JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003508

Special Section:
Machine learning application to Earth system modeling

An Ensemble of Neural Networks for Moist Physics Processes, Its Generalizability and Stable Integration

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

[1]Department of Earth System Science, Tsinghua University, Beijing, China, [2]Scripps Institution of Oceanography, La Jolla, CA, USA

Key Points:

- An ensemble of deep convolutional residual neural networks is used to reduce the uncertainty in moist physics emulations
- The ensemble of the neural networks trained on data from a present-day climate simulation generalizes well to a +4K warm climate offline
- A multi-year stable online integration is achieved in a real-geography GCM with reasonable results

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

Received 10 NOV 2022
Accepted 13 SEP 2023

Author Contributions:

Conceptualization: Guang J. Zhang
Data curation: Yilun Han, Yong Wang
Formal analysis: Yilun Han, Guang J. Zhang
Funding acquisition: Guang J. Zhang, Yong Wang

Abstract With the recent advances in data science, machine learning has been increasingly applied to convection and cloud parameterizations in global climate models (GCMs). This study extends the work of Han et al. (2020, https://doi.org/10.1029/2020MS002076) and uses an ensemble of 32-layer deep convolutional residual neural networks, referred to as ResCu-en, to emulate convection and cloud processes simulated by a superparameterized GCM, SPCAM. ResCu-en predicts GCM grid-scale temperature and moisture tendencies, and cloud liquid and ice water contents from moist physics processes. The surface rainfall is derived from the column-integrated moisture tendency. The prediction uncertainty inherent in deep learning algorithms in emulating the moist physics is reduced by ensemble averaging. Results in 1-year independent offline validation show that ResCu-en has high prediction accuracy for all output variables, both in the current climate and in a warmer climate with +4K sea surface temperature. The analysis of different neural net configurations shows that the success to generalize in a warmer climate is attributed to convective memory and the 1-dimensional convolution layers incorporated into ResCu-en. We further implement a member of ResCu-en into CAM5 with real world geography and run the neural-network-enabled CAM5 (NCAM) for 5 years without encountering any numerical integration instability. The simulation generally captures the global distribution of the mean precipitation, with a better simulation of precipitation intensity and diurnal cycle. However, there are large biases in temperature and moisture in high latitudes. These results highlight the importance of convective memory and demonstrate the potential for machine learning to enhance climate modeling.

Plain Language Summary The representation of storms and clouds through empirical algorithms known as parameterizations in global climate models (GCMs) is one of the main sources of biases in the simulation of rainfall and atmospheric circulation. Here an ensemble of 8 deep neural networks are used to replace the conventional parameterization of atmospheric moist physics processes. They are trained on data sampled from 1-year present-day climate simulation by a "superparameterized" climate model, which uses a two-dimensional cloud-scale model to explicitly simulate convection and clouds inside each GCM grid box. On ensemble averaging, the neural nets produce highly accurate predictions of precipitation characteristics including global distribution and intensity. Furthermore, the machine-learned emulator trained on data in the current climate also represents convection and precipitation extremely well in a warmer climate. A member of the ensemble of the neural nets is implemented into a GCM. The model is then integrated for 5 years, producing reasonable results.

1. Introduction

Convection and cloud parameterization schemes used in global climate models (GCMs) are a major source of many biases in the simulation of climate and its variability. These include biases in the Intertropical Convergence Zone (ITCZ) (Zhang et al., 2019), intraseasonal variability such as Madden Julian Oscillation (MJO) (Cao & Zhang, 2017; Zhang & Mu, 2005) and diurnal cycle of precipitation (Cui et al., 2021; Xie et al., 2019). They are also the main causes of uncertainties in GCM-simulated response of cloud radiative forcing and precipitation to global warming (Stevens & Bony, 2013). Current convection parameterization schemes (e.g., Arakawa & Schubert, 1974; Tiedtke, 1989; Zhang & McFarlane, 1995 and many more) were developed based on limited observations and simplified or heuristic models. Although some incremental progress has been made in climate simulations by improving the parameterization schemes (e.g., Bechtold et al., 2014; Neale et al., 2008; Song & Zhang, 2018; Wang et al., 2016; Xie et al., 2019; Zhang & Mu, 2005), conventional convection and cloud parameterization has reached a deadlock (Gentine et al., 2018; Randall et al., 2003), and other alternatives have been actively explored.

Investigation: Yilun Han, Guang J. Zhang
Methodology: Guang J. Zhang
Project Administration: Guang J. Zhang, Yong Wang
Resources: Guang J. Zhang, Yong Wang
Software: Yilun Han
Supervision: Guang J. Zhang
Validation: Yilun Han
Visualization: Yilun Han, Guang J. Zhang
Writing – original draft: Yilun Han, Guang J. Zhang
Writing – review & editing: Yilun Han, Guang J. Zhang

One of the alternatives is to embed a cloud resolving model (CRM) into each GCM grid box to replace the conventional convection parameterization scheme, the so called superparameterization approach. Khairoutdinov et al. (2005) developed the superparameterized National Center for Atmospheric Research (NCAR) Community Atmosphere Model (SPCAM). SPCAM performs better in the simulation of convection at different scales such as the eastward propagating mesoscale convective systems, the diurnal cycle of convection, and MJO (Jiang et al., 2015; Pritchard & Somerville, 2009).

Data-driven machine learning (ML) has been actively explored for parameterizing subgrid-scale convection and cloud processes in the last few years (Beucler et al., 2021a, 2021b; Brenowitz & Bretherton, 2018, 2019; Brenowitz et al., 2020; Gentine et al., 2018; Han et al., 2020; Irrgang et al., 2021; Rasp et al., 2018; Wang et al., 2022; Yuval & O'Gorman, 2020; Yuval et al., 2021). Gentine et al. (2018) used deep learning to emulate convection and radiation processes simulated by SPCAM. Rasp et al. (2018) coupled a neural network (NN) to a 3-D aqua-planet GCM. Brenowitz and Bretherton (2019) trained a neural network parameterization scheme using coarse-grained global CRM simulation results and realized a multi-day online simulation in a coarse-resolution GCM. Yuval and O'Gorman (2020) developed a random forest-based ML parameterization with simulation results from a high-resolution 3-D model run on an idealized beta plane. They reproduced the climate of the high-resolution model in a coarse-resolution model with this parameterization. Later, Yuval et al. (2021) used neural networks and obtained similar results with less computational memory. The above studies all used the aqua-planet configuration of the GCMs.

Recently, studies have emerged on ML parameterization schemes under real geography. Han et al. (2020), hereafter H20, accurately emulated convective heating, drying, cloud water and ice concentrations in a realistically configured SPCAM by applying a 1-D residual convolution neural network (ResNet) with powerful nonlinear fitting ability, and tested it offline and in a single column model. Mooers et al. (2021) optimized a fully connected neural network with a sophisticated auto-learning technique to emulate convection under real land-ocean distribution and used the neural-network emulated fields to force an offline land surface model with some success.

The performance of ML-based parameterizations has been improved in the last few years. However, not much attention has been paid to their uncertainties. Brenowitz and Bretherton (2019) first noted that the training bias fluctuates significantly from one training epoch to another, and thus determining when to stop the training can lead to considerable uncertainties. Furthermore, individual predictions from a deep learning model can contain sizable uncertainties even though the model performs well on average (Gawlikowski et al., 2021; Pearce et al., 2018). The prediction uncertainty from NN-based parameterizations can come from two major sources: aleatoric and epistemic uncertainties. Aleatoric uncertainty is the intrinsic uncertainty within the target data. For the superparameterization simulations, this uncertainty is mainly from losing many degrees of freedom when the CRM-domain fields are coarse-grained to the GCM grid. On the other hand, epistemic uncertainty is due to limited data and knowledge of the ML models. To speed up the training process of an NN-based parameterization, we only use "limited" training data, which is often an arbitrarily selected subset in space and/or time of a sampling pool from a high-resolution model simulation. Since cloud and convection processes are highly complex and nonlinear, an NN emulator is not "perfect" with 100% fitting accuracy. In practice, the epistemic uncertainty comes from the process of training, which involves randomly initializing the weights and biases in the NNs first, and then training them with data in mini-batches, which randomly distribute the data into numerous subsets and shuffle the subset sequence after every training iteration. The algorithms to optimize the weights and biases during the training are stochastic or are related to stochastic processes, such as Stochastic Gradient Descent, Root Mean Squared Propagation, and Adaptive Moment Estimation (Adam) (Kingma & Ba, 2014). As a result, all the randomness involved in the training process contributes to the prediction uncertainty, which cannot be ignored in developing NN-based parameterizations.

A challenging issue for an NN-based parameterization trained on one climate is to generalize it to another, unseen climate as it requires the neural net to fit out-of-distribution data. Several studies have tested the ability of their NN-based parameterizations to generalize to different climates (Beucler et al., 2021b; Clark et al., 2022; O'Gorman & Dwyer, 2018; Rasp et al., 2018). They found that the NN-based parameterizations trained with data from the current climate degraded seriously in accuracy when directly used in warmer climates. To achieve a better generalization, Rasp et al. (2018) and Clark et al. (2022) included the warm climate simulation output in the training data. Beucler et al. (2021b) rescaled the NN's input and output variables to keep the probability distribution unchanged across climates.

Besides generalization to different climates, making a stable model integration using NN-based parameterizations is another great challenge (Irrgang et al., 2021). Several recent studies have explored the prognostic performance of ML parameterization schemes in 3D real-geography GCMs. Wang et al. (2022) emulated the moist physics and radiation processes in SPCAM with a group of deep neural networks. They succeeded in a 5-year online integration, but with significant climate biases in high latitudes. Bretherton et al. (2022) used machine learning of nudging tendencies as functions of the atmospheric state to correct the physical parameterization tendencies and ran a NOAA global forecasting model for 40 days. Clark et al. (2022) tested this ML-learned tendency correction approach and ran the model for more than 5 years as well as for different climates.

In this study, we use an ensemble of 8 refined deep NNs based on the ResNet in H20 to reduce the uncertainties in NN predictions, similar to Krasnopolsky et al. (2013). We then test its generalizability to a +4K SST warmer climate and explore different attributes of the NN in this regard. Finally, we attempt to carry out a multi-year online integration to assess whether a stable long-term integration is achievable with reasonable results. The organization of the paper is as follows. Section 2 presents the details of the data generation and NN design. Section 3 shows the results of offline validations. Section 4 tests the generalization of the NN to a warmer climate and the roles of the NN architecture and input variables in its generalization ability. Section 5 performs a prognostic online simulation in a 3D real-geography GCM. A summary and discussion are given in Section 6.

2. An Ensemble of Neural Networks

2.1. Selection of Training Data

Same as in H20, we use a year-long simulation from the NCAR SPCAM (Khairoutdinov et al., 2005). It includes a coupled land model CLM 4.0 and is run with prescribed monthly mean climatological sea surface temperature (SST) and sea ice for lower boundary conditions (Hurrell et al., 2008). The model is run for three and a half years with a timestep of 20 min, and we use subsets of year two simulation output for training the NN. To speed up the training, we select 800 points out of the total of 13,824 (96 × 144) grid points in the 2.5° × 1.9° horizontal resolution model. Instead of selecting 800 fixed points as in H20, for data from each day of the year we select 800 points with each grid point randomly chosen from three latitude zones in proportion to their relative surface area. The three latitude zones are the tropics (30°S–30°N), midlatitudes (60°S–30°S and 30°N–60°N), and high latitudes (90°S–60°S and 60°S–90°N). Therefore, we have 56,700 (800 points × 3 timesteps/hr × 24 hr) training samples each day and nearly 21 million samples in total. The new method of data selection ensures that all regions on the globe are represented in the training data set. This training data selection procedure is repeated for training each NN.

2.2. Input and Output

The input variables for the NN are mostly the same as those in H20. These include the GCM grid-scale state variables and tendencies that are used to force the CRM in SPCAM. They are profiles of temperature (T), specific humidity (q_v), large-scale temperature and moisture tendencies $\left(\frac{\partial T}{\partial t}\right)_{ls}$ and $\left(\frac{\partial q_v}{\partial t}\right)_{ls}$ from the dynamic core of SPCAM's host CAM5 and planetary boundary layer (PBL) diffusion, surface sensible and latent heat fluxes (SSHF/c_p and SLHF/L_v) and surface pressure (P_s). We also consider convective memory as in H20, but with some modification. In H20, we considered 4 GCM timesteps. In the sensitivity test in H20, it was found that including 2 timesteps would suffice to account for the effect of the history of convection. Thus, here for convective memory we only consider the following variables in the previous 2 timesteps: the GCM grid-scale T, q_v, $\left(\frac{\partial T}{\partial t}\right)_{ls}$ and $\left(\frac{\partial q_v}{\partial t}\right)_{ls}$, temperature and moisture tendencies from moist physics $\frac{\partial T}{\partial t}$ and $\frac{\partial q_v}{\partial t}$, and cloud water q_c and cloud ice q_i predicted by the CRM. The output variables are also the same as those in H20: GCM grid averaged diabatic temperature and moisture tendencies $\frac{\partial T}{\partial t}$ and $\frac{\partial q_v}{\partial t}$, cloud water and cloud ice contents q_c and q_i. Precipitation is diagnosed from the vertically integrated moisture tendency in the output.

