Extrapolation and distributional robustness for climate downscaling

Gianfranco Basile [*1] **Jan Schlegel** [*1] **Maybritt Schillinger** [†1] **Nicolai Meinshausen** [‡1]

Abstract

Climate downscaling is essential for translating coarse global climate model (GCM) outputs into high-resolution predictions that inform local decision-making. This work investigates the robustness and transferability of statistical climate downscaling methods when applied to unseen GCMs. Building on generative frameworks, such as the Coarse-from-Super and Corrector Diffusion models, we evaluate their performance in extrapolation challenges across diverse GCMs. Our experiments demonstrate that stochastic methods outperform deterministic baselines in energy score loss while maintaining comparable mean squared error performance. The lower energy scores highlight the potential of generative downscaling techniques for modeling variability inherent in climate systems. Notably, straightforward two-step generative approaches can produce realistic downscaling outputs even under temporal and cross-GCM extrapolation conditions.

1. Introduction

As the average global temperature continues to rise and extreme weather events occur more frequently, the impact of climate change becomes increasingly apparent. This phenomenon underscores the need for reliable and accurate climate information to develop effective mitigation strategies. Over recent decades, the scientific community has invested heavily in global climate models (GCMs), which are physics-based approximations of the Earth's climate system (Ling et al., 2024). However, Collins et al. (2013) note that significant biases exist in GCM outputs compared to observations. Moreover, Volosciuk et al. (2015) find that the coarse spatial resolution renders GCMs unsuitable for local climate analysis.

The challenge of deriving regional-scale information from larger-scale climate models is known as *downscaling*. Here one differentiates between dynamic and statistical downscaling approaches. Dynamic downscaling techniques utilize fine-grained regional climate models (RCMs), which ingest boundary conditions from GCMs and use high-resolution physics-based simulations to produce regional climate information, resulting in significant computational expenses (Gutowski et al., 2020). Furthermore, Liang et al. (2008) show that RCMs inherit biases from the parent GCMs, which propagate into the downscaled predictions. In contrast, statistical downscaling techniques learn empirical relationships between coarse and high-resolution climate data, offering a computationally efficient alternative (Ling et al., 2024). Classical machine learning methods, such as random forests or support vector machines, have long provided solid baselines for statistical downscaling. Building on these foundations, recent studies underline the potential of advanced deep learning methods, including convolutional neural networks (e.g. Rodrigues & Netto, 2018) or UNets (e.g. Lin et al., 2023), to further enhance the accuracy and computational performance of downscaling.

However, RCM and deterministic statistical downscaling approaches neglect to quantify underlying uncertainties in climate systems, which are pivotal for precisely understanding extreme weather events and climate change. Instead of expensive ensemble inference methods to address this limitation, inherently probabilistic generative models, such as generative adversarial networks (GANs) or diffusion models, offer a compelling alternative. Building upon recent advancements in generative models for climate downscaling, this work investigates how transferable these learned mappings are to new, unseen GCMs.

In our experimental study, we observe that stochastic approaches outperform deterministic procedures when considering the energy score loss from Section 2.2, while maintaining comparable mean squared error performance. We contribute to the literature by demonstrating that two-step generative downscaling methods produce reliable outcomes in cross-GCM extrapolation tasks. Moreover, we showcase that multilayer perceptrons (MLPs) augmented with stochastic noise injections can match or exceed the performance of more computationally intensive diffusion model approaches on climate downscaling extrapolation problems. Furthermore, our experiments demonstrate that the two-step approach of Mardani et al. (2024) can generate high-quality

[*]Equal contribution [†] Advisor[‡] Co-Advisor [1]Department of Mathematics, Federal Institute of Technology Zurich, Zurich, Switzerland.

downscaling samples, even when employing a simplified architecture that combines a reduced-size diffusion U-Net with a basic MLP for deterministic regression.

2. Methodology

Let $\mathcal{X} \subseteq \mathbb{R}^{c_{\text{in}} \times 36 \times 20}$ denote the space of low-resolution input observations and $\mathcal{Y} \subseteq \mathbb{R}^{c_{\text{out}} \times 128 \times 128}$ the space of high-resolution target random variables. Throughout our experiments, we set $c_{\text{in}} = 2$ (target variable and pressure variable) and $c_{\text{out}} = 1$ (target variable) with the target variable being one of $\{\texttt{temperature}, \texttt{precipitation}\}$. Although individual observations may originate from different GCM-RCM combinations, we omit explicit indices indicating their source for notational simplicity. We denote by $\boldsymbol{X}$ and $\boldsymbol{Y}$ the low and high resolution random vectors, taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, with an unknown joint distribution. Additionally, we define $\boldsymbol{Z}$ as a coarse regional representation obtained by applying average pooling with a kernel size of 16 to $\boldsymbol{Y}$. Given a collection of N paired observations $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, where $\forall i \in \{1, \dots, N\} : \boldsymbol{x}_i \in \mathcal{X},\ \boldsymbol{y}_i \in \mathcal{Y}$, we seek to model the conditional distribution $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$. In practice, we augment the global representations $\{\boldsymbol{x}_i\}_{i=1}^N$ with sinusoidal encodings of the forecast date and one-hot encodings of the GCM-RCM combination to improve model performance.

2.1. Coarse form Super Model

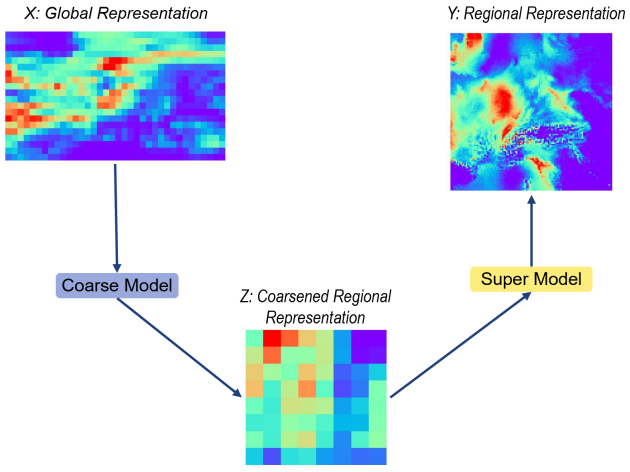


Figure 1. **The workflow for Coarse-form-Super model**. The coarse model first predicts the coarsened regional representation followed by the super model generating the regional representation.

