AGU ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2025GL117397

Generative Downscaling and Bias Correction of Multivariable Earth System Model Simulations

Haijie Li[1,2,3,4], Ya Wang[1,2], Gang Huang[1,4], Weichen Tao[3], and Pengfei Lin[1]

[1]State Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [2]Earth System Numerical Simulation Science Center, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [3]State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [4]University of Chinese Academy of Sciences, Beijing, China

Key Points:

- We propose a generative model for enhancing the spatial resolution of climate simulations while performing simultaneous bias correction
- Our model achieves a reduction of over 70% in climatological bias of three variables, while significantly enhancing the representation of extremes
- The model effectively mitigates the common westward bias in ENSO-related sea surface temperature anomalies

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Y. Wang, G. Huang and W. Tao,
wangya@mail.iap.ac.cn;
hg@mail.iap.ac.cn;
twc@mail.iap.ac.cn

Citation:

Li, H., Wang, Y., Huang, G., Tao, W., & Lin, P. (2025). Generative downscaling and bias correction of multivariable Earth system model simulations. *Geophysical Research Letters*, *52*, e2025GL117397. https://doi.org/10.1029/2025GL117397

Received 31 JUL 2025
Accepted 4 SEP 2025

Abstract The Earth system model is central to understanding climate change and informing policy decisions. However, current models, with their coarse resolution and inherent biases, limit the utility of climate simulations. In this study, we introduce a multivariate generative downscaling model (MVGDM) that downscales global climate simulations from a 100 km to a 25 km resolution, while simultaneously correcting climate simulation biases. The MVGDM provides accurate simulations by refined resolution and addressing biases in three variables: sea surface temperature (SST), 2-m temperature (T2M), and 500-hPa geopotential height (Z500), reducing climatological biases by 72%, 79%, and 71%, respectively. Additionally, the MVGDM mitigates the common westward bias in El Niño-Southern Oscillation (ENSO)-related SST anomalies and significantly improves the simulation of the Indian Ocean Dipole. With high-resolution simulations and improved representation of multi-scale physical processes, the MVGDM substantially enhances the simulation of climate extremes, and shows some potential for improving future climate projections.

Plain Language Summary We need high-resolution and accurate climate simulations to project future climate changes, which can further guide human production and lifestyle choices. In recent years, the rapid development of deep learning technology has provided a new paradigm for advancement across various fields. In this study, we have developed a deep learning-based multivariate generative downscaling model (MVGDM) that can downscale global climate simulations from a 100 km resolution to 25 km, while simultaneously correcting biases in the data. Through quantitative evaluation of several climate simulation indicators, we have demonstrated that MVGDM can provide more refined and accurate reference data for future climate projections.

1. Introduction

Climate model simulations are essential for investigating climate change, providing crucial insights into its mechanisms, informing adaptation policies, and mitigating climate risks. However, two key challenges hinder the progress of climate models: inherent biases and insufficient spatial resolution. Model biases introduce inaccuracies in both global and regional climate simulations and projections, especially in simulating precipitation and extreme weather events. Limited resolution restricts the representation of medium to meso- and micro-scale processes, such as convection and tropical cyclones, compromising predictions of extreme events and the reliability of regional climate analyses.

Traditional bias correction and downscaling approaches face fundamental limitations that constrain their effectiveness in climate model refinement. Quantile mapping (QM), while widely adopted for distributional adjustments at individual grid points (Themeßl et al., 2011), fails to preserve spatial coherence by neglecting inter-grid relationships and risks altering inherent climate change signals (Dosio et al., 2012). Regional Climate Models offer physical consistency through dynamical downscaling but remain computationally prohibitive and inherit input model biases through boundary condition dependencies (Maraun et al., 2010). Statistical downscaling methods, though computationally efficient, exhibit limited skill (Chen et al., 2016). These intrinsic constraints persist across conventional techniques, motivating exploration of alternative approaches.

Deep learning presents transformative potential for climate data refinement, with generative models demonstrating particular promise. Generative models like generative adversarial network (Goodfellow et al., 2014) have

shown superiority over QM in precipitation correction (Hess et al., 2022) and SST bias adjustment (Y Wang et al., 2024). However, existing implementations predominantly focus on single-variable, regional-scale applications with coarse resolutions, leaving global multivariate high-resolution generation (e.g., ≤25 km) largely unexplored.

In this study, we present a multivariate generative downscaling framework (MVGDM) that achieves concurrent global-scale bias correction and fourfold resolution enhancement (100–25 km) for three dynamically coupled variables. When applied to climate model outputs, the MVGDM systematically addresses biases in climatological means, interannual variability, and extremes while preserving cross-variable dynamical coherence—a critical yet previously neglected capability.

2. Data, Methods

2.1. Data

Our analysis employs ECMWF Reanalysis v5 (ERA5; Hersbach et al., 2020) as the observational benchmark, leveraging its high-resolution (0.25° grids) and physically consistent representations of coupled atmosphere-ocean dynamics. For model inputs, we use GFDL-ESM4 outputs from Coupled Model Intercomparison Project phase 6 (CMIP6; Dunne et al., 2020). Three dynamically linked variables—sea surface temperature (SST), 2-m air temperature (T2M), and 500-hPa geopotential height (Z500)—are selected to capture essential air-sea interactions across scales (1° resolution for SST; 1° × 1.25° grid for T2M and Z500). The training period (1950–1990) and testing window (1995–2014) are strategically separated to minimize temporal autocorrelation artifacts, ensuring robust evaluation of interannual variability and extreme event statistics. The GFDL outputs are bilinearly interpolated to the ERA5 grid before training.

