JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2025MS005231

A Decadal Hybrid GCM Simulation Using Deep-Learning-Based Cloud and Convection Parameterization Generalized to a Warm Climate

Yilun Han[1], Guang J. Zhang[1], Yong Wang[2], and Hui Wan[3]

[1]Scripps Institution of Oceanography, La Jolla, CA, USA, [2]Shanghai Key Laboratory of Ocean-land-atmosphere Boundary Dynamics and Climate Change, Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai, China, [3]Atmospheric, Climate, and Earth Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Abstract A critical challenge for machine-learning (ML) parameterization in global climate models (GCMs) is to achieve stable, accurate simulations under climates not seen during training. Previous studies have demonstrated promising offline performance and year-long online stability in aquaplanet simulations but have encountered difficulties in real geography and under climate warming. Here we report that a GCM with real geography configuration using neural-network-based cloud and convection parameterization, trained exclusively with present-day climate data, successfully performs a stable, decade-long simulation of a warm climate with +4 K sea surface temperature (SST). The neural network (NN) is based on Han et al. (2023, https://doi.org/10.1029/2022ms003508) with additional inputs. The simulation captures the global precipitation distribution, surface temperatures, vertical atmospheric structures, and extreme precipitation very well, closely matching simulations from both the superparameterized CAM (SPCAM) and the conventional CAM5 in the warm climate without accuracy degradation compared to those in the baseline climate. Moreover, it produces a climate response to +4 K SST in atmospheric thermodynamic states and circulations similar to those from SPCAM and CAM5. Prognostic ablation tests on NN input variables show that the NN without convective memory as input suffers from numerical instability, and the NN without considering radiative variables and land fraction as input, or with reduced training samples produce less accurate results. To our knowledge, this is the first time an ML parameterization successfully achieves online extrapolation to a warm climate without using additional warm-climate data for training. It demonstrates the potential of ML-driven parameterizations for credible long-term climate projections.

Plain Language Summary Machine learning (ML) has been used to improve climate models, but current ML algorithms struggle to simulate climates they were not trained on, such as future warming scenarios. We present results from a hybrid global climate model simulation for a warm climate in a real land-ocean geography setting using an advanced ML algorithm to represent clouds and storms that are not resolved by the climate model. The ML model uses a convolutional neural network (NN) with shortcuts and is trained on data from a high-resolution cloud model embedded in a low-resolution global model under current climate conditions. After the trained NN is coupled with the global model, it is able to integrate numerically for 10 years stably in a warmer climate with accurate results. The predicted rainfall, temperatures, and extreme weather reproduce well those from the high-resolution cloud resolving model. Further online attribution tests show that incorporating the storm history and information on land surface and radiation in the NN, as well as increasing the training sample size help greatly in achieving these results. This success in extrapolation to a warm climate brings us a step closer to using ML for trustworthy future climate projections.

Key Points:

- A decade-long hybrid global climate model (GCM) simulation for a warm climate with real geography is achieved using neural network (NN) trained on present-day climate
- The trained NN generalizes well to a warm climate in the online integration, with performance comparable to or better than CAM5
- Climate responses simulated by the hybrid GCM agree well with reference results across different key fields and evaluation metrics

Supporting Information:

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

Correspondence to:

G. J. Zhang,
gzhang@ucsd.edu

Citation:

Han, Y., Zhang, G. J., Wang, Y., & Wan, H. (2025). A decadal hybrid GCM simulation using deep-learning-based cloud and convection parameterization generalized to a warm climate. *Journal of Advances in Modeling Earth Systems*, *17*, e2025MS005231. https://doi.org/10.1029/2025MS005231

Received 13 MAY 2025
Accepted 22 NOV 2025

Author Contributions:

Conceptualization: Guang J. Zhang, Yong Wang
Data curation: Yilun Han
Formal analysis: Yilun Han, Guang J. Zhang, Hui Wan
Funding acquisition: Guang J. Zhang, Hui Wan
Investigation: Yilun Han, Guang J. Zhang, Hui Wan
Methodology: Yilun Han, Guang J. Zhang
Project administration: Guang J. Zhang, Hui Wan

1. Introduction

A growing body of research has applied machine learning (ML) to emulate subgrid-scale convection and cloud processes in global climate models (GCMs) (Irrgang et al., 2021). Earlier studies typically drew on aquaplanet configurations for training data and offline testing, leveraging either superparameterized GCMs which embed a cloud-resolving model (CRM) in each coarse grid cell (Khairoutdinov et al., 2005) or global storm-resolving models that partially resolve convection on 3–4 km meshes (Satoh et al., 2019). Experiments in these idealized "water-worlds" demonstrated that NNs can accurately reproduce convective heating and drying tendencies from the high-resolution models in offline mode (Beucler et al., 2021; Gentine et al., 2018).

Resources: Guang J. Zhang
Software: Yilun Han
Supervision: Guang J. Zhang, Hui Wan
Validation: Yilun Han
Visualization: Yilun Han, Guang J. Zhang
Writing – original draft: Yilun Han, Guang J. Zhang, Hui Wan
Writing – review & editing: Yilun Han, Guang J. Zhang, Yong Wang, Hui Wan

Motivated by these aquaplanet successes, subsequent work extended ML-based parameterizations to real geographic conditions. Han et al. (2020; hereafter H20) proposed a one-dimensional residual convolutional network (ResNet) that emulated superparameterized CAM's (SPCAM's) moist physics processes with very high accuracy by incorporating both current atmospheric states and past convective history as input and validated it in a single-column model. Mooers et al. (2021) found that simpler, fully connected NNs generally gave lower accuracy in such complex land-sea contexts unless extensively hyperparameter-tuned.

Meanwhile, moving from offline validations to online embedding of ML schemes into coarse-resolution GCMs presented new challenges. Earlier attempts often encountered instabilities because NNs lack explicit physical constraints (Brenowitz & Bretherton, 2019; Brenowitz et al., 2020). Stable online integrations were eventually achieved in aquaplanet GCMs using fully connected NNs (Lin et al., 2025; Ott et al., 2020; Rasp et al., 2018; Yuval et al., 2021) or random forests (O'Gorman & Dwyer, 2018; Yuval & O'Gorman, 2020). For real-geography GCMs, X. Wang et al. (2022) emulated moist physics and radiation in SPCAM with multiple deep networks. Under a trial-and-error pipeline, they produced stable 5-year simulations but with notable high-latitude biases. Bretherton et al. (2022) applied ML "nudging tendency corrections" to a NOAA global forecasting model for up to 40 days. Clark et al. (2022) expanded that to multi-year runs under different climate and Kwa et al. (2023) achieved further improvements in land temperature and precipitation accuracy, though circulation biases persisted. Pushing the boundaries further, Watt-Meyer et al. (2024) trained an NN physics directly from a realistic, geographical global storm-resolving simulation and conducted online testing within a coarse-grid GCM. Hu et al. (2024) introduced a comprehensive ML parameterization, including condensate and wind tendencies, supporting 5-year hybrid models based on high-fidelity simulations from the "superparameterized" Energy Exascale Earth System Model (E3SM) using the ClimSim data set (Yu et al., 2023). Additionally, a hybrid NeuralGCM couples a learned NN parameterization with a differentiable dynamical core, producing competitive weather and climate forecasts (Kochkov et al., 2024).

Besides stability, generalizing to unseen climates is another major challenge for NN-based parameterizations, which is central to whether they can be used for future climate projections (Beucler et al., 2024). Offline tests have revealed that most NNs and random forests extrapolate poorly unless input variables are "climate-invariant" (Beucler et al., 2024). Poor generalization is also true in online integration. Even under aquaplanet setup, Rasp et al. (2018) found that a fully connected NN trained using present-day climate incurred large errors at +4 K sea surface temperature (SST) although it performed well at +2 K SST. O'Gorman and Dwyer (2018) reported severe tropical biases in a +6.5 K climate using random forests. Although some studies added high-resolution warm-climate simulation data to the training data set to remedy the poor extrapolation problem (Clark et al., 2022; Rasp et al., 2018), robust out-of-distribution generalization remains elusive. This points to a critical question of whether ML merely learns the statistical relationships between input and output variables of NNs or it can learn the physics behind such relationships. If the latter is the case, a well-trained NN parameterization could perform well in extrapolating to an unseen climate without resorting to training data from other climates.

However, realizing long-term stable simulations in an unseen climate, under real-geographic conditions and with performance comparable to present-day climate simulation, is exceptionally challenging. Only a handful of published works have sustained 5+ years of stable runs under present-day climate for which their NN parameterizations are trained (Clark et al., 2022; Han et al., 2023; Hu et al., 2024; X. Wang et al., 2022), and none exists for an unseen climate, as far as the authors know. For example, NeuralGCM (Kochkov et al., 2024) still encounters NN instabilities in its integrations in both the present-day and warm climates, having a severe drift of global mean temperature in a +4 K SST setting.

An even greater challenge is to obtain a physically consistent climate-change response, namely, a hybrid model that, when driven by present-day SST and warmer SST under global warming, would produce changes of temperature and humidity profiles, large-scale circulation, clouds, and precipitation, etc. that are comparable to those from GCMs with conventional parameterizations. No hybrid framework in published works has yet passed this test; the only attempt, the NeuralGCM, failed, producing unrealistic temperature and zonal-wind responses. Yet such a skill is imperative if hybrid GCMs are to be used for climate change studies. Conventional GCMs suffer from large, persistent uncertainties in their climate-change signals (Stevens & Bony, 2013). By contrast, SPCAM with resolved subgrid cloud physics reproduces realistic cloud and precipitation changes under $4\times CO_2$ forcing (Kooperman et al., 2014, 2016). Another data set, the ClimSim (Yu et al., 2023), also provides a high-

fidelity, superparameterized Energy Exascale Earth System Model (E3SM) simulation data set. If an NN parameterization trained on these data could reproduce such responses, it would represent a major breakthrough.

Building on H20, Han et al. (2023; hereafter H23) demonstrated that their NN extrapolated with high accuracy offline to +4 K SST simulation without including the warm-climate data in the training process. Their offline sensitivity tests highlighted the role of 1D convolutional layers and use of convective memory as inputs in more accurate climate extrapolation, consistent with the well-known property that convolutional networks reduce overfitting via parameter sharing and local feature extraction (Krizhevsky et al., 2012; LeCun et al., 2015). From convective dynamics perspective, the inclusion of convective memory is theoretically justified, as memory is known to play an important role in the evolution of convection (Colin & Sherwood, 2021; Davies et al., 2009; Kuang, 2024). The idea of incorporating convective memory has been adopted in other ML parameterizations since H20 (Hu et al., 2024; Lin et al., 2025). However, no study has directly tested memory-based NN parameterizations in online hybrid warm-climate simulations.

In this paper, we report that we are able to achieve a stable decade-long integration using the hybrid NCAR CAM5 model under +4 K SST with our ML-based parameterization trained on present-day climate data only. To our knowledge, this is the first time a decadal-scale hybrid GCM simulation is achieved using ML-based convection and cloud parameterization in a real-geography GCM for a substantially warmer climate. The remainder of the paper is organized as follows: Section 2 describes our deep learning moist-physics scheme. Section 3 presents the results from the hybrid GCM simulations for both the current climate and a +4 K SST warm climate, including climate mean states and higher-order statistics, and compares them with those from SPCAM and conventional CAM5. Section 4 examines their responses to the +4 K SST climate warming. Section 5 performs prognostic ablation tests to identify the roles of key components of the NN in its successful generalization. Section 6 summarizes the results and discusses the broader implications of these findings.

2. The Neural Network and Online Integrations

2.1. Neural Network Setup

The NN we use in this study is based on the one we developed previously using ResNet and convolutional NN architecture, which we refer to as ResCu (H20; H23), with some modifications. The training data is from the output of the NCAR SPCAM (Khairoutdinov et al., 2005). SPCAM (Khairoutdinov et al., 2005) embeds a 2-D CRM in each grid column of the host CAM5.2 to replace all convection and cloud processes (collectively known as moist physics), including deep and shallow convection schemes, as well as microphysics and macrophysics parameterizations. The embedded 2D CRM has 32 subgrid columns aligned in the zonal direction, with a horizontal resolution of 4 km and a timestep of 20 s. There are 30 vertical levels that are the same as the host model CAM5.2, which has a horizontal resolution of 2.5 deg × 1.9 deg, with a timestep of 30 min. The SPCAM is run for 3 years, with climatologically seasonally varying SST and sea ice extent as the lower boundary conditions (Hurrell et al., 2008), and an interactive land surface model Community Land Model 4.0 (CLM4.0; Oleson et al., 2010).

