Global Climate Modeling with Improved Precipitation Characteristics by Learning Physics (GRIST-MPS v1.0) from Global Storm-Resolving Modeling

Yiming Wang[a], Yi Zhang[b], Yilun Han[c,d], Wei Xue[a], Yihui Zhou[e], Xiaohan Li[b], Haishan Chen[b]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[b] State Key Laboratory of Climate System Prediction and Risk Management (CPRM)/Key Laboratory of Meteorological Disaster, Ministry of Education/School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Jiangsu, China.
[c] Department of Earth System Science, Tsinghua University, Beijing, China.
[d] Scripps Institution of Oceanography, La Jolla, CA, USA.
[e] State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, Beijing, China

Correspondence to: Yi Zhang (yizhang@nuist.edu.cn)

Abstract. This study develops a machine learning (ML)-based physics parameterization suite trained on 80-day global storm-resolving model (GSRM) simulation data (5km), attempting to replace all conventional physics tendencies in a general circulation model (GCM, 120km) for real-world simulations with realistic surface topography. The GSRM data are generated using the Global–Regional Integrated Forecast System (GRIST) and subsequently coarse-grained, after which the residual method is applied to derive the corresponding GCM physics tendencies. The resulting workflow relies on standardized pressure-level variables as input features, enabling the GCM—through physics–dynamics coupling—to effectively emulate the multiscale flow interactions captured by the GSRM. This ML-enhanced GCM achieves stable Atmospheric Model Intercomparison Project (AMIP)-type simulations over six years, surpassing its conventional counterpart with improved precipitation performance, especially over the Inter-Tropical Convergence Zone (ITCZ) region. It effectively mitigates the biases of excessively-strong rainbands and overly-wide width of the ITCZ in the conventional configuration, when compared with the Global Precipitation Measurement (GPM) data. Moreover, the hybrid ML-GCM better captures precipitation frequency, notably mitigating the overproduction of light tropical rainfall. Sensitivity experiments using different neural network architectures (ResNet, CNN, DNN) demonstrate that all configurations can maintain long-term simulation stability, with ResNet showing superior capability in the simulation accuracy. This work presents a transferable framework that leverages km-scale GSRM data to enhance GCM performance via ML integration, offering a potential route to reduce the gaps between two modeling paradigms.

1 Introduction

Weather and climate modeling, which both embodies our understanding of the atmosphere and deepens it, currently operates within two distinct paradigms: (i) highly parameterized general circulation models (GCMs), which are extensively utilized in global climate change research initiatives such as the Coupled Model Intercomparison Project (Eyring et al., 2016); and (ii) global storm-resolving models (GSRMs) with kilometer-scale resolutions that can explicitly resolve (deep) convective processes (Satoh et al., 2019). These two modeling paradigms remain operationally decoupled due to the lack of a unified discretization approach that enables seamless resolution transitions (Yu et al. 2019; Brunet et al. 2023; Miura et al. 2023). A major challenge in bridging this gap lies in the representation of moist physical processes, which govern scale interactions across different modeling paradigms. GCMs rely on cumulus parameterization schemes that approximate the bulk effect of interactions between moist convection and large-scale circulation, a well-known source of climate modeling uncertainties (Arakawa 2004; Lin et al. 2022). GSRMs explicitly resolve the coupling between atmospheric dynamics and microphysics, and support multiscale flow, hopefully yielding more physically realistic cumulus convection and multiscale interactions. When incorporated into GCMs, these interactions may replace sub-grid eddy effects relative to the GCM's grid box, alongside representations of heating and cooling effects due to phase changes, radiative transfer, and friction.

Machine learning (ML) algorithms have been increasingly applied to facilitate this integration (Schneider et al., 2023; Eyring et al., 2024), raising the prospect of constructing hybrid ML–physics models (Krasnopolsky and Belochitski 2020). The physical tendencies can be learned separately either to replace an individual scheme (e.g., Chen et al. 2023; Heuer et al. 2024; Morcrette et al. 2025), or to replace the entire tendency from the physics suite. This study focuses on the latter approach. Currently, several methods exist for constructing hybrid ML–physics models using this approach. The online learning strategy, which leverages differentiable numerical solvers to match model outputs with reference/observation datasets, has demonstrated promise in generating reasonably realistic climate simulations (Kochkov et al. 2024). A challenge lies in interpreting the nature of the learned physics in this approach. It remains unclear whether the learned tendencies stem purely from real physical processes (e.g., phase change, eddy effect, friction, radiative heating, etc), or if they also incorporate certain additional components such as the nudging tendency, which can be independently learned (Bretherton et al. 2022); or like state correction, which combines conventional numerical models with ML model (Arcomano et al. 2022). The physical meanings of these tendency terms are different.

Another approach is to directly learn physical tendencies from their generating sources. These sources may include high-resolution process models (e.g., large-eddy simulations, cloud-resolving models) or observational datasets (Zhu et al. 2022; Bracco et al. 2025). For instance, ML schemes trained on physical tendencies derived from super-parameterized GCMs (e.g., Rasp et al., 2018; Gentine et al., 2018; Han et al., 2020) have demonstrated the ability to retain the physical fidelity of super-parameterized modeling while significantly reducing the computational cost. Several operational implementations of such models have achieved multi-year simulation stability in realistic configurations (e.g., Han et al. 2023; Mooers et al., 2021; Wang et al., 2022; Chen et al. 2025). In contrast, GSRMs do not impose artificial scale separation, and learning physics

tendencies from GSRMs presents a unique advantage by allowing for more physically consistent multiscale flow interactions that closely align with real-world atmosphere. Brenowitz and Bretherton (2018) used neural network-based parameterizations and coarse-grained GSRM data, demonstrating multi-year simulation stability in low-resolution aqua-planet scenarios. Yuval and O'Gorman (2020) employed random forests trained on three-dimensional cloud-resolving model outputs to emulate fine-scale processes in coarse-grid model systems. Yuval et al. (2021) refined this approach by leveraging neural networks, achieving comparable predictive accuracy while reducing memory requirements by a factor of 1,900. These advancements have primarily been tested in idealized aqua-planet configurations, raising critical questions about their applicability to realistic climate modeling. Watt-Meyer et al. (2024) developed a GCM physics parameterization suite trained on coarse-grained GSRM data under realistic surface boundary conditions, enabling stable 35-day simulations while significantly reducing mean-state precipitation and temperature errors. While this approach has not demonstrated very significant advantages in real-world modeling with respect to certain utilitarian metrics (e.g., multiyear climate state error), it has the potential to reconcile scale disparities from a physically orientated training way.

In this study, we develop a ML-based Physics parameterization Suite (MPS: a column model). It is then used to generate temperature and humidity tendencies online for a realistic GCM (Global-Regional Integrated Forecast System; GRIST). We propose an integrated workflow that enables the GCM—through physics and dynamics coupling—to emulate the multiscale flow interactions represented by the GSRM, along with other processes such as phase changes and radiative heating. We have experimented with several neural network architectures, including Residual Neural Networks (ResNet), convolutional and deep neural network (CNN, DNN). A sensitivity analysis uncovers that different network architectures produce divergent equilibrium climate states despite identical training data and hyperparameter configurations are used. The optimal outcome is from the ResNet. It achieved long-term stable Atmospheric Model Intercomparison Project (AMIP)-type climate simulations more than six years, and produces simulations comparable or better than those produced by a conventional physics suite (CPS). We focus particularly on precipitation, because it requires a faithful representation of multiscale flow interactions, and accurate reproduction of large-scale state variables does not necessarily translate into improved precipitation performance (e.g., Chen et al. 2025), making its simulations particularly challenging. Therefore, precipitation provides an informative metric for assessing the effectiveness of learning from GSRMs.

The remainder of this paper is organized as follows. Section 2 presents the data and methods. Section 3 presents the simulation results and discusses sensitivity of neural networks. Section 4 gives a summary and outlook.

2 Model, Data and Methods

2.1 Model description and high-resolution GSRM data

The hybrid modeling framework is based on the GRIST model. The features of dynamical core framework are detailed in Zhang et al. (2019), Zhang et al. (2020) and Zhang et al. (2024). The baseline physics suite is described in Li et al. (2023),

with some improved schemes given by e.g., Li et al. (2022), and Li et al. (2024). For this study, we adopt the weather physics (PhysW) suite as the basis of model development (see Li et al. 2023 for details).

GRIST is employed in two configurations: (i) a high-resolution (5 km) GSRM-style setup for generating training data for the MPS, and (ii) a coarse-resolution (120 km) GCM-style setup for applying and evaluating the MPS. Both configurations feature 30 vertical layers. The GSRM setup uses the nonhydrostatic dynamical core with explicit convection, in which cumulus scheme is disabled, following the approach of Zhang et al. (2022). Obviously, the quality of the GSRM data is critical for the effective development of the MPS. In Zhang et al. (2022), the model successfully captured the multiscale interactions between moist convection and large-scale circulation. Their simulations demonstrated that the time-averaged characteristics of these interactions are comparable to those produced by the GRIST-GCM configuration with conventional cumulus parameterization, but supports better transient features (e.g., extreme rainfall intensity). While GRIST-GSRM exhibits slightly higher mean-state precipitation biases, it shows superior skill in reducing systematic errors, for example, reducing the excessive frequency of light tropical rainfall and increasing the frequency of intense rainfall. This underscores the importance of replicating the GSRM-resolved multiscale interactions for developing an effective MPS applicable to GCMs.

The GCM configuration follows the setup described in Zhang et al. (2021), using the hydrostatic dynamical core coupled with the conventional parameterization suite (CPS), where the cumulus parameterization in PhysW is enabled. All other physics schemes—including microphysics, boundary layer, radiation, surface layer, and land surface model—are identical between the GSRM and GCM configurations, thereby ensuring maximum consistency. Some other details of the two configurations are provided in Table 1.

To enhance the representativeness of the training data, we select four 20-day periods (Table 2) that span different seasons and capture key phases of the El Niño–Southern Oscillation (ENSO) and Madden–Julian Oscillation (MJO). These periods collectively ensure comprehensive seasonal coverage—January (boreal winter), April (boreal spring), July (boreal summer), and October (boreal autumn)—and systematically represent the dominant ENSO–MJO interaction regimes that drive climate variability. A 7:1 ratio was used to divide the dataset into training and validation sets. For each day, 12.5% of the time points were randomly allocated to the validation set, and the remaining 87.5% were used for training. This temporal sampling strategy supports a reliable assessment of the model's out-of-sample performance. The current choice of 80 days reflects a practical limitation due to computational and resource constraints, but it already allows essential atmospheric physical processes to be effectively sampled using a limited set of time windows. That said, increasing the number of training samples may further enhance the performance of the MPS.