2.2. Methods

2.2.1. Cycle-Consistent Generative Adversarial Networks

In this study, we employ the Cycle-consistent generative adversarial networks (CycleGAN) model (Zhu et al., 2017) as the core of MVGDM to downscale GFDL-ESM4 low-resolution data (1° × 1°) to a high resolution (0.25° × 0.25°). Climate model downscaling constitutes an unpaired domain translation task, as simulated outputs and observational data inherently lack point-to-point correspondence. CycleGAN addresses this by learning a probabilistic mapping between the two distributions. In our task, as shown in Figure 1, the generator G_X maps samples x from the GFDL low-resolution data set ($x \in X$) to samples y in the ERA5 high-resolution data set ($y \in Y$). The discriminator D_Y evaluates whether the generated samples $G_X(x)$ belong to the distribution of Y. The generator and discriminator engage in an adversarial process, mutually improving each other. Ultimately, a well-trained generator G_X is obtained to downscale low-resolution data to high-resolution data effectively.

To ensure that the generated data not only conforms to the target domain distribution but also retains critical features of the source domain, CycleGAN introduces an additional generator G_Y. This generator G_Y maps the high-resolution data sample $G_X(x)$ generated by G_X back to x, forming a cycle $x \to G_X(x) \to G_Y(G_X(x)) \approx x$. To further improve the training of the generators and discriminators, we also use the previous generator G_Y to map y back to x, incorporating an additional discriminator D_X to evaluate the generated samples accordingly.

Generators are built on a fully convolutional U-Net backbone (Ronneberger et al., 2015), consisting of one input layer, two downsampling layers, nine residual blocks with skip connections, two upsampling layers and one output layer. This encoder-decoder design extracts multiscale features and reconstructs high resolution details (see Figure S1a in Supporting Information S1). Discriminators use convolutional downsampling to produce feature maps for real versus generated field classification (see Figure S1b in Supporting Information S1).

The loss function corresponding to the $X \to Y$ mapping is:

$$\mathcal{L}_{\mathrm{GAN}}(G_X, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\mathrm{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}\left[\log\left(1 - D_Y(G_X(x))\right)\right], \quad (1)$$

and similarly,

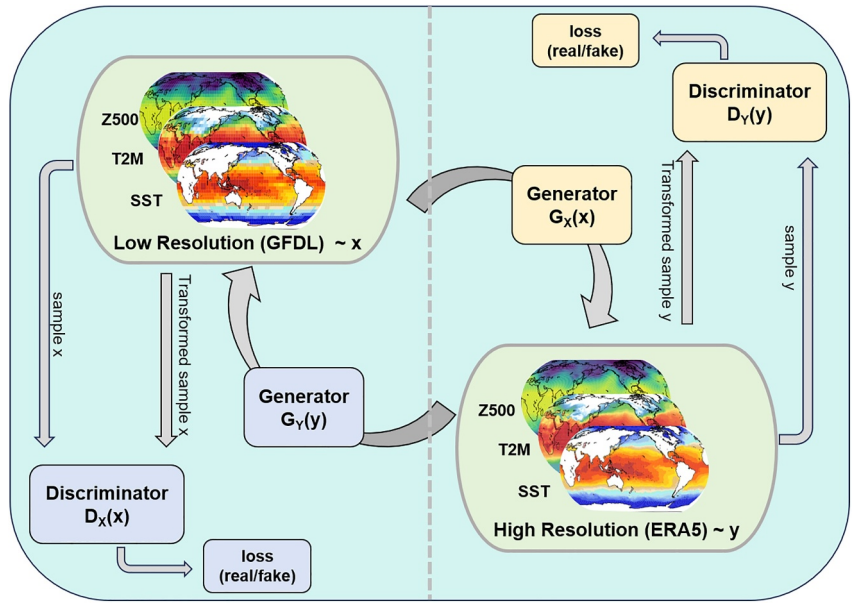

Figure 1. Schematic of the multivariate generative downscaling model. The model consists of two generators and two discriminators, which learn to downscale and correct biases in low-resolution data through adversarial training. The generators aim to establish mapping relationships between data domains, while the discriminators evaluate the authenticity of the generated results, thereby jointly optimizing the model's performance.

$$\mathcal{L}_{\mathrm{GAN}}(G_Y, D_X, Y, X) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}[\log D_X(x)] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\log\left(1 - \mathrm{D}_X(G_Y(y))\right)\right] \tag{2}$$

The cycle consistency loss is defined as:

$$\mathcal{L}_{\mathrm{cyc}}(G_X, G_Y) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\|G_Y(G_X(x)) - x\|_1\right] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\|G_X(G_Y(y)) - y\|_1\right]. \tag{3}$$

Additionally, to preserve the internal variability of the GFDL simulation and reduce bias, we follow the approach of Zhu et al. (2017) and introduce the identity loss:

$$\mathcal{L}_{\mathrm{id}}(G_X, G_Y) = \mathrm{E}_{y\sim p_{\mathrm{daxa}}(y)}\left[\|G_X(y) - y\|_1\right] + \mathrm{E}_{x\sim p_{\mathrm{data}}(x)}\left[\|G_Y(x) - x\|_1\right]. \tag{4}$$

The total loss function is defined as:

$$\mathcal{L}(G_X, G_Y, D_X, D_Y) = \mathcal{L}_{\mathrm{GAN}}(G_X, D_Y, X, Y) + \mathcal{L}_{\mathrm{GAN}}(G_Y, D_X, Y, X) + \lambda_1\mathcal{L}_{\mathrm{cyc}}(G_X, G_Y) + \lambda_2\mathcal{L}_{\mathrm{id}}(G_X, G_Y), \tag{5}$$

where λ_1 and λ_2 are used to control the relative importance of the consistency loss and identity loss. Our ultimate objective is to optimize:

$$G_X^*, G_Y^* = \arg\min_{G_X, G_Y}\max_{D_X, D_Y}\mathcal{L}(G_X, G_Y, D_X, D_Y). \tag{6}$$

This distribution-matching loss function, unlike the point-wise loss used in supervised models, enhances the simulation of extreme values. The weighting factors λ_1 and λ_2 for the cycle consistency loss and identity loss are configured as 10 and 5, respectively. We optimize the model using the Adam optimizer (learning rate = 5×10^{-5}) and scale inputs to (–1, 1) for stability. Training is performed with batch size = 1 for 50 epochs on a single NVIDIA A100 GPU (40 GB), taking approximately six days. The checkpoint with the lowest validation loss is selected for inference.