In total, the input layer consists of 20 vectors with a length of 33 and the output layer consists of four vectors with a length of 30. All input and output variables are normalized with normalization factors the same as those in H20 to ensure that they are of order of magnitude O(1) before they are input into the deep neural network for training and testing.

2.3. Loss Function Accounting for Moist Static Energy Conservation

In moist physics, the atmospheric moist static energy (h) is conserved in the absence of ice phase processes. As in H20, we customize the loss function to include h conservation by adding the mean square error between column-integrated h change from the neural net and that from SPCAM in the form of $\left\| \frac{1}{g}\int_{pt}^{pb} \frac{\partial h_{\mathrm{SP}}}{\partial t} dp - \frac{1}{g}\int_{pt}^{pb} \frac{\partial h_{\mathrm{NN}}}{\partial t} dp \right\|_2$ as a penalty term in our loss function to make the integrated h tendencies from deep learning model approach those from SPCAM. Thus, the loss function is written as

$$\mathrm{loss} = \|\hat{y} - y\|_2 + \lambda \left\| \frac{1}{g}\int_{pt}^{pb} \frac{\partial h_{\mathrm{SP}}}{\partial t} dp - \frac{1}{g}\int_{pt}^{pb} \frac{\partial h_{\mathrm{NN}}}{\partial t} dp \right\|_2, \quad (1)$$

where y is the target fields from SPCAM, $\hat{y}$ is the output of our neural network model, and λ is a Lagrangian multiplier to simultaneously enforce accuracy and h conservation.

2.4. Deep ResNet

In H20, a moist physics parameterization was developed using a 1-D residual convolutional neural network (ResNet), referred to as ResCu for short. We continue to use the same NN construct here, but with the following modifications: (a) extend the number of layers from 22 to 32; (b) add a batch normalization layer after each convolutional layer except the last one; and (c) remove the activation function in the last layer. The first modification is to further improve the accuracy of the neural network. The last two modifications are based on the sensitivity tests of H20. Batch-normalization helps improve the accuracy and robustness when added after each layer, since it normalizes the output of the layer with a running average and a running standard deviation (Ioffe & Szegedy, 2015). With multiple activation functions in previous layers for nonlinear representation, the last layer activation does not add much further improvement in the accuracy of the output in a deep neural network.

After increasing the depth of the NN from 22 to 32 layers, the RMSE of the fitting (the first term on the rhs of Equation 1) is significantly smaller, which makes the h conservation penalty (the second term on the rhs of Equation 1) a dominant term in the loss function. Thus, the original value of the multiplier $\lambda = 5 \times 10^{-7}$ is too large, which affects the convergence of the NN training. We set λ to a new value of 1×10^{-7} for the optimal balance between h conservation and prediction accuracy. This makes the penalty term from h conservation account for about 6% of the total loss. With some preliminary trial tests, we find that 32 layers are optimal for prediction accuracy as well as h conservations, reducing the total loss by 1.7% compared to the 22-layer NN (Figure S1a in Supporting Information S1). Further increasing the depth of the NN (in our case, to 42 layers) does not lead to a further decrease in the total loss function.

This deep ResNet applies 1-D convolutional layers with 128 feature vectors (1-D feature maps) and 128 corresponding filter banks with a kernel size of 3. It contains 15 Resunits, including 32 convolutional layers in total, with approximately 1.5 million trainable parameters and 40,000 untrainable parameters (running averages and standard deviations in the batch normalization layers). The activation algorithms inside each Resunit are Rectified Linear Activations (ReLUs), with no activation in the output layer.

As mentioned earlier, to reduce the prediction uncertainty from the NN, we use an ensemble of 8 NNs, referred to as NN-1 to NN-8 and the ensemble referred to as ResCu-en (Figure S1b in Supporting Information S1). Using the 32-layer ResNet, we independently trained each of the 8 NNs with different random seeds for initialization and selected training subsets (see Section 2.1 above). All 8 neural networks are identical in input and output variables and NN architecture, and trained over 100 epochs, using the Adam optimizer that has an initial learning rate of 3×10^{-4}.

3. Offline Validation for Current Climate

To evaluate the performance of ResCu-en, we compare the ensemble mean predictions from it with SPCAM simulations using the independent third-year testing data. As described in Section 2.1, this target simulation is forced with the climatological mean SST, which we refer to as the baseline simulation hereinafter. We test the performance in multiple aspects: moist static energy conservation, annual mean of the predicted variables, and precipitation frequency distribution. Since the training data are from an SPCAM simulation under the present-day

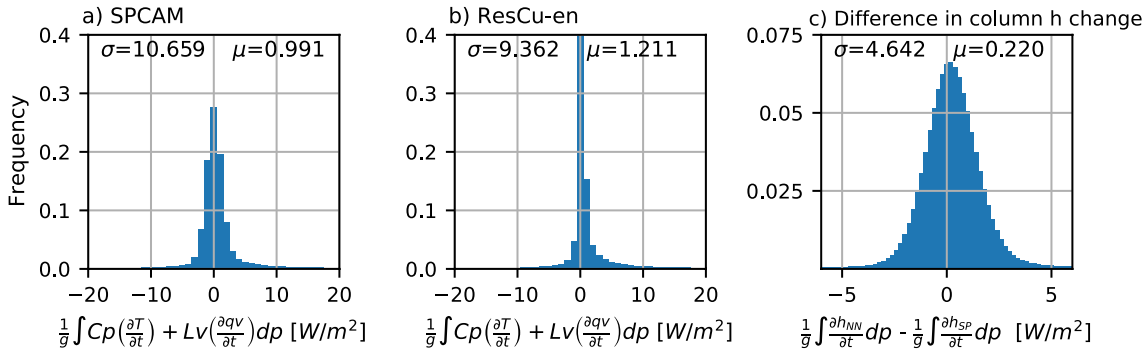

Figure 1. The histograms of probability density function (PDF) of column-integrated moist static energy change for (a) SPCAM, (b) ResCu-en, and (c) the differences between prediction from ResCu-en and SPCAM at each GCM grid column. The standard deviation (σ) and mean (μ) for each PDF are shown at the top of each plot.

climate conditions, an important question is whether the trained NN can be used in a warmer climate. To test the capability of ResCu-en generalization to a warm climate, we perform an SPCAM simulation with +4K SST (Bretherton et al., 2014), that is, we add 4K uniformly on top of the monthly mean global SST distribution as the boundary condition. Then we use the simulated fields from the +4K simulation as input into ResCu-en, which is trained with the present-day climate simulation data, to diagnose the moist physics tendencies and precipitation.

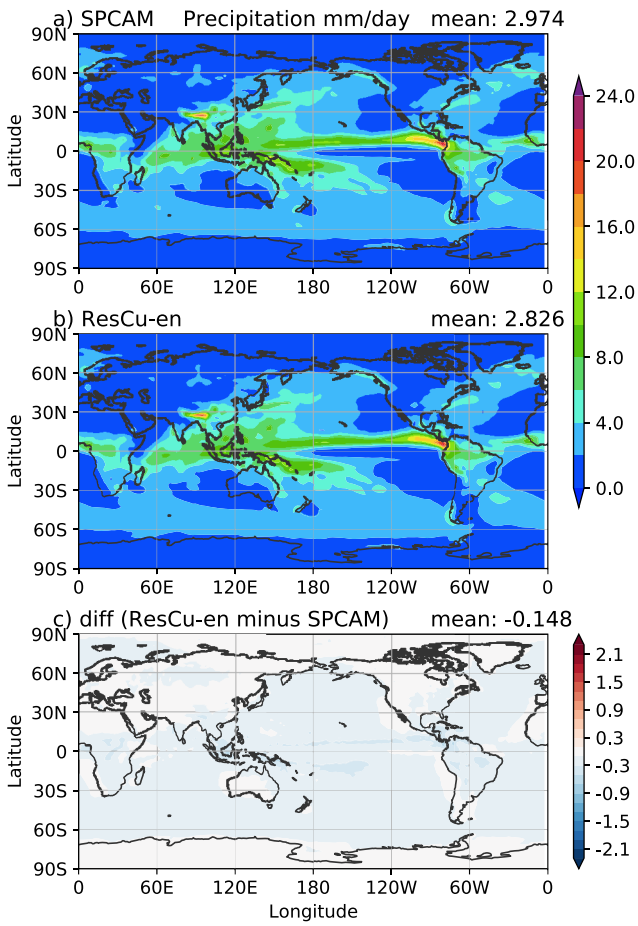

Figure 2. Global distribution of the annual mean precipitation for the baseline climate in (a) SPCAM simulation, (b) offline test by ResCu-en, and (c) their differences (ResCu-en minus SPCAM). Note that the color intervals for the differences is 5% of that for the mean to provide a better visualization of the differences.

First, we check the accuracy of moist static energy conservation in ResCu-en. For moist physics, the column integrated heating and drying or h tendencies $\frac{L_v}{g}\int_{pt}^{pb}\frac{\partial q}{\partial t}dp + \frac{C_p}{g}\int_{pt}^{pb}\frac{\partial T}{\partial t}dp$ should be equal to the net freezing heating and melting cooling associated with ice phase change of hydrometeors in the column. Figure 1 shows the histogram of column-integrated h tendencies from SPCAM, ResCu-en and their differences. The SPCAM simulation shows a mean $\mu = 0.99$ W/m^2 and a standard deviation $\sigma = 10.66$ W/m^2 (Figure 1a). The histogram of the column-integrated h change predicted by ResCu-en is remarkably close to that of SPCAM with a mean of 1.21 W/m^2 and a standard deviation of 9.62 W/m^2 (Figure 1b). The difference plot (Figure 1c) shows the histogram of the differences between column integrated NN-predicted h tendencies and the corresponding SPCAM simulated values at each GCM grid column and time step for all data used in the test. There is only a small systematic positive bias of 0.22 W/m^2 and a difference spread (standard deviation) of 4.71 W/m^2. Note that the temperature and moisture tendencies from the moist physics processes in the NN are predicted independently and their column-integrated values are on the order of 1,000–4,000 W/m^2 (cf. Figure 2 in H20). Thus, this demonstrates that ResCu-en is remarkably accurate in h conservation even though the requirement of h conservation only contributes 5% to the total loss function (Figure S1a in Supporting Information S1). Note that past neural-network-based emulators struggled to maintain strict column-integrated h conservation, with larger standard deviation (Rasp et al., 2018) or imbalances (Brenowitz & Bretherton, 2018). On the other hand, a random-forest-based emulator developed by Yuval and O'Gorman (2020) has a much better h conservation, with only a small bias of 0.0001 W/m^2. This is because random forest by design conserves energy whereas neural networks do not obey energy conservation a priori.