2.2 Coarse graining and data preprocessing

We extract multiscale flow interactions in the GSRM using a thermodynamic framework following Yanai et al. (1973), in which the apparent heat source (Q_1) and apparent moisture sink (Q_2) serve as mathematical representations of these interactions. These quantities are derived from *coarse-grained GSRM data* (at 0.25° resolution) using the residual method (e.g., Zhang and Chen 2016). Specifically, the gradient operator is achieved via the center difference method applied to the coarse-

grained fields, following the governing equations shown in Figure 1 (the middle section of the left panel). While the present study coarse-grains GSRM data to a fixed resolution, the residual method is inherently adaptable. It can seamlessly bridge models of arbitrarily high resolution to GCM target scales. Establishing a robust physical correspondence between GSRMs and GCMs not only allows GCMs to mimic certain behaviors of GSRMs, but also opens the door to unified simulations of atmospheric processes within a single modeling framework—enhancing both theoretical understanding and predictive skill across multiple timescales. This architecture-agnostic framework offers two advantages: (i) ability to enable the transfer of scale interactions represented in GSRMs to a target GCM resolution, and (ii) interoperability with the broader modeling community using standard pressure-level atmospheric variables. Several key design choices are further highlighted below.

Choice of Large-Scale Variables. Some preliminary tests identified the optimal set of input features to include temperature (T), specific humidity or mixing ratio (q), horizontal wind components (U and V), and surface pressure (P). Although the inclusion of vertical velocity (ω) is theoretically advantageous, it was found to introduce numerical instabilities in regions with complex topography—a result consistent with previous studies (Clark et al. 2022, Rasp et al. 2018 and Watt-Meyer et al. 2024). All prognostic variables were normalized using min–max scaling, based on their extrema within the 80-day training dataset.

Vertical coordinate alignment. For machine learning training, it is desirable to use the model's native hybrid coordinate, which avoids topographic distortion during runtime. Calculating Q_1/Q_2 in the residual method requires first obtaining the advection tendencies. However, directly computing advection tendencies offline on the hybrid vertical coordinate is inaccurate because the generalized vertical velocity cannot be reliably reconstructed from coarse-grained data. It would require the generalized vertical velocity to be explicitly saved during the online integration, which is currently not available. More importantly, we prefer to confine our training workflow to standardized pressure-level variables as inputs, ensuring that the workflow has the potential to be consistently applicable to non-GRIST GSRM datasets.

To reconcile this discrepancy, we implement a two-step procedure. In Step I, GSRM variables on the hybrid model coordinate are interpolated to pressure levels for the sole purpose of computing advection tendencies. In Step II, the resulting advection tendencies are interpolated back to the model's hybrid coordinate, where Q_1 and Q_2 are then derived. Ultimately, all training inputs (U, V, T, q, P) and outputs (Q_1 and Q_2) are defined on the model's hybrid vertical coordinate, ensuring compatibility with the runtime model structure while preserving physical accuracy in the derivation process.

Temporal resolution alignment. To enhance temporal resolution, we applied linear interpolation to convert hourly coarsened GSRM model outputs into 20-minute interval data, effectively tripling the temporal sampling frequency. This refinement is crucial for improving stability and accuracy of online model integration, as it increases the time samples and better aligns the temporal characteristics of the training data with the time step of the target GCM (see Section 2.4).

2.3 Training the MPS

The MPS leverages residual neural network architecture by default, with tailored modifications for atmospheric column physics. Central to the design are one-dimensional convolutional layers that explicitly resolve vertical couplings in temperature

and humidity profiles, particularly during deep convective events where multi-level interactions dominate subgrid energy transfer. To balance representational capacity with computational efficiency, the network employs five optimized residual units (ResNet5, Figure 1)—a depth empirically determined to preserve most validation accuracy of deeper architectures while saving a lot of training time and resources. We used the Adam optimizer with a constant learning rate of 3×10^{-4} and a weight decay of 10^{-6} per epoch. The mean absolute error (MAE) loss was selected over the mean squared error (MSE) loss as the loss function, as it demonstrated superior performance during initial training phases.

To optimize computational efficiency while maintaining global representativeness, we implemented a stratified spatiotemporal sampling strategy. Each temporal snapshot (20-minute interval) extracts 86,400 grid columns distributed across key climate regimes: 50% from tropical latitudes (30°S–30°N) where convective processes dominate, 30% from mid-latitudes (60°S–30°S and 30°N–60°N) capturing baroclinic eddy activity, and 20% from polar regions (90°S–60°S and 60°N–90°N) resolving radiative-polar amplification feedbacks. This geographic weighting generates 497,664,000 training samples (80 days × 24 hours × 3 samples/hour × 86,400 columns). The network underwent 100 training epochs with the batch sizes of 1024 with early stopping (patience=5 epochs, δval_loss<0.5%) to ensure full data utilization without overfitting.

Rigorous offline evaluation is important for transitioning ML physics into an operational tool. We quantify emulation fidelity through two complementary metrics: (i) domain-averaged mean squared error (MSE < 1×10^{-4}) and (ii) vertical-latitude cross-sections of the coefficient of determination ($R^2 > 0.3$ across most of tropical and midlatitude tropospheric grid points; Figure 2), which collectively verify process-level skill in moisture-convection coupling. Networks satisfying both thresholds proceeded to online testing. This dual-criterion screening prevents numerically stable but physically implausible models from entering computationally intensive online integration phases.

Superior offline performance alone does not guarantee online stability, as the effects of physics-dynamics coupling cannot be purely grasped through offline training. To address this, the shortlist of networks that meet our predefined offline criteria was the first step, then we subject them to online testing. The final selection of our optimal MPS is based on a dual evaluation: satisfying offline performance benchmarks and demonstrating stability in online integration, which must maintain stable online integration for more than 3 months. For the NNs that meet the above criteria, we will continue their integration until collapse. Among the 8 NNs, 2 NNs can integrate stable for more than 6 years, and we have selected the one with better performance as the optimal MPS.

2.4 Importance of using balanced spatiotemporal sample and temporal resolution alignment

During model development, we identified two key factors that significantly improve the stability and accuracy of the MPS. The first is achieving a more balanced spatiotemporal sample. Initial experiments using the full spatial samples (1440 x 720 grid columns per timestep) combined with coarse temporal sampling (hourly data) led to numerical instabilities during online integration. This instability stemmed from an extreme space–time sampling ratio, which caused the neural network to overfit spatial patterns while failing to adequately learn temporal evolution. To address this issue, we adopted a stratified spatiotemporal subsampling approach: at each timestep, only 86,400 geographically distributed columns were randomly

selected, and the temporal resolution was increased to 20-minute intervals via linear interpolation. This strategy balanced spatial and temporal dimensionality while effectively increasing the number of training samples, encouraging the network to focus on both the time evolution of atmospheric processes and static spatial features.

The second key aspect is aligning the temporal resolution of the data with the model's integration time step. As noted earlier, we refined the temporal resolution of the large-scale variables by linearly interpolating hourly data to 20-minute intervals prior to computing Q_1 and Q_2 tendencies. This refinement offers two primary benefits. First, augmenting the dataset by a factor of three provides a regularization effect, which is known to improve model stability and generalization (Bishop et al. 1995). Second, the use of linear interpolation is justified for large-scale state variables, which typically evolve quasi-linearly over sub-hourly timescales ($\Delta t < 1$ hr). However, this assumption does not hold as well for Q_1 and Q_2, which exhibit stronger spatiotemporal nonlinearity. As such, performing interpolation only on the input variables—rather than generating full 20-minute GSRM outputs—achieves a 2/3 data compression ratio compared to storing the full-resolution dataset. Linear interpolation is not the only means of generating more data; one may alternatively choose to directly sample the model state at finer, aligned timesteps, but this is more expensive. While temporal resolution alignment is the key, our results demonstrate that the linear interpolation of large-scale state variables serves as an effective ecomonical alternative to finely sampled model output.

Altogether, these two methods enhance the stability and the accuracy of the simulations. The systematic evaluation of training strategies (Table 3) highlights the critical role of spatiotemporal data optimization in governing model performance. In the baseline experiment (EXP1), which employed neither spatial subsampling nor temporal refinement, the model-maintained stability for only three years. Introducing spatial subsampling alone (EXP2) extended stable integration to six years. Further incorporating 20-minute temporal interpolation in EXP3—i.e., full spatiotemporal optimization—maintained six-year stability while substantially reducing the tropical precipitation RMSE by 42% (2.78 mm/day vs. 4.81 mm/day in EXP1). Compared to EXP2, EXP3 yielded a 10% reduction in six-year mean precipitation RMSE (2.81 mm/day vs. 3.12 mm/day), demonstrating the additive benefit of temporal refinement beyond spatial subsampling alone. This underscores that careful data curation, without modifications to the AI-model architecture, can effectively address key challenges in ML–physics integration.

2.5 Online GCM simulation workflow with the MPS

The ML-physics-hybrid GCM builds upon the GRIST framework, with the control experiment (CPS) replicating the configuration described in Zhang et al. (2021) (Table 1, GRIST-CPS). To interface the Fortran-based GRIST model code with the PyTorch-formatted MPS, we implemented bidirectional coupling through the Ftorch library—a framework enabling real-time tensor exchange between the dynamical core and pretrained neural networks while maintaining operational efficiency.

The online implementation (Figure 1, right panel) adopts a modular architecture, in which the GRIST-GCM dynamical core iteratively transfers atmospheric state tensors to the MPS. The MPS, interfaced via Ftorch, returns Q_1 and Q_2 tendencies, while legacy CPS diagnostic modules—such as radiation and land-surface coupling—remain unmodified. By restricting

replacements to the physical tendency generation components and preserving the native diagnostic workflow, the framework mirrors the CPS substitutions and ensures fully backward compatibility. The replaced CPS components include tendencies from the cumulus parameterization, cloud microphysics, boundary layer scheme, and radiative transfer. The radiation module—the most computationally expensive element in the CPS—is still activated to generate surface fluxes for the land surface and may require a special training in future. The surface layer and land surface models are also retained in their original form, consistent with standard CPS configurations. Surface precipitation flux (*Prec*; unit: $kg.m^{-2}.s^{-1}$) is diagnosed by the MPS via vertically integrated moisture tendency equation plus the evaporation flux ($Evap$), calculated using: $Prec = -\frac{1}{g}\int (Q_2/L)dp + Evap$. The evaporation term is included solely for diagnostic purposes; the precipitation input provided to the land surface model actually excludes this term, as a tuning procedure.