To further reduce residual mean bias, the CycleGAN-downscaled output is post-processed using QM (Fulton et al., 2023; Hess et al., 2025), adjusting the empirical distributions to match ERA5 without altering spatial coherence.

2.2.2. Baselines for Comparison

Our primary baseline is the Adaptive Fourier Neural Operator Networks (AFNONet), which is the backbone of FourCastNet by NVIDIA (Kurth et al., 2023). AFNONet leverages Fourier transforms to learn nonlinear dynamics and has shown state-of-the-art skill in weather forecasting. Following the supervised framework of F. Wang and Tian (2022), we pair daily GFDL inputs with same-day ERA5 targets and train with a batch size of 16, a learning rate of 5×10^{-4}. We apply the same QM post-processing at inference to ensure an equitable comparison. We also evaluated additional supervised architectures (e.g., U-Net and Vision Transformer; Dosovitskiy et al., 2020), but they performed on par with or worse than AFNONet, and are therefore not reported. By comparing against these supervised methods, we aim to assess across multiple evaluation dimensions whether their generated samples are reliable and to identify any limitations.

Our secondary baseline uses conventional bilinear interpolation followed by QM, as in Fulton et al. (2023). GFDL outputs are first interpolated to the target grid and then bias-corrected via QM. The quantile bins are 500, which we find the best performance during training phase.

3. Results

3.1. The Finer Characteristic of the MVGDM Results

Figures 2a–2e compare the spatial patterns of T2M across different results on 29 December 1997. MVGDM, AFNONet and QM all exhibit substantially enhanced spatial detail compared to GFDL outputs, particularly in capturing topographic temperature characteristics over the Tibetan Plateau and coastal temperature features along eastern China.

However, the AFNONet T2M field exhibits low fidelity to the original GFDL pattern and instead overly conforms to the ERA5, indicating that fully supervised downscaling may overwrite inherent model characteristics. In contrast, MVGDM and QM retain the broad scale features of the GFDL simulation. To further check how well internal variability is kept, we look at a day with an El Niño event in GFDL. On that day the GFDL SST anomaly (SSTA) clearly shows the El Niño warm pattern in the equatorial Pacific, whereas ERA5 does not. This difference is due to phase mismatch, not a bias (Figures 2f and 2j). MVGDM preserves the El Niño pattern while reducing bias, achieving a pattern correlation coefficient (PCC) of 0.86 with GFDL (Figure 2g). AFNONet fails to capture the El Niño anomaly and has low PCC with GFDL (Figure 2h). QM leaves the SST distribution almost unchanged from GFDL (PCC = 0.95) because it rescales each grid point independently (Figure 2i); this preserves internal variability but does little to correct the spatial bias. MVGDM avoids inappropriate over-correction toward observational states, which is a fundamentally ill-posed adjustment given the non-synchronized nature of internal climate variability between models and reality.

To verify the role of the cycle structure of MVGDM in preserving internal variability, we conduct ablation experiments comparing models with and without the cycle-consistency constraint. The results show that including the cycle structure greatly improves the model's ability to retain the original signal of GFDL (Figure S2 in Supporting Information S1).

3.2. The Climatological Bias

In this section, we assess systematic differences in climatology, interannual variability, and extremes. Figure 3 compares the mean biases in SST, T2M and Z500 for GFDL and the three post-processing methods. The GFDL

