in the atmospheric sciences, no. volume 91 in *International geophysics series*, Elsevier, 2nd edn., ISBN 978-0-12-751966-1, 2006.
- Wilks, D. S.: Indices of Rank Histogram Flatness and Their Sampling Properties, *Monthly Weather Review*, 147, 763–769, <https://doi.org/10.1175/MWR-D-18-0369.1>, 2019.