Due to the MPS's coarser vertical resolution in the lower troposphere (Δz exceeding 200 m below 850 hPa), we retain CPS-derived temperature tendencies (Q_1) in the *lowest four model levels* and moisture tendencies (Q_2) in the *lowest two model levels*. This selective preservation, validated through sensitivity experiments, serves as a stability-enhancing mechanism. Meanwhile, as in prior studies (Brenowitz and Bretherton, 2019; Clark et al., 2022; Watt-Meyer et al., 2024), we apply vertical truncation of the MPS-predicted Q_1/Q_2 tendencies above 300 hPa, effectively excluding the top 12 model layers from machine-learned physics. Because the dominant portion of the total physical tendencies is supplied by the ML-based physics, the conventional parameterizations, although still active during the model integration for the diagnostic purpose, primarily respond to the prognosed grid-scale states but exert little direct influence on it (except at the specific levels noted above).

This hybrid replacement strategy demonstrates that partial physics–ML integration can achieve climate fidelity comparable to a full replacement and mitigating numerical instability. Furthermore, when extended to higher resolutions, it reduced computational costs by over 30% with a similar configuration (Duan et al., 2025) when optimizations have been carried out. This is attributed to the more optimizable computational structures of ML models (convolution, matrix multiplication), which are clearly difficult to achieve in conventional schemes.

3 Results

3.1 Real-world climate simulations

Two six-year AMIP-style simulations (2001-2006) were conducted at 120 km horizontal resolution: a control experiment with the CPS and an ML-enhanced counterpart with the MPS. We evaluate the zonal mean vertical structures of long-term mean temperature (T), specific humidity (q), and zonal wind (U) represent direct prognostic targets of the MPS (through Q_1 and Q_2 tendencies), while U emerges as a dynamically constrained diagnostic variable reflecting momentum redistribution. ERA5 reanalysis data (Hersbach et al., 2020) serve as the observational benchmark, with all model outputs regrided to 1°×1° resolution using conservative remapping.

Figure 3 demonstrates close alignment between GRIST-MPS and GRIST-CPS in simulating zonal-mean vertical structures. Both models exhibit temperature deviations (shading) within ±5 K from ERA5 reanalysis, demonstrating consistent cold biases in the polar lower stratosphere and warm biases in the tropical upper troposphere. Specific humidity profiles (black contours) display nearly identical vertical distributions between configurations. The structure of the zonal wind (U) form a wedge-like structure with the humidity, showing little differences in midlatitude jet core positions.

Precipitation is evaluated against the Global Precipitation Measurement (GPM) Product (Huffman et al., 2019). Both configurations realistically capture the boreal summer (June-July-August; JJA) precipitation dipole—the Intertropical Convergence Zone (ITCZ, 0°N–15°N) and South Pacific Convergence Zone (SPCZ, 5°S–15°S) with maximum rates exceeding 12 mm/day over the Bay of Bengal and western Pacific warm pool (Figures 4a-c).

During JJA months (Table 4), GRIST-MPS produces a more realistic ITCZ in terms of both strength and width, despite exhibiting a slightly lower pattern correlation coefficient (PCC: 0.86) than GRIST-CPS (0.94). Following established practice, we quantify ITCZ strength by the maximum zonal-mean tropical precipitation rate (Wodzicki and Rapp, 2016) and define its width as the tropical latitudinal extent with precipitation exceeding 5 mm day^{-1} (Byrne et al., 2018). The MPS accurately captures the ITCZ strength (8.69 mm day^{-1}), closely matching the GPM estimate (8.54 mm day^{-1}). By contrast, the CPS produces an excessively strong (10.09 mm day^{-1}) and overly broad (9.60°) rain band, compared with the observed width of 7.44° and the MPS width of 7.76°.

Extratropical performance remains comparable, with both models capturing most of observed midlatitude storm-track variance (55°N–65°N). The MPS slightly underestimates precipitation over the southern oceans (30°S–70°S) at 2.19 mm/day against the observed 2.67 mm/day, whereas the CPS overestimates it (3.25 mm/day), with the bias extending to 70°S.

During boreal winter (December-January-Feburary; DJF), GPM observations reveal a meridionally contracted state of tropical rainbands and intensified midlatitude storm-track precipitation (45°N–60°N, Figure 4d). Both configurations capture this seasonal transition (Figure 4e, f), with GRIST-MPS demonstrating enhanced fidelity over the equatorial Pacific and South American through a 4% RMSE reduction (1.96 mm/day versus 2.04 mm/day, not shown). In particular, over the South American region, the spatial correlation coefficient of MPS precipitation (0.95) exceeds that of the CPS (0.93).

Meanwhile, residual biases persist in GRIST-MPS: a 15%–20% overestimation of summer tropical Indian ocean rainfall (10°S–10°N, 65°E–95°E), a 4-6 mm/day overestimation of Northern Equatorial Pacific and a systematic 1-3 mm/day underestimation of Southern Ocean (50°S–60°S) and Maritime Continent (5°S–5°N, 95°E–150°E) precipitation across seasons.

Both configurations accurately reproduce the observed seasonal migration of tropical precipitation maxima (Figure 5), with boreal summer peaks centered near 5-10°N aligned with the northward-migrating ITCZ. However, systematic discrepancies emerge in the meridional range of precipitation representation: GRIST-CPS overestimates the central precipitation intensity, generating strengthened rainfall distributions of overactive convective initiation in cumulus parameterizations. GRIST-MPS demonstrates slightly overestimation of the precipitation range throughout the seasonal cycle.

To systematically assess the performance of GRIST-MPS in characterizing complex atmospheric systems, we employ the East Asian Monsoon as our case study. Our analysis utilizes established East Asian monsoon index (EAMI) from prior studies

as benchmark metrics (Zhu et al. 2005). The EAMI takes the influence of the annual cycle of the meridional and zonal sea-land thermal differences into account in the East Asia-Pacific region and reasonably describes the characteristics of the annual cycle of the transition between the East Asian winter and summer monsoons, which is defined as:

$$EAMI = (U_{850hPa} - U_{200hPa}) *_{(100-130^{\circ}E,0-10^{\circ}N)} + (SLP_{160^{\circ}E} - SLP_{110^{\circ}E}) *_{(10-50^{\circ}N)} \quad (1)$$

where U represents area-averaged $(100-130^{\circ}E, 0-10^{\circ}N)$ monthly mean zonal winds (dimensionless), SLP denotes averaged monthly sea level pressure $(10-50^{\circ}N)$ (dimensionless), and the asterisk (*) operator indicates variable standardization through mean removal and unit-variance scaling ($X = (X - \mu)/\sigma$), where X is the corresponding variable (U, SLP), where μ represents the mean of variable X and σ represents the standard deviation of variable X. This enables a quantitative assessment of the model's ability to capture both the seasonal and interannual variability characteristics of monsoon dynamics.

We computed the EAMI for monthly variables and derived its climatological seasonal cycle across a six-year period (Figure 6). Both GRIST-CPS and GRIST-MPS successfully replicate the observed seasonal monsoon phase, capturing the July maximum and February minimum. While GRIST-CPS simulations align closely with observations, GRIST-MPS exhibits a systematic bias: it overestimates monsoon intensity prior to July and underestimates it post-July. This indicates that GRIST-MPS could simulate the annual cycle of the East Asian monsoon, even though the training data only includes 80 days. This outcome strongly motivates a further refinement of MPS for extended climate applications.

The intensity–frequency distribution of precipitation reflects intrinsic model characteristics that remain stable over the course of a simulation. To assess whether the MPS faithfully captures the behavior of the GSRM, we conducted parallel experiments with the MPS and CPS using time periods aligned with the GSRM (i.e., the four cases listed in Table 2). Focusing on tropical precipitation (10°S–10°N), we categorize rainfall into four intensity ranges: light (0.1–10 mm day^{-1}), moderate (10–25 mm day^{-1}), heavy (25–50 mm day^{-1}), and extreme (>50 mm day^{-1}). As shown in Fig. 7a, relative to GRIST-CPS, the GSRM exhibits reduced total precipitation frequency and a lower frequency of light rainfall. GRIST-MPS consistently reproduces these features, with both total and light precipitation frequencies lower than in GRIST-CPS. Furthermore, comparing Figs. 7a and 7b reveals that both GRIST-CPS and GRIST-MPS display similar frequency characteristics in the GSRM-aligned experiments and the long-term free-run integrations, underscoring the robustness of these model behaviors.

Besides GPM observations, the ensemble means values of 11 CMIP6 models (CESM2, CESM2-WACCM, CMCC-CM2-SR5, E3SM-2-0, E3SM-2-0-NARRM, EC-Earth3, EC-Earth3-AerChem, GFDL-CM4, MRI-ESM2-0, SAM0-UNICON, TaiESM1; hereafter CMIP6-ENS) are included. In relative to GPM data, both CMIP6-ENS and GRIST-CPS overestimate total precipitation occurrence by 54% and 34%, respectively (Figure 7b)—consistent with earlier documented biases (Fu et al., 2024). The MPS reduces this discrepancy to 31%. It reduces light and heavy rain overprediction by 10% and 5%, respectively, while preserving observed extreme precipitation frequencies. This demonstrates that MPS effectively mitigates persistent precipitation distribution errors without compromising heavy-precipitation event statistics. Meanwhile, neither the CPS nor the MPS indicates a long-term artificial declining trend in precipitation (figure not shown).

3.2 A sensitivity analysis of different neural networks

Besides ResNet, we have also integrated two alternative neural network architectures—Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN)—to examine the sensitivity of online simulations to neural networks. The three networks are trained by identical datasets and preprocessing procedures. The switch of each network during the GRIST-MPS runtime only needs to change the NN file which contains the weights and structures of each NN.

Comparative analysis of neural architecture reveals distinct thermodynamic fidelity characteristics (Figure 8). ResNet architecture demonstrates superior temperature profile reconstruction, maintaining deviations < 5 K from ERA5 reanalysis throughout the troposphere. In contrast, CNN and DNN architectures exhibit systematic warm biases (5-10 K) between 300–600 hPa, while DNN exhibit warm biases at both North and South pole. Humidity simulations further highlight architectural divergence: while CNN/DNN architectures compress moisture profiles toward lower altitudes (peaking at 850 hPa with about 50% faster moisture decay rates above 500 hPa), ResNet and DNN preserves physically consistent specific humidity gradients up to 300 hPa, a capability enabling enhanced representation of upper-tropospheric moist processes. Wind field simulations demonstrate architectural invariance, indicating dynamical core constraints predominantly govern momentum balance regardless of physics parameterization. These findings indicate that neural network selection significantly influences thermodynamic fidelity which is a critical design consideration for developing ML-based parameterizations.