The predicted annual mean precipitation by ResCu-en is in excellent agreement with the SPCAM simulation, with no significant localized biases but a slight underestimation on global average (Figure 2). The differences between individual NN and SPCAM (Figure S2 in Supporting Information S1) simulation are relatively larger, highlighting the advantage of using an ensemble

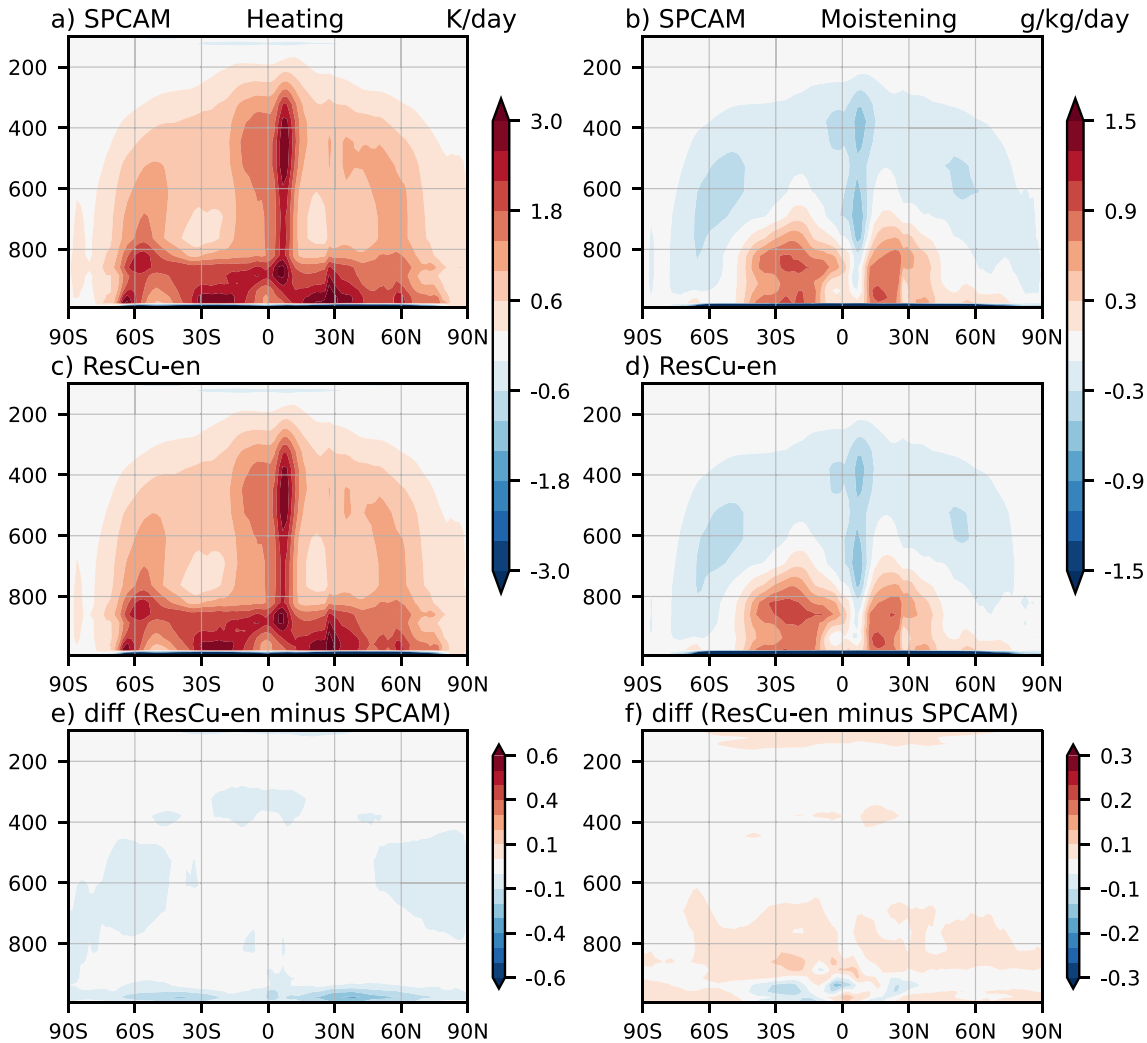

Figure 3. Latitude-pressure cross sections of annual and zonal average heating (left) and moistening (right) from moist physics for (a and b) SPCAM simulation, (c and d) offline test by ResCu-en, (e and f) and their differences (ResCu-en minus SPCAM).

of NNs. In H20, ResCu can reproduce the target precipitation with high accuracy already, except with some noticeable overestimation over the Tibetan Plateau and underestimation in the ITCZ and SPCZ. These biases are either almost completely gone or less evident in the individual NNs in Figure S2 of the Supporting Information S1, indicating a clear improvement owing to the use of a deeper NN (32 layers here vs. 22 layers in H20).

The diabatic heating and drying rates in SPCAM from the CRM simulated convection and condensation processes are also reproduced to a high degree of accuracy by ResCu-en. In the pressure-latitude cross section of the annual and zonal mean, the SPCAM simulation (Figures 3a and 3b) shows the typical climatological features: a deep tropospheric heating and corresponding condensational drying in the tropics from deep convection, heating and moistening in the lower troposphere in the subtropics from shallow convection and stratiform processes, and heating and drying in the mid- and low troposphere by midlatitude cyclones. These features are well captured by ResCu-en (Figures 3c and 3d), with biases no larger than 5% of the SPCAM simulated values (Figures 3e and 3f). Even for the strong cooling and drying near the surface, which are the CRM responses to the PBL forcing, ResCu-en reproduces them accurately. The individual NNs that constitute ResCu-en, on the other hand, have relatively larger biases (Figure S3 in Supporting Information S1). We also computed the RMSE of heating rate relative to the SPCAM values at each GCM grid point using data from every time step and averaged the RMSE over the tropics (20°S, 20°N) following the method of Beucler et al. (2021b). Figure S4 in Supporting Information S1 shows the vertical profiles of RMSE and model layer thickness-weighted MSE for each member of

ResCu-en. The RMSE is 2–4 K/day in the lower and middle troposphere. A more direct comparison with Beucler et al. (2021b) is the thickness-weighted MSE, which has a maximum of about 800 W^2/m^4. This compares to about 2,000 W^2/m^4 for the climate-invariant NN in Beucler et al. (2021b), indicating that even members of ResCu-en are quite accurate. Similar accuracies are found for ResCu-en predicted cloud water and cloud ice distributions, with differences between the ResCu-en prediction and SPCAM simulation less than 0.2 mg/kg everywhere (Figure S5 in Supporting Information S1). These again demonstrate the superiority of using an ensemble of NNs with deeper NNs.

In addition to the annual mean fields, we also examine the frequency of precipitation, one of the essential precipitation characteristics conventional parameterization schemes often fail to represent (Wang et al., 2016; Xie et al., 2019). Figure 4 shows the frequency distribution of daily averaged precipitation for SPCAM simulation and ResCu-en prediction. To show the land-sea contrast, the model grid points are divided into ocean (land fraction less than 0.1) and land (land fraction greater than 0.95). We also present the latitudinal differences by showing the results in the tropics (20°S–20°N), northern hemisphere mid-latitudes (20°N–50°N), and northern hemisphere high latitudes (50°N–90°N). For comparison, we also plot the precipitation frequency for a simulation under the global warming scenario to be discussed in the next section and from TRMM observations for reference. In all regions, the precipitation intensity pdf from SPCAM is very well captured by ResCu-en. Compared to the TRMM observations, SPCAM underestimates the frequency of occurrence of heavy precipitation. Consequently, ResCu-en has the same deficiency.

To summarize, an ensemble of neural networks, ResCu-en, obeys moist static energy conservation very well, with little systematic bias. It accurately reproduces the annual mean heating and drying from moist physics processes in SPCAM. For precipitation, ResCu-en reproduces the mean and the frequency of occurrence distribution with high accuracy.

4. Offline Test of Generalization to a Warmer Climate

4.1. Performance of Generalization to +4K SST Simulation

ResCu-en is trained with an SPCAM simulation under current SST conditions. Can it be extrapolated to represent moist processes in warmer climates? Rasp et al. (2018) tested their DNN parameterization against aquaplanet SPCAM simulations under a warm climate with uniformly increased 4K SST. They showed that the DNN resulted in large errors when it was not trained with the warm climate simulations, including overestimating heavy rainfall rates and having large diabatic heating biases in the tropical lower troposphere, possibly due to out-of-distribution data in the warmer climate. Recently, Beucler et al. (2021b) developed a climate-invariant rescaling approach to help machine learning better generalize to climates different from that used in the training. They showed that when moisture was rescaled with relative humidity and temperature was rescaled with plume buoyancy the NN trained using simulation data from one climate could generalize well to another climate in offline tests. Here we also evaluate ResCu-en in a warm climate simulation by SPCAM with +4K SST added uniformly to the prescribed present-day climatological SST. The SPCAM is run for 2 years with a timestep of 20 min and the second year is used for the ResCu-en offline validation.

In the warm climate with +4K SST, the global average rainfall simulated by SPCAM is increased by about 11% (Figure 5a). Even though ResCu-en is trained using simulation data for the current climate, it can still accurately reproduce the global annual mean precipitation distribution under +4K SST conditions (Figure 5b), with a slight overestimation in ITCZ, SPCZ and the western tropical Indian Ocean, and a slight underestimation over midlatitude oceans (Figure 5c). For precipitation intensity frequency, the SPCAM simulates a significant shift of precipitation occurrence frequency toward higher precipitation rates in the warm climate over the oceans (Figure 4), but no obvious shifts over land in the tropics and midlatitudes for daily average precipitation. ResCu-en accurately reproduces the same shift as the SPCAM in all regions examined. Since precipitation is derived from the vertical integral of moisture tendencies from moist physics, we show in Figure 6 the pressure-latitude cross section of temperature and moisture tendencies from SPCAM, ResCu-en and their differences to further demonstrate the ability of ResCu-en to generalize to a warmer climate. Clearly, ResCu-en reproduces the SPCAM temperature and moisture tendencies with high accuracy, with biases generally less than 5% of the maximum heating and moistening in SPCAM. The differences from individual ensemble member are somewhat larger than the ensemble mean (Figure S6 in Supporting Information S1). They are also only slightly larger than those for the baseline simulation

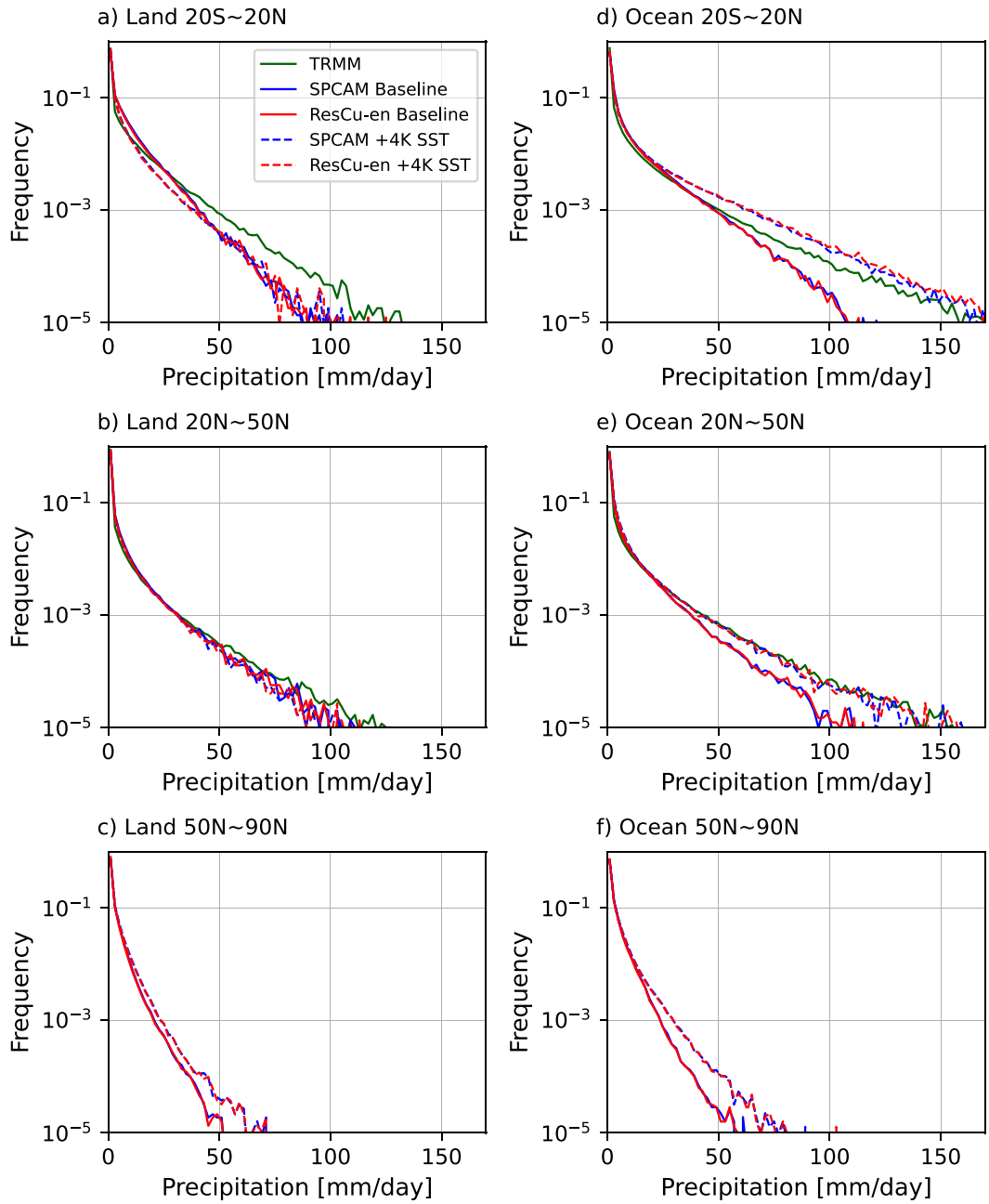

Figure 4. The probability distribution function (PDF) of rainfall intensity for both baseline climate (solid lines) and +4K SST warm climate (dashed lines) for different regions: (a and d) tropics (20°S–20°N), (b and e) northern hemisphere extratropical regions (20–50°N), and (c and f) northern hemisphere high latitudes (50–90°N). The left column is for land regions and the right column is for oceans. The TRMM 3B42 rainfall product (green solid line) is included for reference. The bin intervals for the PDFs are 2 mm/day.