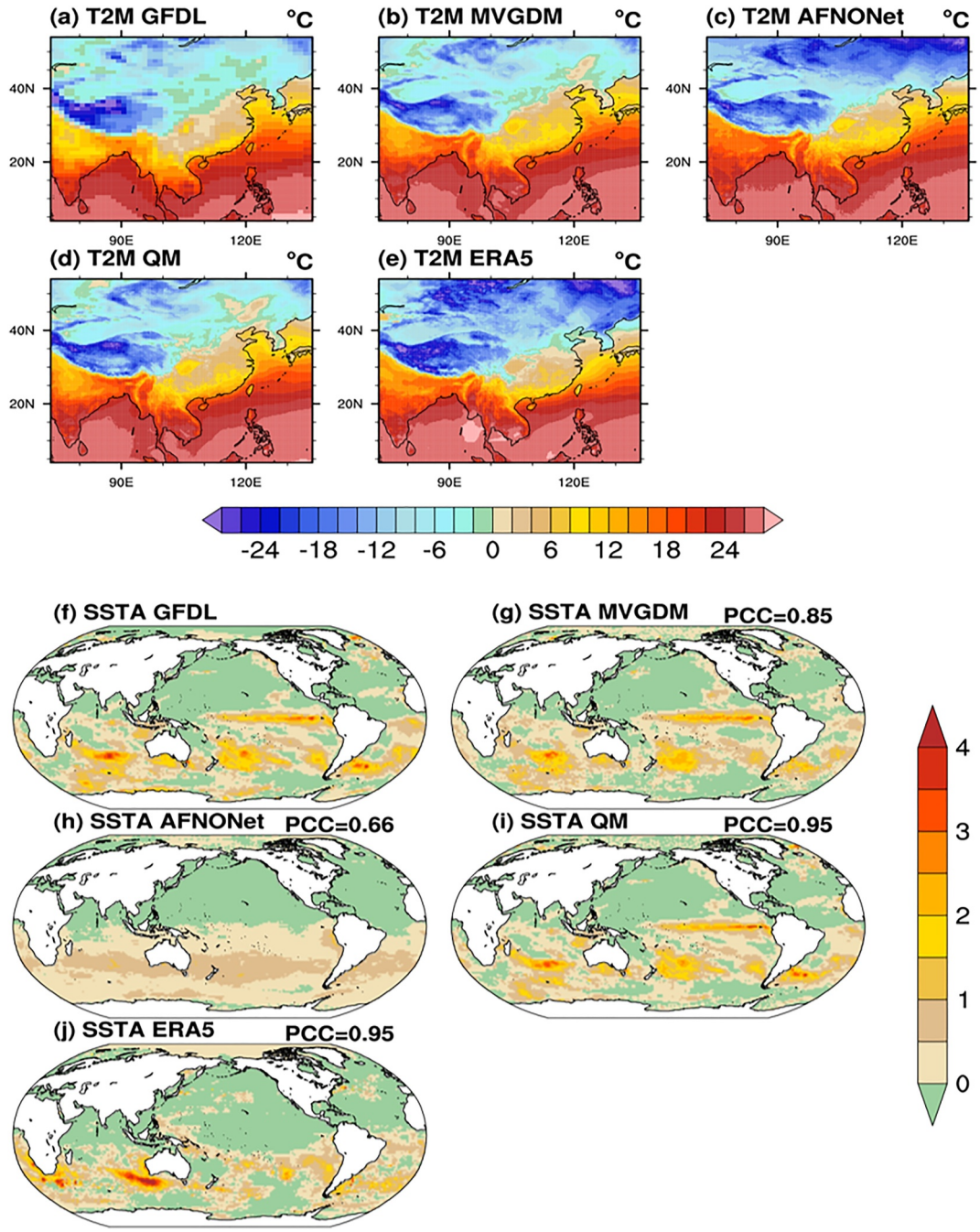

Figure 2. (a–e) T2M (°C) in East Asia from coarse-resolution GFDL outputs (a), MVGDM-generated high-resolution outputs (b), AFNONet-generated high-resolution outputs (c), QM-generated high-resolution outputs (d), and ERA5 (e) on the same date (29 December 1997); (f–j) sea surface temperature anomaly from coarse-resolution GFDL outputs(f), MVGDM-generated high-resolution outputs (g), AFNONet-generated high-resolution outputs (h), QM-generated high-resolution outputs (i), and ERA5 (j) on the same date (29 December 1999).

exhibits pronounced SST biases, featuring warm anomalies (>1.2°C) along eastern boundary currents (e.g., Humboldt and Benguela systems) and cold biases (<−0.8°C) in western boundary currents including the Kuroshio Extension and Gulf Stream (Figure 3a). These errors may stem from GFDL's inherent resolution constraints in resolving mesoscale eddy dynamics (Dennis et al., 2010), compounded by a persistent 0.9°C warm bias in Southern Hemisphere mid-latitudes. Downscaling substantially mitigates these biases, with MVGDM,