Neural architecture selection induces substantial discrepancy in precipitation simulations, particularly in tropical convective organization (Figure 9). During boreal summer, the CNN architecture overestimates western Pacific and tropical Indian Ocean precipitation relative to observations, generating an excessively broad ITCZ with spurious drizzle artifacts across subtropical highs. The DNN exhibits systematic 15%–20% underestimation globally (3.06 versus CNN's 5.08 mm/day) while maintaining comparable spatial PCC (0.88 versus CNN's 0.78) to observations (Table 4).

Winter simulations of CNN reveal pronounced biases: precipitation over ITCZ and SPCZ exhibits large (more than 20%) overestimation relative to observations. The DNN's underestimation persists at about 5% magnitude but shows improved spatial pattern alignment with ResNet. The ResNet architecture consistently outperforms other configurations in maintaining a small deviation across seasons. These systematic discrepancies highlight how architectural inductive biases—specifically, the CNN's excessive sensitivity to localized features compared to the DNN's global feature integration—substantially influence precipitation simulations. This underscores the critical need for architecture-specific uncertainty quantification in machine learning-driven climate modelling, as model design disparities directly shape predictive outcomes.

Seasonal precipitation migration patterns reveal distinct architectural sensitivities (Figure 10). While all architectures capture fundamental north–south displacement of tropical precipitation maxima, CNN simulations exhibit 30% greater meridional spread, consistent with documented overestimation of tropical precipitation (Figures 9b,9e). Conversely, DNN systematically underestimates peak precipitation intensities on summer while overestimates the precipitation center of 10° S on winter by about 1 mm/day, a deficiency attributable to its limited capacity in resolving nonlinear moisture-convection

feedback inherent to fully connected architectures. ResNet maintains the closest fidelity to observed seasonal progression (<5% phase error in ITCZ migration timing).

The frequency-intensity distribution of precipitation (Figure 11) reveal neural architectural influences on precipitation distribution characteristics: CNN more than doubled the frequency of light precipitation occurrence compared with conventional GCMs. Conversely, the DNN achieves the closest alignment with observed frequency distributions despite systematically underestimating heavy precipitation (>50 mm/day). This apparent paradox originates from DNN's inherent regularization properties, its fully connected architecture preferentially attenuates extreme convective events while better constraining pervasive light precipitation (1–10 mm/day) that dominates tropical rainfall occurrence (accounting for >78% of events). ResNet demonstrates intermediate performance.

4 Summary and outlook

This study establishes a new ML-physics hybrid modeling framework through seamless integration of neural networks trained on high-resolution GSRM data into the GCM model, achieving stable six-year climate simulations with enhanced process-level fidelity. The major conclusions are given below.

Major achievement. The GRIST-MPS exhibits strong thermodynamic consistency, closely replicating ERA5 vertical profiles of temperature (T bias < 5 K) and specific humidity (q bias < 1.5 g/kg), while improving tropical precipitation—primarily through improved representation of convective–diabatic processes. Key improvements include more accurate ITCZ strength and width, phase-aligned midlatitude storm tracks, and improved precipitation frequency, particularly the improved light rainfall frequency (0.1–10 mm/day). Crucially, the framework preserves long-term numerical stability and accuracy through architectural innovations and an optimized spatiotemporal sampling strategy, all embedded within a workflow built on standardized pressure-level input variables. These results demonstrate that ML–physics integration has the potential to overcome long-standing trade-offs in conventional parameterizations, offering a transformative pathway for next-generation climate modeling. Moreover, leveraging GSRM-driven learning to construct ML–physics hybrid GCMs offers distinct advantages: GSRMs inherently capture multiscale atmospheric interactions without imposing artificial scale separation, while allowing flexible resolution specifications. Furthermore, community-standardized GSRM datasets based on common state variables promote reproducibility and interoperability. We contend that this modeling paradigm paves the way toward unifying GSRM and GCM scales by harnessing the synergy of ML and high-fidelity data, offering a scalable and physically grounded foundation for future Earth system modeling.

Remaining challenges. The current training is limited to an *80-day only* GSRM dataset, future extensions are expected to enhance model generalization and fidelity. One limitation of the present framework is the absence of momentum feedback in the ML architecture, which may lead to systematic biases in upper-tropospheric jet stream positioning (e.g., U bias > 5 m/s at 200 hPa). Additionally, raw GSRM-derived multiscale interactions may require constraints. Despite these limitations, our results demonstrate that GSRM-trained ML–physics suites can achieve simulation stability (over six years) and high physical

fidelity (e.g., ITCZ positional refinement within 2° latitude). This establishes a strong foundation for scalable and physically consistent next-generation multiscale climate modeling paradigms.

Interdisciplinary implications. The ML–physics model introduces a novel computational framework that has interdisciplinary implications. The MPS relies heavily on matrix multiplication, a computational pattern well-suited for optimization techniques (e.g., reduced precision) that align with recent advances in high-performance computing (Chen et al. 2024). In terms of computational efficiency, the current unoptimized GRIST-MPS shows limited advantage over GRIST-CPS, primarily due to the activation of diagnostic modules (which can be optimized), and lower-resolution CPS does not present significant overhead. However, targeted optimizations reveal its inherent scalability advantages on the new Sunway architecture: Duan et al. (2025) successfully deployed an earlier version of the MPS suite on the new Sunway supercomputer, significantly accelerating global 1km GRIST-GSRM. Xu et al. (2025) extended the work of Duan et al. (2025) by integrating the model into a fully coupled earth system model, and further improved the computational performance through code optimizations. This demonstrates that while the baseline MPS performance is constrained by auxiliary computational overhead, its architectural design enables superior acceleration potential when leveraging platform-specific optimizations. The software framework presented can serve as a platform for testing ML-trained physics suites within hybrid AI-Physics GCMs.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant U2242210), the National Youth Talent Project (grant no. 2021), the Startup Foundation for Introducing Talent of NUIST (2025r096), the National Natural Science Foundation of China (42305169) and the Basic Research Fund of CAMS (2023Y001). Editors and reviewers are thanked for their comments and handling of this paper.

Code and Data availability

Frozen model code, including the MPS, a manual, configuration files and input data, training and plotting scripts and plotting data are available at https://doi.org/10.5281/zenodo.15853268 (GRIST-Dev, 2025). GPM data may be downloaded at: https://gpm.nasa.gov/data/directory. ERA5 data may be downloaded at: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.

Author contribution

YW performed ML training and online model integration, with inputs from all authors. All the authors discussed this work and contributed to the final manuscript version.

Competing interests

None.

References

Arakawa, A.: The Cumulus Parameterization Problem: Past, Present, and Future, Journal of Climate, 17, 2493–2525, doi:10.1175/1520-0442(2004)017<2493:RATCPP>2.0.CO;2, 2004.

Arcomano, T., Szunyogh, I., Wikner, A., Pathak, J., Hunt, B. R., and Ott, E.: A Hybrid Approach to Atmospheric Modeling That Combines Machine Learning With a Physics-Based Numerical Model, Journal of Advances in Modeling Earth Systems, 14, e2021MS002712, doi:https://doi.org/10.1029/2021MS002712, 2022.

Bracco, A., Brajard, J., Dijkstra, H. A., Hassanzadeh, P., Lessig, C., and Monteleoni, C.: Machine learning for the physics of climate, Nature Reviews Physics, 7, 6–20, doi:10.1038/s42254-024-00776-3, 2025.

Brenowitz, N. D. and Bretherton, C. S.: Prognostic Validation of a Neural Network Unified Physics Parameterization, Geophysical Research Letters, 45, 6289–6298, doi:10.1029/2018GL078510, 2018.

Brenowitz, N. D. and Bretherton, C. S.: Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining, Journal of Advances in Modeling Earth Systems, 11, 2728–2744, doi:https://doi.org/10.1029/2019MS001711, 2019.

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., Perkins, W. A., Clark, S. K., and Harris, L.: Correcting Coarse-Grid Weather and Climate Models by Machine Learning From Global Storm-Resolving Simulations, Journal of Advances in Modeling Earth Systems, 14, e2021MS002794, doi:https://doi.org/10.1029/2021MS002794, 2022.

Brunet, G., Parsons, D. B., Ivanov, D., Lee, B., Bauer, P., Bernier, N. B., Bouchet, V., Brown, A., Busalacchi, A., Flatter, G. C., Goffer, R., Davies, P., Ebert, B., Gutbrod, K., Hong, S., Kenabatho, P. K., Koppert, H.-J., Lesolle, D., Lynch, A. H., Mahfouf, J.-F., Ogallo, L., Palmer, T., Petty, K., Schulze, D., Shepherd, T. G., Stocker, T. F., Thorpe, A., and Yu, R.: Advancing Weather and Climate Forecasting for Our Changing World, Bulletin of the American Meteorological Society, 104, E909–E927, doi:https://doi.org/10.1175/BAMS-D-21-0262.1, 2023.

Byrne, M. P., Pendergrass, A. G., Rapp, A. D., and Wodzicki, K. R.: Response of the Intertropical Convergence Zone to Climate Change: Location, Width, and Strength, Current Climate Change Reports, 4, 355–370, doi:10.1007/s40641-018-0110-5, 2018.

Chen, G., Wang, W.-C., Yang, S., Wang, Y., Zhang, F., and Wu, K.: A Neural Network-Based Scale-Adaptive Cloud-Fraction Scheme for GCMs, Journal of Advances in Modeling Earth Systems, 15, e2022MS003415, doi:https://doi.org/10.1029/2022MS003415, 2023.

Chen, J., Zhang, M., Zhang, T., Lin, W., and Xue, W.: Stable Simulation of the Community Atmosphere Model Using Machine-Learning Physical Parameterization Trained With Experience Replay, Journal of Advances in Modeling Earth Systems, 17, e2024MS004722, doi:https://doi.org/10.1029/2024MS004722, 2025.

Chen, S., Zhang, Y., Wang, Y., Liu, Z., Li, X., and Xue, W.: Mixed-precision computing in the GRIST dynamical core for weather and climate modelling, Geosci. Model Dev., 17, 6301–6318, doi:10.5194/gmd-17-6301-2024, 2024.

Chen, T., Zhang, Y., Wang, Y., and Yuan, W.: Impact of Lateral Boundary Flows on Regional Convection-Permitting Simulations Over the Tibetan Plateau: A Global-Regional Integrated Modeling Study, Journal of Geophysical Research: Atmospheres, 130, e2024JD042952, doi:https://doi.org/10.1029/2024JD042952, 2025.