The Coarse-from-Super model is composed of two sub-models:

1. **Coarse Model:** A stochastic multi layer perceptron (MLP) with six layers arranged as three residual blocks. In each block, a noise vector sampled from $\mathcal{N}(0, \mathbf{I})$ is concatenated with the features, injecting stochasticity at every stage. The first block (input layer) maps the high-dimensional flattened input to a 64-dimensional hidden representation using ReLU activations. The intermediate block enhances these features without altering their dimensionality, and the final block produces the intermediate target representation.

2. **Super Model:** This model refines the intermediate coarse representation $\boldsymbol{Z}$ into a high-resolution output using a similar stochastic residual design. Configured with six layers (three residual blocks), it employs a larger hidden dimension of 500. It directly maps the intermediate representation to the final output, preserving the structural and statistical details needed for accurate downscaling.

Our training workflow for the Coarse-from-Super model is designed as a two-stage process (Figure 1). Initially, the super model is trained to map coarse regional representations to high-resolution outputs. Once the super model has been trained its parameters are frozen. The coarse model then takes low-resolution inputs $\boldsymbol{X}$ and generates coarse regional representations $\hat{\boldsymbol{Z}}$. These are first evaluated against the corresponding true coarse representations $\boldsymbol{Z}$ using an energy score loss, defined in Section 2.2. The coarse output is subsequently fed into the fixed super model to produce refined regional predictions $\hat{\boldsymbol{Y}}$, which are again compared to the true high-resolution targets $\boldsymbol{Y}$ using the energy loss, defined in Section 2.2. The final loss for training the coarse model is computed as a weighted sum of these two energy losses, controlled by the parameter λ_{coarse}. For more details about the Coarse-from-Super model, we refer to Appendix A.

2.2. Energy Score Loss

In our framework, to train the Coarse-from-Super model we minimize an energy loss defined as the negative of the energy score. The energy score introduced by Gneiting & Raftery (2007) is a popular scoring rule that generalizes the continuous ranked probability score (CRPS) to multivariate settings (Szekely, 2003; Matheson & Winkler, 1976).

For two independent samples $\boldsymbol{Y}_1,\ \boldsymbol{Y}_2$ drawn from a distribution P and an observation $\boldsymbol{y}$, the energy score is defined as

$$\text{ES}(P, \boldsymbol{y}) = \frac{1}{2}\mathbb{E}_P \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\| - \mathbb{E}_P\|\boldsymbol{Y}_1 - \boldsymbol{y}\| \qquad (1)$$

where $\|\cdot\|$ denotes the Euclidean norm. The energy score is a strictly proper scoring rule, attaining its minimum expected value if and only if P corresponds to the true underlying distribution of the target variable (Szekely, 2003). By minimizing the empirical energy loss, the model is encouraged to generate samples that differ from each other, capturing

the variety of possible climate scenarios while remaining close to the observed values.

2.3. Corrector Diffusion Model

Recognizing that plain conditional diffusion models yield low-quality regional samples and slow convergence when directly modeling the distribution $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$, Mardani et al. (2024) propose a two-step approach with inherent variance reduction, termed corrector diffusion (CorrDiff), to mitigate these issues. As displayed in Figure 2, in this framework we decompose the generation procedure into

$$\boldsymbol{Y} = \underbrace{\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]}_{\text{deterministic regression } \boldsymbol{\mu}(\boldsymbol{X})} + \underbrace{\boldsymbol{R}}_{\text{stochastic residual}}, \qquad (2)$$

where the residual is defined as $\boldsymbol{R} := \boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X})$ and is modeled by a diffusion process.

Therefore, we first predict the conditional mean $\boldsymbol{\mu}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$ using a deterministic regression model. Contrary to the original CorrDiff formulation in Mardani et al. (2024), we do not employ a large U-Net in light of computational resource restrictions, but instead adopt a deterministic (i.e. without noise injection) MLP $f_\theta : \mathcal{X} \to \mathcal{Y}$ similar to the super component of the Coarse-from-Super architecture in Section 2.1. The parameters θ are learned by minimizing

$$\mathcal{L}_{\text{reg}}(\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{\text{data}}}\left[\|\boldsymbol{y} - f_\theta(\boldsymbol{x})\|_2^2\right],$$

yielding an estimate of the conditional mean $\boldsymbol{\mu}(\boldsymbol{X})$. This enables the computation of the empirical residuals

$$\widehat{\boldsymbol{r}}_i = \boldsymbol{y}_i - f_\theta(\boldsymbol{x}_i),\ i \in \{1, \dots, N\}.$$

By decoupling the estimation of the conditional mean from the modeling of the residual distribution, the function $f_\theta(\cdot)$ can be learned in isolation, simplifying and stabilizing the optimization procedure.

To more accurately capture extreme weather events that the deterministic model does not explain, Mardani et al. (2024) suggest employing an Elucidated diffusion model (cf. Karras et al. 2022) to sample from the conditional residual distribution $\mathbb{P}(\boldsymbol{R}|\boldsymbol{X})$. Therein, in the forward process, noise is progressively added to the residual. Following Karras et al. (2022), this forward process for some sample residual $\boldsymbol{r}$ is governed by the forward stochastic differential equation (SDE)

$$d\boldsymbol{r} = \sqrt{2\dot{\sigma}(t)\sigma(t)}d\boldsymbol{W}_t,$$

where $\sigma(t)$ denotes a prescribed noise schedule, typically $\sigma(t) \propto \sqrt{t}$, $\dot{\sigma}(t)$ the derivative of $\sigma(t)$ with respect to the diffusion time $t \in [0, T]$, and $d\boldsymbol{W}_t$ represents an increment of a Wiener process. The corresponding backward SDE governs the sample generation and is given by

$$d\boldsymbol{r} = -2\dot{\sigma}(t)\sigma(t)\nabla_{\boldsymbol{r}} \log p(\boldsymbol{r}|\sigma(t), \boldsymbol{c})dt + \sqrt{2\dot{\sigma}(t)\sigma(t)}d\bar{\boldsymbol{W}}_t, \qquad (3)$$