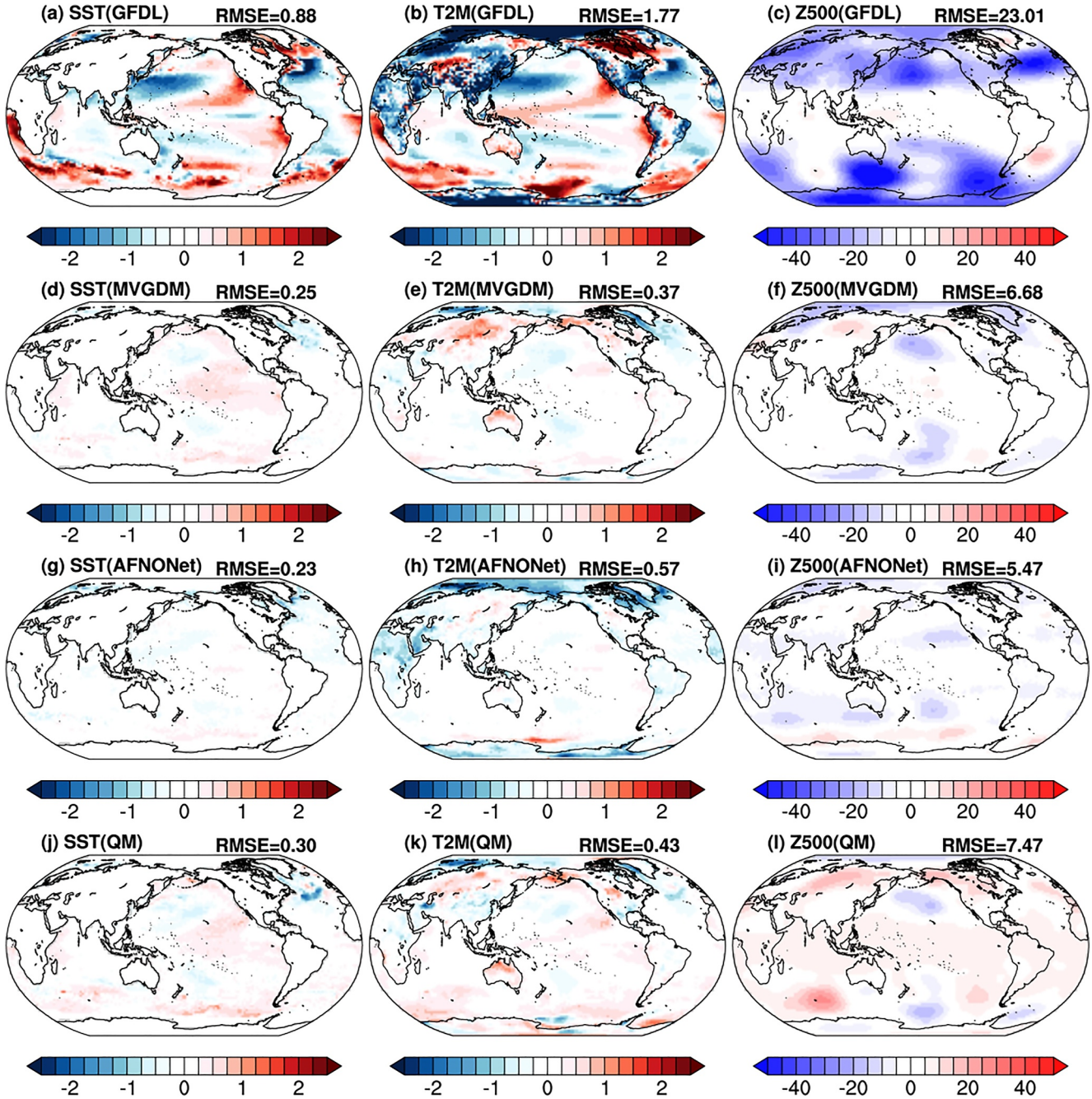

Figure 3. Climatological biases in annual mean sea surface temperature, T2M, and Z500 for various models (GFDL, multivariate generative downscaling model (MVGDM), AFNONet, quantile mapping (QM)) compared to the ERA5 reanalysis are shown in panels (a–l): GFDL results in (a–c), MVGDM results in (d–f), AFNONet results in (g–i), and QM results in (j–l).

AFNONet, QM reducing basin-wide SST root-mean-square errors (RMSE) from 0.88°C to 0.25°C, 0.23°C and 0.30°C, respectively (Figures 3d, 3g, and 3j, and Figure S3a in Supporting Information S1).

The T2M bias pattern mirrors SST patterns over oceans, with additional Arctic cold biases (−2.1°C). Over continental regions, alternating warm/cold biases (peak ±1.8°C) cluster in topographically complex regions (Tibetan Plateau, Cordilleras, Antarctica), reflecting GFDL's inadequate representation of fine-scale topographic complexity at coarse resolution. Downscaling substantially improves land temperature simulations, reducing RMSE from 1.77°C to 0.37°C (MVGDM), 0.57°C (AFNONet) and 0.43°C (QM) (Figures 3e, 3h, and 3k).

GFDL's Z500 simulations show pronounced zonal structure of bias, particularly at mid-high latitudes (RMSE = 23.01 m; Figure 3c). Downscaling reduces the RMSE to 6.68 m (MVGDM), 5.47 m (AFNONet) and 7.47 m (QM), effectively constraining the zonal mean biases (Figures 3f, 3i, and 3l, and Figure S3c in Supporting Information S1).

Overall, MVGDM, AFNONet and QM each reduce the climatological bias of GFDL by more than 65% across these three variables. MVGDM and AFNONet outperform QM for all three fields. SST correction is comparable between MVGDM and AFNONet MVGDM achieves the best improvement in T2M, and AFNONet shows the greatest reduction in Z500 bias. However, as noted above, AFNONet cannot preserve the model's internal variability, so its superior climatological performance does not guarantee suitability for tasks requiring faithful representation of internal variability.

In addition, power spectral analysis shows that MVGDM and QM capture the mean low-frequency spatial variability of each field, whereas AFNONet struggles in this regard (Figure S4 in Supporting Information S1).

3.3. Interannual Variability

Beyond climatological means, interannual dynamical variability serves as a critical metric for assessing climate models. Variabilities such as the ENSO and Indian Ocean Dipole (IOD) directly or indirectly influence regional and global climate variability.

In the raw GFDL simulation, the ENSO SSTA pattern from December to February of the following year is reproduced with high performance, albeit exhibiting a CMIP-common excessive westward extension of equatorial Pacific warming (Figure 4b; Jiang et al., 2021; Tao et al., 2014). MVGDM corrects this bias by recentering the equatorial warming and reducing the westward extension (Figure 4c). The QM method leaves the westward extension nearly unchanged (Figure 4e). AFNONet not only fails to reduce the bias but yields lower PCC and higher RMSE than the original GFDL simulation (Figure 4d).

The IOD results further highlight MVGDM's strength. Figures 4f–4j show the SSTAs associated with IOD from September to November. IOD-related cold SSTAs are centered near 90°E, which is too much westward in GFDL, whereas MVGDM shifts the center eastward to 100°E, raising PCC from 0.82 to 0.87 and lowering RMSE from 0.26 to 0.21 (Figures 4f–4h). QM produces only slight metric gains and retains the westward bias (PCC = 0.83, RMSE = 0.25; Figure 4j). AFNONet amplifies the westward cold bias, degrading performance to PCC = 0.75 and RMSE = 0.33 (Figure 4i).

We also compare the coupling relationship across different variables, for example, the Z500 response to ENSO. MVGDM better preserves the spatial pattern of ENSO-driven geopotential height anomalies than AFNONet, indicating that MVGDM maintains the dynamical link between SST and Z500 that AFNONet distorts (Figure S5 in Supporting Information S1).

Overall, MVGDM effectively corrects interannual SST variability and addresses common model errors such as excessive westward SSTA extension in ENSO and IOD. QM gives moderate improvements but is less effective than MVGDM, while AFNONet fails to resolve these biases and may even worsen them.

3.4. Extreme Values

In addition to interannual variability, previous studies have extensively investigated events such as extreme cold, heatwaves, and droughts, among others (Gong et al., 2014; Ma et al., 2021; Mishra & Singh, 2010). In this study, the 95th percentile for each variable is selected to represent extreme high-value events. The GFDL markedly overestimates SST in the central-eastern Pacific and western Atlantic, as well as the mid-to-high latitudes of the Southern Ocean, resulting in an RMSE of 1.20°C (Figure 5c). MVGDM-generated corrections significantly reduce these biases by 66.9%, lowering the RMSE to 0.40°C (Figure 5e). This is particularly crucial for the equatorial central-eastern Pacific, where significant SST anomalies are commonly linked to ENSO events that can trigger extreme global or regional weather events. QM performs less well (RMSE = 0.48°C; Figure 5i). AFNONet achieves a moderate reduction (RMSE = 0.75°C) but introduces a new cold bias between North and South America and along the equatorial central eastern Pacific (Figure 5g).