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., Watt-Meyer, O., Bretherton, C. S., and Harris, L. M.: Correcting a 200 km Resolution Climate Model in Multiple Climates by Machine Learning From 25 km Resolution Simulations, Journal of Advances in Modeling Earth Systems, 14, e2022MS003219, doi:https://doi.org/10.1029/2022MS003219, 2022.

Duan, X., Zhang, Y., Xu, K., Fu, H., Yang, B., Wang, Y., Han, Y., Chen, S., Zhou, Z., Wang, C., Huang, D., An, H., Ju, X., Huang, H., Liu, Z., Xue, W., Liu, W., Yan, B., Hou, J., Yu, M., Chen, W., Li, J., Jing, Z., Liu, H., and Wu, L.: An AI-Enhanced 1km-Resolution Seamless Global Weather and Climate Model to Achieve Year-Scale Simulation Speed using 34 Million Cores, Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming, Las Vegas, NV, USA, 2025.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, doi:10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C. S., Christensen, H. M., Dagon, K., Gagne, D. J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M. C., Meehl, G. A., Molina, M. J., Monteleoni, C., Mueller, J., Pritchard, M. S., Rolnick, D., Runge, J., Stier, P., Watt-Meyer, O., Weigel, K., Yu, R., and Zanna, L.: Pushing the frontiers in climate modelling and analysis with machine learning, Nature Climate Change, doi:10.1038/s41558-024-02095-y, 2024.

Fu, Z., Zhang, Y., Li, X., and Rong, X.: Intercomparison of Two Model Climates Simulated by a Unified Weather-Climate Model System (GRIST), Part I: Mean State, Climate Dynamics, 2024.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock?, Geophysical Research Letters, 45, 5742–5751, doi:https://doi.org/10.1029/2018GL078202, 2018.

Han, Y., Zhang, G. J., Huang, X., and Wang, Y.: A Moist Physics Parameterization Based on Deep Learning, Journal of Advances in Modeling Earth Systems, 12, e2020MS002076, doi:https://doi.org/10.1029/2020MS002076, 2020.

Han, Y., Zhang, G. J., and Wang, Y.: An Ensemble of Neural Networks for Moist Physics Processes, Its Generalizability and Stable Integration, Journal of Advances in Modeling Earth Systems, 15, e2022MS003508, doi:https://doi.org/10.1029/2022MS003508, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, doi:https://doi.org/10.1002/qj.3803, 2020.

Heuer, H., Schwabe, M., Gentine, P., Giorgetta, M. A., and Eyring, V.: Interpretable Multiscale Machine Learning-Based Parameterizations of Convection for ICON, Journal of Advances in Modeling Earth Systems, 16, e2024MS004398, doi:https://doi.org/10.1029/2024MS004398, 2024.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG final precipitation L3 half hourly 0.1 degree x 0.1 degree V06, Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA, doi:10.5067/GPM/IMERG/3B-HH/06, 2019.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural general circulation models for weather and climate, Nature, doi:10.1038/s41586-024-07744-y, 2024.

Krasnopolsky, V. and Belochitski, A. A.: Using Machine Learning for Model Physics: an Overview, Atmospheric and Oceanic Physics, Atmospheric and Oceanic Physics (physics.ao-ph); Machine Learning (stat.ML), 2020.

Li, X., Zhang, Y., Peng, X., Chu, W., Lin, Y., and Li, J.: Improved Climate Simulation by Using a Double-Plume Convection Scheme in a Global Model, Journal of Geophysical Research: Atmospheres, 127, e2021JD036069, doi:https://doi.org/10.1029/2021JD036069, 2022.

Li, X., Zhang, Y., Peng, X., Zhou, B., Li, J., and Wang, Y.: Intercomparison of the weather and climate physics suites of a unified forecast–climate model system (GRIST-A22.7.28) based on single-column modeling, Geosci. Model Dev., 16, 2975–2993, doi:10.5194/gmd-16-2975-2023, 2023.

Li, X., Chu, W., Zhang, Y., and Wang, Y.: Extending a dry-environment convection parameterization to couple with moist turbulence and a baseline evaluation in the GRIST model, Quarterly Journal of the Royal Meteorological Society, 150, 3368–3384, doi:https://doi.org/10.1002/qj.4763, 2024.

Lin, J., Taotao, Q., Peter, B., Georg, G., J., Z. G., Ping, Z., R., F. S., Hannah, B., and and Han, J.: Atmospheric Convection, Atmosphere-Ocean, 60, 422–476, doi:10.1080/07055900.2022.2082915, 2022.

Miura, H., Suematsu, T., Kawai, Y., Yamagami, Y., Takasuka, D., Takano, Y., Hung, C.-S., Yamazaki, K., Kodama, C., Kajikawa, Y., and Masumoto, Y.: Asymptotic Matching between Weather and Climate Models, Bulletin of the American Meteorological Society, 104, E2308–E2315, doi:https://doi.org/10.1175/BAMS-D-22-0128.1, 2023.

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., and Gentine, P.: Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models With Real-Geography Boundary Conditions, Journal of Advances in Modeling Earth Systems, 13, e2020MS002385, doi:https://doi.org/10.1029/2020MS002385, 2021.

Morcrette, C., Cave, T., Reid, H., da Silva Rodrigues, J., Deveney, T., Kreusser, L., Van Weverberg, K., and Budd, C.: Scale-Aware Parameterization of Cloud Fraction and Condensate for a Global Atmospheric Model Machine-Learned From Coarse-Grained Kilometer-Scale Simulations, Journal of Advances in Modeling Earth Systems, 17, e2024MS004651, doi:https://doi.org/10.1029/2024MS004651, 2025.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, doi:doi:10.1073/pnas.1810286115, 2018.

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., and Düben, P.: Global Cloud-Resolving Models, Curr Clim Change Rep, 5, 172–184, doi:10.1007/s40641-019-00131-0, 2019.

Schneider, T., Behera, S., Boccaletti, G., Deser, C., Emanuel, K., Ferrari, R., Leung, L. R., Lin, N., Müller, T., Navarra, A., Ndiaye, O., Stuart, A., Tribbia, J., and Yamagata, T.: Harnessing AI and computing to advance climate modelling and prediction, Nature Climate Change, 13, 887–889, doi:10.1038/s41558-023-01769-3, 2023.

Wang, X., Han, Y., Xue, W., Yang, G., and Zhang, G. J.: Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes, Geosci. Model Dev., 15, 3923–3940, doi:10.5194/gmd-15-3923-2022, 2022.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., Harris, L., and Bretherton, C. S.: Neural Network Parameterization of Subgrid-Scale Physics From a Realistic Geography Global Storm-Resolving Simulation, Journal of Advances in Modeling Earth Systems, 16, e2023MS003668, doi:https://doi.org/10.1029/2023MS003668, 2024.

Wodzicki, K. R. and Rapp, A. D.: Long-term characterization of the Pacific ITCZ using TRMM, GPCP, and ERA-Interim, Journal of Geophysical Research: Atmospheres, 121, 3153–3170, doi:https://doi.org/10.1002/2015JD024458, 2016.

Xu, K., Yu, M., Chen, Y., Gao, J., Wang, S., Song, J., Duan, X., Wei, J., Yu, J., Liu, H., Jiang, J., Zhang, Y., Lin, P., Wang, T., Wang, P., Zheng, W., Xie, J., Zhang, J., Liu, Z., Jin, X., Wei, J., Chang, Q., Lin, Q., Zhou, Y., Liu, W., Xue, W., Li, Y., Fu, H., Yu, Y., Chi, X., and Wu, L.: Kilometer-Scale AI-Powered and Performance-Portable Earth System Model (AP3ESM) to Achieve Year-Scale Simulation Speed on Heterogeneous Supercomputers, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2025.

Yanai, M., Esbensen, S., and Chu, J.-H.: Determination of Bulk Properties of Tropical Cloud Clusters from Large-Scale Heat and Moisture Budgets, Journal of the Atmospheric Sciences, 30, 611–627, doi:10.1175/1520-0469(1973)030<0611:DOBPOT>2.0.CO;2, 1973.

Yu, R., Zhang, Y., Wang, J., Li, J., Chen, H., Gong, J., and Chen, J.: Recent Progress in Numerical Atmospheric Modeling in China, Advances in Atmospheric Sciences, 36, 938–960, doi:10.1007/s00376-019-8203-1, 2019.

Yuval, J. and O'Gorman, P. A.: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions, Nature Communications, 11, 3295, doi:10.1038/s41467-020-17142-3, 2020.

Yuval, J., O'Gorman, P. A., and Hill, C. N.: Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision, Geophysical Research Letters, 48, e2020GL091363, doi:https://doi.org/10.1029/2020GL091363, 2021.

Zhang, Y. and Chen, H.: Comparing CAM5 and Superparameterized CAM5 Simulations of Summer Precipitation Characteristics over Continental East Asia: Mean State, Frequency-Intensity Relationship, Diurnal Cycle, and Influencing Factors, Journal of Climate, 29, 1067–1089, doi:10.1175/JCLI-D-15-0342.1, 2016.

Zhang, Y., Li, J., Yu, R., Zhang, S., Liu, Z., Huang, J., and Zhou, Y.: A Layer-Averaged Nonhydrostatic Dynamical Framework on an Unstructured Mesh for Global and Regional Atmospheric Modeling: Model Description, Baseline Evaluation, and Sensitivity Exploration, Journal of Advances in Modeling Earth Systems, 11, 1685–1714, doi:https://doi.org/10.1029/2018MS001539, 2019.

Zhang, Y., Li, J., Yu, R., Liu, Z., Zhou, Y., Li, X., and Huang, X.: A Multiscale Dynamical Model in a Dry-Mass Coordinate for Weather and Climate Modeling: Moist Dynamics and Its Coupling to Physics, Monthly Weather Review, 148, 2671–2699, doi:10.1175/mwr-d-19-0305.1, 2020.

Zhang, Y., Yu, R., Li, J., Li, X., Rong, X., Peng, X., and Zhou, Y.: AMIP Simulations of a Global Model for Unified Weather-Climate Forecast: Understanding Precipitation Characteristics and Sensitivity Over East Asia, Journal of Advances in Modeling Earth Systems, 13, e2021MS002592, doi:https://doi.org/10.1029/2021MS002592, 2021.

Zhang, Y., Li, X., Liu, Z., Rong, X., Li, J., Zhou, Y., and Chen, S.: Resolution Sensitivity of the GRIST Nonhydrostatic Model From 120 to 5 km (3.75 km) During the DYAMOND Winter, Earth and Space Science, 9, e2022EA002401, doi:https://doi.org/10.1029/2022EA002401, 2022.

Zhang, Y., Liu, Z., Wang, Y., and Chen, S.: Establishing a limited-area model based on a global model: A consistency study, Quarterly Journal of the Royal Meteorological Society, 150, 4049–4065, doi:https://doi.org/10.1002/qj.4804, 2024.