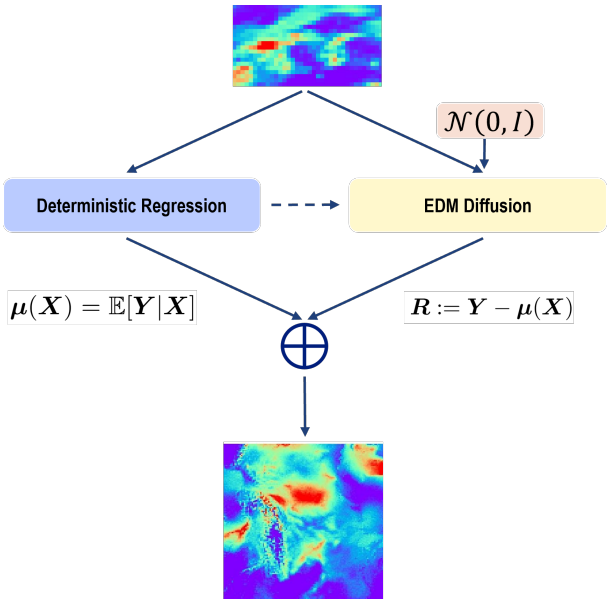


Figure 2. **The workflow for CorrDiff probabilistic downscaling**. A regression model first predict the conditional mean $\boldsymbol{\mu}(\boldsymbol{X})$ from low-resolution GCM input, followed by an Elucidated Diffusion model generating the residual $\tilde{\boldsymbol{r}}$. Adding together the estimated conditional mean and $\tilde{\boldsymbol{r}}$ yields a high-resolution sample. Inspired by a similar graphic in Mardani et al. (2024)

where $\nabla_{\boldsymbol{r}} \log p(\boldsymbol{r}|\sigma, \boldsymbol{c})$ is the so-called score function, a vector field indicating the directions of increasing data density at a specified noise level σ, and $d\bar{\boldsymbol{W}}_t$ denotes increments of a reverse-time Wiener process (Karras et al., 2022). Note that we add the option to guide the diffusion process by including potential conditioning information $\boldsymbol{c}$ in the score function $\nabla_{\boldsymbol{r}} \log p(\boldsymbol{r}|\sigma, \boldsymbol{c})$. In the CorrDiff framework of Mardani et al. (2024), this conditioning information $\boldsymbol{c}$ consists of the (upsampled) global representation $\boldsymbol{x}$ and the encoded time-point for which samples should be drawn, $\psi(t_{\text{target}})$, where $\psi(\cdot)$ is a learnable map.

For sample generation, we have to solve the reverse SDE in Equation (3), requiring the estimation of the score function $\nabla_{\boldsymbol{r}} \log p(\boldsymbol{r}|\sigma, \boldsymbol{c})$. The detailed algorithm for this stochastic sampling procedure is documented in Algorithm 1, and is based on Algorithm 2 of Karras et al. (2022). As the score function operates independently of the intractable normalizing constant, we can estimate the score function using a denoising neural network $D_\phi(\cdot)$. Moreover, if it holds that

$$\nabla_r \log p(\boldsymbol{r}|\sigma, \boldsymbol{x}, \psi(t_{\text{target}})) = \frac{D_\phi(\boldsymbol{r}, \sigma, \boldsymbol{x}, \psi(t_{\text{target}})) - \boldsymbol{r}}{\sigma^2},$$

the loss function can be written in the simple form

$$\mathcal{L}_{\text{diff}}(\phi) = \mathbb{E}_{\boldsymbol{r}\sim p_{\boldsymbol{r}}} \mathbb{E}_{\sigma\sim p_\sigma} \mathbb{E}_{\boldsymbol{\eta}\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\|D_\phi(\boldsymbol{r}+\boldsymbol{\eta}, \sigma, \boldsymbol{x}, \psi(t_{\text{target}})) - \boldsymbol{r}\|_2^2\right],$$

Realize that the expectation is also taken over noise levels σ

to ensure that the score function can be accurately estimated across various scales of noise corruption. Furthermore, the conditioning of the model is done by concatenating $\boldsymbol{x}$ with the noise $\boldsymbol{\eta}$ along the channel dimension. In addition, for $D_\phi(\cdot)$, we follow Mardani et al. (2024) and employ the U-Net of the DDPM++ architecture of Song et al. (2021). For a detailed breakdown of the architecture and the relevant hyperparameters used in CorrDiff training, we refer to Appendix B.2.

Finally, at test time, we leverage Equation (2) and obtain the high-resolution output for a new global representation $\boldsymbol{x}_{\text{new}}$ by combining the deterministic mean prediction with a residual sample $\tilde{\boldsymbol{r}}_{\text{new}}$ generated by the EDM using Algorithm 1:

$$\boldsymbol{y}_{\text{new}} = f_\theta(\boldsymbol{x}_{\text{new}}) + \tilde{\boldsymbol{r}}_{\text{new}}.$$

In case multiple samples are desired (e.g. for quantile estimation), one can generate M samples $\{\tilde{\boldsymbol{r}}_{\text{new},1}, \ldots, \tilde{\boldsymbol{r}}_{\text{new},M}\}$ from the EDM to arrive at

$$\boldsymbol{y}_{\text{new},j} = f_\theta(\boldsymbol{x}_{\text{new}}) + \tilde{\boldsymbol{r}}_{\text{new},j} \quad \text{for } j \in \{1, \ldots M\}.$$

