Towards a Climate Counterfactual Autoencoder

Frieder Loer*
Institute for Meteorology
Leipzig University
Germany, 04103
frieder.loer@uni-leipzig.de

Sebastian Sippel
Institute for Meteorology
Leipzig University
Germany, 04103
sebastian.sippel@uni-leipzig.de

Abstract

Separating forced climate change from internal climate variability is a fundamental challenge in climate attribution. In the emerging field of extreme event attribution, climate models are used to produce so-called 'storyline simulations' of climate extreme events: those events are simulated under a constrained atmospheric circulation in factual (=present-day) and counterfactual (=without anthropogenic forcing) climate conditions in order to attribute the thermodynamic effects of anthropogenic climate change. However, traditional approaches for producing such circulation-conditioned counterfactuals are computationally costly, cannot be directly transferred to observations, and cannot be easily transferred to other climate conditions than the ones simulated. Here we show that deep learning offers large potential to generate highly versatile climate counterfactuals: we use a Variational Autoencoder to predict counterfactual European winter temperatures by providing only the global mean warming level (i.e., the background climate) and the atmospheric circulation state as inputs. The results are benchmarked against traditional nudged-circulation climate model simulations. The deep learning based counterfactuals are shown to perform extremely well and can be applied in any background climate state, thus providing a versatile climate counterfactual generator. Future work could target counterfactual climate states based on observed weather states. Accurate climate counterfactuals could strongly support climate adaptation and communication efforts.

1 Introduction

Climate attribution aims to establish the causal drivers of observed changes in the climate system. This requires separating effects that originate from outside of the climate system (externally forced) and those of internal climate variability which exists due to the chaos in atmospheric dynamics [1].

Hence, a fundamental goal in climate attribution is to estimate how forced climate change contributed to certain (extreme) weather events (as opposed to internal variability, for example). The so-called 'storyline method' [2] asks how a specific event would have unfolded in the absence of climate change. Hence, it aims to compare an observed ('factual') climate event to a hypothetical, 'counterfactual' climate, that is without anthropogenic forcing (or a different level of forcing). It approaches this question by comparing similar atmospheric dynamic conditions under different thermodynamic (factual and counterfactual) climate conditions [2]. Because forced changes in atmospheric dynamics are highly uncertain [3], reflecting to a very large extent internal variability in individual events, storyline attribution focuses on forced changes in thermodynamical conditions conditional on a certain atmospheric circulation state. Both statistical and model simulation techniques exist that aim to reconstruct the events conditional on the atmospheric circulation state [4–6].

*Corresponding author.

Tackling Climate Change with Machine Learning workshop at NeurIPS 2025

Deep learning techniques are now being studied for the creation of counterfactual storylines to overcome the limitations of traditional methods like computational costs and difficult transferability to observations and other climate conditions. For example, convolutional neural networks are used to create counterfactual estimates of regional mean temperatures [7] or counterfactuals are created from deep learning driven weather forecasts [8].

Here, we evaluate whether deep learning based variational autoencoders could generate accurate climate counterfactuals for storyline-based climate attribution. We train variational autoencoders to reconstruct spatial temperature fields, thus implicitly inferring the regional climate change signal. Subsequently, we evaluate our variational autoencoder predictions for counterfactual climate conditions against nudged-circulation climate simulations that represent counterfactual storylines generated via numerical climate models. Specifically, we use the Latent Linear Adjustment Autoencoder framework (LLAAE, [9]) to predict European counterfactual temperatures from a proxy for atmospheric circulation and a covariate on forced global mean temperature. The circulation state defines the weather event of interest and is used to translate it across different climates, answering the question: 'What temperature (field) would this specific circulation state produce in a different, counterfactual climate?'. This method paves the way towards efficient creation of counterfactuals and could ultimately create observational counterfactuals for any background climate state. The goal of this study is to highlight the potential of deep learning, specifically, of the LLAAE architecture, for creating highly versatile climate counterfactuals.