Zhu, C., Lee, W.-S., Kang, H., and Park, C.-K.: A proper monsoon index for seasonal and interannual variations of the East Asian monsoon, Geophysical Research Letters, 32, doi:https://doi.org/10.1029/2004GL021295, 2005.

Zhu, Y., Zhang, R.-H., Moum, J. N., Wang, F., Li, X., and Li, D.: Physics-informed deep-learning parameterization of ocean vertical mixing improves climate simulations, National Science Review, 9, nwac044, doi:10.1093/nsr/nwac044, 2022.

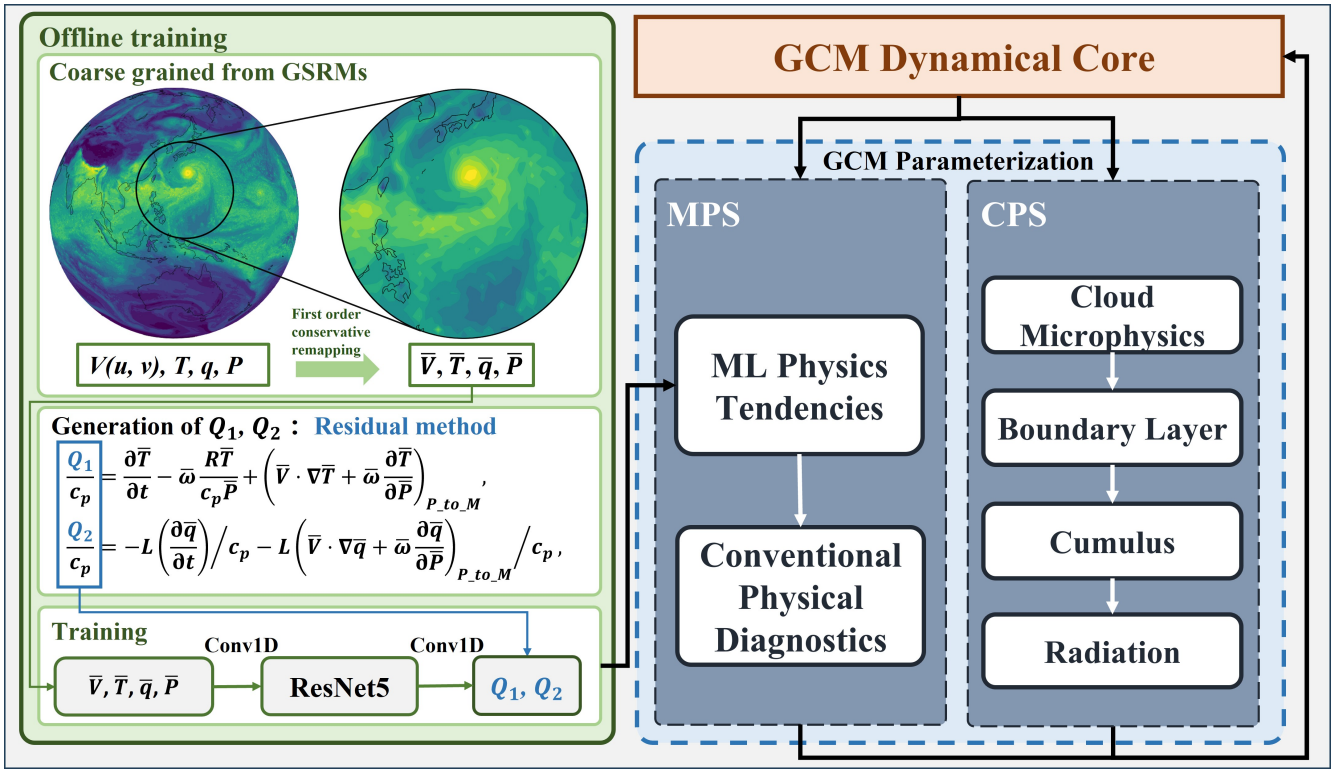


Figure 1. The workflow of offline training of MPS (Machine-Learning Physics Suite; Section 2.1-2.3) and online simulation of the GCM with ML-physics (Section 2.5). In the equations, T represents temperature, q specific humidity, V horizontal wind components (zonal u and meridional v), ω vertical velocity, R the gas constant for dry air, P the atmospheric pressure at all vertical levels, c_p the specific heat at constant pressure, and L latent heat of evaporation or condensation. The notation $\overline{(\cdot)}$ represents the horizontally coarse graining operator, from 5 km to 30 km in this study. The subscript P_to_M represents the conversion from pressure coordinate to the model level, after the calculation of advection terms on the pressure level.

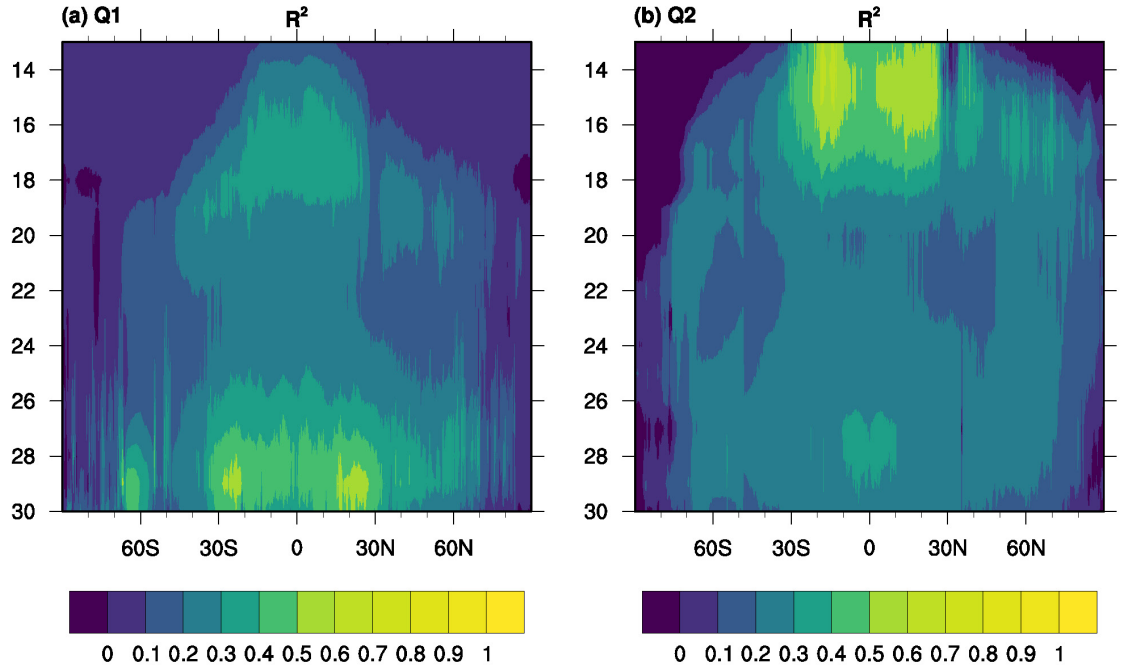
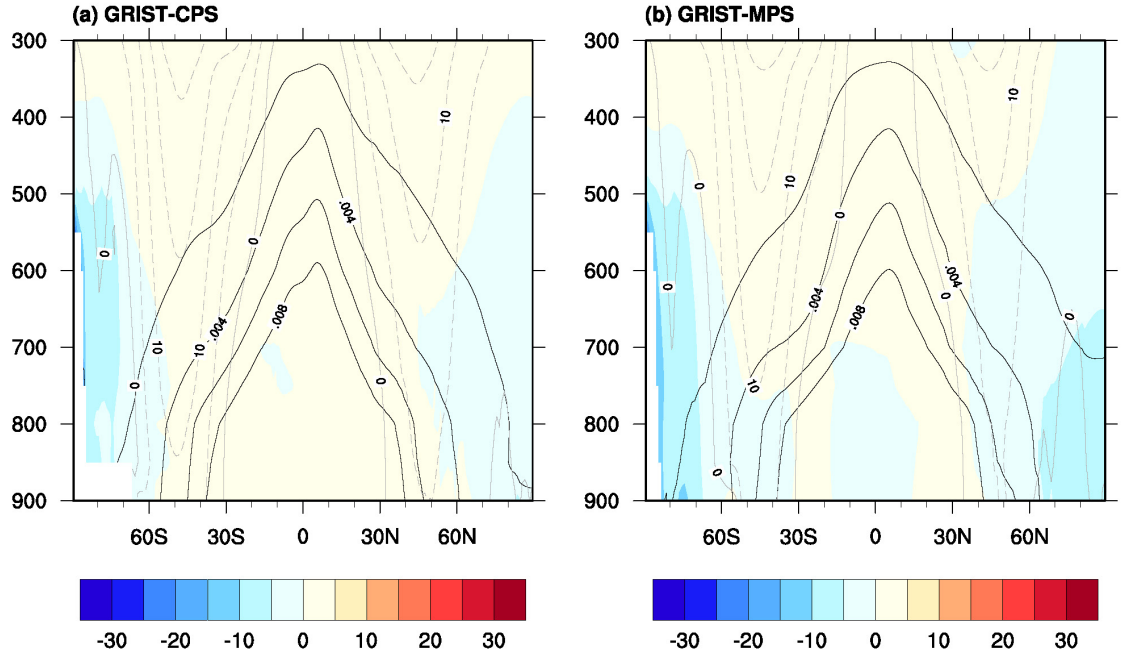
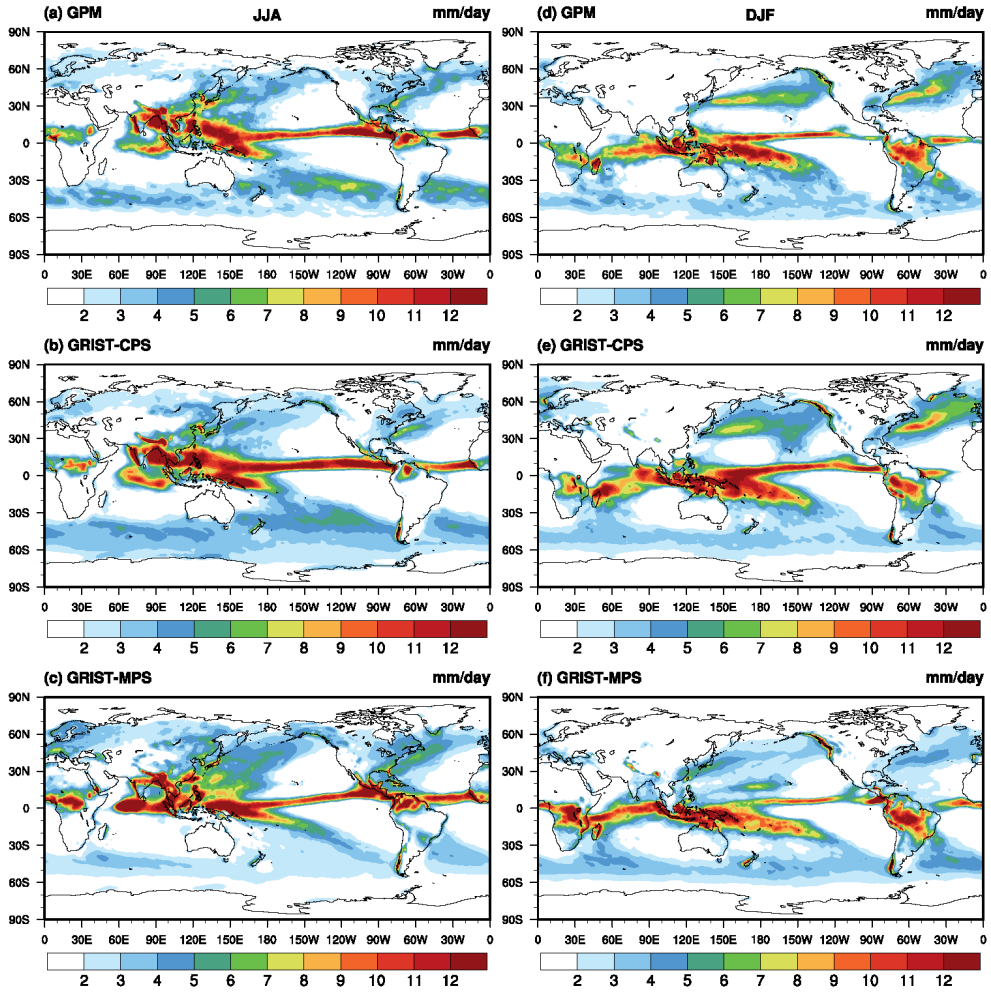