Extreme T2Ms are directly linked to severe terrestrial events such as heatwaves and droughts. The GFDL exhibits a warm bias of approximately 2°C in Oceania, South America, southern North America, and the Indian

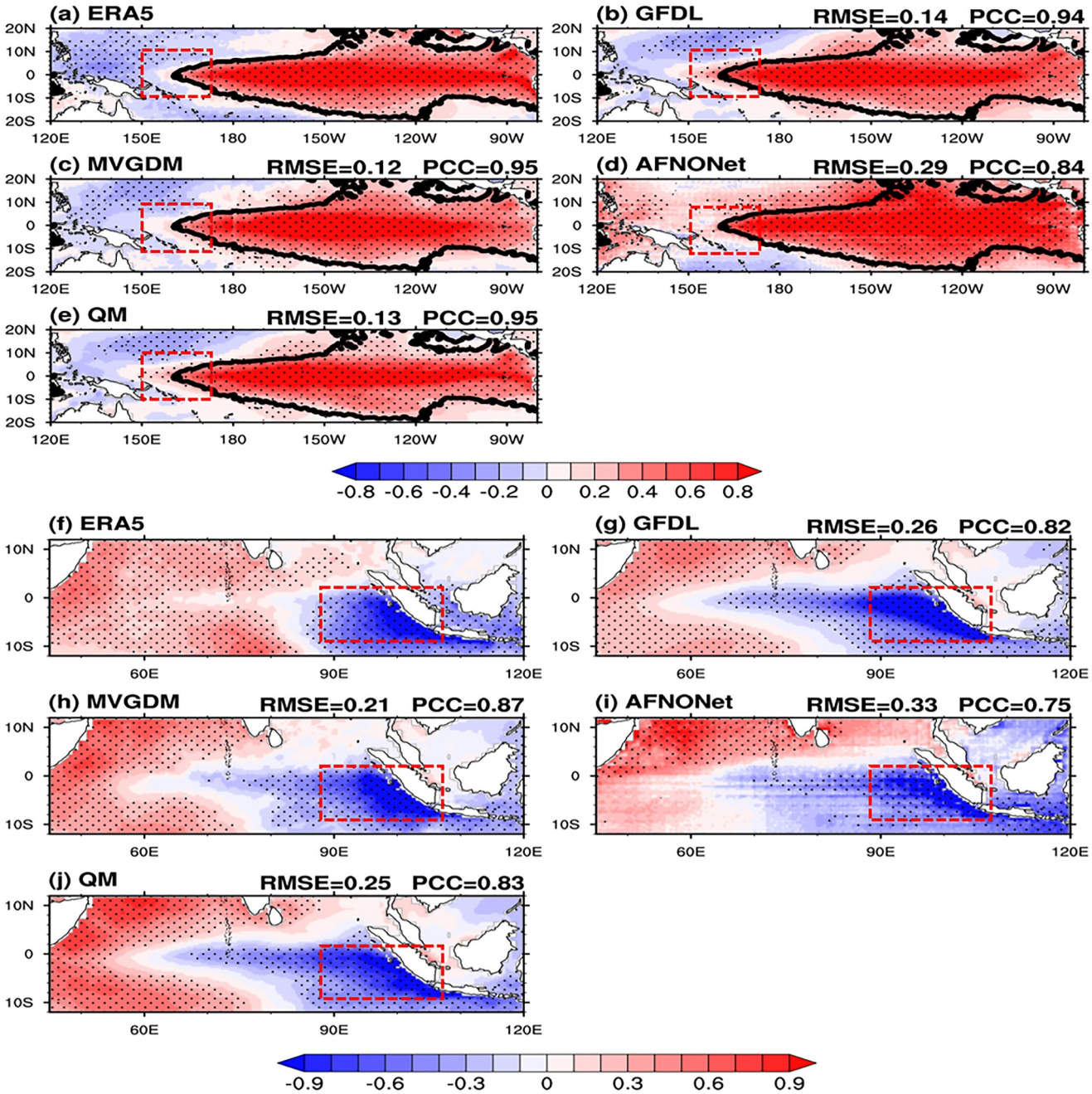

Figure 4. ENSO-related SSTAs in (a) ERA5, (b) GFDL, (c) multivariate generative downscaling model (MVGDM), (d) AFNONet and (e) quantile mapping (QM). The dotted areas indicate where the regression is significant at the 95% confidence level, and the black contours represent +0.2°C in ERA5. IOD-related SSTAs in (f) ERA5, (g) GFDL, (h) MVGDM, (i) AFNONet and (j) QM. The dotted areas indicate where the regression is significant at the 95% confidence level.

subcontinent, while a cold bias of nearly 2°C is observed over the Tibetan Plateau, resulting in an average RMSE of 1.67°C (Figure 5d). After MVGDM downscaling, these biases in complex terrain regions are greatly reduced and match observations closely, with an average RMSE of 0.42°C (Figure 5f). QM performs less well, with an RMSE of 0.48°C, while AFNONet performs worst, yielding an RMSE of 1.38°C (Figures 5j and 5h).