ResCu is a deep residual convolutional NN designed to emulate the GCM grid-averaged heating (dT/dt) and drying (dq/dt) effects of convection and clouds, cloud water (q_c) and cloud ice (q_i) simulated by the embedded CRM in SPCAM. ResCu's input variables include profiles of temperature (T), specific humidity (q), and their large-scale tendencies from the dynamic core and PBL processes, as well as surface sensible and latent heat fluxes and surface pressure. As the SPCAM's embedded CRM carries convective memory across GCM timesteps, ResCu also incorporates convective memory by taking data from two previous GCM timesteps as input variables, including T, q, large-scale tendencies of T and q, CRM-predicted temperature and moisture tendencies from convection and clouds, cloud water and cloud ice, as well as surface sensible and latent heat fluxes and surface pressure at those previous two timesteps. The output variables from ResCu are temperature tendency, moisture tendency, cloud water and cloud ice contents from convection and cloud processes, with 30 elements for each variable. Surface precipitation is diagnosed by vertically integrating the moisture tendencies predicted by ResCu.

With the above setup, which is the same as in H20 and H23, we make some additional changes. First, we add the solar insolation at the top of the model (SOLIN) and the upward longwave radiation from the surface (LWUP) of the current timestep, as well as land fraction, to the input variables. The radiation-related variables were used in other NN emulators and were thought to help the land-atmosphere coupling (Hu et al., 2024; Yu et al., 2023).

Table 1
Input and Output Variables and Their Normalization Factors

Input variables	Normalization factors	Output variables
$dT(t_{-2},t_{-1},z)$	2.5×10^{-3} (K s^{-1})	$dT(t_0,z)$
$dq(t_{-2},t_{-1},z)$	2×10^{-6} (kg kg^{-1} s^{-1})	$dq(t_0,z)$
$qc(t_{-2},t_{-1},z)$	1.1×10^{-3} (kg kg^{-1})	$qc(t_0,z)$
$qi(t_{-2},t_{-1},z)$	3.5×10^{-4} (kg kg^{-1})	$qi(t_0,z)$
$dT_{ls}(t_{-2},t_{-1},t_0,z)$	2.5×10^{-3} (K s^{-1})	
$dq_{ls}(t_{-2},t_{-1},t_0,z)$	2×10^{-6} (kg kg^{-1} s^{-1})	
$T(t_{-2},t_{-1},t_0,z)$	325 (K)	
$q(t_{-2},t_{-1},t_0,z)$	2.5×10^{-2} (kg kg^{-1})	
$\frac{SHF}{Cp}(t_{-2},t_{-1},t_0,s \rightarrow z)$	6.5×10^{-1} (K s^{-1})	
$\frac{LHF}{Lv}(t_{-2},t_{-1},t_0,s \rightarrow z)$	3×10^{-4}(g kg^{-1} s^{-1})	
$Ps\ (t_{-2},t_{-1},t_0,s \rightarrow z)$	1.05×10^5 (Pa)	
SOLIN $(t_0,s \rightarrow z)$	1,413 (W m^{-2})	
LWUP $(t_0,s \rightarrow z)$	732 (W m^{-2})	
landfrac $(s \rightarrow z)$	1	

Note. The variables with "z" are 30-layer vertical profiles, and those with "$s \rightarrow z$" are scalars copied 30 times to fill the 30-layer vertical profiles. Altogether, this yields 32 channels of 30 elements each for all input variables, and 4 channels for the output. Each data field is normalized by dividing by the corresponding factor listed in the center column.

Together with the land fraction information, we expect them to help improve the simulation of land-atmosphere coupling and the atmospheric state over land. Second, for the structural configuration of the variables in the NN, instead of stacking scalar variables below each profile as in H20 and H23, here we separate each of these scalars from the profiles and let them form their own (constant) profile by replicating those scalars (surface sensible and latent heat fluxes and surface pressure as well as the three newly added scalars SOLIN, LWUP, and land fraction) 30 times. The last step that makes each of the scalars a 30-element vector is purely for the convenience of algebraic matrix operation. This method was also used in Hu et al. (2024). These, together with the vertical profile vectors, yield a total of 32 1-D vectors or channels. Table 1 lists the input and output variables and their associated information. Third, we add the GCM-grid-averaged CRM-grid radiative heating (both shortwave and longwave) (Khairoutdinov et al., 2005), which is calculated outside the CRM, to the large-scale temperature forcing $\left(\frac{\partial T}{\partial t}\right)_{ls}$. This ensures that all external forcings relevant to the CRM convection and clouds in SPCAM are accounted for. All input and output variables are normalized with normalization factors in Table 1 to ensure that they are of order of magnitude O(1) before they are input into the deep NN for training and testing.

The ResCu NN shares the same main architecture as in H23, using 1-D convolutional layers with 128 corresponding filter banks of kernel size 1×3 and batch normalization after each convolutional layer except the last layer (Figure S1 in Supporting Information S1). All ResUnits employ rectified linear activations (ReLUs; Glorot et al., 2011), with no activation function in the output layer. In H23, a 32-layer network improved offline testing accuracy, but in online simulations coupled with the model dynamics, which introduce added uncertainty (Lin et al., 2025), with higher computational cost the benefit is less obvious. Therefore, we revert to the 22-layer structure used in H20, but with everything else remaining the same as in H23.

With the above modifications, we re-trained the neural net. We use only the second simulation year of the SPCAM baseline climate, discarding the first year as spin-up. Due to high horizontal variation in topography over land, we doubled our overall training sample size globally compared to H23 to better capture highly spatially varying precipitation over land. Specifically, we randomly select 2,400 grid cells (instead of 800 grid cells as in H23) out of the total of 13,824 (for a 96×144 grid mesh at $2.5° \times 1.9°$ resolution) and sample the data at a daily frequency. Since the SPCAM used in this study has a timestep of 30 min instead of 20 min as used in H23, this gives roughly 42 million total samples (vs. ∼21 million samples in H23). During the training, we shuffle all samples each epoch, use a mini-batch size of 2,400, and train for 100 epochs with the Adam optimizer (Kingma & Ba, 2014). Examination of the validation loss as a function of training epochs indicates no overfitting (not shown).

Same as in H20 and H23, the NN-predicted heating and drying rates are constrained by the moist static energy (h) conservation, with the Lagrangian multiplier λ set to 1×10^{-7} for an optimal balance between h conservation and prediction accuracy in the NN training, which makes the penalty term from h conservation account for about 6% of the total loss. H20 and H23 showed that in offline tests h conservation is accurately obeyed by the NN when compared against SPCAM. The same degree of accuracy is seen in online simulations compared to SPCAM and CAM5 (Figure S2 in Supporting Information S1) for both the baseline climate and a +4 K SST climate to be presented in Section 3.

2.2. Hybrid Modeling Setups and Simulations

After the training, we implement the NN into CAM5 and carry out a baseline present-day climate simulation and a +4 K SST warm climate simulation, with atmospheric states initialized using the standard present-day (year 2000) climatology (Neale et al., 2012). The +4 K SST for the lower boundary conditions is obtained by adding 4 K uniformly to the present-day monthly mean global climatological SST distribution, an experimental design that has been used in many previous studies (e.g., Beucler et al., 2024; Bretherton et al., 2014; Rasp et al., 2018).

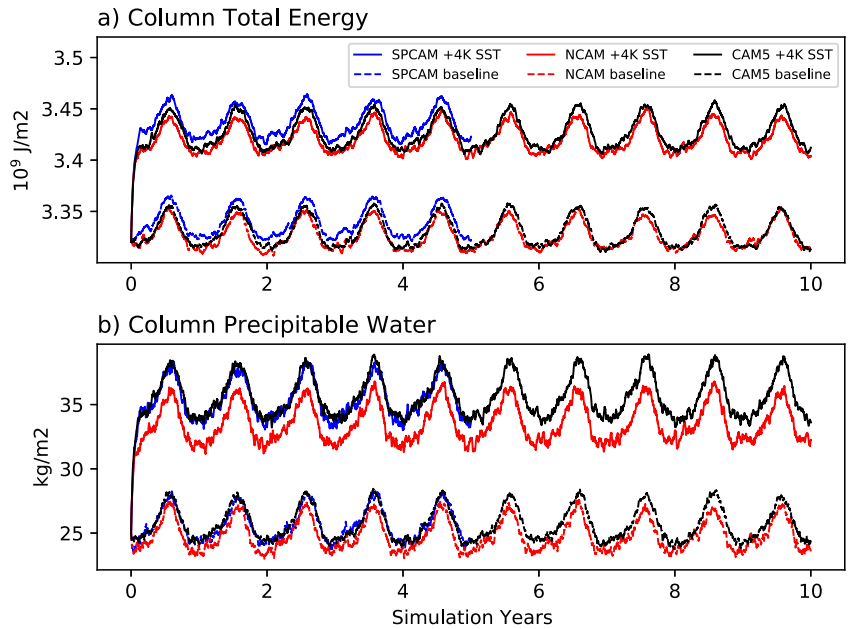

Figure 1. Time evolution of globally averaged, column-integrated (a) total energy and (b) total precipitable water. The data are from simulations using superparameterized CAM (SPCAM), NCAM, and CAM5 under different climate conditions: SPCAM's 5-year simulations for both +4 K sea surface temperature (SST) warm climate (solid blue line) and baseline climate (dashed blue line); NCAM's 10-year simulation for both +4 K SST warm climate (solid red line) and baseline climate (dashed red line); and CAM5's 10-year simulation for both +4 K SST warm climate (solid black line) and baseline climate (dashed black line).

We refer to this setup of the hybrid CAM5 as NCAM, where the NN emulator, ResCu, replaces the moist diabatic heating and drying, vertical transport, and the cloud liquid and ice water contents that would typically be provided by the default conventional parameterization schemes in CAM5 for deep convection (Zhang & McFarlane, 1995), shallow convection (Park & Bretherton, 2009), and cloud microphysics (Morrison & Gettelman, 2008). The conventional cloud parameterization schemes remain in place to supply quantities that the NN does not predict but are needed for the radiative transfer scheme in CAM5, such as cloud liquid and ice number concentrations and cloud fraction. As in H23, in online integration, moist heating and drying tendencies and cloud water and ice contents in previous two time steps needed as part of convective memory in ResCu use NN predicted values, instead of the SPCAM values, which are not available. For the initial two timesteps their values are supplied by the conventional convection and cloud parameterizations.

Using NCAM, a 10-year stable simulation is achieved for both the present-day climate and the +4 K SST warm climate, even though the NN is trained with the current climate condition. For comparison and for examining the response to +4 K SST warming, we also conduct a 10-year simulation each using the standard CAM5 under the same present-day climatological SST and the +4 K SST conditions. Due to the high computational cost, the target SPCAM is integrated for 5 years under each condition. Both CAM5 and SPCAM have been used extensively in climate warming simulations to provide results under different warming scenarios (Bretherton et al., 2014; Kooperman et al., 2014, 2016; Zhou & Khairoutdinov, 2017). Thus, they can serve as a benchmark for assessing how NCAM performs in a warming climate. For computational speed, on 208 Intel CPU cores, NCAM achieves 5.2 simulated years per day (SYPD), SPCAM is more than 10 times slower (0.49 SYPD), and CAM5 is 5 times faster, running at 26 SYPD.

3. Results of +4 K SST Simulation

In this section, we first want to determine whether ResCu-driven NCAM can provide a stable and non-drifting simulation for the intended duration in both warm and baseline climates. Figure 1 shows the time evolution of the global mean column-integrated total energy (sum of internal energy, geopotential, and kinetic energy) and precipitable water from these six simulations. The NCAM is stable for both the current and +4 K warm climates,