Figure 2. Offline skill of the coefficient of determination (R^2) for Q_1 and Q_2, as functions of latitude and model level.



585

Figure 3. (a) Latitude–pressure cross section of the time averaged zonal mean temperature differences (shaded), climatology specific humidity (black lines) and climatology zonal winds (gray lines) with GRIST-CPS. (b) as in (a) but for GRIST-MPS simulation. The simulation period for all of the models was from 2001 to 2006.



590

Figure 4. The mean precipitation rate (unit: mm/day) averaged from 2001 to 2006 for June–July–August (a, b, c) and December–January–February (d, e, f) by (a, d) GPM, (b, e) GRIST-CPS, and (c, f) GRIST-MPS.

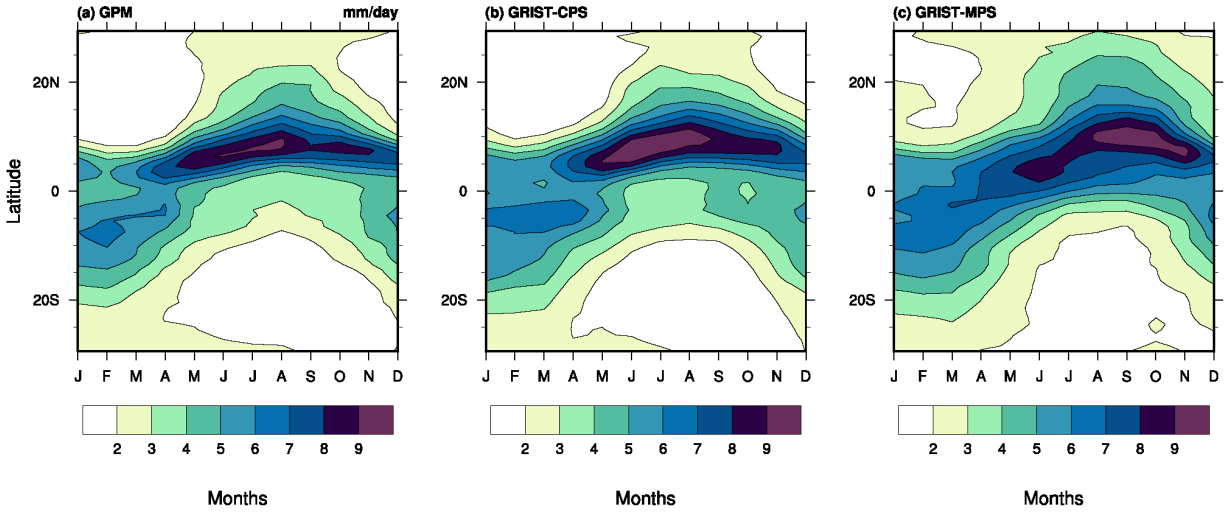


Figure 5. Seasonal evolution of tropical precipitation from 2001-2009 for observation from (a) GPM, (b) GRIST-CPS, and (c) GRIST-MPS (unit: mm/day).

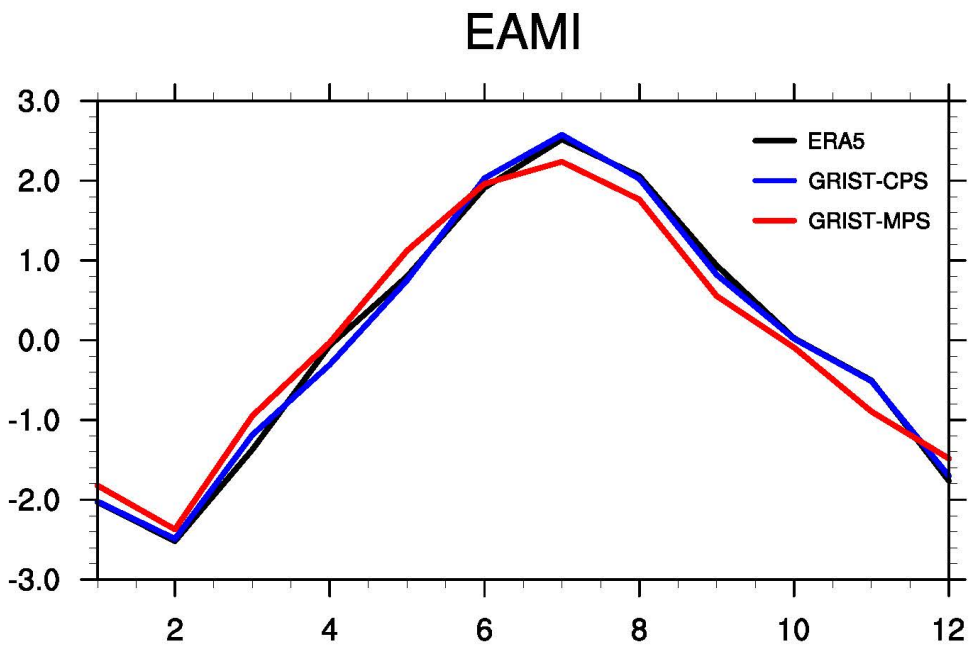


Figure 6: The East Asian Monsoon Index (EAMI) of GPM (black line), GRIST-CPS (blue line) and GRIST-MPS (red line).

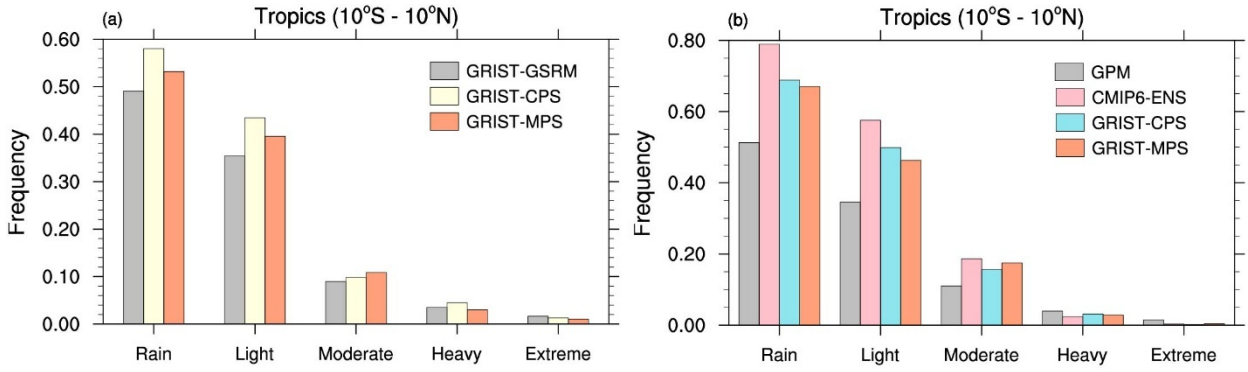


Figure 7. (a) The frequency distributions of tropical daily precipitation corresponding to the 80day-GSRM period, obtained from GSRM (gray boxes), GRIST-CPS (yellow boxes) and GRIST-MPS (orange boxes). (b) As in (a) but for precipitation frequency from 2001-2006, obtained from GPM (gray boxes), 11 CMIP6 models ensemble mean (CMIP6-ENS; pink boxes), GRIST-CPS (blue boxes) and GRIST-MPS (orange boxes).