3. Results

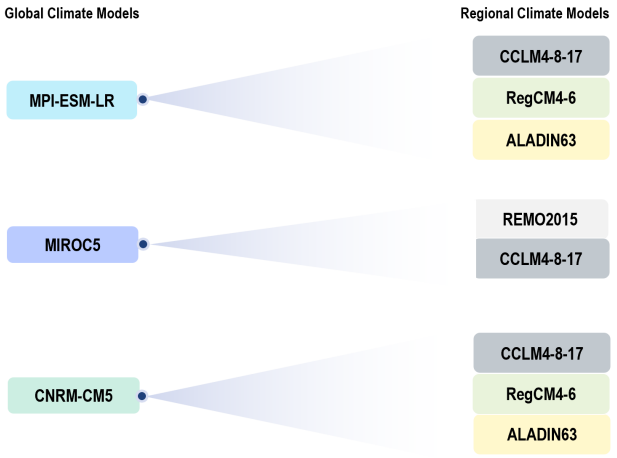


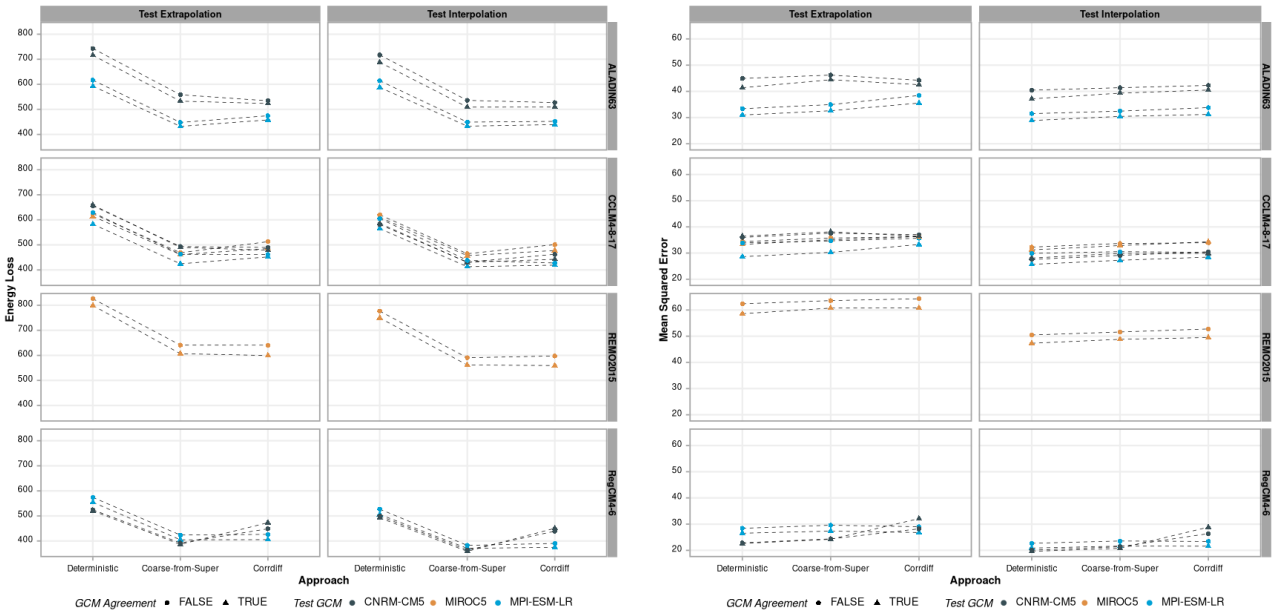
Figure 3. **Global and Regional Climate Model Combinations.** The figure depicts the three global climate models (MPI-ESM-LR, MIROC5, and CNRM-CM5) and their associated regional climate models (CCLM4-8-17, ALADIN63, RegCM4-6, and REMO2015). Not all global models are paired with every regional model.

We use three GCMs in our study, namely MPI-ESM-LR, MIROC5, and CNRM-CM5, simulate the global climate system at a coarse resolution of approximately 100 km (Kaltenborn et al., 2023). Each GCM is paired with one or more regional climate models (RCMs) that refine the large-scale output to provide high-resolution climate details. The available RCMs include CCLM4-8-17, ALADIN63, RegCM4-6, and REMO2015. However, as shown in Figure 3, not every GCM is associated with every RCM, as the availability of regional representations varies between global models.

We evaluate three downscaling approaches in terms of their extrapolation capabilities: (i) the Coarse-from-Super model introduced in Section 2.1, (ii) a deterministic (i.e. without any noise injections) MLP based on the super constituent of the Coarse-from-Super model, that is trained using mean-squared error loss, and (iii) the CorrDiff model established in Section 2.3. As some RCMs differ considerably in complexity and can be more challenging to learn, we stratify by the RCM at test time. Concretely, we train on all available GCM–RCM pairings from a subset consisting of two GCMs, leaving one GCM completely out of the training set. We then evaluate on all RCMs associated with this unseen GCM to assess how well the models extrapolate to novel GCM boundary conditions. As a baseline, the models are trained using all RCM–GCM combinations from a single GCM, and their performance is then assessed on test data from the same GCM. This enables us to compare GCM extrapolation (unseen GCM) and GCM interpolation (known GCM) performance directly.

We consider the time period from 1971–2099. The years 2030–2039 (*test interpolation*) and 2090–2099 (*test extrapolation*) are exclusively used for assessing model generalization to unseen time windows, while the remaining years (1971–2029 and 2040–2089) are available for training and validation. To ensure comparability, we use 60,000 training and 10,000 validation subsamples across all experiments, drawn randomly without regard to temporal order. For both the stochastic Coarse-from-Super and the deterministic networks, we opt for the model weights that yield the lowest validation score; for CorrDiff, we reuse the selected deterministic regressor and train the EDM component for a fixed 30 epochs. Detailed overviews of the training regimes for the Coarse-from-Super and CorrDiff models are provided in Appendix A.2 and Appendix B.2, respectively. Finally, on the temporal test interpolation (2030–2039) and temporal test extrapolation (2090–2099) datasets, we evaluate the three downscaling routines in terms of mean squared error and energy loss. Note that the loss values are computed based on the RCM outputs instead of the ground truth observations.

Figure 4 presents a comparative analysis of precipitation prediction performance across different RCMs using the three downscaling approaches. Panel (a) demonstrates that probabilistic models (Coarse-from-Super and CorrDiff) consistently achieve lower energy scores compared to the deterministic model in both temporal extrapolation and interpolation scenarios, and regardless of the particular GCM–RCM pairing or whether the test GCM matches the GCMs utilized during training (GCM agreement). This superior performance likely derives from their ability to model variability



(a) Energy Score

(b) MSE

Figure 4. **Absolute downscaling extrapolation performance for precipitation.** Panel (a) shows the energy score, while Panel (b) displays the mean squared error (MSE) for precipitation. The performance of three downscaling approaches, CorrDiff, deterministic, and Coarse-from-Super, is compared across different test regional climate models (RCMs) and experimental setups (test extrapolation vs. test interpolation). Colors denote the test global climate model (GCM) and shapes indicate GCM agreement between training and testing; dashed lines connect observations from the same training GCM.

inherent in climate systems. While performing on par with or better than CorrDiff, the Coarse-from-Super model demands fewer memory resources and less runtime throughout training and sample generation. However, the fact that Coarse-from-Super directly utilizes the energy score loss objective during training likely contributes to its low energy score in evaluation.