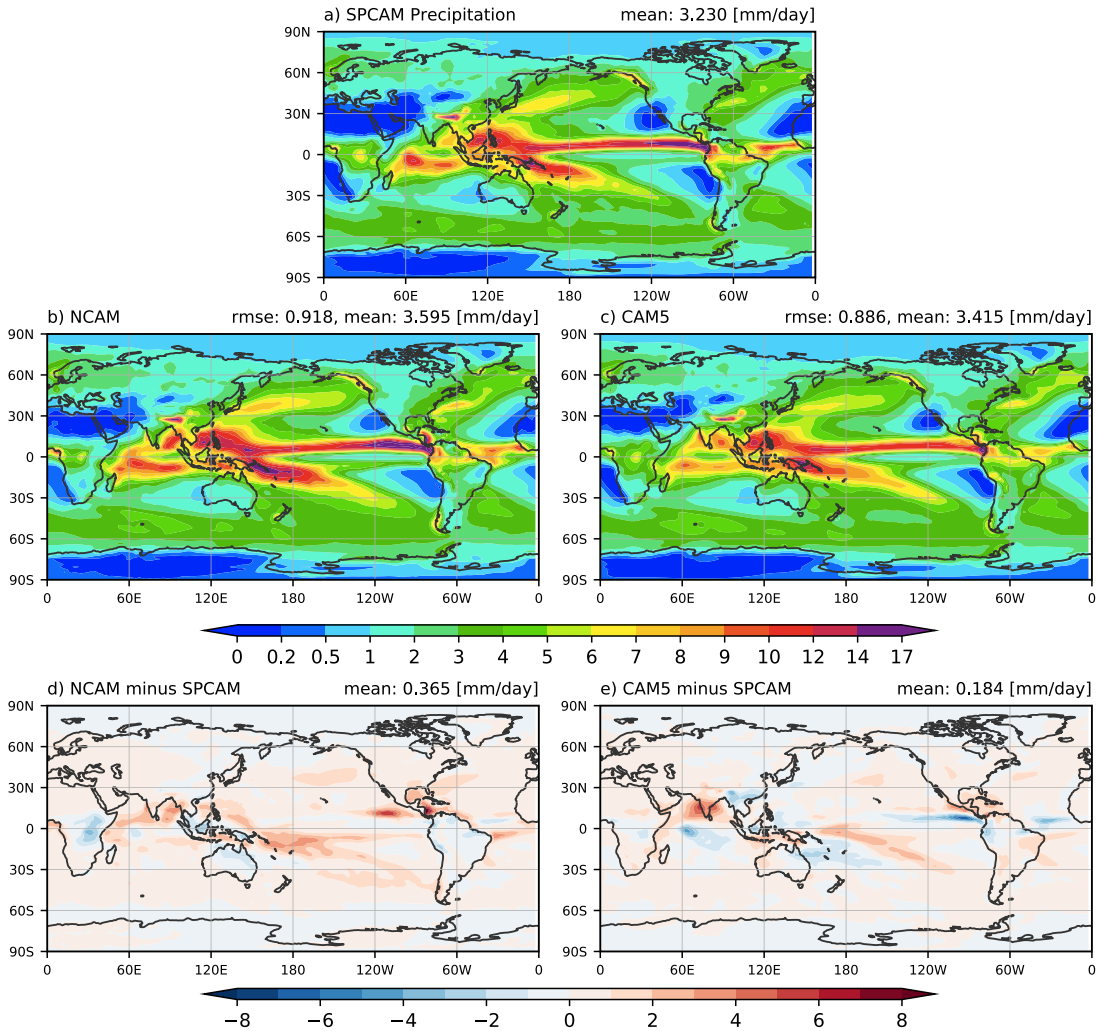

Figure 2. Global distribution of annual mean precipitation for the +4 K sea surface temperature warm climate, depicting: (a) 5-year averages from superparameterized CAM (SPCAM); (b) 10-year averages from NCAM; and (c) 10-year averages from CAM5. Panels (d) and (e) present the annual mean differences relative to SPCAM for NCAM and CAM5, respectively. In panels (a–c), the global mean precipitation appears in the top right corner, while in (d, e) this value shown is the global mean difference. For (b) and (c), the root mean squared error relative to SPCAM is also listed next to the mean value.

with no drift. It even survives the initial shock as it transitions from present-day initial states to an equilibrium +4 K environment, while maintaining physically realistic behavior throughout. Nonetheless, NCAM has some systematic negative biases compared to SPCAM simulations both in total energy and in precipitable water (red lines in Figure 1). Interestingly, the CAM5 simulation is closer to NCAM in total energy but closer to SPCAM in precipitable water. The precipitable water differences between the simulations are amplified in the +4 K SST simulations compared to their current climate counterparts.

3.1. Surface Fields

With this background information, next we compare the 10-year mean surface precipitation from NCAM for the +4 K SST simulation with that from SPCAM simulation in Figure 2, along with the 10-year CAM5 simulation. As mentioned earlier, due to the computational cost we only ran SPCAM for 5 years. However, as will be seen shortly, the difference in the simulation length does not affect the comparison. In the +4 K SST warmer climate, SPCAM simulates strong precipitation in the intertropical convergence zones over tropical oceans compared to current climate (Figures 2a and Bretherton et al., 2014). Generally, the 10-year mean NCAM-simulated precipitation (Figure 2b) captures all major global precipitation systems in the SPCAM warm climate run, with a root-mean-square error (root mean squared error) comparable to that from CAM5 (0.92 vs. 0.89 mm/day). In the

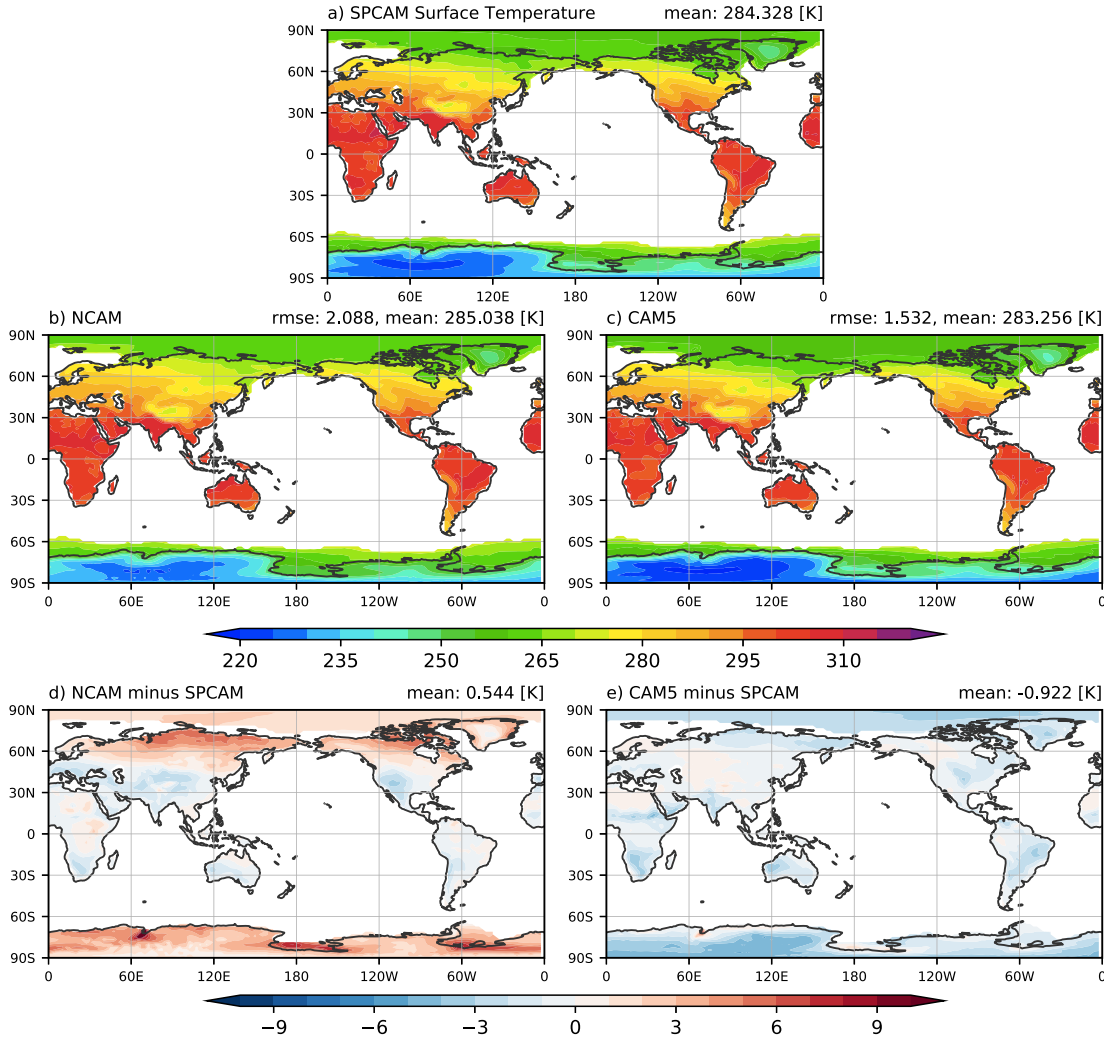

Figure 3. Same as Figure 2 but for surface temperature over land and sea ice in the +4 K sea surface temperature warm climate.

difference plots, NCAM simulates slightly higher tropical precipitation in the South Pacific Convergence Zone (SPCZ), the eastern Intertropical Convergence Zone (ITCZ) and northern Indian Ocean compared to SPCAM, by about 3 mm/day (Figure 2d), whereas CAM5 simulates less precipitation in these regions, by 2 mm/day (Figure 2e). In midlatitudes, NCAM simulates the northern hemisphere Pacific storm track well, with intensity matching those from SPCAM, while the Pacific storm track in CAM5 is slightly weaker, by less than 1 mm/day.

The surface temperatures over land and ice (Figure 3) in the +4 K SST warm climate in the NCAM simulation agree with the SPCAM simulation reasonably well, to within about 5 K (Figure 3d). Most of the differences from SPCAM are in polar regions and are positive. Negative differences are relatively small and appear over sub-tropical and midlatitude land. These differences are comparable to those in Clark et al. (2022), whose NN included warm-climate samples during training to achieve similar performance. On the other hand, the differences between CAM5 and SPCAM surface temperatures are smaller in magnitude (≤3 K) but are systematically negative across all regions (Figure 3e), leading to a larger global mean difference in CAM5 (−0.92 K) than in NCAM (+0.54 K).

To confirm that the different simulation length does not affect the results when comparing SPCAM simulation (5-year long) with NCAM simulation (10-year long), Figure S3 in Supporting Information S1 shows the difference of precipitation and surface temperature between the first 5-year mean of NCAM and the 5-year SPCAM simulation. These differences are almost indistinguishable from those in Figures 2d and 3d, suggesting that

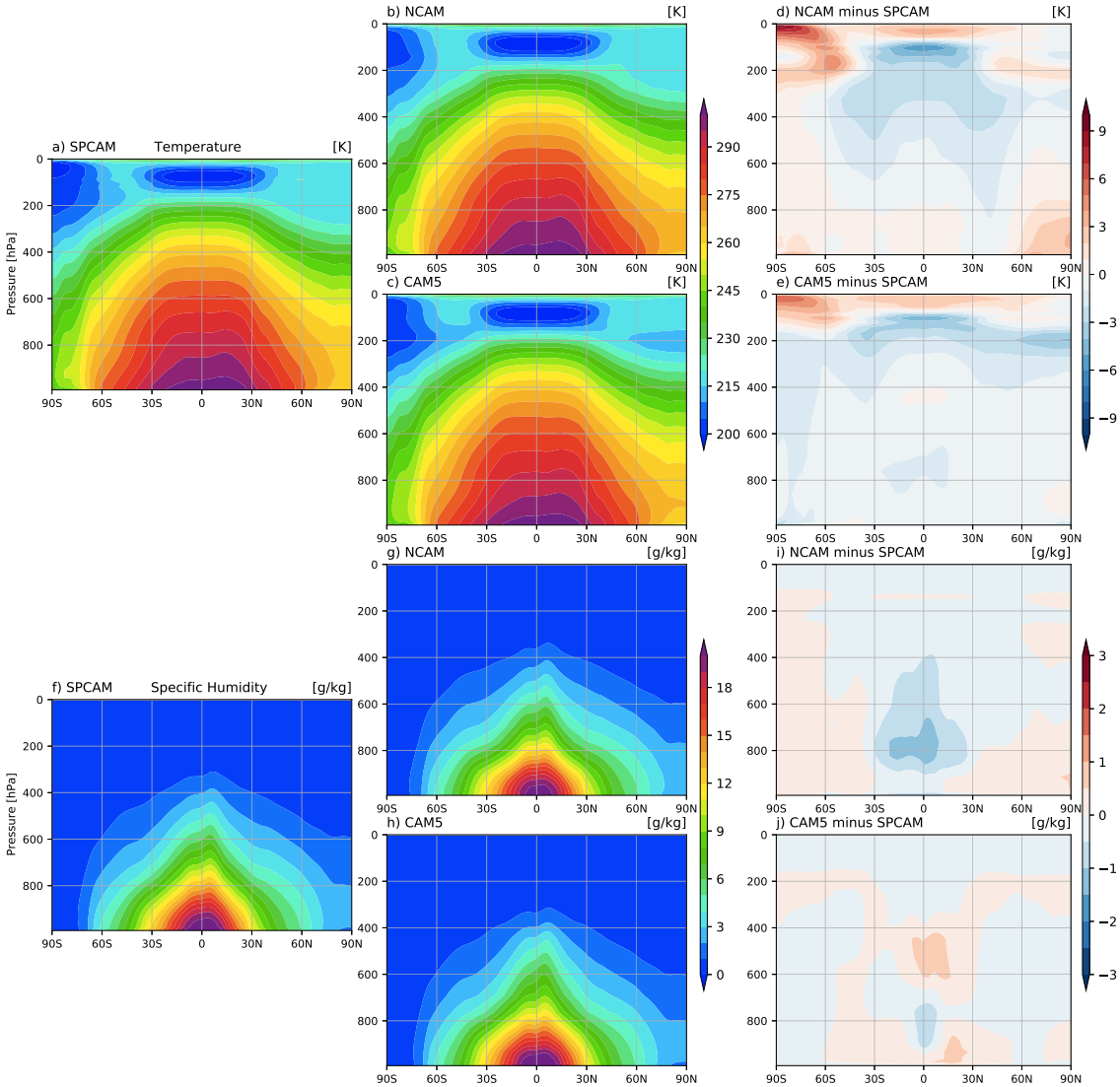

Figure 4. Latitude-pressure cross-sections of the zonal-mean (a–e) temperature and (f–j) specific humidity in the +4 K sea surface temperature warm climate. For temperature, panels (a–c) show 5-year averages from superparameterized CAM (SPCAM), 10-year averages from NCAM, and 10-year averages from CAM5, respectively. Panels (d) and (e) present the corresponding differences relative to SPCAM for NCAM and CAM5. Similarly, for specific humidity, panels (f–h) show 5-year averages from SPCAM, 10-year averages from NCAM, and 10-year averages from CAM5, while panels (i) and (j) show the differences relative to SPCAM for NCAM and CAM5, respectively.

comparing the NCAM simulation with the shorter-length SPCAM simulation is justifiable and does not change the conclusions.