2 Datasets and methods

2.1 Community Earth System Model 2

For training the LLAAE architecture, we use the Community Earth System Model 2 Large Ensemble (CESM2-LE) consisting of 100 members [10]. Members differ only by minimally perturbed initial conditions, thus each simulation represents one physically plausible evolution of the climate system from 1850-2100. The ensemble thus captures the distribution of possible climate system trajectories [10, 11]. Importantly, the initial condition large ensemble allows for obtaining the deterministic global warming signal (forced response) as the ensemble mean, averaging out the contributions from internal climate variability [11]. Additionally, we use six simulations of the CESM2 climate model (CESM2-ETH ensemble) that were simulated individually with equivalent configurations as the CESM2-LE [12]. Further, we test the LLAAE-predicted counterfactuals against three nudged-circulation preindustrial CESM2 control simulations [12, 13]. In these simulations, the horizontal wind field of a preindustrial climate simulation is driven ('nudged') towards the wind field of a corresponding historical simulation undergoing forced climate change [14]. Nudged-circulation climate simulations provide the counterfactuals that our method seeks to generate, and thus an ideal benchmark.

2.2 Methods

Latent Linear Adjustment Autoencoder The LLAAE extends a traditional variational autoencoder by linearly regressing the latent space onto a set of predictors [9]. During training, the LLAAE encodes and decodes fields of European surface air temperature while simultaneously predicting the latent space from a proxy for atmospheric circulation. This is a well-suited predictor, because temperature variability in mid-latitude winter is driven to a large extent by atmospheric circulation variability [15]. We add the covariate on forced global mean temperature (fGMT) to the set of predictors, which accounts for the level of forced global warming. By adjusting this covariate, the fitted model is expected to construct SAT fields from a specific circulation state in different levels of forced climate change. We anticipate that the model's learning objective allows sufficiently reducing the representation of the SAT field under a changing climate into the lower-dimensional latent space such that it can be reasonably approximated by a linear model. The LLAAE combines the flexible power of a Variational Autoencoder, representing generative machine learning methods, with the interpretability of a linear model and predicts a coherent spatial field.

Model setup We predict European surface air temperature (T) from the sea level pressure (SLP) field (covering the North Atlantic and Europe), a proxy of atmospheric circulation (Figure S1). The CESM2-LE forced response serves as the global warming covariate fGMT. The LLAAE was tuned for

latent dimensionality and number of included SLP predictors, and trained with the Adam optimizer. [16]. Further experimental details, tuning results and training parameter values are given in the appendix.

Model evaluation Factual and counterfactual LLAAE predictions are evaluated against factual simulations and counterfactual, nudged-circulation simulations, respectively. We use the coefficient of determination R^2 and the RMSE to evaluate our model predictions. Both are calculated on the grid cell level, thereby spatially resolving model performance. The RMSE quantifies the absolute prediction errors, complementing the variability and correlation information conveyed by the R^2 value.

Table 1: Evaluation metrics of LLAAE-predicted temperature fields; shown are the mean value (μ) and standard deviation (σ) of the corresponding spatial field (Figures S2 and S3). Metrics per grid cell are computed between true and LLAAE-predicted time series and averaged over three test members.

	$\mu(R^2)$	$\sigma(R^2)$	μ(RMSE)	σ(RMSE)
Factual	0.91	0.02	0.65	0.19
Counterfactual	0.84	0.05	0.82	0.31

3 Results: Factual and counterfactual predictions

We begin by predicting and evaluating factual temperature maps with the LLAAE. The LLAAE shows overall high skill, as reflected by the evaluation metrics in Table 1. These indicate high similarity between predicted and true time series. This is furthermore supported by the high explainability of atmospheric circulation for short-term cold-season temperature variability, as is confirmed by earlier studies [17]. Predictions of individual factual temperature fields are shown in column two of Figure 1. These results demonstrate the overall skill of the model in generating temperature fields before intervening on the value of the forced response to obtain counterfactual temperature estimates.

The forced response covariate is set to zero to translate weather patterns into the preindustrial climate for counterfactual predictions. Table 1 still shows high skill in predicting counterfactual temperatures, when evaluated against the respective nudged-circulation simulation (compare Figure S4). Modeling accurate counterfactuals is a key challenge for the LLAAE. This is because the test distribution of nudged-circulation temperatures entails a distributional shift from the training data due to the physical interventions of prescribed circulation and the absence of anthropogenic forcings.