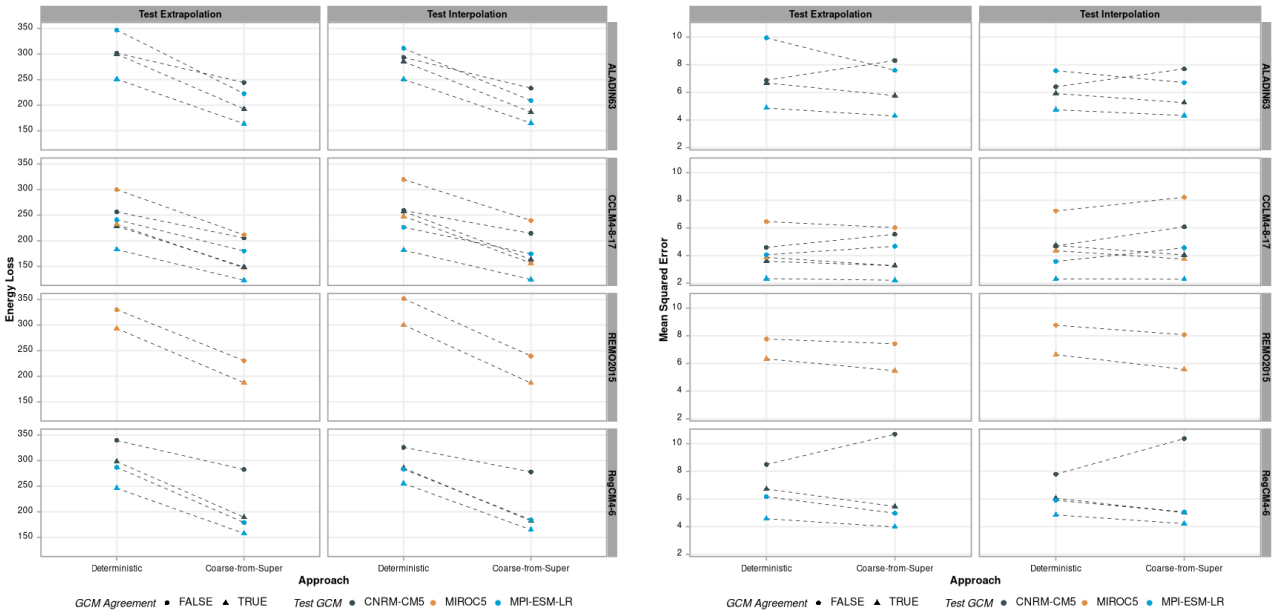
Panel (b) reveals relatively uniform mean squared error across all methodologies, suggesting the models maintain similar bias characteristics despite their architectural differences. Notably, experiments where training and testing employ the same GCM consistently exhibit reduced absolute energy score and MSE. However, this challenge of cross-GCM generalization is tackled better than we initially expected and the mappings appear to be quite robustly transferable to new GCMs. Moreover, the metrics achieved on the extrapolation test period resemble those measured on the interpolation test period, demonstrating the temporal robustness of all evaluated models.

As for the temperature response, it generally exhibits higher autocorrelation in both space and time, making it easier to predict than precipitation. Consequently, it represents a less challenging downscaling problem than precipitation. Given computational resource constraints, only the deterministic and Coarse-from-Super models are trained for this task. As is visualized in Figure 5(b), the Coarse-from-Super model achieves considerably lower energy scores compared to the deterministic approach. In contrast, with respect to mean squared error, both models display similar performance, with the deterministic version occasionally achieving a lower MSE loss than its stochastic Coarse-from-Super counterpart.

In addition to the absolute metrics from above, we perform a detailed relative analysis by normalizing the metrics with respect to the baseline. A comprehensive discussion of these relative comparisons is provided in Appendix C.

4. Discussion

The experimental results reveal that the generative downscaling methods under consideration consistently achieve lower energy scores across temperature and precipitation compared to a deterministic approach. In contrast, the mean squared errors remain similar for all three architectures under evaluation. Notably, our experiments indicate that the mappings that are learned on a collection of GCM-RCM combinations transfer effectively to unseen GCMs, demonstrating strong cross-GCM generalizability. Similarly, we observe comparable metrics on the interpolation and ex-



(a) Energy Score

(b) MSE

Figure 5. **Absolute downscaling extrapolation performance for temperature.** Panel (a) shows the energy score, while Panel (b) displays the mean squared error (MSE) for temperature. The performance of two downscaling approaches, deterministic, and Coarse-from-Super, is compared across different test regional climate models (RCMs) and experimental setups (test extrapolation vs. test interpolation). Colors denote the test global climate model (GCM) and shapes indicate GCM agreement between training and testing; dashed lines connect observations from the same training GCM.

trapolation test time periods, indicating temporal stability. Interestingly, the simple Coarse-from-Super architecture from Section 2.1 yields similar or improved energy scores (and means squared errors) compared to the CorrDiff architecture, offering a computationally less resource-intensive alternative to more complex generative approaches.

Finally, this study does not quantify the impact of the number of subsampled training datapoints on predictive performance, which warrants further investigation. Moreover, future work could also explore the influence of the choice of the deterministic model within the CorrDiff framework, focusing on performance and computational resource usage. Instead of the simple residual MLP that we employ, one might consider a U-Net as in Mardani et al. (2024) or novel foundation models for climate such as Prithvi WxC from Schmude et al. (2024). Additionally, a more thorough investigation of the impact of different hyperparameter settings in the CorrDiff framework suggests itself. Furthermore, one can explore alternative extrapolation scenarios, such as spatial transferability, for an enhanced understanding of model robustness in climate downscaling tasks.

5. Conclusion

In summary, our work demonstrates that generative two-step approaches are well-suited for tackling complex climate downscaling problems. Particularly, we observe remarkable extrapolation capabilities to out-of-distribution GCMs. Moreover, we reveal that simple stochastic MLPs, augmented with noise injection, can match or surpass the performance in terms of energy score and mean squared error compared to more computationally demanding diffusion models. This finding challenges the common belief in the literature that complex architectures are always necessary for climate downscaling. Additionally, by employing a smaller deterministic regression model in the CorrDiff framework by Mardani et al. (2024) that is nevertheless capable of generating high-quality high-resolution samples, we pave the way for more robust and scalable downscaling techniques.

References

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A., and Wehner, M. Long-term climate change: Projections, commitments and irreversibility. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Gutowski, W. J., Ullrich, P. A., Hall, A., Leung, L. R., O'Brien, T. A., Patricola, C. M., Arritt, R. W., Bukovsky, M. S., V., C. K., Feng, Z., Jones, A. D., Kooperman, G. J., Monier, E., Pritchard, M. S., Pryor, S. C., Qian, Y., Rhoades, A. M., Roberts, A. F., Sakaguchi, K., Urban, N., and Zarzycki, N. The ongoing need for high-resolution regional climate models: Process understanding and stakeholder information. *Bulletin of the American Meteorological Society*, 101(5):E664 – E683, 2020. doi: 10.1175/BAMS-D-19-0113.1. URL https://journals.ametsoc.org/view/journals/bams/101/5/bams-d-19-0113.1.xml.