(compare Figure S3 with Figure S6 in Supporting Information S1). The RMSE and thickness-weighted MSE are also larger (Figure S4 in Supporting Information S1). The MSE for the warmer climate can also be compared with that in Buecler et al. (2021b). For all members of ResCu-en, the maximum MSE is about 1,300 W^2/m^4, which is smaller compared with 2,000–4,000 W^2/m^4 in Buecler et al. (2021b).

The performance of ResCu-en is further evaluated in terms of the geographical distribution of the coefficient of determination R^2 for precipitation for both current and +4K climates (Figure 7), which measures how accurately ResCu-en emulates the time series of the target precipitation at each grid point. Most regions have high accuracy

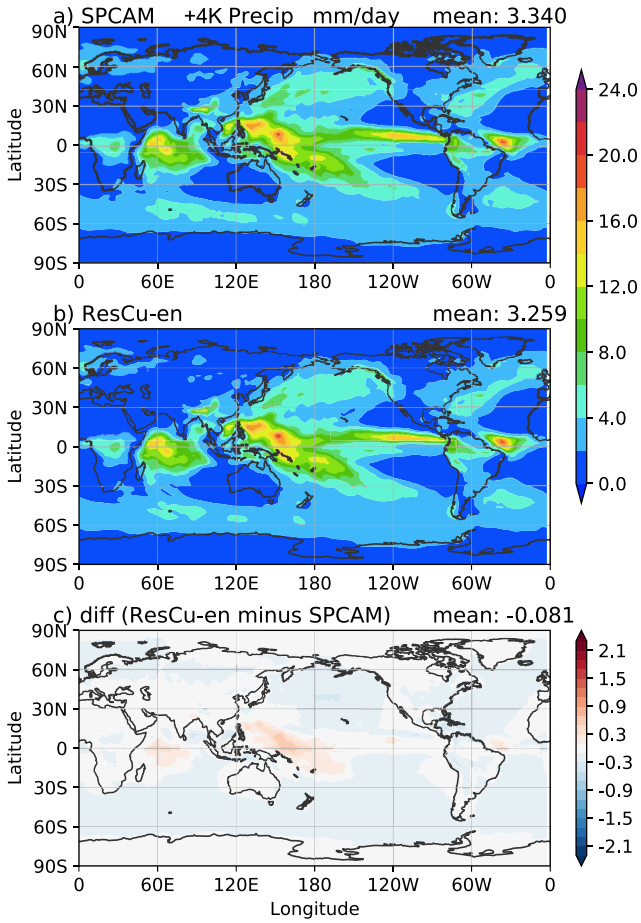

Figure 5. Same as Figure 2 but for the warm climate with 4K increase in SST.

with R^2 greater than 0.9 (Figure 7a). Some areas in tropical and subtropical oceans and land regions have low R^2 values, especially in subtropical eastern Pacific and Atlantic, and to some extent in the central equatorial Pacific and the Sahara Desert. All these low R^2 regions have low precipitation rates. The R^2 distribution for the +4K SST simulation is similar to that in the baseline simulation, except in the Sahara Desert where the R^2 values are much lower.

To have a more intuitive feel on how well ResCu-en performs in both current and warm climates, we compare the precipitation time series from ResCu-en with those from SPCAM at two representative model grid points. We select one grid point in the ITCZ region (5°N, 180°E) where R^2 is about 0.8 and another in the subtropical southeastern Pacific (20°S, 90°W) where R^2 is below 0.5. For a one-month-long precipitation time series (Figure 8), ResCu-en can reproduce the timing and magnitude of the heavy rainfall at the ITCZ grid point extremely well for both the baseline and +4K SST simulations (Figures 8a and 8b). For the southeastern Pacific grid point with low rainfall rates, ResCu-en generally underestimates the peak rainfall rates, but it can still capture the timing accurately for both the baseline and +4K SST simulations despite the low R^2 values (Figures 8c and 8d).

All these results from the +4K SST simulation demonstrate that ResCu-en is capable of generalizing to a warmer climate with remarkable accuracy. In the next subsection, we will investigate what properties of our ResCu-en are responsible for this.

4.2. Why Is ResCu-en Able to Generalize to a Different Climate?

The ability of a neural network to generalize to a different climate is an important attribute as it can then be used in global warming simulations. In this subsection, we investigate what attributes of ResCu-en make it generalizable to a warmer climate by testing different input variables and NN constructs. In doing so, we note that each member of the NN ensemble, when applied individually to the +4K SST SPCAM simulation offline is also able to reproduce the SPCAM results well (Figure S6 in Supporting Information S1). Thus, the use of an ensemble of NNs is not among the factors responsible for ResCu-en's ability to generalize to a warmer climate. As such, we will use a single member of ResCu-en for this purpose.

4.2.1. Convective Memory

The first factor we examine is convective memory since it is unique to ResCu. All NN's developed by other researchers use current atmospheric state variables as input to their NNs. To this end, we developed an NN, ResCu-t0-ls, using only the current step temperature, humidity states and advective forcings as inputs. We also trained two deep fully connected neural networks: DNN-mem, which uses all the input variables as in ResCu, and DNN-t0-ls, which uses only the current step states and forcings. DNN-mem has 10 layers of 512 nodes, the same as DNN-t0-ls. Table 1 lists the NN training experiments used in both this and next subsections. We train all three neural networks (ResCu-t0-ls, DNN-mem, and DNN-t0-ls) on one subset of the data described in Section 2 and evaluate them on 1-year independent data sets from both the current and +4K warm climates, as described in Section 3. We measure the accuracy of the NN predictions using R^2 of the zonally averaged diabatic heating, which is a frequently used metric in previous studies (Gentine et al., 2018; Mooers et al., 2021; Wang et al., 2022). These experiments allow us to evaluate the roles of convective memory and architecture on ResCu's generalization capability.

Figure 9 shows R^2 for moist diabatic heating from ResCu-en, ResCu-t0-ls, DNN-mem and DNN-t0-ls for the baseline and +4K SST climate, respectively. ResCu-en demonstrates remarkable generalization capability, with almost no drop in accuracy from the current climate to the +4K SST warm climate (Figures 9a and 9b), consistent with Figures 3 and 6. Without convective memory (ResCu-t0-ls), the NN is less accurate in the entire troposphere over the tropics and subtropics compared to ResCu-en for the current climate (Figure 9a vs. Figure 9c). There is

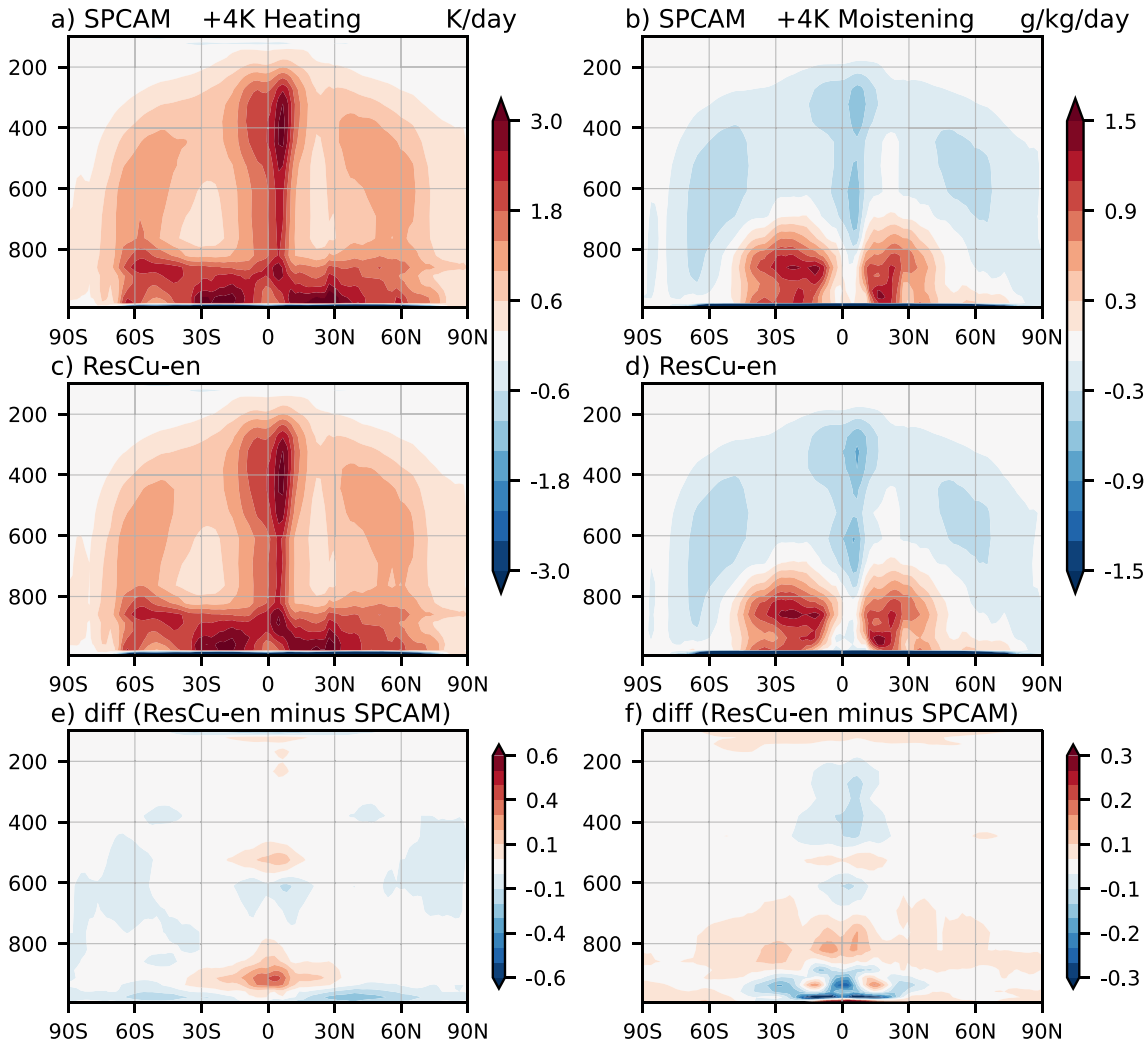

Figure 6. Same as Figure 3 but for the warm climate with 4K increase in SST.

noticeable deterioration in R^2 (Figures 9c and 9d) from the baseline to +4K SST climate in the tropical lower and mid-troposphere. When fully connected NNs are used, DNN-mem performs well in both climates (Figures 9c and 9d), while DNN-t0-ls experiences a significant accuracy drop in the tropical mid and lower troposphere in the warm climate (Figures 9g and 9h), even more so than ResCu-t0-ls. Note that DNN-t0-ls is a fully connected NN with current time step variables as input. It is similar to the NN used in Beucler et al. (2021b) without physical rescaling. Consistent with their findings, the generalizability to a different climate is poor (Figure 9g vs. Figure 9h). The use of convective memory as input alleviates this deficiency markedly, and the use of residual convolution neural net further improves the accuracy and generalizability in the absence of convective memory as input (compare Figures 9c and 9d and Figures 9g and 9h).

4.2.2. NN Architectures

In this subsection, we further explore the impact of different neural network architectures within the framework of ResCu-en on the generalization capability of the NN-based parameterization. We present 4 different NNs with different combinations of the 3 architectures (1D convolution, residual shortcuts, and batch-normalization): ResCu, the first member of ResCu-en, with all three architectures, ResCNN with 1D convolution and residual shortcut, but no batch-normalization, CNN with 1D convolution, but neither residual shortcuts nor batch normalization, and ResDNN, a residual fully connected neural network with no batch normalization, in which all 1D convolution layers in ResCNN are replaced with fully connected layers (Table 1). All four NNs use the same input and output variables as in ResCu-en, and we evaluate their generalization capability in the same way as in Section 4.2.1.

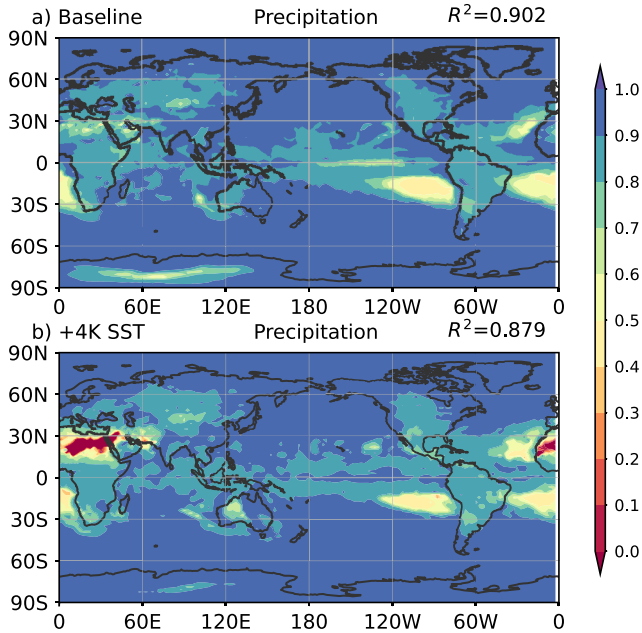

Figure 7. Global distribution of coefficient of determination (R^2) for precipitation predictions by ResCu-en under different climates: (a) the baseline climate for the present-day (b) the warm climate with 4K increase in SST. R^2 is calculated using $R^2 = 1 - \frac{\text{MSE}}{\text{var}}$, where MSE is the mean squared error between ResCu-en predictions and SPCAM targets and var is the variance of the SPCAM targets.