Overall, the results show that the LLAAE successfully removes the underlying forced warming signal. This demonstrates that the model has satisfactorily learned the influence of the forced response covariate, as all other predictors remain constant compared to the factual predictions. The result suggests that the linear model constrains the latent space well enough, such that the same atmospheric circulation states are separated in the latent space due to the forced response covariate. Future work could analyze the latent space structure, as shown in recent studies on European heat extremes [18] and on the identification of impactful weather regimes [19].

Individual counterfactual temperature fields are shown in Figure 1. Looking for example at the third row, a factual test sample in column one is compared with the corresponding nudged-circulation simulation in column three, which is overall colder due to the missing climate change signal but retains the same circulation-driven temperature pattern. The LLAAE counterfactual in column four closely matches this truth, though with reduced spatial detail. Overall, the LLAAE counterfactuals compare well to the true fields. The LLAAE-predicted counterfactuals perform better than counterfactuals predicted by a principal component linear regression benchmark model over most parts of the European domain (Figure S5). The linear benchmark is a standard method in earlier studies [20], which we extended with a predictor for the forced response to have identical predictors as the LLAAE.

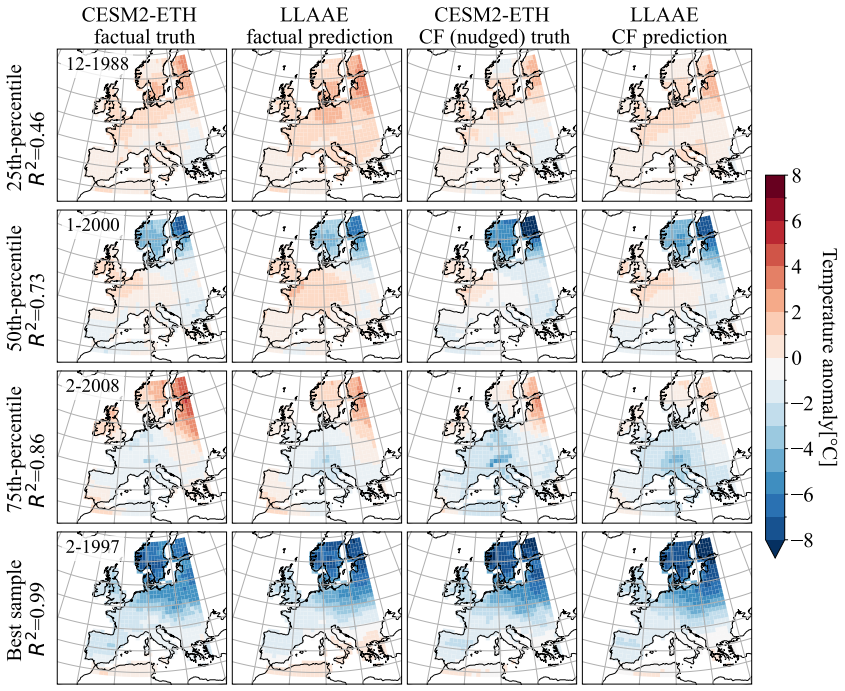

Figure 1: Temperature samples chosen from the R^2 distribution of LLAAE-predicted counterfactuals. The first column shows factual simulations with factual LLAAE predictions next to it in column two. Columns three and four show true (nudged) and predicted counterfactuals (CF) respectively.

4 Outlook and impact

Our results show that generative machine learning, specifically, a variational autoencoder, provides a powerful approach towards creating versatile and computationally efficient counterfactual storylines. The generated temperature fields agree well with nudged-circulation climate simulations, for which our method provides a data-driven equivalent.

There is a physical limit to the fraction of variability that can be explained by SLP and the forced response in a target variable. Other drivers like surface albedo, soil moisture, and evapotranspiration influence local temperatures. While including more such predictors will likely increase the quality of factual temperature predictions, their uncertain counterfactual values would make corresponding LLAAE counterfactuals unreliable. By design, the LLAAE answers the introductory question with the expected average temperature under the given atmospheric circulation. While this is valuable, even greater insights would come from the temperature distribution around this mean, reflecting contributions from other temperature-driving factors. These distributions would be highly valuable future work for making comprehensive attribution statements, including those about extreme events. We envision constructing such distributions with distributional autoencoders in the future.