Kaltenborn, J., Lange, C. E. E., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., and Rolnick, D. Climateset: A large-scale climate model dataset for machine learning, 2023. URL https://arxiv.org/abs/2311.03721.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Liang, X.-Z., Kunkel, K. E., Meehl, G. A., Jones, R. G., and Wang, J. X. Regional climate models downscaling analysis of general circulation models present climate biases propagation into future change projections. *Geophysical research letters*, 35(8), 2008.

Lin, H., Tang, J., Wang, S., Wang, S., and Dong, G. Deep learning downscaled high-resolution daily near surface meteorological datasets over east asia. *Scientific Data*, 10 (1):890, 2023.

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(1):131, 2024. URL https://www.nature.com/articles/s41612-024-00679-1.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

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, 2024. URL https://arxiv.org/abs/2309.15214.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629907.

Rodrigues, E.R., O. I. C. R. and Netto, M. Deepdownscale: A deep learning strategy for high-resolution weather forecast. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 415–422, 2018. doi: 10.1109/eScience.2018.00130.

Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., Kuehnert, J., Ankur, K., Gupta, A., Phillips, C. E., Kienzler, R., Szwarcman, D., Gaur, V., Shinde, R., Lal, R., Silva, A. D., Diaz, J. L. G., Jones, A., Pfreundschuh, S., Lin, A., Sheshadri, A., Nair, U., Anantharaj, V., Hamann, H., Watson, C., Maskey, M., Lee, T. J., Moreno, J. B., and Ramachandran, R. Prithvi wxc: Foundation model for weather and climate, 2024. URL https://arxiv.org/abs/2409.13598.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Szekely, G. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, pp. 3(05):1–18, 2003.

Volosciuk, C., Maraun, D., Semenov, V. A., and Park, W. Extreme precipitation in an atmosphere general circulation model: Impact of horizontal and vertical model resolutions. *Journal of Climate*, 28(3):1184 – 1205, 2015. doi: 10.1175/JCLI-D-14-00337.1. URL https://journals.ametsoc.org/view/journals/clim/28/3/jcli-d-14-00337.1.xml.

A. Coarse-from-Super Model

A.1. Training Loss

Given $\boldsymbol{X}$ as the low-resolution input, $\boldsymbol{Z}$ as the true coarse regional representation, and $\boldsymbol{Y}$ as the true high-resolution (regional) representation, our framework first generates an intermediate coarse representation using the coarse model, i.e., $\hat{\boldsymbol{Z}} = f_{\text{coarse}}(\boldsymbol{X}, \boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. This intermediate representation is then refined by the pre-trained (and frozen) super model to produce the final high-resolution output, $\hat{\boldsymbol{Y}} = f_{\text{super}}(\hat{\boldsymbol{Z}}, \boldsymbol{\eta})$, where $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I})$. We compute energy losses $\mathcal{L}_{\text{energy}}(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$ and $\mathcal{L}_{\text{energy}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ for the coarse and super outputs, respectively. The final loss function, which is used to train the coarse model while keeping the super model fixed, is defined as a convex combination of these individual energy losses:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) = \lambda_{\text{coarse}}\, \mathcal{L}_{\text{energy}}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) + (1 - \lambda_{\text{coarse}})\, \mathcal{L}_{\text{energy}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) \tag{4}$$

A.2. Experimental Details

The Coarse-from-Super model is composed of two different sub-models as explained in Section 2.1. In the following, the experimental details for the coarse model and for the super model are described.

Coarse-from-Super

- **Data Configuration:**
  - **Target Variable:** Precipitation (`pr`) or Temperature (`tmp`).
  - **Normalization:** Input data are normalized using a scalar method, while outputs are normalized piecewise. Square root transformations are applied to both input and output.
  - **Data Split:** The dataset comprises 60,000 training samples and 10,000 validation samples.
  - **Low-Resolution Processing:** Low-resolution prediction is enabled with an average pooling kernel size of 16.
- **Training Strategy:**
  - **Duration and Batch Size:** Training is conducted over 300 epochs using a batch size of 512.
  - **Learning Rate:** Optimization is performed with the AdamW (cf. Loshchilov & Hutter, 2019) optimizer at an initial learning rate of 3×10^{-4}.
  - **Scheduler:** A decay-based (i.e. exponential) learning rate scheduler is employed, multiplying the current learning rate by a factor of 0.99 after every epoch.
  - **Lambda Mixing Factor:** The λ_{coarse} mixing factor is 0.9

coarse Model

- **Model Architecture:**
  - **Network Structure:** The coarse model is implemented as a multilayer perceptron (MLP) with 6 layers and a hidden dimension of 64.
  - **Stochasticity:** A 10-dimensional noise vector (with a noise standard deviation of 1.0) is injected to account for uncertainty.

Super Model

- **Model Architecture:**
  - **Network Structure:** The super model is implemented as a multilayer perceptron (MLP) with 6 layers.
  - **Hidden Dimensions and Layer Shrinkage:** The model features a hidden dimension of 500 and a layer shrinkage factor of 16.
  - **Stochasticity:** A 100-dimensional noise vector (with a noise standard deviation of 1.0) is injected to account for uncertainty.

B. Elucidated Diffusion Model

B.1. Stochastic Sampler

Algorithm 1 details the stochastic sampling procedure used in the CorrDiff framework. Adapted from Algorithm 2 in Karras et al. (2022), this algorithm is modified for consistent notation across this work. Here, $\boldsymbol{x}$ represents the (nearest neighbor) upsampled coarse GCM input, $\{t_i\}_{i=0}^{N}$ denotes a sequence of predefined noise levels, and $\psi(t_{\text{target}})$ corresponds to learnable embeddings of the forecast date. Additionally, we define a sequence $\{\gamma_i\}_{i=0}^{N-1}$ of non-negative factors for adjusting the noise level at every diffusion time step, where

$$\gamma_i = \begin{cases} \min\left(\frac{S_{\text{churn}}}{N}, \sqrt{2}-1\right), & \text{if } t_i \in [S_{\text{tmin}}, S_{\text{tmax}}] \\ 0, & \text{otherwise.} \end{cases}$$

In accordance with Karras et al. (2022), we choose $S_{\text{tmin}} = 0$ and $S_{\text{tmax}} = \infty$. Moreover, we empirically established that $N = 40$ steps are sufficient to generate high-quality precipitation samples. Finally, we define $D_\phi(\cdot)$ as a denoising neural network that takes the high-resolution output and its corresponding noise level at a specific time point as input, and that is conditioned on the coarse GCM input and the time embeddings. This parametrized function leverages the DDPM++ architecture of Song et al. (2021).