Figure 10 shows R^2 of moist diabatic heating for the NN architectures described above for both the current climate and the +4K SST warm climate. ResCu, ResCNN, and CNN are all able to generalize well to the warm climate. The 1D convolutional layer is the shared architecture in all three, suggesting that the 1D convolution layer plays a major role in the generalization capability of the NN. This is further demonstrated by comparing ResCNN and ResDNN (Figures 10e and 10f vs. Figures 10g and 10h). These results are consistent with the work of Molina et al. (2021) who found that convolutional neural networks have a better generalization capability. Without the convolutional layers, ResDNN has noticeable degradation in R^2 from the current climate to +4K SST climate in tropical mid-troposphere. Batch normalization (compare ResCu and ResCNN) does not affect much the generalization capability of ResCu. While residual shortcuts help improve the prediction accuracy of ResCu for both current and warm climates, their impact on the generalizability of ResCu is not significant.

It is noted that while the degradation of R^2 for DNN-t0-ls is substantial going from the current climate to a warmer climate (e.g., Figures 9g and 9h), it is not as drastic as reported in Beucler et al. (2021b). Out of curiosity, we conducted three additional tests on DNNs using the most basic input variables: temperature (T), specific humidity (q_v), and surface sensible and latent heat fluxes. We have DNN-10 with 10 layers of 512 nodes, which is as deep as a DNN can go and much wider than a typical DNN is, DNN-7 with 7 layers of 128 nodes, which has the same NN architecture as in Beucler et al. (2021b) without physical rescaling, and DNN-7-nc without the moist static energy conservation penalty in the loss function on the basis of DNN-7 (Table 1, bottom three rows). Since the DNN with no rescaling in Beucler et al. (2021b) did not have energy conservation constraints, it is the closest to DNN-7-nc here. All DNNs with the basic input variables perform reasonably well in the current climate (Figure 11). However, for the warmer climate, DNN-7-nc has the poorest generalization capability throughout the entire tropical troposphere, to a similar extent to that reported in Beucler et al. (2021b) for their NN without rescaling. DNN-7 with the MSE conservation penalty recaptures some accuracy in the tropical upper troposphere. The generalizability of DNN-10 is further improved. Therefore, within the DNN architecture, a wider and deeper neural network and the use of moist static energy conservation in the loss function contribute to the generalization capability.

In summary, ResCu-en is capable of generalizing to a warm climate. When evaluated in the +4K SST warm climate that is not included in the training data, ResCu-en successfully predicts the global precipitation distribution and heating and moistening by moist physics processes with high accuracy. Higher order statistics of precipitation, such as intensity increase and occurrence frequency shift toward heavier precipitation over oceans simulated by SPCAM in the +4K SST simulation are also captured by ResCu-en.

The use of convective memory as input is the most important attribute to the generalization capability of ResCu-en to a warmer climate. The 1D convolutional layers further boost its warm-climate generalizability. The residual shortcuts also help improve the generalizability of ResCu-en, while the benefit of batch normalization is not noticeable. For fully connected neural networks, while the generalizability is poor, relatively speaking, the use of moist static energy conservation has the most impact on improving the DNN's generalizability. A wider and deeper net also improves it.

5. Stable Online Integration

The ultimate test of an NN-based parameterization is its performance in online GCM integration. Attempts from past studies to make online integrations stable using their NN-based parameterizations were not successful until recently, particularly in real land-ocean geography GCMs (Bretherton et al., 2022; Clark et al., 2022; Wang et al., 2022). Wang et al. (2022) emulated the moist physics and radiation processes in SPCAM with a group of deep neural networks, each for a different process. They succeeded in a 5-year online integration through trial and error; some were successful in stable integration and others failed. No definitive answer was offered to explain

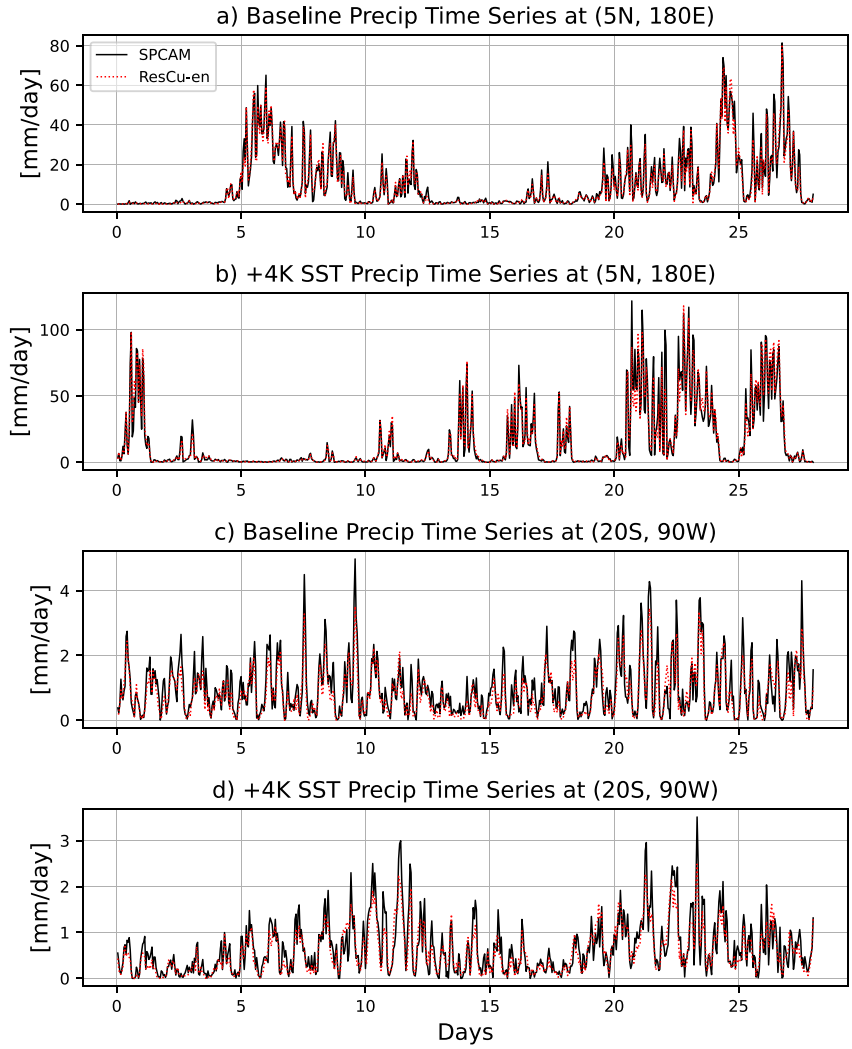

Figure 8. Time series of precipitation from SPCAM simulations (black lines) and ResCu-en predictions (red dotted lines) at selected locations under (a and c) the baseline climate and (b and d) the warm climate. A grid point in the northern ITCZ region is selected (a and b) for heavy precipitation where R^2 is high and a grid point in the southeastern Pacific stratus region is selected (c and d) for light precipitation where R^2 is low.

this different model integration behavior though. Bretherton et al. (2022) took a different approach by learning the nudging tendencies as functions of the atmospheric state and then using these tendencies to correct the physical parameterization biases in a NOAA global weather forecasting model. They were able to integrate the model for 40 days. Clark et al. (2022) extended this work and were able to integrate the model for more than 5 years and for different climates. They applied input ablation and output tapering for the top 25 model levels (levels above ~200 hPa) to maintain stability and to prevent the model from drifting. In this section, we implement our neural network into the NCAR CAM5. The main objective is to demonstrate its ability to perform long-term stable integration consistently.

5.1. The Implementation of ResCu

Due to computational cost (see details below), we only implement one member of ResCu-en into CAM5 (ResCu, i.e., member NN-1 of ResCu-en) instead of the ensemble of 8 members in this initial exploratory online implementation. ResCu replaces the moist diabatic heating and drying and cloud liquid and ice water contents from the conventional parameterization schemes for moist physical processes, including deep convection, shallow convection, and microphysics. The conventional cloud parameterization schemes are still used to provide

Table 1
List and Description of Neural Networks Used in the Offline Generalization Test

	1D convolution layers	Fully connected layers	Residual shortcuts	Batch-normalization	Convective memory inputs	MSE penalty in loss function
ResCu-t0-ls	32 layers of 128 3 × 1 kernels	No	Yes	Yes	No	Yes
DNN-mem	No	10 layers of 512 nodes	No	No	Yes	Yes
DNN-t0-ls	No	10 layers of 512 nodes	No	No	No	Yes
ResCu	32 layers of 128 3 × 1 kernels	No	Yes	Yes	Yes	Yes
ResCNN	32 layers of 128 3 × 1 kernels	No	Yes	No	Yes	Yes
CNN	32 layers of 128 3 × 1 kernels	No	No	No	Yes	Yes
ResDNN	No	32 layers of 512 nodes	Yes	No	Yes	Yes
DNN-10	No	10 layers of 512 nodes	No	No	No	Yes
DNN-7	No	7 layers of 128 nodes	No	No	No	Yes
DNN-7-nc	No	7 layers of 128 nodes	No	No	No	No

Note. The configurations include 1D convolutional layers, fully connected layers, residual shortcuts, batch-normalization, convective memory as inputs, and moist static energy conservation penalty in the loss function.

quantities that are not predicted by our neural network but needed by the radiative transfer scheme in CAM5, such as cloud liquid and ice number concentrations and cloud fraction. We refer to this configuration as NCAM.

Before going into the online model integration, we should point out two technical details of the implementation of ResCu into CAM5. First, recall that ResCu includes convective memory as input. In the training and offline test of ResCu, moist physics heating and drying as well as cloud water and ice contents at two previous time steps, as part of convective memory, were taken from SPCAM. In online model integration, no such "ground truth" is available for representing convective memory. A natural substitute for them would be the predicted values at the previous time steps by the NN itself. This approximation will lead to some degradation in accuracy because the neural net is trained on SPCAM data. To estimate the impact of this approximation, we use the same trained neural network ResCu and test it offline using SPCAM data, except replacing the SPCAM values

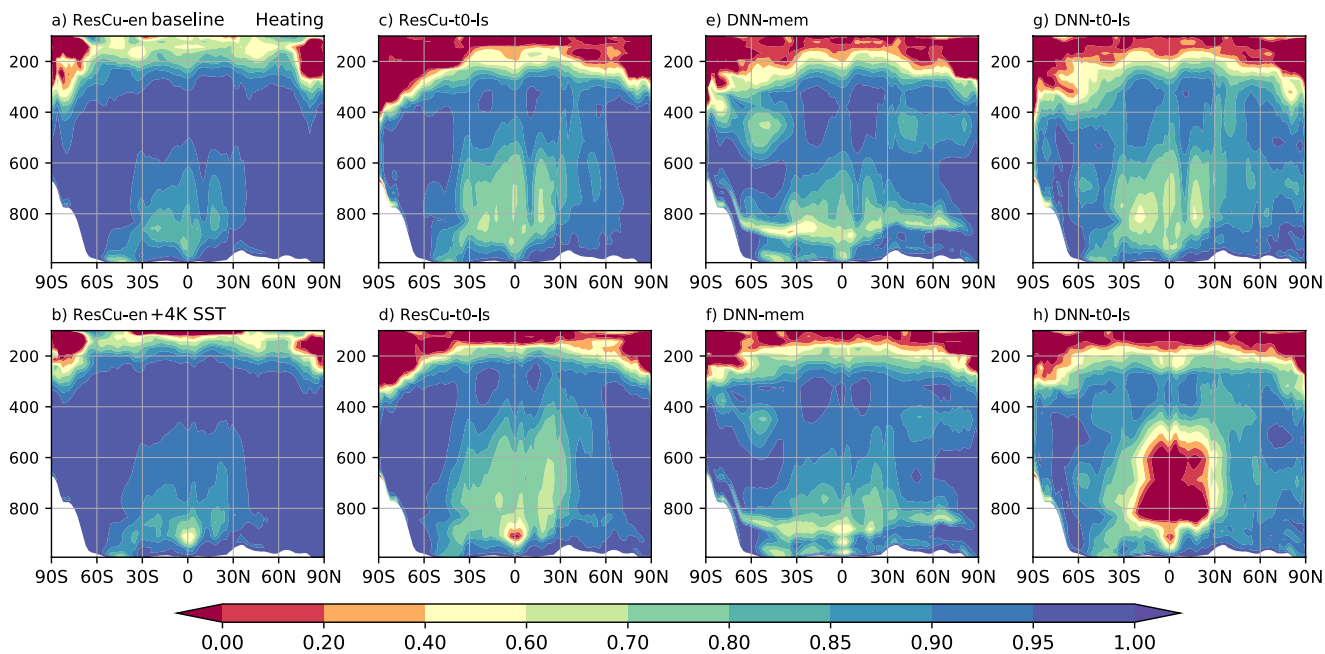

Figure 9. The coefficient of determination (R^2) of the zonally averaged heating for different NNs in the current climate (top row) and the +4K SST warm climate (bottom row): (a and b) ResCu-en, (c and d) ResCu-t0-ls, (e and f) DNN-mem, and DNN-t0-ls (g and h). Note that ResCu-en and DNN-mem are trained with full input variables including convective memory, while ResCu-t0-ls and ResCu-t0-ls are only trained on input variables T, q_v, $dT_{l.s.}$, and $dq_{vl.s.}$ at current timestep.