To put the +4 K NCAM simulation in proper context as compared to the current baseline climate, we show in Figure S4 in Supporting Information S1 the 10-yr mean precipitation and in Figure S5 in Supporting Information S1 surface temperature for the current baseline climate from NCAM and CAM5, but 5 years for SPCAM. Again, the simulated precipitation and surface temperature from SPCAM are well captured by both NCAM and CAM5, with the spatial patterns of the differences very similar to those in the +4 K simulation, suggesting that the NN is able to learn the physical relationships across different climates.

3.2. Vertical Structures

After presenting the surface fields, we examine the fundamental atmospheric states of the NCAM simulations in the +4 K warm climate, namely, temperature, specific humidity, relative humidity (RH), and zonal winds.

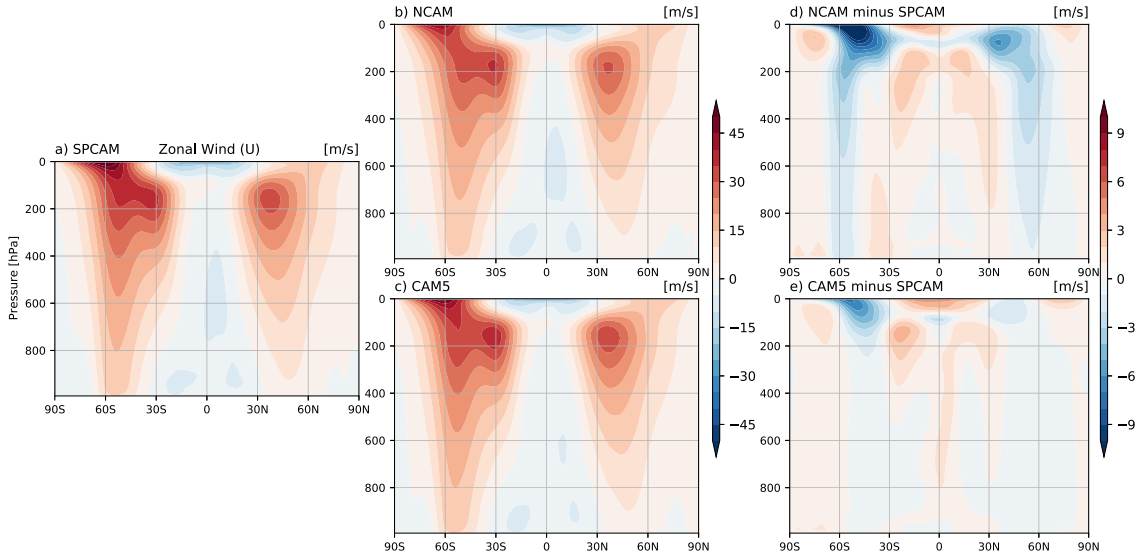

Figure 5. Latitude-pressure cross-sections of the zonal-mean zonal wind (u) in the +4 K sea surface temperature warm climate: (a) 5-year averages from superparameterized CAM (SPCAM), (b) 10-year averages from NCAM, and (c) 10-year averages from CAM5. Panels (d) and (e) show the differences from SPCAM for NCAM and CAM5, respectively.

Figure 4 shows the time-zonal-mean latitude–pressure cross sections of simulated temperature and moisture. Compared to previous studies that have reported large temperature biases at high latitudes in NN-embedded GCMs, both in present-day climate (X. Wang et al., 2022; H23) and warm climate (Rasp et al., 2018), our neural net ResCu, despite being trained on current climate, produces much smaller errors. Compared to SPCAM, biases in the NCAM-simulated temperatures are relatively small (less than 3 K) below 200 hPa, with negative biases in the tropics and midlatitudes, and positive biases in high latitude lower troposphere (Figure 4d). The biases become larger above 200 hPa. In comparison, CAM5 shows negative temperature differences from SPCAM across all regions below 200 hPa and exhibits similar patterns to NCAM above 200 hPa. The moisture biases in NCAM are mostly negative in the lower and middle tropical troposphere, by about −1 g/kg (Figure 4i), which is small by the current standard for hybrid simulations. The differences between CAM5 and SPCAM are small (<1 g/kg) and positive in mid-troposphere but negative in the lower troposphere in the tropics (Figure 4j).

Recent studies have highlighted relative humidity (RH) as a more stringent diagnostic for evaluating moist physics emulators (Beucler et al., 2024; Lin et al., 2025; Watt-Meyer et al., 2024), as small specific humidity and temperature errors at high latitudes and upper levels can translate to large RH biases. RH was also used as a climate-invariant input variable in Beucler et al. (2024) to train their NN. A constant RH hypothesis has been used for climate change since Manabe and Wetherald (1967) and is supported by observations in Douville et al. (2022). Figure S6 in Supporting Information S1 compares RH in SPCAM, NCAM, and CAM5 for both the +4 K SST and baseline simulations. Both SPCAM and CAM5 show high RH in the tropics and high latitudes, with relatively low RH in the subtropics. NCAM captures this spatial structure well. The difference between CAM5 and SPCAM shows a higher RH between 400 and 200 hPa in CAM5, especially in the polar regions in both hemispheres, by up to 30%. The difference between NCAM and SPCAM has a similar pattern, but with a larger magnitude, by up to 50%. In most regions below 400 hPa, the RH biases for CAM5/NCAM versus SPCAM are within ±10%. Comparison of the RH differences between the +4 K SST and baseline simulations shows that despite the large increase in specific humidity from the current climate to the +4 K SST climate (see Figure 8 below in Section 4), RH in NCAM remains largely unchanged or "climate-invariant," as in SPCAM and CAM5 (Figures S6i–S6k in Supporting Information S1). This result again suggests that the physical relationships are learned by our NN.

The zonal wind field is also generally well reproduced by NCAM (Figure 5). Because of the weakened meridional temperature gradients due to the tropical cold biases, the midlatitude westerly jets are weaker and are shifted equatorward in NCAM (Figure 5d). CAM5 shows similar patterns of biases, but with a smaller magnitude (Figure 5e). The same level of performance of NCAM is also seen for the current climate baseline simulation (Figures S6–S8 in Supporting Information S1). When we compare the mean state differences of NCAM and

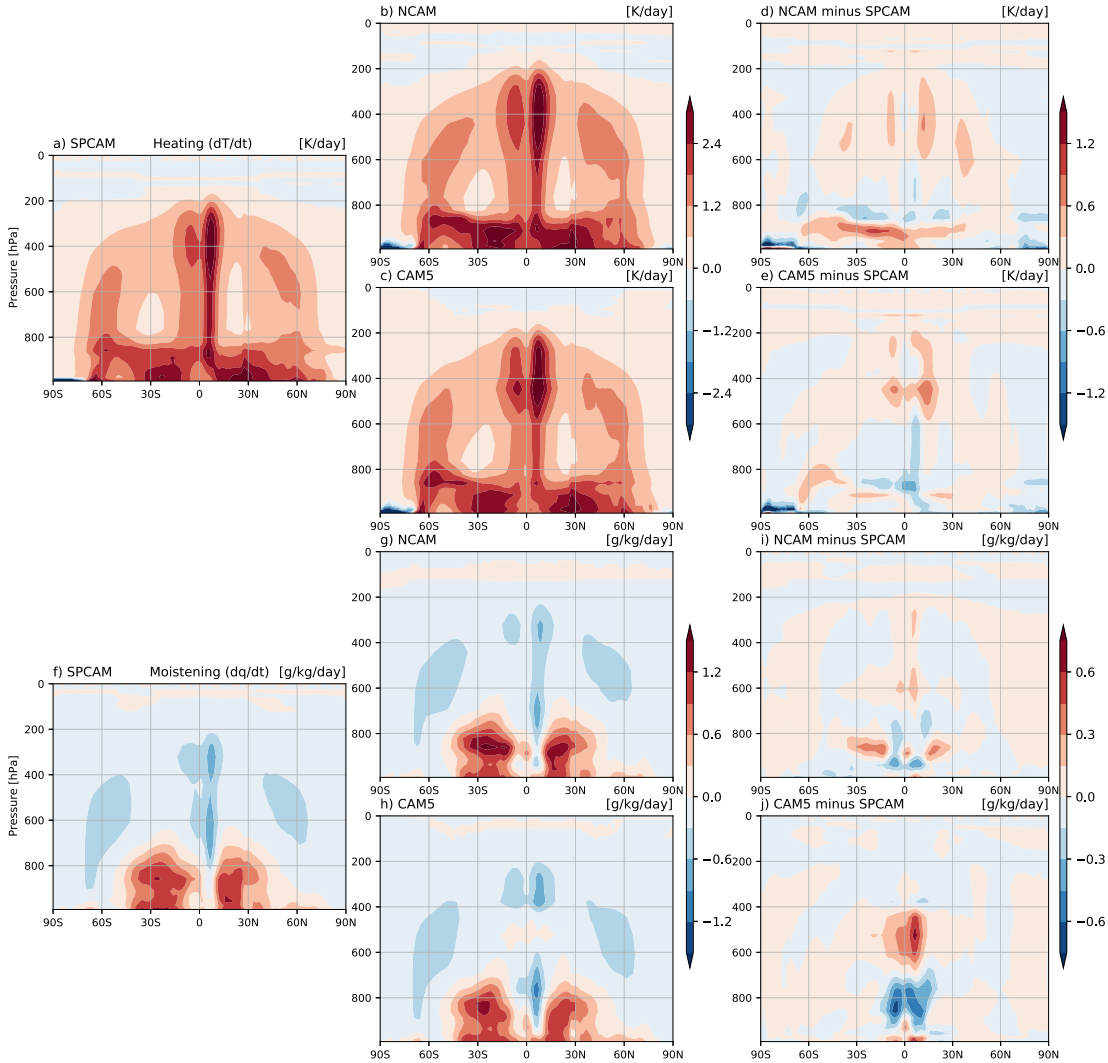

Figure 6. Same as Figure 4 but for the latitude-pressure cross-sections of the zonal-mean (a–e) heating and (f–j) moistening rate in the +4 K sea surface temperature warm climate.

CAM5 from SPCAM in the baseline climate with those in the +4 K SST climate, both NCAM and CAM5 exhibit the same patterns in temperature, specific humidity, RH, and zonal wind. This suggests that ResCu in NCAM can maintain a physically consistent performance comparable to the conventional parameterizations used in CAM5 across climates.

As the neural net predicts temperature and moisture tendencies from subgrid convection and cloud processes, it is natural to compare their simulation in NCAM with that in SPCAM (Figure 6). Again, they agree very well, particularly considering that the NCAM's performance is evaluated in an online simulation spanning 10 years, which is much more difficult to achieve than offline performance. NCAM captures the major heating and drying peaks in the tropics from deep convection, midlatitude heating and drying associated with extratropical cyclones and subtropical heating and moistening from shallow convection and stratocumulus. The heating centers in the NCAM are slightly stronger than those in SPCAM in the mid-troposphere and the boundary layer. Conversely, CAM5 deviates more from SPCAM due to failing to reproduce the deep penetrative drying in the tropics, resulting in two drying centers at low levels and high levels, likely linked to the separation of deep and shallow convection parameterizations, a common issue in GCMs (Zhang & Song, 2009).