We also assess each model's skill in reproducing extreme low values (fifth percentile) of SST and T2M (Figure S6 in Supporting Information S1). In the raw GFDL output, the SST and T2M yield RMSEs of 1.03°C and 3.15°C, respectively. MVGDM downscaling reduces these errors to 0.31°C for SST and 0.64°C for T2M, indicating a marked improvement in cold event representation. QM achieves RMSEs of 0.40°C (SST) and 0.90°C (T2M),

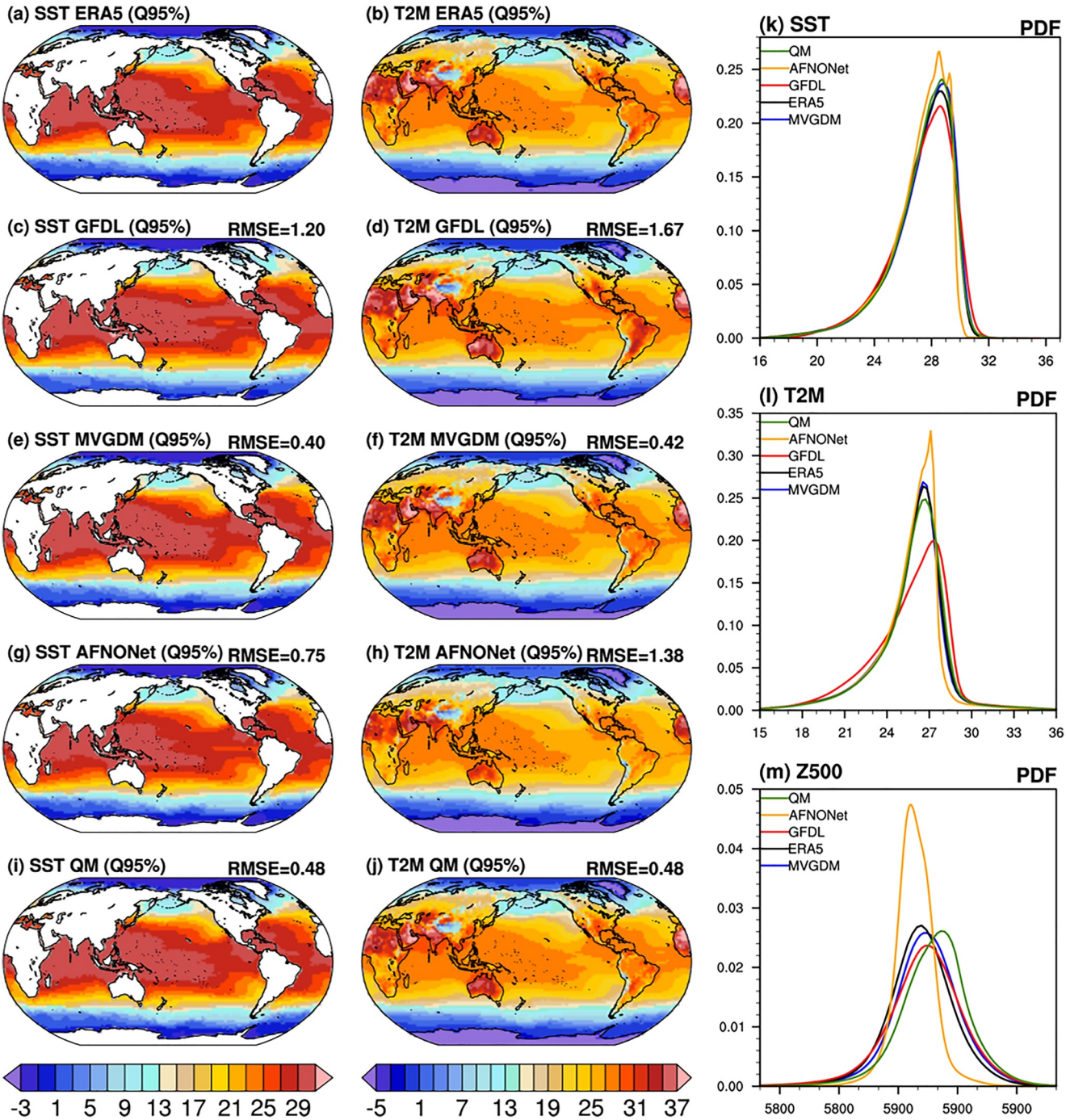

Figure 5. The 95th percentile of sea surface temperature (SST) (a, c, e, g, i), and T2M (b, d, f, h, j). PDF of Tropical SST (k), T2M (l), and Z500 (m) for the outputs of GFDL (red line), ERA5 (black line), multivariate generative downscaling model (blue line), AFNONet (yellow line) and quantile mapping (green line).

showing moderate correction but lagging behind MVGDM. The AFNONet approach lowers SST bias to an RMSE of 0.45°C yet leaves T2M errors high at an RMSE of 2.98°C.

Overall, MVGDM most faithfully reproduces the full probability distributions of SST and T2M, reducing biases by roughly 65% for extremes and outperforming both AFNONet and conventional QM. This indicates that MVGDM is highly effective at improving the simulation of high frequency variability at both ends of the

distribution. MVGDM also more accurately captures the joint distribution of SST and Z500, further demonstrating its capability to preserve the dynamic coupling between key variables (Figure S7 in Supporting Information S1).

4. Discussion

While MVGDM shows strong performance in historical simulations, a key question is whether it generalizes to future projections, especially in preserving large-scale warming signals. This issue is common across many machine learning-based downscaling and forecasting models (e.g., Hess et al., 2022), which tend to underestimate long-term trends.

We observed that directly applying MVGDM to future GFDL simulations resulted in reduced global-mean warming. To address this, we introduced a global-mean constraint post-adjustment as that in Hess et al. (2022). We first compute the global mean values of both the input and output fields of MVGDM. The output field of MVGDM is then multiplied by a scaling factor (i.e., the ratio of the global mean of the input field to that of the output field), thereby ensuring the downscaled fields maintain consistency with the original model's large-scale trends. This correction retains fine-scale detail while preserving the global warming signal, enabling high-resolution projections (Figure S8 in Supporting Information S1). However, we acknowledge that this approach does not correct any bias that may already exist in the model's global mean, leaving room for further improvement. Additionally, because regional and latitudinal warming rates differ, this type of global constraint can only partially mitigate the challenges of out-of-sample application. Therefore, the corrected results should still be interpreted with caution. These findings underscore the need to enhance model generalization in climate downscaling, e.g., by incorporating physics-informed learning or scenario-augmented training.