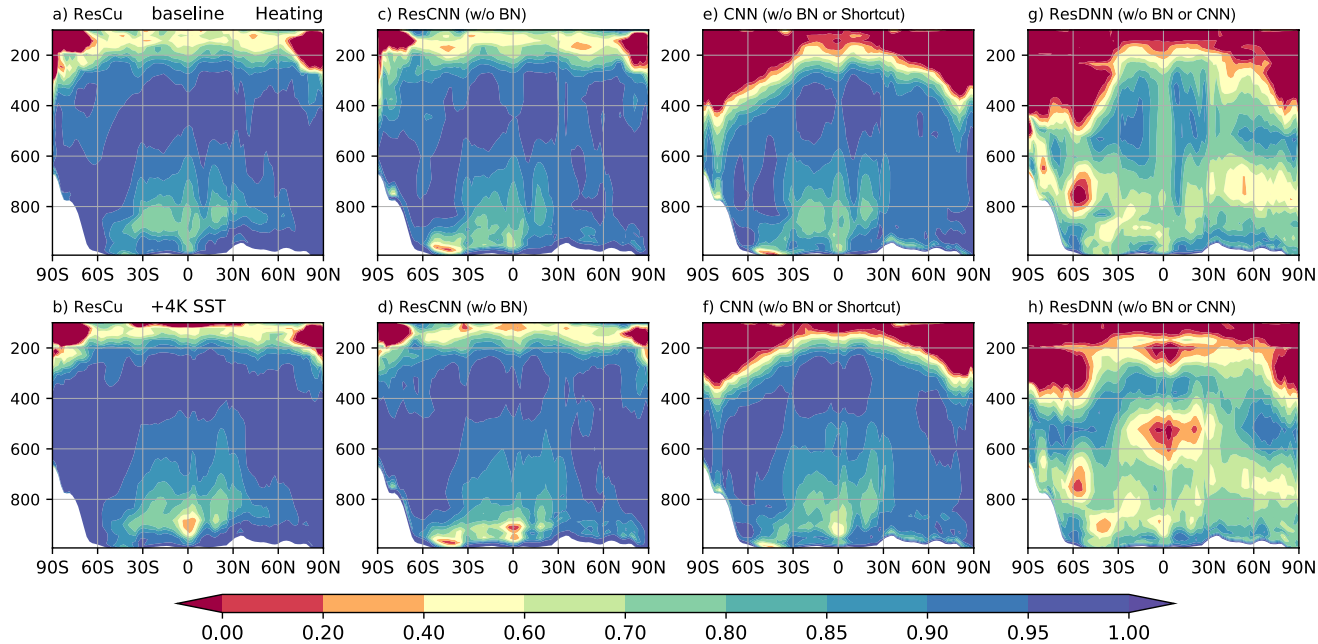

Figure 10. Same as Figure 9 but for NNs trained on full input variables including convective memory: (a and b) ResCu, (c and d) ResCNN, which is ResCu without batch normalization, (e and f) CNN, which is ResCu without batch normalization or residual shortcuts, and (i and j) ResDNN, which is ResCu without batch normalization or convolution layers.

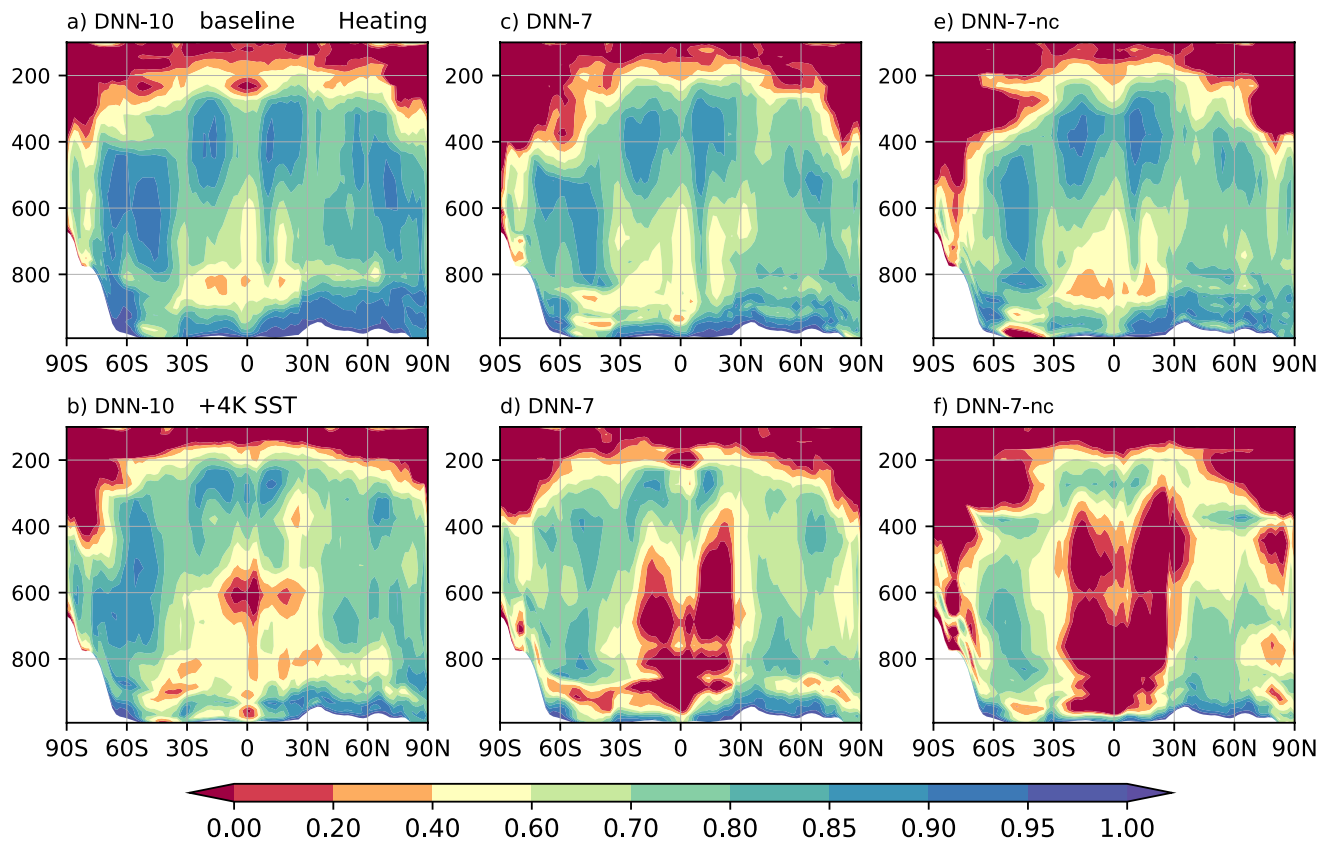

Figure 11. Same as Figure 9 but for fully connected NNs only trained on input variables T and q_v at current timestep: (a and b) DNN-10, (c and d) DNN-7, and (e and f) DNN-7-nc without moist static energy conservation.

with ResCu-predicted values at past time steps for convective memory. Figure S7 in Supporting Information S1 shows that there is some degradation in NN-predicted precipitation in the Intertropical Convergence Zone, by up to 1.5 mm/day locally. While this is a significant increase in prediction biases, compared to the differences between typical GCM simulations and observations, which are often as much as 3–4 mm/day in tropical oceans (Kooperman et al., 2016; Rasch et al., 2019; Xie et al., 2012), this difference is still small. Therefore, in our implementation of ResCu into CAM5, we use the ResCu-predicted values at past time steps for convective memory.

Second, similar to Brenowitz and Bretherton (2019) and Clark et al. (2022), we ablate the heating and drying rates from moist physics above the CAM5 model level close to 120 hPa from the NN. The reason for doing so is that near the tropopause and above moist heating and drying values in GCMs (and the real world too) are very small due to low moisture content. Although the NN also predicts small values, the relative errors are large, as can be seen from the low R^2 values in the last section. In our initial tests without ablating the heating and drying tendencies above 120 hPa, these errors cause the model integration to drift due to their effects on radiation although the integration remains stable.

For computational cost, using 200 intel CPU cores, NCAM with a 32-layer deep neural net can reach 3.8 simulation years per day (SYPD), which is 10 times faster than SPCAM (0.37 SYPD), but 6 times slower than the default CAM5 (23.5 SYPD). This computational speed can be improved in the future since the Fortran implementation of the neural network, which contains excessive use of loops, has not been optimized.

5.2. Online Simulation Results

In addition to NCAM, we also run the standard CAM5 and SPCAM for the same period for comparison to put NCAM simulation in context. We succeeded in conducting a 5-year NCAM stable simulation from January 1998 to December 2002 without encountering any integration instability, with no systematic drift in global mean total energy (Figure S8a in Supporting Information S1) and precipitable water (Figure S8b in Supporting Information S1), although there are some systematic biases. We also tested all other 7 members of the ResCu-en ensemble for a shorter period, and none experienced any integration instability either. Figure 12 shows the 5-year average boreal summer (June-July-August) and winter (December-January-February) precipitation for TRMM 3B42 observations, SPCAM, NCAM, and CAM5, respectively. Comparing against TRMM observations, NCAM can capture the major features in precipitation distribution including the ITCZ and the South Pacific Convergence Zone (SPCZ) in the tropics and midlatitude storm tracks. Interestingly, in the western Pacific warm pool region both SPCAM and CAM5 underestimate the precipitation in JJA, a well-known problem in the NCAR model, while NCAM simulation is much better. However, it underestimates precipitation over tropical land compared to both TRMM observations and SPCAM/CAM5 simulations in JJA and DJF.

Although the simulated precipitation in NCAM is realistic, the simulated temperature and moisture in high latitudes have much larger biases than those in CAM5 when compared against ECMWF Reanalysis - Interim (ERA-Interim) (Dee et al., 2011) (Figure 13; Figure S9 in Supporting Information S1). These high latitude biases are probably caused by inadequate representation of cloud-radiation interaction due to inconsistencies between NN-based parameterization and conventional cloud microphysics and macrophysics parameterizations. For instance, cloud fraction is parameterized by a conventional macrophysics scheme. Cloud water and ice number concentrations as well as snow mass and number concentrations are parameterized by the conventional Morrison and Gettelman (2008) cloud microphysics scheme. The mismatch between cloud ice and water contents from ResCu and their number concentrations from conventional microphysics scheme will affect cloud droplet and ice crystal sizes, thereby affecting cloud-radiation interaction. These issues show that there is still a long way to go before NN-based parameterization can replace the conventional physics parameterization schemes.

In offline validation, we showed that ResCu-en predicts the pdf of precipitation intensity extremely well (Figure 4). In the online simulation, the ResCu-predicted precipitation pdf is not as close to that from SPCAM, as shown in Figure 14 for tropical oceans. However, SPCAM itself underestimates the occurrence frequency of precipitation intensity greater than 50 mm/day compared to TRMM observations. In this regard, the ResCu-predicted precipitation pdf is actually closer to TRMM observations, especially for high intensity precipitation greater than ∼70 mm/day. In comparison, CAM5 shows the well-known "too much light rain and too little heavy rain" problem (Wang et al., 2016).