The simulated cloud water and cloud ice (Figure S9 in Supporting Information S1) show that NCAM generally reproduces the spatial pattern of cloud water from SPCAM but shows a significant overestimation, especially in

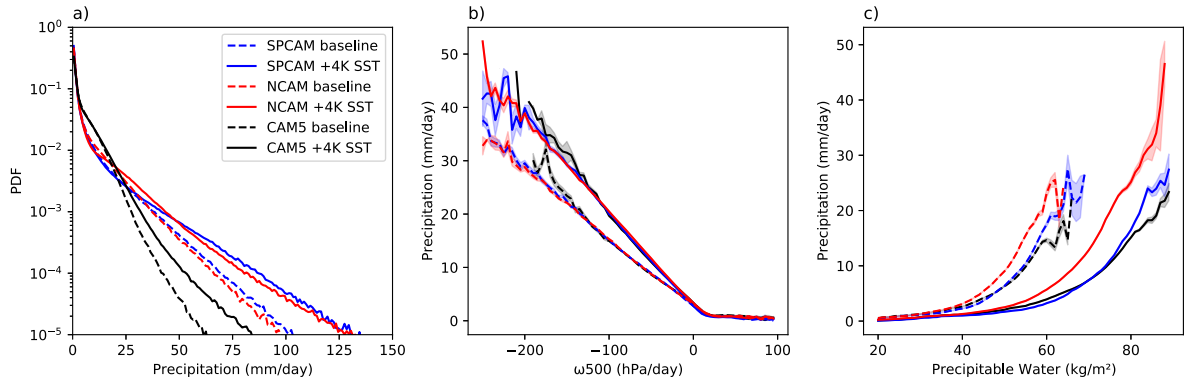

Figure 7. (a) Probability density functions of daily precipitation; (b) mean precipitation as a function of 500 hPa vertical velocity (ω500); and (c) mean precipitation as a function of column-integrated precipitable water vapor. Data are from superparameterized CAM (blue lines), NCAM (red lines), and CAM5 (black lines) simulations under both +4 K sea surface temperature warm climate (solid lines) and baseline climate (dashed lines), averaged over the 20°S–20°N latitude belt. Panel (a) utilizes daily outputs, while panels (b) and (c) analyze monthly outputs. Shaded areas in panels (b) and (c) represent the standard error of the mean for each bin.

mid and high latitudes below 800 hPa. On the other hand, the CAM5-simulated cloud water is much lower in both the tropical/subtropical region and the northern hemisphere polar region compared to SPCAM. It is unclear why NCAM performs relatively poorly for cloud water compared to all other variables examined. The cloud ice is better simulated than cloud water by NCAM compared to SPCAM, with small positive biases in the tropical upper troposphere and negative biases in the southern hemisphere high latitudes. Interestingly, the differences between CAM5 and SPCAM are comparable to those between NCAM and SPCAM.

3.3. Higher-Order Statistics

Not only does NCAM simulate the climate mean states well under +4 K warming, but it also reproduces well the precipitation intensity statistics and its response to the host model's dynamic and thermodynamic fields. Figure 7 shows the precipitation intensity frequency and precipitation rates as functions of 500 hPa vertical velocity (ω500) and column integrated water vapor (precipitable water) in the latitude belt of 20°S–20°N, which were often used to measure the tropical tropospheric large-scale circulation and thermodynamic states (Bony et al., 2004; Bretherton et al., 2004; Peters & Neelin, 2006).

The daily tropical precipitation probability density function for both the baseline and +4 K SST climates (Figure 7a) reveals that NCAM accurately reproduces the precipitation intensity distributions from SPCAM not only for the baseline as also found in X. Wang et al. (2022) and H23, but also for the warm climate, especially for the extreme precipitation at the tail end of the distribution. However, it slightly overestimates the frequency of relatively low-intensity precipitation of ~25 mm/day for both the baseline and the +4 K SST simulation compared to SPCAM. CAM5 suffers from the well-known problem of too much light rain and too little heavy rain (Y. Wang et al., 2016, 2017). This is true for both the baseline and +4 K SST climates.

The relationships between precipitation and ω500 are almost indistinguishable between SPCAM and NCAM for both the baseline and +4 K simulations, except for the very strong ascent regime where the sample size is small. As ω500 becomes more negative (stronger ascent), precipitation rises sharply, underlining the strong link between upward motion and convective rainfall. As expected, under the +4 K warming the same strength of upward motion produces more precipitation because of the increased moisture content in the atmosphere. Turning to the thermodynamic relationships between precipitation and precipitable water (Figure 7c), precipitation follows a power-law relationship with precipitable water (Bretherton et al., 2004; Peters & Neelin, 2006). In the baseline climate (dashed lines), precipitation picks up at lower precipitable water values than in the +4 K warm climate (solid lines), a well-known observation in previous studies due to convection onset at higher SST and tropospheric humidity (Neelin et al., 2009; Sherwood et al., 2010). Again, NCAM captures the relationships between precipitation and precipitable water well for both climates, although the precipitation is overestimated in both cases for given precipitable water, consistent with too dry an atmosphere in the NCAM shown in Figure 1. CAM5 generally simulates similar relationships between precipitation and ω500 to those in SPCAM and NCAM except

for some deviations for ω500 in the range of −150 to −200 hPa/day for some unknown reason. It captures the precipitation-precipitable water relationship in SPCAM better than NCAM, although high precipitation rates are underestimated for both the baseline and +4 K climates.

4. Responses to Climate Warming

With NCAM successfully benchmarked in both present-day and +4 K SST climates, we can ask a question rarely posed to hybrid models: does its neural-network moist physics deliver a physically consistent climate change response? The bar is high. Conventional GCMs, with their subgrid unresolved convection and clouds parameterized, have long shown large, model-to-model spreads in cloud radiative forcing and precipitation changes under warming (Knutti & Sedláček, 2013; Stevens & Bony, 2013). Cloud-resolving approaches, GCRMs and SPCAM, reduce those uncertainties and capture robust signals such as "wet-get-wetter, dry-get-drier" (Bretherton, 2015; Kooperman et al., 2016). Yet their computational cost precludes century-scale ensembles, motivating efforts to train NNs on CRM output and then run GCMs at a fraction of the expense (Gentine et al., 2018; Rasp et al., 2018).

Previous research by O'Gorman and Dwyer (2018) encountered unrealistic temperature and precipitation responses when applying their baseline climate-trained random forest parameterization in a warm climate. A more recent study, NeuralGCM, also attempted to simulate climate change responses using its NN parameterization trained only on baseline climate. In AMIP runs with uniform SST increases, it remained stable and reasonable at +1 and +2 K, but at +4 K it developed pronounced climate drift, with global-mean surface temperature drifting back toward its control value in the current baseline climate (Kochkov et al., 2024). Hence, despite NeuralGCM's skills at weather-scale prediction and historical-climate reproduction, it failed to generalize to a much warmer world.

By contrast, our neural net ResCu-embedded NCAM simulation demonstrates excellent generalizability, as evidenced by the fact that the zonal mean annual mean responses to +4 K SST forcing simulated by NCAM closely resemble the responses simulated by SPCAM and CAM5 for key atmospheric quantities including temperature, specific humidity, and zonal wind (Figure 8), whereas the previously tested NeuralGCM exhibited substantial deviations in temperature and zonal wind responses compared to conventional GCMs like CESM (Kochkov et al., 2024). Notably, NCAM exhibits no drift in global mean total energy and precipitable water throughout the entire 10-year integration of the +4 K SST climate (Figure 1). NCAM reproduces the well-known observed warming features, including the upper-tropospheric amplification (Figures 8a–8c) (Fu et al., 2011) and the poleward shift of the southern hemisphere jet (Figures 8g–8i) (Vallis et al., 2015).

Changes in mean and extreme precipitation are crucial indicators of climate responses (Donat et al., 2016), especially since identifying regions likely to become wetter or drier, and quantifying future extreme rainfall hazards remain challenging due to large uncertainties inherent in traditional GCMs. In Figure 9a, by explicitly resolving convection, SPCAM simulates wide spread precipitation intensification along the ITCZ, SPCZ, and monsoon corridors. Conversely, reduced precipitation is simulated over land areas and subtropical oceanic subsidence zones, particularly in Africa and the eastern Pacific Ocean, due to decreased drizzle. These compensating effects yield only modest global mean precipitation increases (+0.4 mm/day). NCAM closely replicates SPCAM's global precipitation response pattern but slightly overestimates the positive precipitation response along the ITCZ and underestimates the negative response over land areas and oceanic subsidence zones (Figure 9b). CAM5, on the other hand, shows a much weaker precipitation response (Figure 9c).

Next, we examine the response of extreme precipitation, which is defined as daily mean precipitation exceeding the 95th percentile at each grid point over the entire simulation period in the baseline climate (R95; Y. Wang et al., 2017). We use the average of these conditionally sampled data at a given location to measure the average intensity of extreme precipitation. For the frequency of occurrence, by definition, it is globally uniformly 5% for the baseline climate. For the +4 K SST simulation, we use the same extreme precipitation rate threshold at each grid point as determined in the baseline climate and compute the average intensity and occurrence frequency similarly. We first examine the baseline simulations (Figures 10a–10c). SPCAM simulates intense precipitation across the ITCZ, SPCZ, monsoon regions, and over tropical land areas including Africa, the Amazon, and central America (Figure 10a), consistent with its known strengths for simulating heavy precipitation (Li et al., 2012; Zhou & Khairoutdinov, 2017). NCAM reproduces the spatial pattern and intensity of the SPCAM simulation very well, including intense rainfall events over land regions (Figure 10b). CAM5 significantly underestimates the

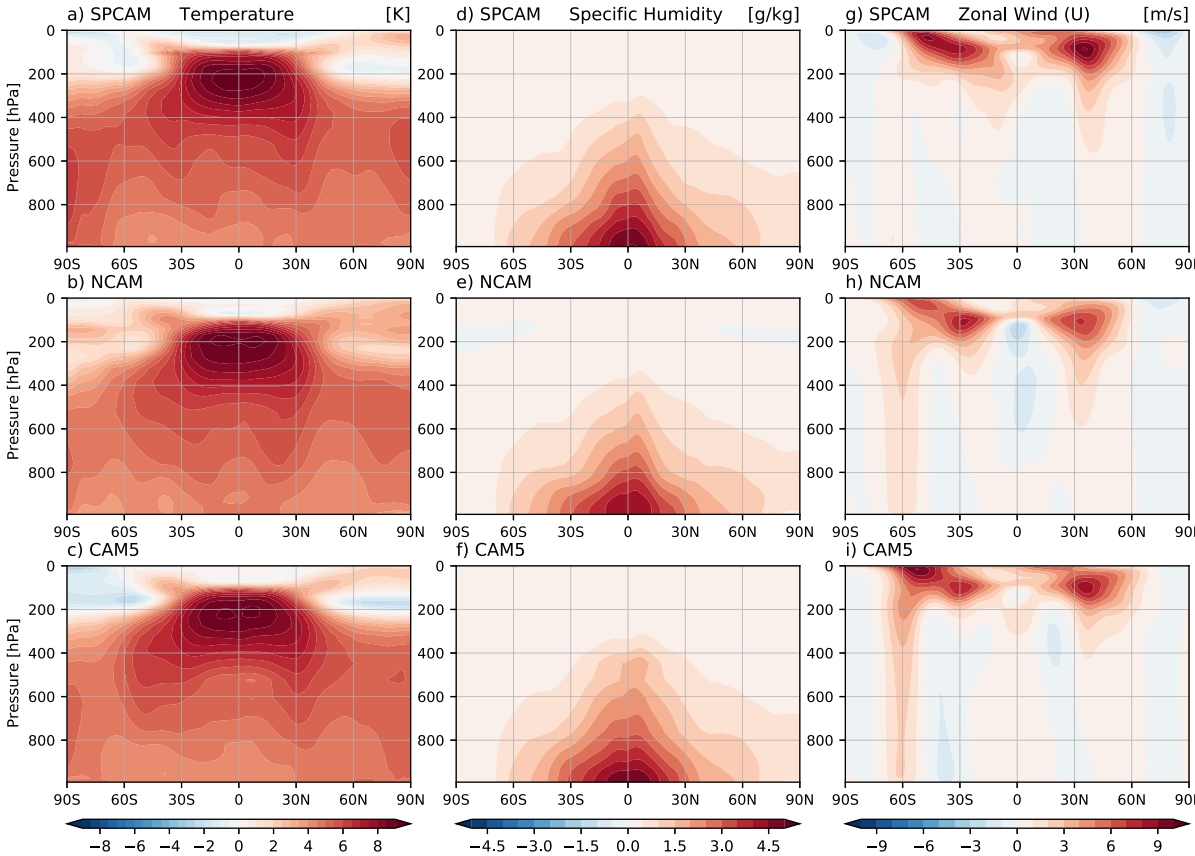

Figure 8. Latitude-pressure cross-section of zonal mean changes of (a–c) temperature, (d–f) specific humidity, and (g–i) zonal wind (u) from the baseline climate to the +4 K sea surface temperature warm climate: (a, d, g) superparameterized CAM; (b, e, h) NCAM, and (c, f, i) CAM5.

mean extreme precipitation intensity, less than half that simulated by NCAM and SPCAM (Figure 10c). The global mean extreme precipitation intensity is 19.5, 19.7, and 15.4 mm/day, respectively, for SPCAM, NCAM, and CAM5. Under the +4 K SST warming, SPCAM shows a sharp increase in heavy rain rates, with the mean rainfall on R95 days increasing by 2.7 mm day^{-1} (Figure 10d). NCAM largely replicates this spatial and global mean response, capturing the enhanced R95 intensity along the ITCZ-SPCZ-monsoon belt (Figure 10e). The primary discrepancy arises in the central-to-eastern equatorial Pacific and equatorial Indian Ocean, where NCAM underestimates the increase in R95 intensity. CAM5, on the other hand, underestimates the intensification of heavy rainfall, exhibiting a greatly weakened increase in R95 intensity in all corresponding regions in SPCAM and NCAM (Figure 10f). The frequency of occurrence of extreme precipitation with respect to the baseline is increased in the ITCZ, SPCZ and high latitude regions, but decreased in subtropical land and oceans in all three simulations. The magnitude of the frequency changes is comparable among the three simulations although the global mean for SPCAM is somewhat smaller (1.47%) compared to NCAM (2.01%) and CAM5 (2.07%).