Algorithm 1 CorrDiff Stochastic Sampler

procedure CORRDIFFSAMPLER($D_\phi(\cdot), \{t_i\}_{i=0}^{N}, \{\gamma_i\}_{i=0}^{N-1}, S_{\text{noise}}$)
  sample $\boldsymbol{r}_0 \sim \mathcal{N}(\mathbf{0}, t_0^2\mathbf{I})$
  for $i = 0, \ldots, N-1$ **do**
    $\widehat{t}_i \leftarrow t_i + \gamma_i t_i$
    sample $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, S_{\text{noise}}^2\mathbf{I})$
    $\widehat{\boldsymbol{r}}_i \leftarrow \boldsymbol{r}_i + \sqrt{\widehat{t}_i^{\,2} - t_i^2}\,\boldsymbol{\eta}_i$
    $\boldsymbol{d}_i \leftarrow \dfrac{\left(\widehat{\boldsymbol{r}}_i - D_\phi\left(\widehat{\boldsymbol{r}}_i, \widehat{t}_i, \boldsymbol{x}, \psi(t_{\text{target}})\right)\right)}{\widehat{t}_i}$
    $\boldsymbol{r}_{i+1} \leftarrow \widehat{\boldsymbol{r}}_i + \left(t_{i+1} - \widehat{t}_i\right)\boldsymbol{d}_i$ ▷ Euler step
    if $t_{i+1} \neq 0$ **then**
      $\boldsymbol{d}'_i \leftarrow \dfrac{\left(\boldsymbol{r}_{i+1} - D_\phi\left(\boldsymbol{r}_{i+1}, t_{i+1}, \boldsymbol{x}, \psi(t_{\text{target}})\right)\right)}{t_{i+1}}$
      $\boldsymbol{r}_{i+1} \leftarrow \widehat{\boldsymbol{r}}_i + \left(t_{i+1} - \widehat{t}_i\right)\frac{1}{2}\left(\boldsymbol{d}_i + \boldsymbol{d}'_i\right)$ ▷ Second-order correction (Heun's method)
    end if
  end for
  return $\boldsymbol{r}_N$
end procedure

Due to computational constraints, we forgo a grid search for the S_{churn} hyperparameter and instead fix $S_{\text{churn}} = 0$. However, Karras et al. (2022) demonstrate that the choice of S_{churn} substantially influences performance in terms of the Fréchet inception distance (FID) on traditional image dataset. Thus, future work should include ablations on S_{churn} in the context of climate downscaling tasks.

B.2. Experimental Details

The deterministic regression component is trained using the same configurations as the super component in Appendix A.2 but utilizing mean-squared error loss instead of the combined energy loss from Appendix A.1. Subsequently, the weights of the deterministic regression map are frozen in the training of the EDM constituent. We use the following hyperparameter settings for learning the EDM model:

- **Training Duration:** The model is trained for a fixed 30 epochs

- **Batch Size:** 32
- **Learning Rate and Scheduler:** We employ the AdamW (cf. Loshchilov & Hutter, 2019) optimizer with an initial learning rate of 3×10^{-4}. A linear warmup over the first 10 epochs followed by cosine annealing is used.
- **U-Net Architecture for Parametrized Score Function:**
  - **Input and Output Resolution:** 128×128
  - **Input and Output Channels:** 3 input channels (noise, target variable and pressure variable), 1 output channel (target variable)
  - **U-Net Downsampling**: The spatial resolution is progressively halved along both spatial dimensions at each encoder stage, before a symmetric upsampling in the decoder. Hence, we get the following resolution levels: $[128 \times 128, 64 \times 64, 32 \times 32, 16 \times 16]$, with the order reversed in the decoder constituent.
  - **Base Channels:** 64 with channel multipliers of [1, 2, 2, 2]
  - **Residual Blocks:** 4 per resolution level
  - **Dropout rate:** 0.1, is applied after each convolutional layer in both the encoder and the decoder
  - **Attention:** Is applied at a spatial resolution of 16×16
  - **Time Embedding:** Sinusoidal encodings, with a label dimension of 5 (for time-only), that are passed through a linear layer before being added to the noise embeddings.
- **Gradient Clipping:** A maximum gradient norm of 10 is applied to ensure training stability.
- **Reverse Diffusion:** 40 reverse diffusion steps for generating samples. Utilize Algorithm 1 with $S_{\text{tmin}} = 0, S_{\text{tmax}} = \infty$, and $S_{\text{churn}} = 0$.
- **Precision:** For training and sampling 32-bit floating point precision is used