The diurnal cycle of precipitation is another rainfall characteristic that is a long-standing challenge in GCMs with conventional parameterizations (Cui et al., 2021; Dai, 2006; Zhang, 2003). The diurnal cycle of rainfall

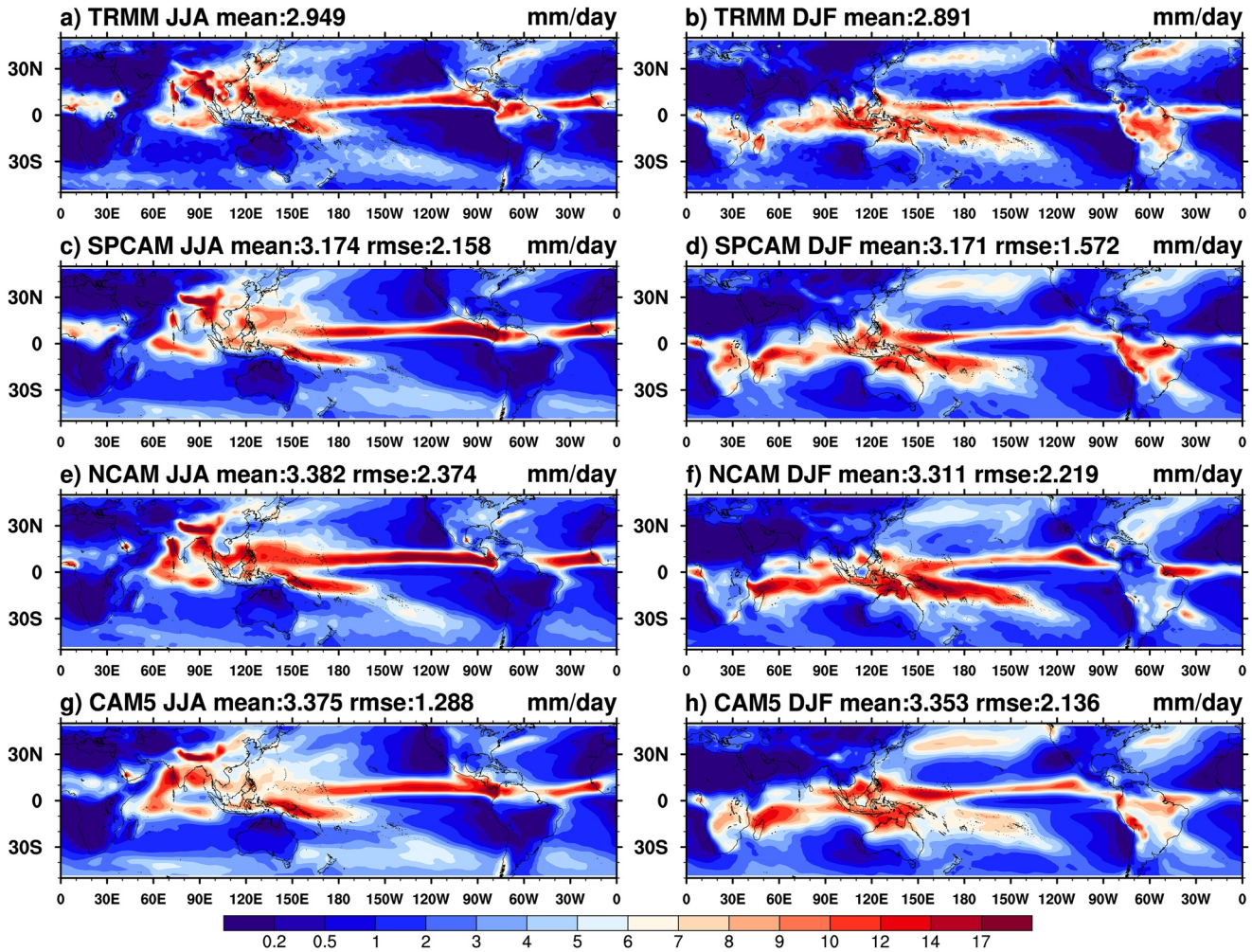

Figure 12. Global distribution of the mean precipitation rate (mm/day) in June-July-August (the left panels) and December-January-February (the right panels) over the years of 1998–2002 for (a and b) TRMM 3B42, (c and d) NCAM, and (e and f) CAM5. The spatial mean and root mean square error to the TRMM 3B42 observations are shown above each frame.

is characterized by the local solar time (LST) of maximum precipitation of the day and the amplitude within the diurnal cycle. We calculate the annually averaged diurnal cycle of rainfall at every grid point globally and then find the LST of the maximum rainfall rate in the day and regard the difference between the maximum and minimum rainfall rate as the amplitude. Figure 15 shows the warm season average (June-July-August for northern hemisphere and December-January-February for southern hemisphere) rainfall diurnal cycle between 45°S and 45°N from the 3h TRMM 3B42 observation and hourly output from SPCAM, NCAM, and CAM5 simulations, respectively. In CAM5, as in many other GCMs, the simulated warm-season precipitation peaks 4–6 hr earlier than observations over land and 2–4 hr earlier over oceans (Dai, 2006), as shown in Figures 15a and 15d. SPCAM only manages to mitigate this delay in some ocean areas (visually below 40%), while similar effects are not observed over land (Figure 15b). However, NCAM alleviates the early precipitation problem remarkably by delaying the peak time by 2–4 hr over tropical land areas and by 2 hr over 50% of the ocean areas (Figure 15c). Moreover, the amplitude of the diurnal cycle over land in most models is weak compared with observations (Dai, 2006; Xie et al., 2019). In CAM5, the amplitude over tropical land area is less than half of that in TRMM observations (Figures 15e and 15h). Both SPCAM and NCAM increase the amplitude by a factor of 2 in many tropical land regions (Figures 15f and 15g).

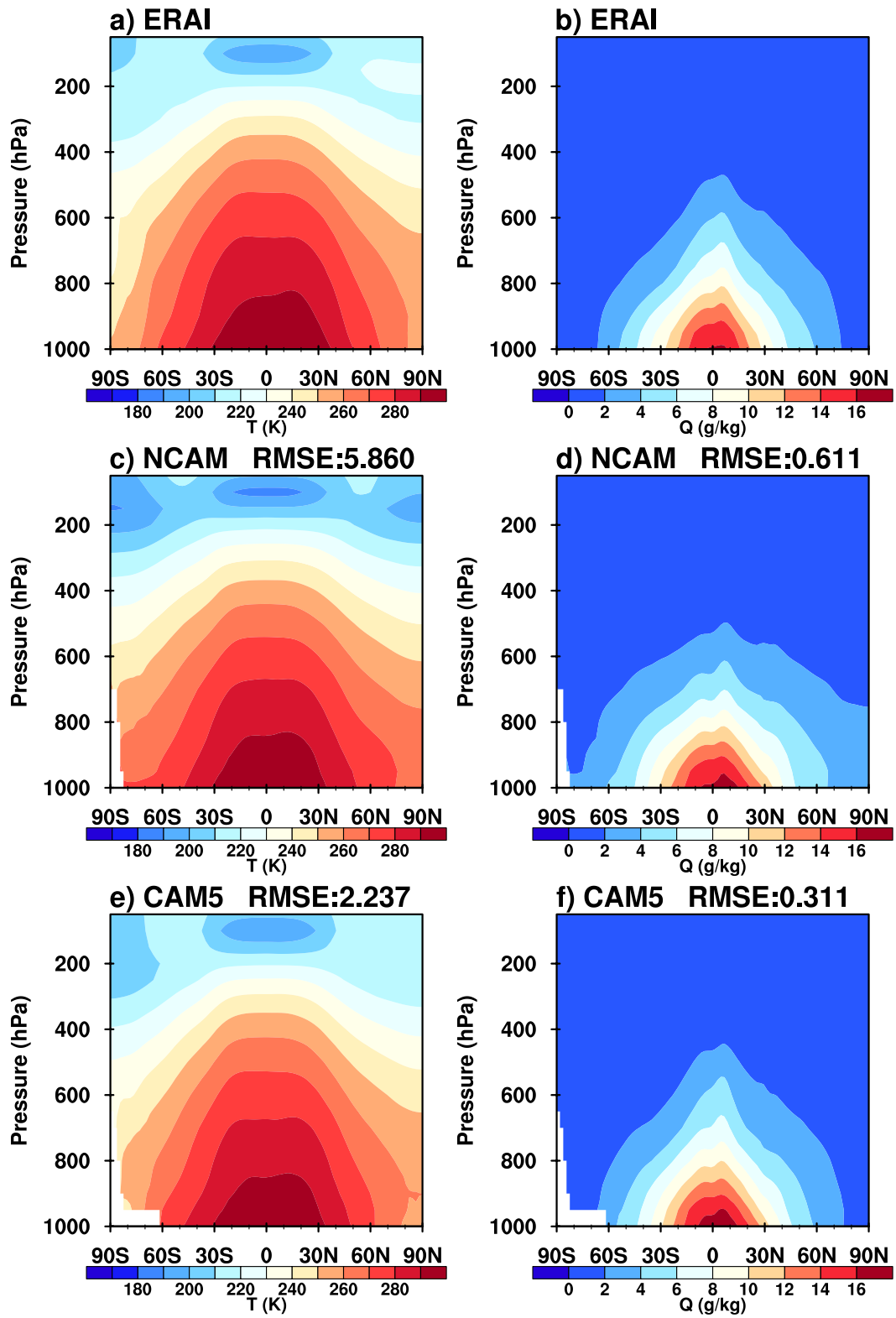

Figure 13. Latitude-pressure cross sections of annual and zonal average temperature (left) and specific humidity (right) over years 1998–2002 for (a and b) ERA-Interim, (c and d) NCAM, and (e and f) CAM5. The root mean square error to ERA-Interim reanalysis is shown above each frame from (c) to (f).

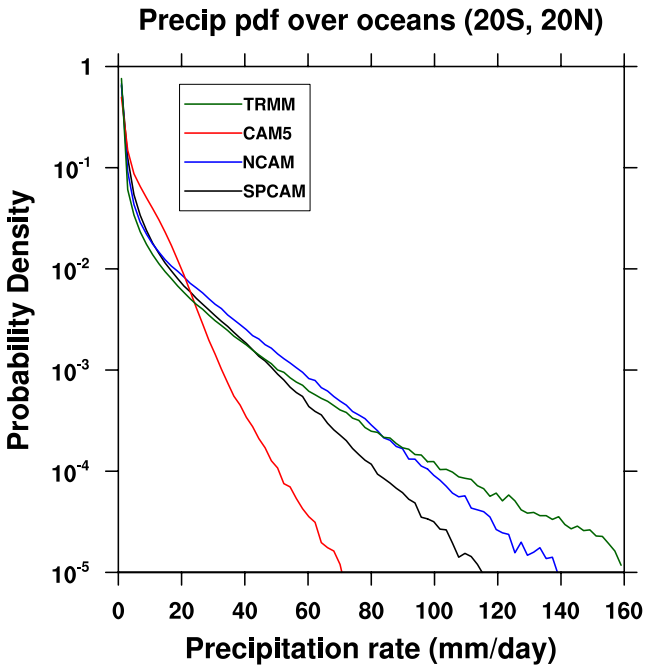

Figure 14. Probability density distribution of the daily mean precipitation in the tropics (20°S−20°N) over oceans from the three model simulations and the TRMM 3B42 daily product. The black, blue, red and green lines are for SPCAM, NNCAM, CAM5, and TRMM 3B42, respectively.

6. Summary and Discussions

This study extends the work of Han et al. (2020) by using an ensemble of 8 neural networks (ResCu-en) to account for the random errors inherent in the NN configuration. The depth of the NN is also increased from 22 layers to 32 layers to improve the accuracy of the predictions, with a batch-normalization layer added after each convolution layer for more robustness. The sampling strategy of the training data is also improved by selecting 800 model grid columns randomly over the globe every day in the SPCAM simulation instead of taking data from 800 fixed model grid columns distributed over the globe. Therefore, ResCu-en has a stronger nonlinear fitting capability from more layers, with reduced uncertainties from the ensemble mean.

In the independent offline test, ResCu-en reproduces all four output variables and the derived precipitation with smaller biases and higher R^2 than ResCu in H20. ResCu-en can also reproduce accurately the SPCAM's rainfall frequency distribution. To assess the ability of ResCu-en trained on data from current climate to emulate convection in a warmer climate, we evaluated ResCu-en in a +4K SST simulation using SPCAM. ResCu-en has an excellent generalizability to a warmer climate when tested offline, with performance in predictions in a warmer climate comparable to that for the current climate. It reproduces very well the precipitation intensity increase and the occurrence frequency shift toward heavier rainfall over oceans.

To understand what factors contribute to the strong generalization capability to warm climates, we conducted a series of tests on input variables and NN architectures that are used in ResCu-en. It is found that the use of convective memory as input variables is the most important factor contributing to ResCu-en's generalization capability, as demonstrated in both deep convolutional NNs and fully connected NNs. Under the framework of ResCu-en, among the NN architectures we tested, 1D convolutional layers were found to be the most important, while residual shortcuts improved accuracy in both current and warm climates. On the other hand, batch normalization did not have a significant impact on generalization. Fully connected NNs performed relatively poorly on generalization, but their performance can be improved by deepening and widening the NNs or by adding a moist static energy conservation penalty to the loss function.

The success of ResCu-en highlights the importance of incorporating convective memory into machine-learning-based parameterization schemes. Several previous studies have noticed the role of convective memory in the prediction of convection. They range from simple theories and toy model simulations (Colin & Sherwood, 2021; Davies et al., 2009) to detailed simulations with CRMs (Colin et al., 2019; Muller & Bony, 2015). In an NN-based parameterization, Shamekh et al. (2023) explored the impact of convective organization and memory on precipitation intensities and extremes.