5. Conclusion

In this study, we propose MVGDM, an unsupervised generative model for downscaling GFDL simulations of SST, T2M, and Z500 from 100 to 25 km resolution, while reducing systematic biases. The model is highly efficient, completing 20-year simulations in under 10 min on a single A100 GPU. The model improves spatial detail, preserves internal variability, and reduces climatological biases by over 70% across key variables. MVGDM also improves interannual variability (e.g., ENSO, IOD) and better captures extremes compared to traditional QM and supervised models like AFNONet. These results suggest that unsupervised domain translation is particularly effective for Earth system downscaling tasks.

In tests with future projections, MVGDM required a global mean constraint to maintain large-scale warming signals, highlighting the importance of trend preservation when applying deep learning to climate projection scenarios. This also points to future opportunities for physics-informed designs and scenario-aware training to improve model generalization.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The CMIP6 outputs are available online on https://esgf-node.llnl.gov/projects/cmip6/. The ERA5 data set (Hersbach et al., 2019; Muñoz-Sabater et al., 2021) can be obtained at https://cds.climate.copernicus.eu/datasets. Codes used in this paper can be found at https://github.com/Haijiepwd/MVGDM.git.

Acknowledgments
We acknowledge the World Climate Research Program, which coordinated and promoted CMIP6 through its Working Group on Coupled Modeling. This work is supported by the National Natural Science Foundation of China (Grant 42141019, 92358302, 42261144687, 42175049, 42475048, and 42405041) and CCF-Baidu Open Fund. Code optimizations also benefited from the PaddlePaddle.

References

Chen, J., Xu, C., Guo, S., & Chen, H. (2016). Progress and challenge in statistically downscaling climate model outputs. *Journal of Water Resources Research*, *5*(4), 299–313. https://doi.org/10.12677/jwrr.2016.54037

Dennis, J. M., Tomas, R., Bryan, F. O., Chelton, D. B., Loeb, N. G., & McClean, J. L. (2010). Frontal scale air–sea interaction in high-resolution coupled climate models. *Journal of Climate*, *23*(23), 6277–6291. https://doi.org/10.1175/2010jcli3665.1

Dosio, A., Paruolo, P., & Rojas, R. (2012). Bias correction of the ENSEMBLES high resolution climate change projections for use by impact models: Analysis of the climate change signal. *Journal of Geophysical Research*, *117*(D17), D17110. https://doi.org/10.1029/2012jd017968

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv e-prints*, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., et al. (2020). The GFDL Earth system model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. https://doi.org/10.1029/2019ms002015

Fulton, D. J., Clarke, B. J., & Hegerl, G. C. (2023). Bias correcting climate model simulations using unpaired image-to-image translation networks. *Artificial Intelligence for the Earth Systems*, *2*(2). https://doi.org/10.1175/aies-d-22-0031.1

Gong, Z., Feng, G., Ren, F., & Li, J. (2014). A regional extreme low temperature event and its main atmospheric contributing factors. *Theoretical and Applied Climatology*, *117*(1), 195–206. https://doi.org/10.1007/s00704-013-0997-7

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS 2014)* (Vol. 27, pp. 2672–2680).

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Dee, D., Horányi, A., et al. (2019). The ERA5 global atmospheric reanalysis at ECMWF as a comprehensive dataset for climate data homogenization, climate variability, trends and extremes. In *EGU general assembly conference abstracts* (p. 10826).

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hess, P., Aich, M., Pan, B., & Boers, N. (2025). Fast, scale-adaptive and uncertainty-aware downscaling of Earth system model fields with generative machine learning. *Nature Machine Intelligence*, *7*(3), 363–373. https://doi.org/10.1038/s42256-025-00980-5

Hess, P., Drüke, M., Petri, S., Strnad, F. M., & Boers, N. (2022). Physically constrained generative adversarial networks for improving precipitation fields from Earth system models. *Nature Machine Intelligence*, *4*(10), 828–839. https://doi.org/10.1038/s42256-022-00540-1

Jiang, W., Huang, P., Huang, G., & Ying, J. (2021). Origins of the excessive westward extension of ENSO SST simulated in CMIP5 and CMIP6 models. *Journal of Climate*, *34*(8), 2839–2851. https://doi.org/10.1175/jcli-d-20-0551.1

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., et al. (2023). FourCastNet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference, pasc 2023*.

Ma, C. S., Ma, G., & Pincebourde, S. (2021). Survive a warming climate: Insect responses to extreme high temperatures. *Annual Review of Entomology*, *66*(1), 163–184. https://doi.org/10.1146/annurev-ento-041520-074454

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*(3), RG3003. https://doi.org/10.1029/2009rg000314

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional networks for biomedical image segmentation*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Tao, W., Huang, G., Hu, K., Qu, X., Wen, G., & Gong, H. (2014). Interdecadal modulation of ENSO teleconnections to the Indian Ocean basin mode and their relationship under global warming in CMIP5 models. *International Journal of Climatology*, *35*(3), 391–407. https://doi.org/10.1002/joc.3987

Themeßl, M. J., Gobiet, A., & Heinrich, G. (2011). Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Climatic Change*, *112*(2), 449–468. https://doi.org/10.1007/s10584-011-0224-4

Wang, F., & Tian, D. (2022). On deep learning-based bias correction and downscaling of multiple climate models simulations. *Climate Dynamics*, *59*(11), 3451–3468. https://doi.org/10.1007/s00382-022-06277-2

Wang, Y., Huang, G., Pan, B., Lin, P., Boers, N., Tao, W., et al. (2024). Correcting climate model Sea surface temperature simulations with generative adversarial networks: Climatology, interannual variability, and extremes. *Advances in Atmospheric Sciences*, *41*(7), 1299–1312. https://doi.org/10.1007/s00376-024-3288-6

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Paper presented at 2017 IEEE international conference on computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2017.244