In summary, NCAM reproduces most of SPCAM's physically grounded state and circulation responses to a uniform +4 K SST warming, including the upper tropospheric warming and the poleward shift of the southern hemisphere jet seen in SPCAM and CAM5. For precipitation, NCAM is able to reproduce some of SPCAM's hallmark precipitation response patterns: heavier rainfall-rate increases along the ITCZ-SPCZ-monsoon belt, accompanied by a large jump in the mean rain on the R95. CAM5, known to produce convection too easily, inflates extreme-precipitation-day counts but fails to intensify them as much, as noted in previous studies (e.g., Kooperman et al., 2014).

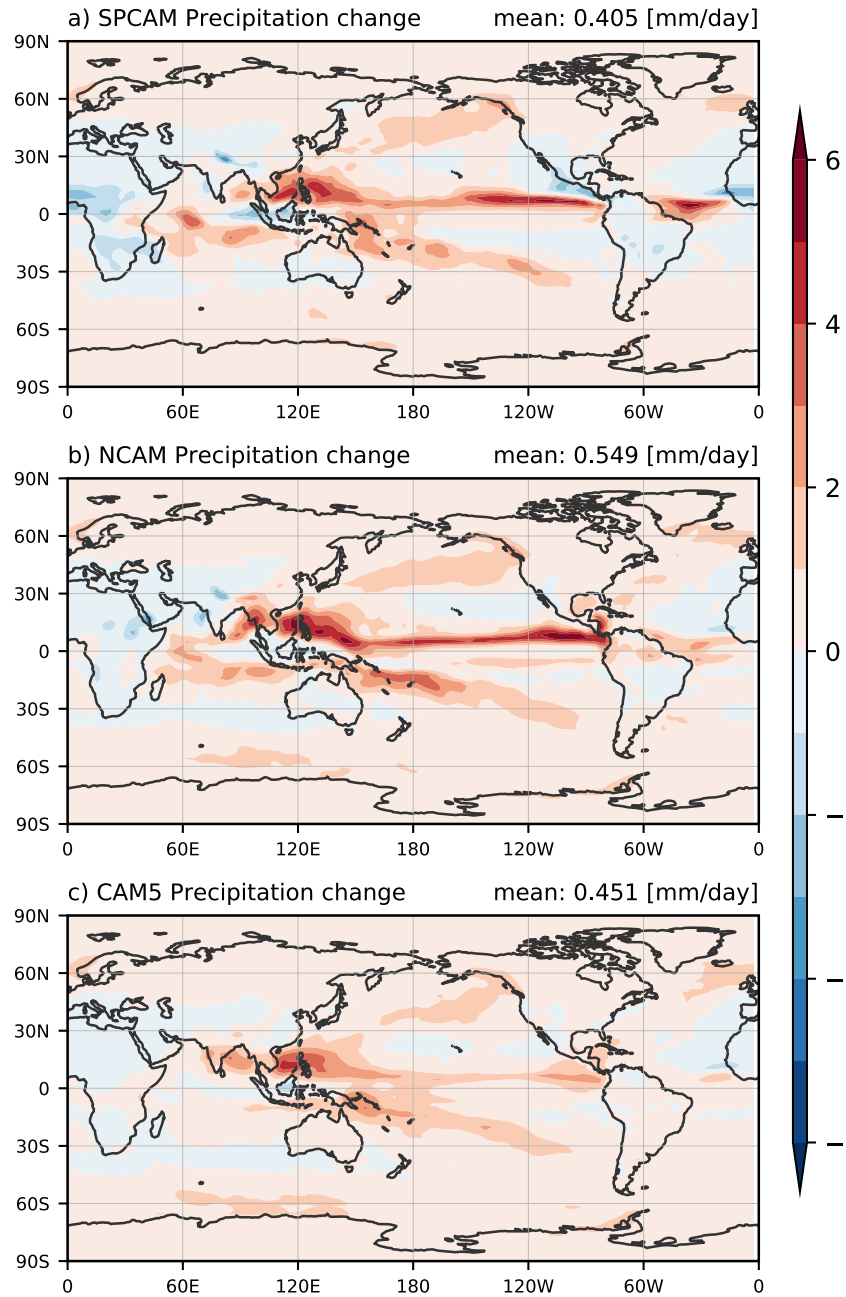

Figure 9. Global distribution of the annual mean precipitation changes from the baseline to the +4 K sea surface temperature climates for (a) superparameterized CAM, (b) NCAM, and (c) CAM5.

5. Prognostic Ablation Sensitive Tests

To further investigate the role of the factors in contributing to the successful generalization of the NN to a warm climate in the stable decade-long simulation, we conduct additional ablation tests in this section. First, in H23 we showed that convective memory played the most important role in our NN's (ResCu) generalizability to warm climate in offline tests. Here we test it online to confirm its role by removing convective-memory-related inputs in the NN training and denote the online simulation NCAM-NoMem. Second, compared to H23 we included radiative variables and land fraction in the NN and its online simulations presented in the previous sections, so that it can better represent the land-atmosphere coupling. To isolate their roles, we exclude them in the NN training and denote the online simulation NCAM-NoRL. Third, compared to H23, we doubled the sample size of the

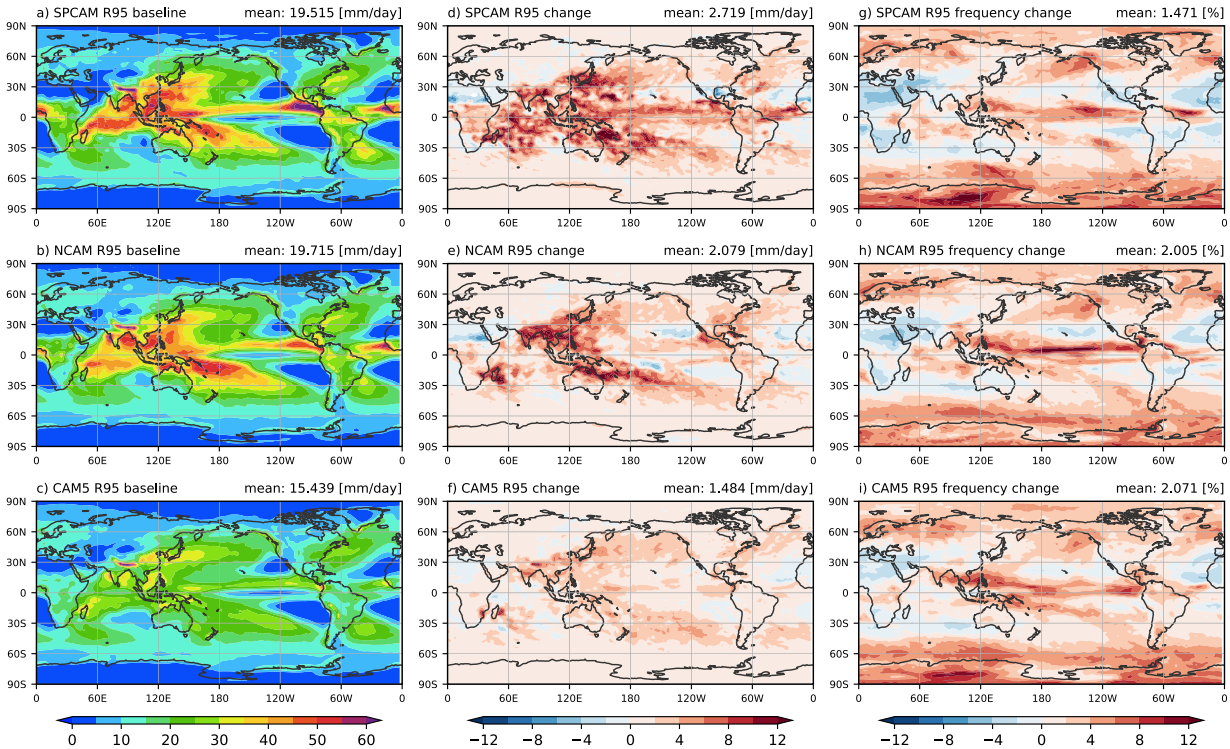

Figure 10. Global distribution for (a–c) annual mean intensity of extreme events that exceed the 95th-percentile threshold in the baseline climate, (d–f) the change in that intensity between the baseline and the +4 K sea surface temperature climate, and (g–i) the corresponding change in the frequency of such extreme events. The domain-averaged value for each field is printed at the upper-right corner of its panel.

training data, so that the NN can better capture the heterogeneity due to topography. To isolate the effect of sample size increase in the training data, we train the NN on only half of the training data by randomly selecting 50% of the samples used in the default NN in this study and denote the simulation NCAM-Half. In each of the test simulations, except the specified changes all else are identical to the default NN. The simulations are conducted for both the baseline current climate and the +4 K SST setups.

The NCAM-NoMem simulation becomes unstable quickly in both the baseline and warm climates. It crashes after 28 days in the baseline climate and after 10 days in the +4 K SST climate. As shown in Figure 11, the precipitable water in the baseline simulation rises quickly after the first few days and deviates greatly from that in SPCAM before the simulation eventually crashes on day 28. For the +4 K SST simulation, the same rapid rise and deviation in precipitable water from SPCAM are seen, and the simulation crashes even sooner, on day 10. A similarly quick rise is observed in column-integrated total energy before its crash in both climates (not shown). The NCAM-NoMem simulations clearly demonstrate that convective memory plays an essential role in stabilizing the hybrid model integration. The NCAM-NoRL simulations are stable throughout the decade, but produce a much drier atmosphere compared to SPCAM in both climates. This indicates that including radiative variables helps improve the accuracy of the model simulations. The NCAM-Half simulations track the SPCAM counterparts much more closely compared to NCAM-NoRL. Comparing to Figure 1b, not much simulation degradation is seen for global mean precipitable water when sample size is reduced. In the rest, we will focus on the +4 K SST simulation.

In the +4 K SST climate, NCAM-NoRL, lacking TOA and surface radiation and land-fraction inputs, performs relatively poorly in reproducing the land-sea precipitation contrast. It substantially overestimates tropical oceanic rainfall and underestimates rainfall over land in Africa and the Amazon, yielding 0.649 mm day^{-1} more global-mean precipitation (Figure 12a). With these inputs restored, but trained on only half of the data, NCAM-Half attains a performance similar to the fully trained NCAM, with somewhat larger (both positive and negative) regional biases than in NCAM, but markedly better than in NCAM-NoRL (Figure 12b). For surface temperature (Figures 12c and 12d), NCAM-NoRL performs reasonably well overall, with only localized warm biases in parts

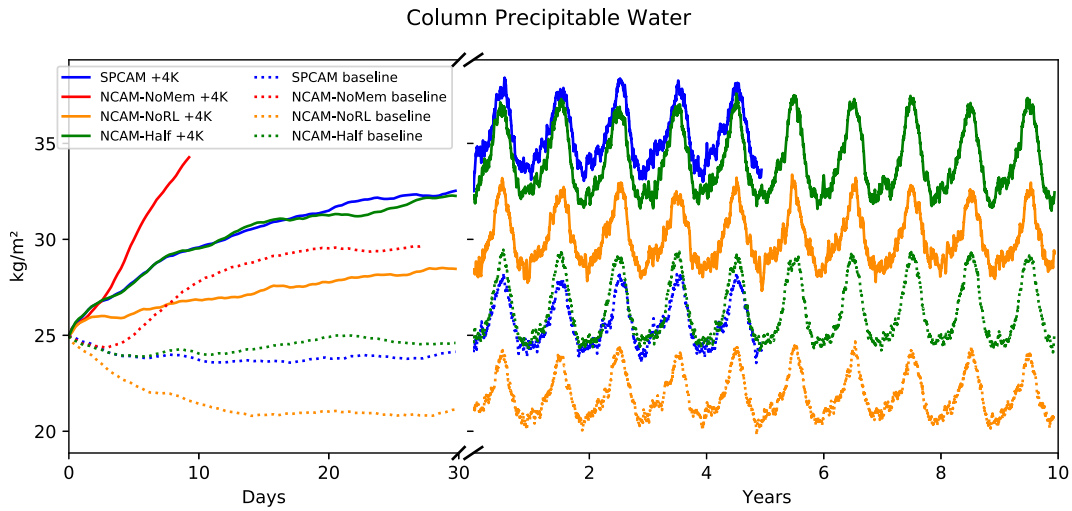

Figure 11. Time evolution of the global-mean, column-integrated precipitable water (kg m^{-2}). NCAM-NoMem in +4 K sea surface temperature (SST) (red solid) and baseline (red dotted); NCAM-NoRL in +4 K SST (orange solid) and baseline (orange dashed); NCAM–Half in +4 K SST (green solid) and baseline (green dashed); superparameterized CAM (SPCAM) 5-year reference in +4 K SST (blue solid) and baseline (blue dashed). The left half shows the first 30 days of each experiment; the right half shows the remaining 5 years of the SPCAM simulations and 10 years of simulations from NCAM-NoRL and NCAM-Half.

of Siberia and Canada and cold biases of −3 to −4 K in tropical and subtropical land as well as polar regions. NCAM-Half has much larger warm biases in both the northern hemisphere high latitudes and Antarctica. In the tropics and subtropics, however, its temperature biases remain within about 2 K, comparable to those in NCAM.