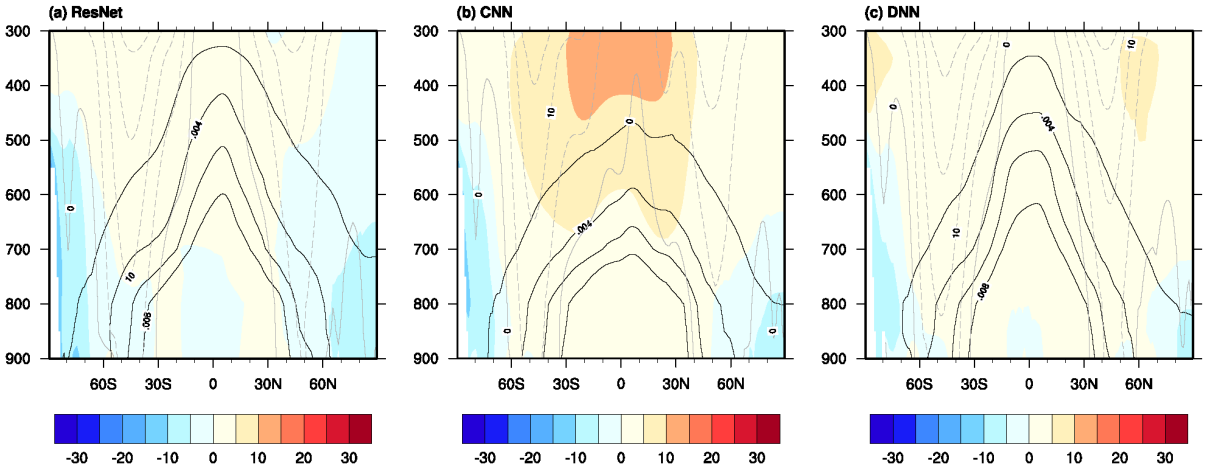
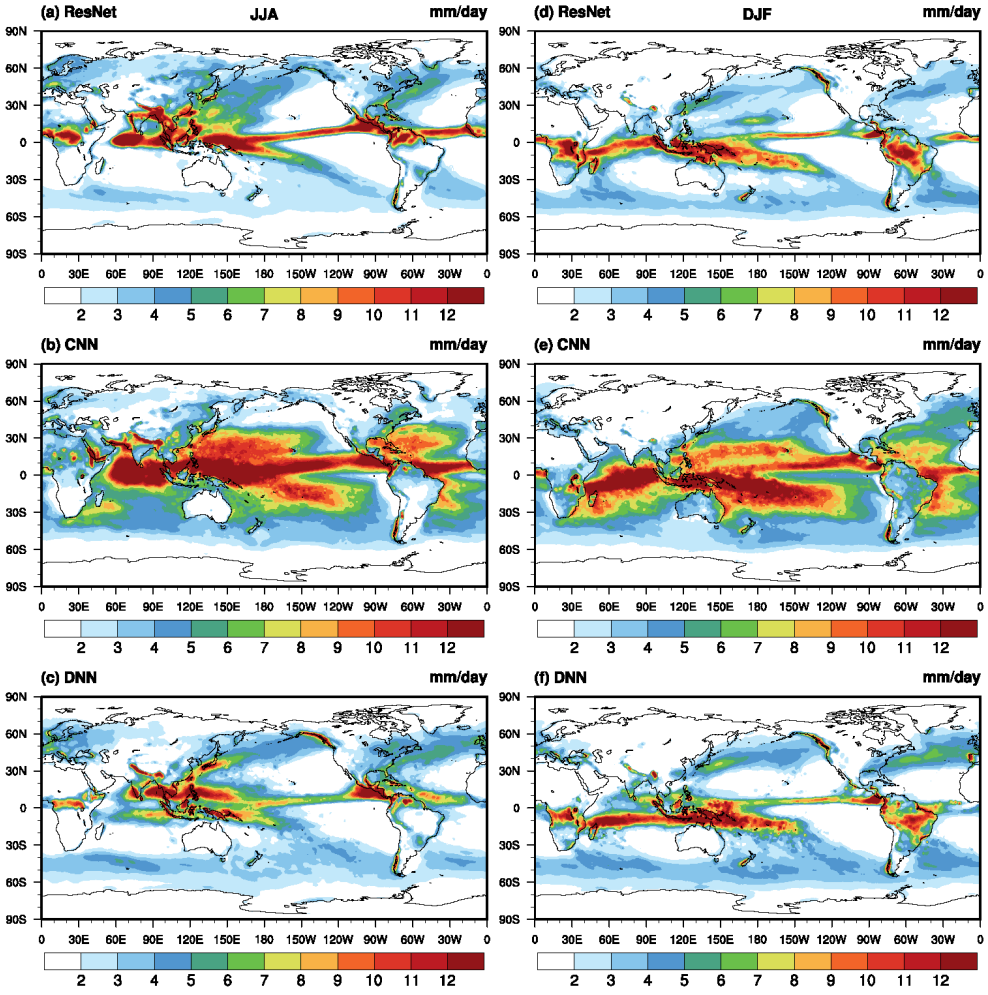


Figure 8. As in Figure 3 but for (a) ResNet, (b) CNN and (c) DNN.



605

Figure 9. As in Figure 3, but for (a) ResNet, (b) CNN, (c) DNN in JJA, (d)-(f) in DJF.

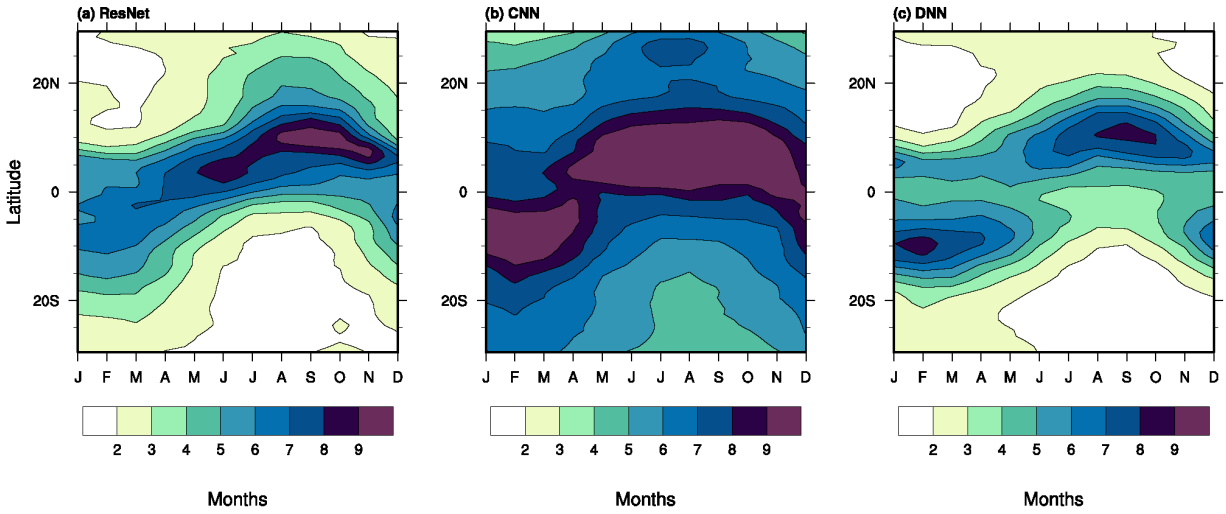
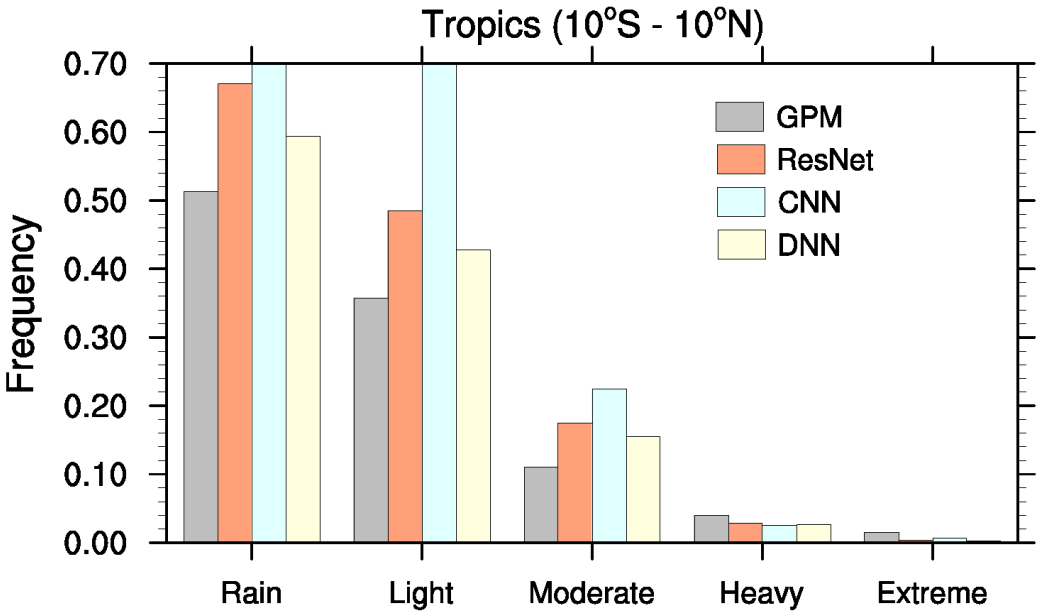


Figure 10. Same as in Figure 5, but for (a) ResNet, (b) CNN, (c) DNN.



610

Figure 11. As in Figure 7b but with CNN (yellow bins) and DNN (blue bins) added.

Table 1 The GSRM and GCM configurations of GRIST for this study.

Set up	Dynamics	Horizontal resolution	Dycore/Tracer/Fast Physics time steps(s)	Square of Smagorinsky Coefficient(C_s^2)	Hyperdiffusion coefficient (m^4/s)
GSRM	Nonhydrostatic	G9B3(5km)	6/30/60	0.005	$1x10^{10}$
GCM	Hydrostatic	G6(120km)	300/600/1200	0.015	$2x10^{14}$

615

Table 2 Selected time periods and climate characteristics.

Experiments	Time period	Oceanic Niño Index	Real-time Multivariate MJO index
1	1-20, Jan, 1998	2.2(El Niño)	0.69 to 1.98
2	1-20, Apr, 2005	0.4(neutral)	2.72 to 3.71
3	10-29, Jul, 2015	-0.4(neutral)	0.17 to 1.05
4	1-20, Oct, 1988	-1.5(La Niña)	0.67 to 2.98

Table 3 The optimal MPS experimental results of each setup.

Experiments	Random points selection	Temporal alignment	resolution	Stable integration time	The RMSE of time-averaged precipitation
EXP1	×	×		3yr	4.81
EXP2	√	×		**6yr**	3.12(3yr)/3.12(6yr)
EXP3	√	√		**6yr**	**2.78**(3yr)/**2.81**(6yr)

620

Table 4 The performance metrics from the six-year simulations for each experiment during summer (June–July–August, JJA) and winter (December–January–February, DJF). The metrics include the spatial pattern correlation (PCC), global-mean precipitation (mm day^{-1}), ITCZ strength (mm day^{-1}), and ITCZ width (degrees). Values in bold and italic indicate the closest match to the observations.

625

Experiments	Season	PCC	Mean	ITCZ strength	ITCZ width
GPM		1	2.85	8.54	7.44
CPS		***0.94***	3.17	10.09	9.60
MPS-ResNet	**JJA**	0.86	3.21	***8.69***	***7.76***
MPS-CNN		0.78	5.27	12.66	17.64
MPS-DNN		0.88	***3.15***	6.50	7.87
GPM		1	2.66	6.48	5.79
CPS		***0.93***	3.05	7.29	7.22
MPS-ResNet	**DJF**	0.89	***3.01***	***6.49***	5.13
MPS-CNN		0.82	5.01	10.33	13.86
MPS-DNN		0.90	3.21	7.42	***5.47***