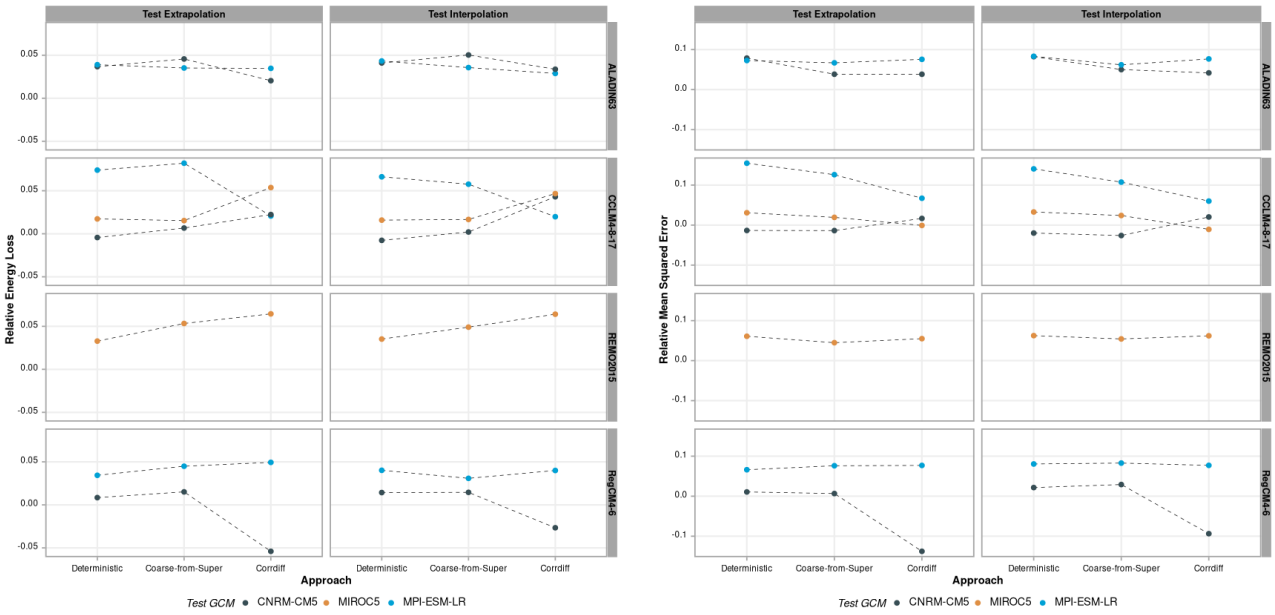
C. Complementary Visualizations

In the appendix figures (Figure 6 for precipitation and Figure 7 for temperature), all reported metrics are expressed relative to a single GCM baseline. In other words, each point indicates how much each model's performance improves upon (negative values) or falls behind (positive values) a baseline model trained and tested on the same GCM.

Turning first to precipitation (Figure 6), the relative energy score and relative MSE vary considerably among GCM-RCM pairs, and there is no consistent pattern that separates probabilistic models (CorrDiff and Coarse-from-Super) from the deterministic approach. In some cases, the Coarse-from-Super or CorrDiff methods achieve slightly more negative values (hence outperforming the baseline), but other points lie above zero, indicating less favorable performance. This variability suggests that the performances in term of relative measure depend on which GCM or RCM is being considered.

A clearer pattern emerges for temperature (Figure 7). Here, only the deterministic and Coarse-from-Super models are compared. The Coarse-from-Super approach tends to show higher positive values for the energy score. That outcome implies that relative to the single-GCM baseline, the generative model sometimes struggles to offer systematic improvements in capturing the distributional aspects of temperature fields. A possible reason for this result stems from the fact that temperature is a relatively easy variable to model.

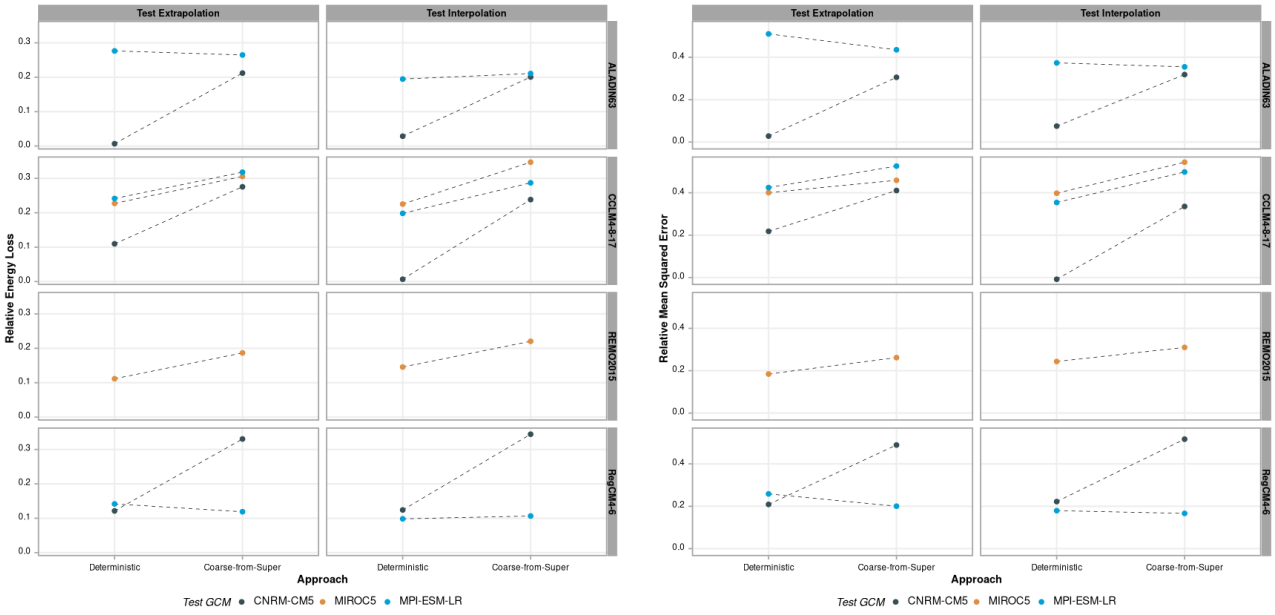
Overall, these relative plots highlight the sensitivity of cross-GCM extrapolation to the baseline choice and underscore that improvements in one metric (e.g., distributional accuracy) may not always translate to uniform gains when viewed as a ratio or difference from a single-GCM reference. The observed swings above and below zero in both figures further reflect the diverse nature of the GCM–RCM combinations, reinforcing that each approach's relative advantage or disadvantage depends on which boundary conditions it encounters.



(a) Energy Score

(b) MSE

Figure 6. **Relative downscaling extrapolation performance for preciptation.** Panel (a) shows the relative energy score, while Panel (b) displays the relative mean squared error (MSE) for precipitation. The performance of three downscaling approaches, CorrDiff, deterministic, and Coarse-from-Super, is compared across different test regional climate models (RCMs) and experimental setups (test extrapolation vs. test interpolation). Colors denote the test global climate model (GCM) and shapes indicate GCM agreement between training and testing; dashed lines connect observations from the same training GCM.



(a) Energy Score

(b) MSE

Figure 7. **Relative downscaling extrapolation performance for temperature.** Panel (a) shows the relative energy score, while Panel (b) displays the relative mean squared error (MSE) for temperature. The performance of two downscaling approaches, Deterministic, and Stochastic, is compared across different test regional climate models (RCMs) and experimental setups (test extrapolation vs. test interpolation). Colors denote the test global climate model (GCM) and shapes indicate GCM agreement between training and testing; dashed lines connect observations from the same training GCM.