We anticipate that our climate counterfactual generation method will add to the attribution toolbox: It provides a powerful, highly versatile climate counterfactual generation technique that will be extended to observations to derive fast, accurate and high-resolution estimates of 'today's (or any other day's) weather in a different climate'. Climate attribution actively supports managing the global climate crisis through improving the understanding of complex climate change signals and dynamics [21]. By estimating and communicating climate change effects, adaptation measures and the support for climate mitigation policies may be enhanced [22].

Acknowledgements Frieder Loer and Sebastian Sippel acknowledge the project 'Economics of Connected Natural Commons: Atmosphere and Biodiversity' (Research Training Group 2939), funded by the German Research Foundation.

References

[1] Clara Deser, Reto Knutti, Susan Solomon, and Adam S. Phillips. Communication of the role of natural variability in future north american climate. *Nature Climate Change*, 2(11):775–779, 2012.

[2] Theodore G. Shepherd. A Common Framework for Approaches to Extreme Event Attribution. *Current Climate Change Reports*, 2(1):28–38, 2016.

[3] Theodore G. Shepherd. Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geoscience*, 7(10):703–708, 2014.

[4] Ed Hawkins, Gilbert P. Compo, and Prashant D. Sardeshmukh. ESD Ideas: Translating historical extreme weather events into a warmer world. *Earth System Dynamics*, 14(5):1081–1084, 2023.

[5] Linda Van Garderen, Frauke Feser, and Theodore G. Shepherd. A methodology for attributing the role of climate change in extreme events: a global spectrally nudged storyline. *Natural Hazards and Earth System Sciences*, 21(1):171–186, 2021.

[6] Marina Baldissera Pacchetti, Liese Coulter, Suraje Dessai, Theodore G. Shepherd, Jana Sillmann, and Bart Van Den Hurk. Varieties of approaches to constructing physical climate storylines: A review. *WIREs Climate Change*, 15(2):e869, 2024.

[7] Jared T. Trok, Elizabeth A. Barnes, Frances V. Davenport, and Noah S. Diffenbaugh. Machine learning–based extreme event attribution. *Science Advances*, 10(34):eadl3242, 2024.

[8] B. Jiménez-Esteve, D. Barriopedro, J. E. Johnson, and R. García-Herrera. AI-Driven Weather Forecasts to Accelerate Climate Change Attribution of Heatwaves. *Earth's Future*, 13(8):e2025EF006453, 2025.

[9] Christina Heinze-Deml, Sebastian Sippel, Angeline G. Pendergrass, Flavio Lehner, and Nicolai Meinshausen. Latent Linear Adjustment Autoencoder v1.0: a novel method for estimating and emulating dynamic precipitation at high resolution. *Geoscientific Model Development*, 14(8):4977–4999, 2021.

[10] Keith B. Rodgers, Sun-Seon Lee, Nan Rosenbloom, Axel Timmermann, Gokhan Danabasoglu, Clara Deser, Jim Edwards, Ji-Eun Kim, Isla R. Simpson, Karl Stein, Malte F. Stuecker, Ryohei Yamaguchi, Tamás Bódai, Eui-Seok Chung, Lei Huang, Who M. Kim, Jean-François Lamarque, Danica L. Lombardozzi, William R. Wieder, and Stephen G. Yeager. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4):1393–1411, 2021.

[11] C. Deser, F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D. E. Horton, J. E. Kay, R. Knutti, N. S. Lovenduski, J. Marotzke, K. A. McKinnon, S. Minobe, J. Randerson, J. A. Screen, I. R. Simpson, and M. Ting. Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 10(4):277–286, 2020.

[12] Ana Bastos, Sebastian Sippel, Dorothea Frank, Miguel D. Mahecha, Sönke Zaehle, Jakob Zscheischler, and Markus Reichstein. A joint framework for studying compound ecoclimatic events. *Nature Reviews Earth & Environment*, 4(5):333–350, 2023.

[13] Jitendra Singh, Sebastian Sippel, Lei Gu, et al. Externally forced circulation changes amplify mid-latitude regional heat extremes in climate model nudged-circulation experiments. *Research Square*, 2025. Preprint (Version 2).

[14] National Center for Atmospheric Research, CAM6.3 User's Guide: Physics Modifications via the Namelist. `https://ncar.github.io/CAM/doc/build/html/CAM6.3_users_guide/physics-modifications-via-the-namelist.html`. Accessed: 2025-08-20.