For the vertical structure of state variables in the warm climate (Figure 13), NCAM-NoRL shows similar vertical and latitudinal distribution of the biases in temperature, specific humidity and zonal winds to those in the default NCAM, but with larger magnitudes. There are large positive temperature biases above approximately 400 hPa and relatively smaller negative biases below. The moisture biases are much more pronounced, as are the zonal wind biases. Despite using only half the training data, NCAM-Half maintains modest tropical and subtropical temperature biases, but its high-latitude temperature biases and midlatitude westerly jet biases exceed those in NCAM.

To summarize, the ablation experiments demonstrate that convective memory is very important for numerical stability in this hybrid CAM5 framework. Including radiation and land fraction inputs (SOLIN, LWUP, LANDFRAC), considered jointly here, improves tropical and subtropical precipitation, particularly the land-sea contrast, and reduces surface and air temperature biases. It also reduces the tropical lower and middle troposphere moisture biases. Finally, training-data volume matters at high latitudes. NCAM and NCAM-NoRL exhibit substantially smaller high-latitude temperature biases than NCAM-Half.

6. Summary and Discussions

Machine learning using NNs has been applied in recent years to emulating subgrid scale processes in GCMs. While offline tests have shown that these NNs can emulate the heating and drying tendencies of subgrid-scale convection and cloud processes very accurately, two challenging issues that prevent their application to hybrid GCM simulations and climate change research are generalizability to unseen climate and numerical instability in model integration. Our neural-net-based parameterization for moist physics, ResCu, addresses both issues.

The hybrid GCM, a neural-net-embedded CAM5 referred to as NCAM, is run stably for a decade in a uniform +4 K SST climate with no drift in total energy or precipitable water, despite the neural net being trained solely on the current climate. The global distribution of the simulated precipitation and surface temperature agrees very well with that of SPCAM's own +4 K SST simulation. The zonal-mean vertical cross sections of temperature, specific humidity, and zonal wind are all reproduced well by NCAM. The RH field from NCAM is also realistic, with only limited pockets of upper-level supersaturation over the polar regions; this addresses a common

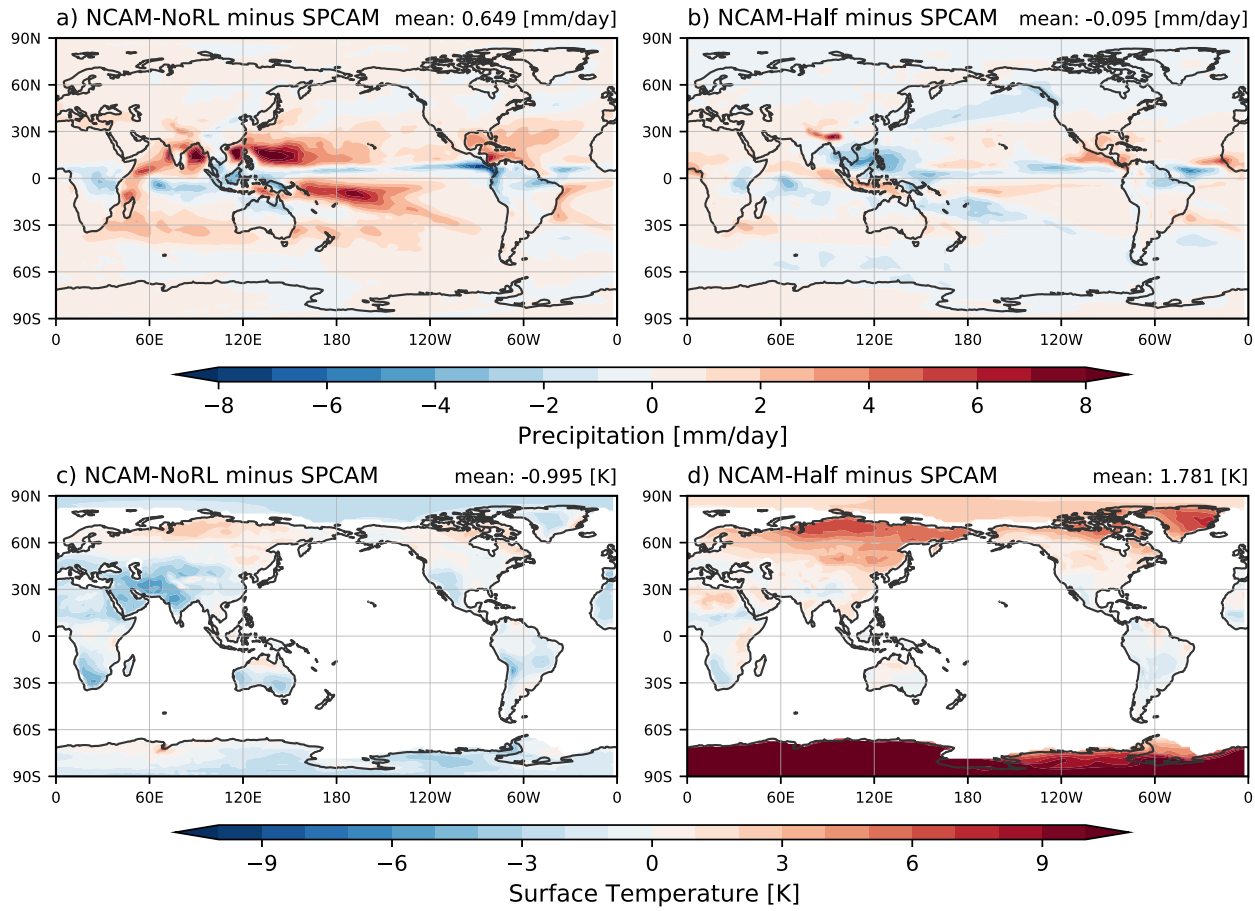

Figure 12. Global distribution of annual mean differences of (a, b) precipitation and (c, d) surface temperature for the +4 K sea surface temperature warm climate relative to 5-year averages from SP CAM, with (a, c) for 10-year averages from NCAM-NoRL to superparameterized CAM (SPCAM) and (b, d) 10-year averages from NCAM-Half to SPCAM. In all panels, global mean differences appear in the top right corner.

weakness highlighted in a recent study (e.g., Watt-Meyer et al., 2024). Furthermore, in the NCAM's +4 K SST simulation, the ResCu-predicted heating and moistening rates also agree well with those from SPCAM. Cloud water is less well simulated compared to other fields, showing large positive biases, whereas cloud ice aligns more closely with that from SPCAM. Overall, the simulation performance of NCAM is comparable to that of the conventional CAM5 for the same +4 K SST simulation.

Not only are the climate-mean variables simulated well by NCAM, the statistics of precipitation intensity and its relationships to the host model's dynamic and thermodynamic fields are also captured. Especially, the precipitation intensity PDFs from NCAM for both the current climate and the +4 K warm climate closely follow those of SPCAM, including the tail end of the distribution for extreme precipitation. As such, NCAM does not have the familiar bias of "too much drizzle, too few downpours", a problem in CAM5 and many other GCMs as well.

More strikingly, NCAM demonstrates an unprecedented capability to predict climate responses to the +4 K SST change. The simulated responses are robustly accurate in the sense that very good agreements with SPCAM results are seen across board for a large variety of key fields (physical quantities) and for different metrics (e.g., both spatiotemporal averages and statistical distributions). The climate responses simulated by NCAM are also physically consistent, in the sense that the SST-induced changes in different atmospheric quantities and climate regimes agree well with results from the physics-based SPCAM model and match our understanding of the underlying physics. For example, NCAM accurately reproduces changes in temperature, moisture, and zonal winds and delivers global precipitation responses in close agreement with SPCAM, especially in intensifying extreme rainfall rates with respect to extreme events in the baseline climate. In contrast, CAM5 struggles across

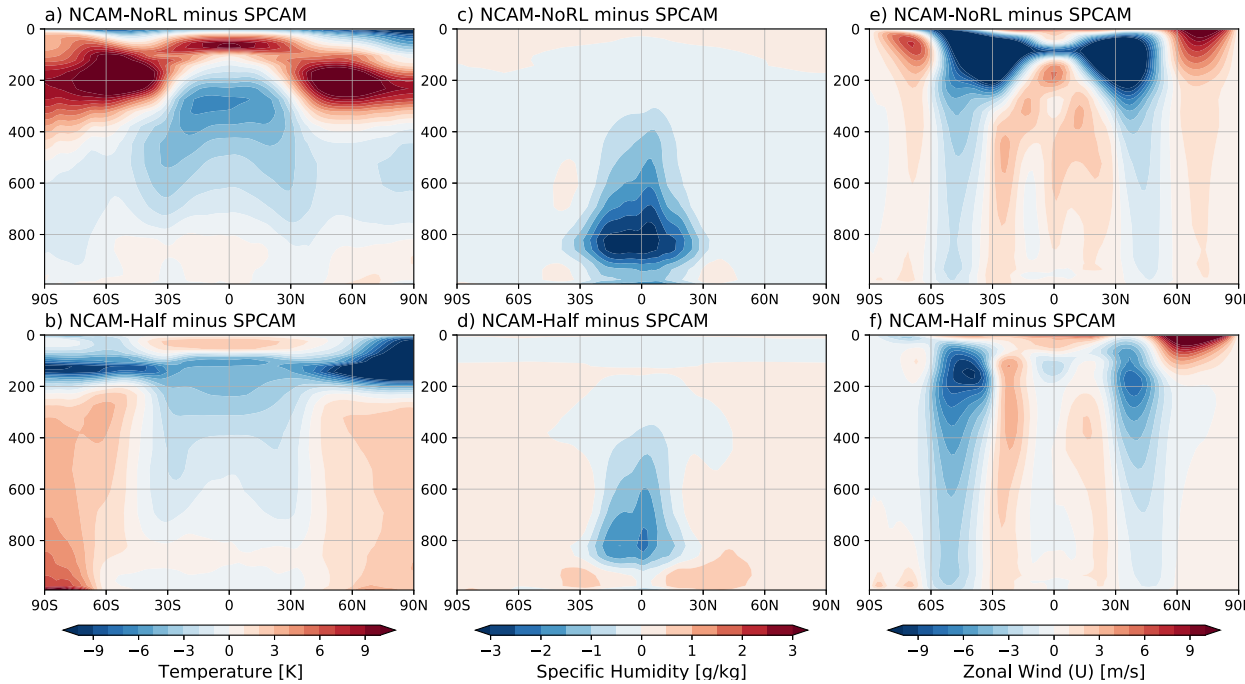

Figure 13. Latitude-pressure cross-sections of the zonal-mean differences of (a, b) temperature and (c, d) specific humidity, and (e, f) zonal wind relative to superparameterized CAM in the +4 K sea surface temperature warm climate for (a, c, e) 10-year averages from NCAM-NoRL and (b, d, f) 10-year averages from NCAM-Half.

these higher-order precipitation metrics. This opens the door for new research using ML hybrid GCM to study how climate changes affect atmospheric states, circulation, and extreme precipitation.

The NN used in this study is based on our earlier work (H20; H23). H23 finds that their NN generalizes very well to warm climate in offline tests, in contrast to findings in several other studies (Beucler et al., 2024; Clark et al., 2022; Rasp et al., 2018). These studies find that the NNs they use fail to extrapolate to unseen climates in offline tests, prompting them to either include warm-climate information in the training data (Clark et al., 2022) or apply "climate-invariant" rescaling (Beucler et al., 2024) to remedy the lack of generalizability of their NNs. In contrast, without using the warm climate data for training or climate-invariant predictors for generalization (Beucler et al., 2024; Clark et al., 2022; Rasp et al., 2018), our results demonstrate successful prognostic generalization using present-climate training alone. To our knowledge, this is the first published decade-scale demonstration of warm-climate stability and realism from a present-climate-trained NN embedded in a real-geography GCM, establishing a benchmark for future hybrid modeling efforts.

To understand the roles of factors contributing to the successful generalization of our NN to the warm climate, we performed several targeted ablation tests: one without convective memory, one without the newly added top-of-model solar insolation, surface outgoing longwave radiation, and land fraction, and one only trained with half of the training data. Without convective memory, the simulations become unstable in both the baseline and +4 K SST climates, indicating the fundamental role of convective memory in stable online integration. Adding radiative fluxes at TOM and surface and land fraction information as input variables results in more realistic simulations of precipitation, surface temperature, and the vertical structures of temperature, humidity, and zonal winds. Increasing the training data volume helps improve the simulation of temperature and zonal winds in mid- and high latitudes. Recently, in a series of input variable sensitivity tests of their NN, Hu et al. (2024) find that including convective memory does not improve the stability and performance of their online integrations using U. S. Department of Energy's Energy Exascale Earth System Model (E3SM) whereas Lin et al. (2025) find including convective memory to be beneficial. Systematic testing of all NN types on a common platform using large ensembles would require significant computational resources. The ClimSim project (Yu et al., 2023) appears to have potential in this area in future endeavors.

The results in this study have broad implications for machine-learning-based hybrid global climate modeling. The fact that NCAM can reproduce the relationships between precipitation and atmospheric circulation and thermodynamic states as well as nearly unchanged RH under both current climate and a warm climate suggests that our NN ResCu not just merely learns the statistical relationships among its input and output variables in the training data but instead is able to learn the fundamental physics behind the statistical relationships. While generalizing NNs to unseen climates remains challenging, as seen in many earlier studies using fully connected architectures that failed to do so, our work suggests that carefully refining NN designs and incorporating critical physical factors for example, convective memory, can substantially improve its generalizability and perform stable hybrid GCM integration for warm climates. The ResNet-based ResCu with convolutional layers was developed 5 years ago in H20. More recent NN frameworks with additional refinements (Hu et al., 2024; Yu et al., 2023) could potentially offer similar capability. Our work and these developments signal a promising future for using deep-learning-based parameterizations in long-term simulations of future climate change, even without incorporating specialized warm-climate training data.

Conflict of Interest

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

Data Availability Statement

The data and neural network codes used in this study (Han & Zhang, 2025) are available via Creative CC-BY-4.0 license from the public data repository (https://doi.org/10.5281/zenodo.17127771). Software Availability Statement: ResCu is developed based on Pytorch v2.2 (Paszke et al., 2019), All statistics are analyzed using Numpy (Harris et al., 2020) with the repository of https://github.com/numpy/numpy; All Figures are plotted using Matplotlib (Hunter, 2007) from https://github.com/matplotlib/matplotlib, and Cartopy (Met Office, 2010–2015).

Acknowledgments
This research was supported by the U.S. Department of Energy's (DOE's) Scientific Discovery through Advanced Computing (SciDAC) program via a partnership between the Office of Biological and Environmental Research (BER) and the Office of Advanced Scientific Computing Research (ASCR). Additional support was provided by BER's Earth System Model Development and Analysis (ESMDA) program under Award Number DE-SC0022064 and National Science Foundation Grant AGS-2054697. Computing resources for this study were provided by the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy. YW was supported by the National Key Research and Development Program of China Grant 2022YFF0802002. We would like to thank the three anonymous reviewers for their constructive comments that have helped improve the manuscript.

References

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., et al. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 098302. https://doi.org/10.1103/PhysRevLett.126.098302

Bony, S., Dufresne, J. L., Le Treut, H., Morcrette, J. J., & Senior, C. (2004). On dynamic and thermodynamic components of cloud changes. *Climate Dynamics*, *22*(2), 71–86. https://doi.org/10.1007/s00382-003-0369-6

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/jas-d-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019MS001711

Bretherton, C. S. (2015). Insights into low-latitude cloud feedbacks from high-resolution models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2054), 20140415. https://doi.org/10.1098/rsta.2014.0415

Bretherton, C. S., Blossey, P. N., & Stan, C. (2014). Cloud feedbacks on greenhouse warming in the superparameterized climate model SP-CCSM4. *Journal of Advances in Modeling Earth Systems*, *6*(4), 1185–1204. https://doi.org/10.1002/2014MS000355

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., et al. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. https://doi.org/10.1029/2021MS002794

Bretherton, C. S., Peters, M. E., & Back, L. E. (2004). Relationships between water vapor path and precipitation over the tropical oceans. *Journal of Climate*, *17*(7), 1517–1528. https://doi.org/10.1175/1520-0442(2004)017<1517:rbwvpa>2.0.co;2

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., et al. (2022). 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*(9), e2022MS003219. https://doi.org/10.1029/2022MS003219

Colin, M., & Sherwood, S. C. (2021). Atmospheric convection as an unstable predator–prey process with memory. *Journal of the Atmospheric Sciences*, *78*(11), 3781–3797. https://doi.org/10.1175/jas-d-20-0337.1

Davies, L., Plant, R. S., & Derbyshire, S. H. (2009). A simple model of convection with memory. *Journal of Geophysical Research*, *114*(D17), 17202-1. https://doi.org/10.1029/2008JD011653

Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A., & Maher, N. (2016). More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, *6*(5), 508–513. https://doi.org/10.1038/nclimate2941

Douville, H., Qasmi, S., Ribes, A., & Bock, O. (2022). Global warming at near-constant tropospheric relative humidity is supported by observations. *Communications Earth & Environment*, *3*(1), 237. https://doi.org/10.1038/s43247-022-00561-z

Fu, Q., Manabe, S., & Johanson, C. M. (2011). On the warming in the tropical upper troposphere: Models versus observations. *Geophysical Research Letters*, *38*(15), 1–6. https://doi.org/10.1029/2011GL048101

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018GL078202

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the 14th international conference on artificial intelligence and statisitics (AISTATS) 2011* (Vol. 15, pp. 315–323).

Han, Y., & Zhang, G. J. (2025). The code and data for the +4K online extrapolation of ResCu [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17127771

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. https://doi.org/10.1029/2020MS002076

Han, Y., Zhang, G. J., & Wang, Y. (2023). An ensemble of neural networks for moist physics processes, its generalizability and stable integration. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2022MS003508. https://doi.org/10.1029/2022MS003508

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hu, Z., Subramaniam, A., Kuang, Z., Lin, J., Yu, S., Hannah, W. M., et al. (2024). Stable machine-learning parameterization of subgrid processes with real geography and full-physics emulation. *arXiv preprint arXiv:2407.00124*.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A new sea surface temperature and sea ice boundary dataset for the community atmosphere model. *Journal of Climate*, *21*(19), 5145–5153. https://doi.org/10.1175/2008JCLI2292.1

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Khairoutdinov, M., Randall, D., & DeMott, C. (2005). Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *Journal of the Atmospheric Sciences*, *62*(7), 2136–2154. https://doi.org/10.1175/JAS3453.1

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*(4), 369–373. https://doi.org/10.1038/nclimate1716

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016). Robust effects of cloud superparameterization on simulated daily rainfall intensity statistics across multiple versions of the Community Earth System Model. *Journal of Advances in Modeling Earth Systems*, *8*(1), 140–165. https://doi.org/10.1002/2015MS000574

Kooperman, G. J., Pritchard, M. S., & Somerville, R. C. J. (2014). The response of US summer rainfall to quadrupled CO_2 climate change in conventional and superparameterized versions of the NCAR community atmosphere model. *Journal of Advances in Modeling Earth Systems*, *6*(3), 859–882. https://doi.org/10.1002/2014MS000306

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Paper presented at the advances in neural information processing systems*.

Kuang, Z. (2024). Linear time-invariant models of a large cumulus ensemble. *Journal of the Atmospheric Sciences*, *81*(3), 605–627. https://doi.org/10.1175/jas-d-23-0194.1

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., et al. (2023). Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003400. https://doi.org/10.1029/2022MS003400

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, F., Rosa, D., Collins, W. D., & Wehner, M. F. (2012). "Super-parameterization": A better way to simulate regional extreme precipitation? *Journal of Advances in Modeling Earth Systems*, *4*(2), M04002. https://doi.org/10.1029/2011MS000106

Lin, J., Yu, S., Peng, L., Beucler, T., Wong-Toi, E., Hu, Z., et al. (2025). Navigating the noise: Bringing clarity to ML parameterization design with $O(100)$ ensembles. *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004551. https://doi.org/10.1029/2024MS004551

Manabe, S., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *Journal of the Atmospheric Sciences*, *24*(3), 241–259. https://doi.org/10.1175/1520-0469(1967)024<0241:teotaw>2.0.co;2

Met Office. (2010–2015). Cartopy: A cartographic python library with a matplotlib interface [Software]. *Github*. Retrieved from https://github.com/SciTools/cartopy

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2021). 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*(5), e2020MS002385. https://doi.org/10.1029/2020MS002385

Morrison, H., & Gettelman, A. (2008). A new two-moment bulk stratiform cloud microphysics scheme in the community atmosphere model, version 3 (CAM3). Part I: Description and numerical tests. *Journal of Climate*, *21*(15), 3642–3659. https://doi.org/10.1175/2008JCLI2105.1

Neale, R. B., Chen, C.-C., Gettelman, A., Lauritzen, P. H., Park, S., Williamson, D. L., et al. (2012). Description of the NCAR community atmosphere model (CAM 5.0). *NCAR Technical Note NCAR/TN-486+ STR*, *1*(1), 1–12.

Neelin, J. D., Peters, O., & Hales, K. (2009). The transition to strong convection. *Journal of the Atmospheric Sciences*, *66*(8), 2367–2384. https://doi.org/10.1175/2009jas2962.1

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018MS001351

Oleson, K. W., Lawrence, D. M., Gordon, B., Flanner, M. G., Kluzek, E., Peter, J., et al. (2010). Technical description of version 4.0 of the Community Land Model (CLM). In *NCAR technical note. NCAR/TN-486+ STR*.

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras deep learning Bridge for scientific computing. *Scientific Programming*, *2020*(1), 8888811–8888813. https://doi.org/10.1155/2020/8888811

Park, S., & Bretherton, C. S. (2009). The university of Washington shallow convection and moist turbulence schemes and their impact on climate simulations with the community atmosphere model. *Journal of Climate*, *22*(12), 3449–3469. https://doi.org/10.1175/2008jcli2557.1

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (Vol. 32).

Peters, O., & Neelin, J. D. (2006). Critical phenomena in atmospheric precipitation. *Nature Physics*, *2*(6), 393–396. https://doi.org/10.1038/nphys314

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., & Düben, P. (2019). Global cloud-resolving models. *Current Climate Change Reports*, *5*(3), 172–184. https://doi.org/10.1007/s40641-019-00131-0

Sherwood, S. C., Roca, R., Weckwerth, T. M., & Andronova, N. G. (2010). Tropospheric water vapor, convection, and climate. *Reviews of Geophysics*, *48*(2), RG2001. https://doi.org/10.1029/2009RG000301

Stevens, B., & Bony, S. (2013). What are climate models missing? *Science*, *340*(6136), 1053–1054. https://doi.org/10.1126/science.1237554

Vallis, G. K., Zurita-Gotor, P., Cairns, C., & Kidston, J. (2015). Response of the large-scale structure of the atmosphere to global warming. *Quarterly Journal of the Royal Meteorological Society*, *141*(690), 1479–1501. https://doi.org/10.1002/qj.2456

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Wang, Y., Zhang, G. J., & Craig, G. C. (2016). Stochastic convective parameterization improving the simulation of tropical precipitation variability in the NCAR CAM5. *Geophysical Research Letters*, *43*(12), 6612–6619. https://doi.org/10.1002/2016GL069818

Wang, Y., Zhang, G. J., & He, Y.-J. (2017). Simulation of precipitation extremes using a stochastic convective parameterization in the NCAR CAM5 under different resolutions. *Journal of Geophysical Research: Atmospheres*, *122*(23), 12875–12891. https://doi.org/10.1002/2017JD026901

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (2024). Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *Journal of Advances in Modeling Earth Systems*, *16*(2), e2023MS003668. https://doi.org/10.1029/2023MS003668

Yu, S., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., et al. (2023). ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation. In *Advances in neural information processing systems* (Vol. 36).

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). 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*(6), e2020GL091363. https://doi.org/10.1029/2020GL091363

Zhang, G. J., & McFarlane, N. A. (1995). Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmosphere-Ocean*, *33*(3), 407–446. https://doi.org/10.1080/07055900.1995.9649539

Zhang, G. J., & Song, X. (2009). Interaction of deep and shallow convection is key to Madden-Julian Oscillation simulation. *Geophysical Research Letters*, *36*(9), L09708. https://doi.org/10.1029/2009GL037340

Zhou, X., & Khairoutdinov, M. F. (2017). Changes in temperature and precipitation extremes in superparameterized CAM in response to warmer SSTs. *Journal of Climate*, *30*(24), 9827–9845. https://doi.org/10.1175/jcli-d-17-0214.1