[15] C. Deser and A. S. Phillips. A range of outcomes: the combined effects of internal variability and anthropogenic forcing on regional climate trends over europe. *Nonlinear Processes in Geophysics*, 30(1):63–84, 2023.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017. arXiv:1412.6980.

[17] Sebastian Sippel, Nicolai Meinshausen, Anna Merrifield, Flavio Lehner, Angeline G. Pendergrass, Erich Fischer, and Reto Knutti. Uncovering the Forced Climate Response from a Single Ensemble Member Using Statistical Learning. *Journal of Climate*, 32(17):5677–5699, 2019.

[18] A. Paçal, B. Hassler, K. Weigel, M.-Á. Fernández-Torres, G. Camps-Valls, and V. Eyring. Understanding european heatwaves with variational autoencoders. *EGUsphere*, 2025. preprint.

[19] Fiona R. Spuler, Marlene Kretschmer, Yevgeniya Kovalchuk, Magdalena Alonso Balmaseda, and Theodore G. Shepherd. Identifying probabilistic weather regimes targeted to a local-scale impact variable. *Environmental Data Science*, 3, 2024.

[20] Claudio Saffioti, Erich M. Fischer, Simon C. Scherrer, and Reto Knutti. Reconciling observed and modeled temperature and precipitation trends over Europe by adjusting for circulation variability. *Geophysical Research Letters*, 43(15):8189–8198, 2016.

[21] Rachel A. James, Richard G. Jones, Emily Boyd, Hannah R. Young, Friederike E. L. Otto, Christian Huggel, and Jan S. Fuglestvedt. *Attribution: How Is It Relevant for Loss and Damage Policy and Practice?*, pages 113–154. Springer International Publishing, Cham, 2019.

[22] Viktoria Cologna, Simona Meiler, Chahan M. Kropf, Samuel Lüthi, Niels G. Mede, and David N. Bresch et al. Extreme weather event attribution predicts climate policy support across the world. *Nature Climate Change*, 15(7):725–735, 2025.

A Appendix

A.1 Model setup

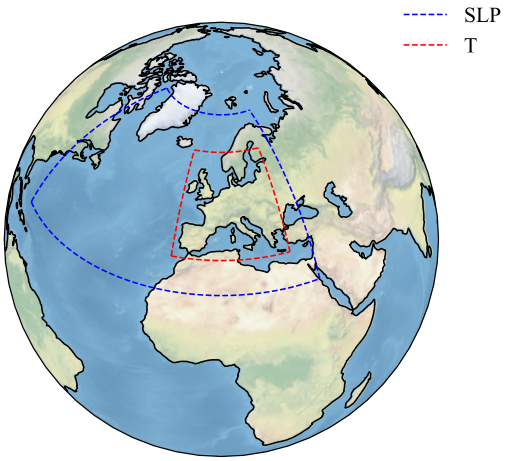

Figure S1: European surface air temperature domain (red box) and North Atlantic Sea Level Pressure domain (blue box).

For simplicity, we use monthly values of the cold season months December, January and February (DJF) in 1850-2100. We use monthly temperature anomalies with respect to the reference period 1850-1900.

Sea level pressure predictors We use SLP anomalies calculated with respect to the large ensemble mean. Instead of inputting the SLP field directly, we project the SLP field of each month onto the (leading 1000) principal components of the SLP dataset to obtain the principal components scores (this could be equivalently done for observations). We compute the SLP principal components using all winter months (DJF) of the CESM2-LE.

Data splitting We use the entire CESM2 LE for training the LLAAE. Three simulations from the CESM2-ETH ensemble are used for model validation, and the three CESM2-ETH runs (1300, 1400, 1500) that have corresponding nudged runs are used for testing. The data splitting is indicated in Table S1.

Table S1: Data splitting into train, validation and test set.

Train	**Validation**	**Test factual**	**Test counterfactual**
100 LE members	CESM2-ETH members 1000, 1100, 1200	factual CESM2-ETH members 1300, 1400, 1500	nudged CESM2-ETH members 1300, 1400, 1500

Tuning results We tuned the number of latent dimensions and the number of included SLP predictors (number of included SLP principal components) to 50 and 1000 respectively.

Training parameters The LLAAE is trained using the Adam optimizer [9, 16]. Here, the Adam optimizer uses the following staircase exponential schedule for adjusting the learning rate during training

$$\eta(t) = \eta_0 \times \gamma^{\frac{t}{n}} \,. \tag{1}$$

η : learning rate at step t, η_0 : initial learning rate
t : step number, n : number of decay steps, γ : decay rate

The step number t refers to the number of processed batches. The training parameters are: (η_0: 10^{-3}, n: 5000, γ: 0.96, batch size: 64).

A.2 Model performance

A.2.1 Factual LLAAE predictions

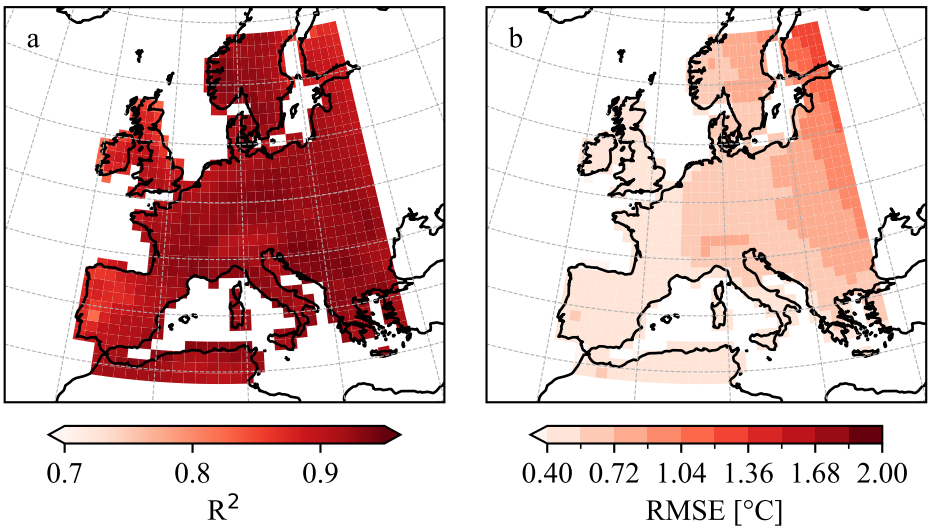

Figure S2: Spatial distribution of R^2 and RMSE. Values are calculated in each grid cell for the temperature time series of factual LLAAE prediction and the corresponding factual climate simulation in the test set. Shown are the mean values per grid cell among the three test members in the period 1950-2100.

A.2.2 Counterfactual LLAAE predictions

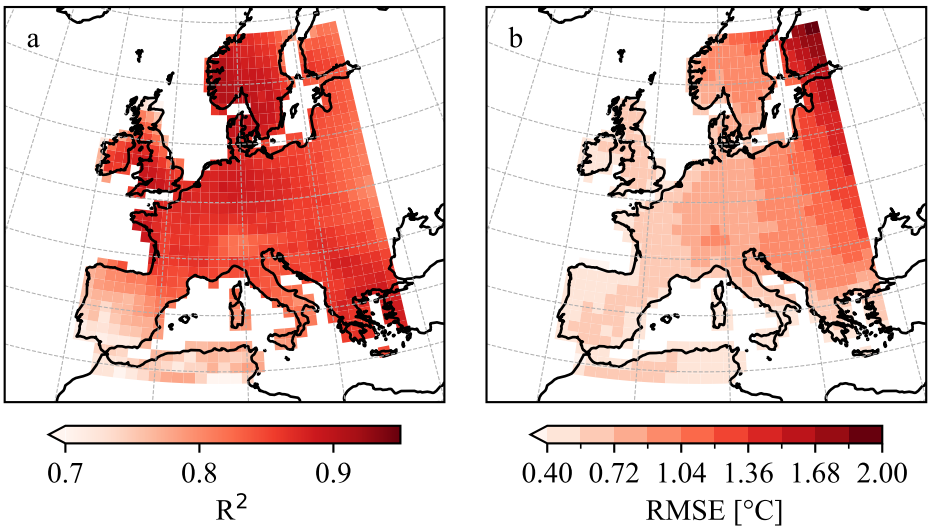

Figure S3: Spatial distribution of R^2 and RMSE. Values are calculated in each grid cell for the temperature time series of LLAAE counterfactual prediction and the corresponding nudged simulation (i.e. truth) in the test set. Shown are the mean values per grid cell among the three test members in the period 1950-2100.

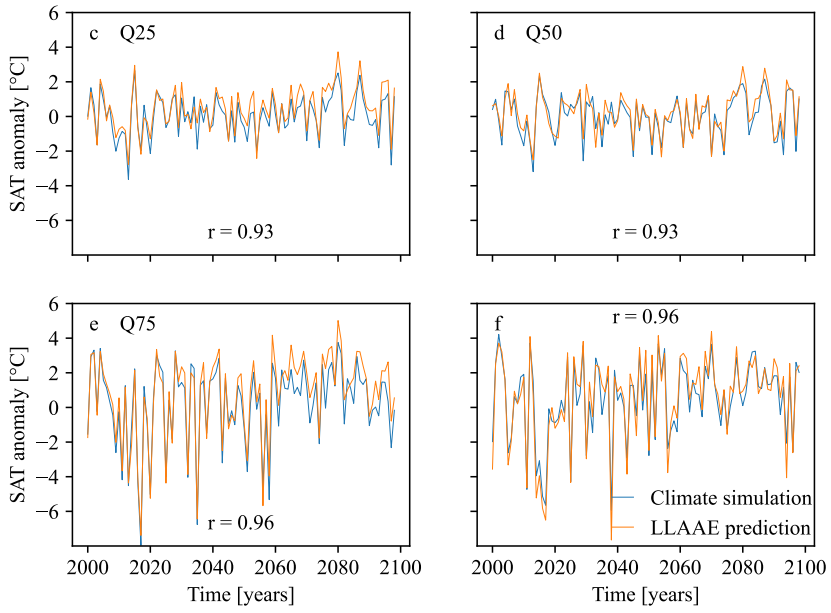

Figure S4: Time series of LLAAE-predicted counterfactuals vs. nudged-circulation climate simulations in test-set grid cells, shown at quartiles and the best R^2 among all land grid cells in Figure S3. Each panel reports the Pearson correlation coefficient between the two time series.

A.2.3 Comparison to benchmark model

Figure S5 resolves the difference in spatial performance measures between the LLAAE and benchmark model predicted counterfactuals. Like before, both fields are computed as the mean of the grid-cell-wise performance measures from the three test members.

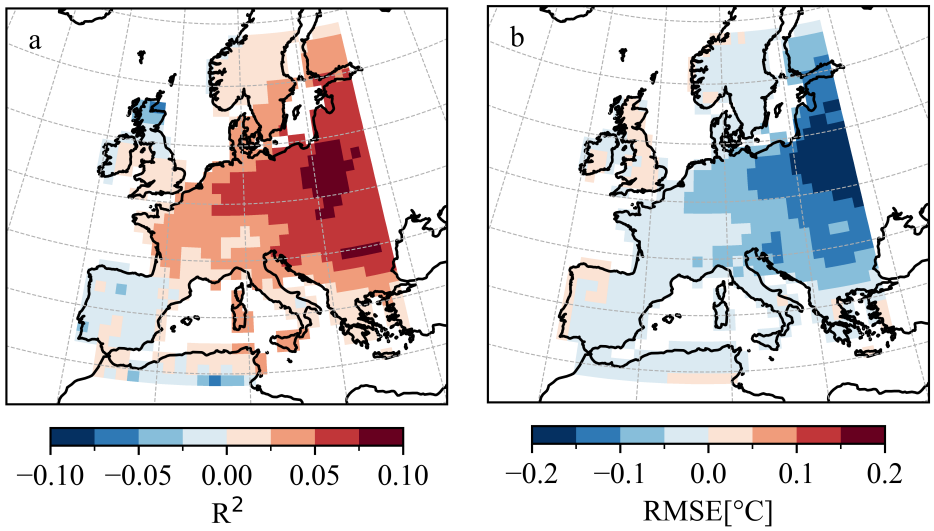

Figure S5: Comparison of performance measures of LLAAE counterfactual predictions and PCA benchmark model counterfactuals (LLAAE - PCA). Shown are differences in mean values of the test set members.