We further implemented a member of ResCu-en into CAM5 with real-world geography, referred to as NCAM, and ran it successfully for 5 years without encountering any model integration stability issue. The simulated 5-year mean precipitation captures the major features of the global precipitation distribution, including the ITCZ in the tropics and the storm tracks in midlatitudes. However, NCAM underestimates precipitation over land and has large biases in temperature and moisture in high latitudes.

NCAM also produces a frequency distribution of precipitation intensity that is closer to TRMM observations than CAM5, with significantly less bias in underestimating heavy precipitation. Additionally, NCAM improves the diurnal cycle of precipitation in CAM5 by delaying the peak time and increasing the diurnal amplitude. All these online simulation results show that our NN-based parameterization is promising for use in future simulations for both current climate and future climate projection studies. However, before this is possible, we must address the issues of large biases in high latitudes, which were also noticed in previous studies. Furthermore, conventional convection and cloud parameterization schemes output many more parameters that NN-based parameterizations do not provide, such as cloud droplet and ice crystal numbers, as well as snow and graupel mass and number concentrations. These variables are needed for aerosol-cloud-radiation interactions and cloud feedbacks that are

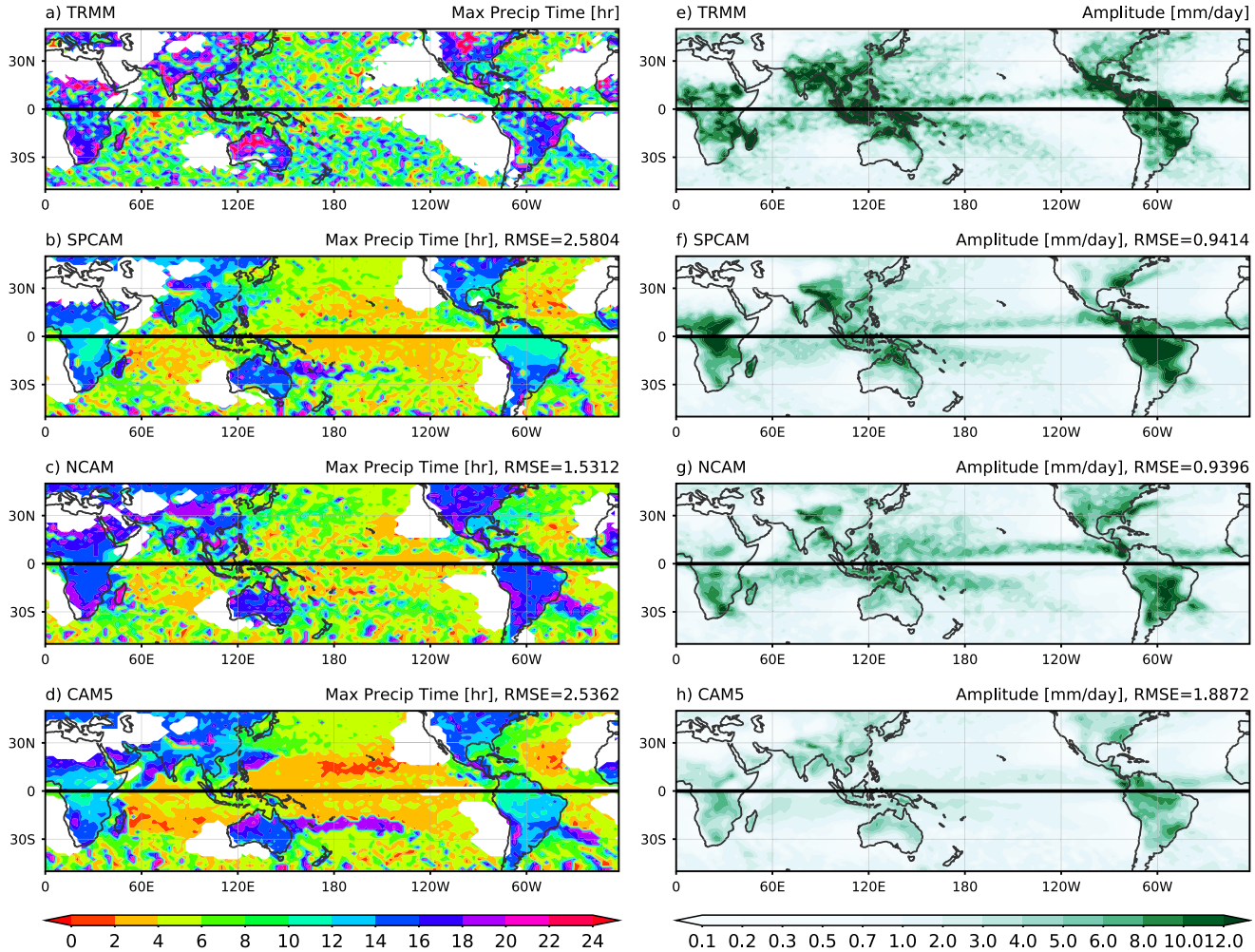

Figure 15. Tropical distribution of warm season averaged diurnal peak time (a)–(d) and amplitude (e)–(h) of the diurnal cycle of precipitation (mm day^{-1}) derived from observations from hourly data of (a and e) TRMM 3B42, (b and f) SPCAM, (c and g) NCAM, and (d and h) CAM5. In (a)–(c), areas with precipitation less than 1 mm/day are masked. The warm season is defined as June-July-August (JJA) for Northern Hemisphere and December–January-February (DJF) for Southern Hemisphere, respectively. The thick black line marks the equator, where the warm season is undefined.

fundamentally important for climate projection studies. Therefore, much more work is needed before it is feasible for NN-based parameterizations to fully replace physically based parameterizations in GCMs.

Acknowledgments
YH and YW were supported by the National Key Research and Development Program of China Grant 2022YFF0802002 and the Tsinghua University Initiative Scientific Research Program Grant 20223080041. GJZ was supported by the U.S. Department of Energy (DOE) Office of Science, Biological and Environmental Research Program (BER) under Award Number DE-SC0022064, the DOE Office of Science's Scientific Discovery through Advanced Computing (SciDAC) program via a partnership in Earth System Model Development between BER and Advanced Scientific Computing Research (ASCR) programs, and the National Science Foundation Grant AGS-2054697.

Data Availability Statement

The data and neural network codes used in this study (Han et al., 2023) are available via Creative CC-BY-4.0 license from the public data repository (https://doi.org/10.5281/zenodo.8228408). *Software Availability Statement*: ResCu-en is developed based on TensorFlow v2.3 (Abadi et al., 2015), available from https://github.com/tensorflow/tensorflow/releases/tag/v2.3.0; 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).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). Tensorflow: Large-scale machine learning on heterogeneous distributed systems [Software]. https://www.tensorflow.org/

Arakawa, A., & Schubert, W. H. (1974). Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *Journal of the Atmospheric Sciences*, *31*(3), 674–701. https://doi.org/10.1175/1520-0469%281974%29031%3C0674%3AIOACCE%3E2.0.CO%3B2

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J., Beljaars, A., & Bormann, N. (2014). Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, *71*(2), 734–753. https://doi.org/10.1175/JAS-D-13-0163.1

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

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021b). Climate-invariant machine learning. *arXiv preprint arXiv:2112.08440*.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018GL078510

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

Brenowitz, N. D. A. H., Brian, Clark, S., Kwa, A., McGibbon, J., Perkins, W. A., et al. (2020). Machine learning climate model dynamics: Offline versus online performance. In *Paper presented at the NeurIPS 2020 workshop on tackling climate change with machine learning*. Retrieved from https://www.climatechange.ai/papers/neurips2020/50

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

Cao, G., & Zhang, G. J. (2017). Role of vertical structure of convective heating in MJO simulation in NCAR CAM5.3. *Journal of Climate*, *30*(18), 7423–7439. https://doi.org/10.1175/JCLI-D-16-0913.1

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., Geoffroy, O., Bony, S., & Fuchs, D. (2019). Identifying the sources of convective memory in cloud-resolving simulations. *Journal of the Atmospheric Sciences*, *76*(3), 947–962. https://doi.org/10.1175/JAS-D-18-0036.1

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

Cui, Z., Zhang, G. J., Wang, Y., & Xie, S. (2021). Understanding the roles of convective trigger functions in the diurnal cycle of precipitation in the NCAR CAM5. *Journal of Climate*, *34*(15), 6473–6489. https://doi.org/10.1175/jcli-d-20-0699.1

Dai, A. (2006). Precipitation characteristics in eighteen coupled climate models. *Journal of Climate*, *19*(18), 4605–4630. https://doi.org/10.1175/jcli3884.1

Davies, L., Plant, R. S., & Derbyshire, S. H. (2009). A simple model of convection with memory. *Journal of Geophysical Research*, *114*(D17), D17202. https://doi.org/10.1029/2008JD011653

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al. (2021). A survey of uncertainty in deep neural networks. *ArXiv, abs/2107.03342*.

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

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). The data and codes for training, testing, and prognostic validation of A ResNet ensemble for moist physics (ResCu-en) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8228408

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

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

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

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

Jiang, X., Waliser, D. E., Xavier, P. K., Petch, J., Klingaman, N. P., Woolnough, S. J., et al. (2015). Vertical structure and physical processes of the Madden-Julian oscillation: Exploring key model physics in climate simulations. *Journal of Geophysical Research: Atmospheres*, *120*(10), 4718–4748. https://doi.org/10.1002/2014JD022375

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*.

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

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, *2013*, 485913. https://doi.org/10.1155/2013/485913

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

Molina, M. J., Gagne, D. J., & Prein, A. F. (2021). A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth and Space Science*, *8*(9), e2020EA001490. https://doi.org/10.1029/2020EA001490

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

Muller, C., & Bony, S. (2015). What favors convective aggregation and why? *Geophysical Research Letters*, *42*(13), 5626–5634. https://doi.org/10.1002/2015GL064260

Neale, R. B., Richter, J. H., & Jochum, M. (2008). The impact of convection on ENSO: From a delayed oscillator to a series of events. *Journal of Climate*, *21*(22), 5904–5924. https://doi.org/10.1175/2008jcli2244.1

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

Pearce, T., Zaki, M., Brintrup, A., & Neely, A. (2018). Uncertainty in neural networks: Bayesian ensembling. *ArXiv, abs/1810.05546*.

Pritchard, M. S., & Somerville, R. C. J. (2009). Empirical orthogonal function analysis of the diurnal cycle of precipitation in a multi-scale climate model. *Geophysical Research Letters*, *36*(5), L05812. https://doi.org/10.1029/2008GL036964

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003). Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547–1564. https://doi.org/10.1175/BAMS-84-11-1547

Rasch, P. J., Xie, S., Ma, P.-L., Lin, W., Wang, H., Tang, Q., et al. (2019). An overview of the atmospheric component of the energy exascale Earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2377–2411. https://doi.org/10.1029/2019MS001629

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

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(20), e2216158120. https://doi.org/10.1073/pnas.2216158120

Song, X., & Zhang, G. J. (2018). The roles of convection parameterization in the formation of double ITCZ syndrome in the NCAR CESM: I. Atmospheric processes. *Journal of Advances in Modeling Earth Systems*, *10*(3), 842–866. https://doi.org/10.1002/2017MS001191

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

Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493%281989%29117%3C1779%3AACMFSF%3E2.0.CO%3B2

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. (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

Xie, S., Ma, H.-Y., Boyle, J. S., Klein, S. A., & Zhang, Y. (2012). On the correspondence between short- and long-time-scale systematic errors in CAM4/CAM5 for the year of tropical convection. *Journal of Climate*, *25*(22), 7937–7955. https://doi.org/10.1175/jcli-d-12-00134.1

Xie, S., Wang, Y.-C., Lin, W., Ma, H.-Y., Tang, Q., Tang, S., et al. (2019). Improved diurnal cycle of precipitation in E3SM with a revised convective triggering function. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2290–2310. https://doi.org/10.1029/2019MS001702

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. (2003). Roles of tropospheric and boundary layer forcing in the diurnal cycle of convection in the U.S. southern great plains. *Geophysical Research Letters*, *30*(24), 2281. https://doi.org/10.1029/2003GL018554

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., & Mu, M. (2005). Simulation of the Madden–Julian oscillation in the NCAR CCM3 using a revised Zhang–McFarlane convection parameterization scheme. *Journal of Climate*, *18*(19), 4046–4064. https://doi.org/10.1175/jcli3508.1

Zhang, G. J., Song, X., & Wang, Y. (2019). The double ITCZ syndrome in GCMs: A coupled feedback problem among convection, clouds, atmospheric and ocean circulations. *Atmospheric Research*, *229*, 255–268. https://doi.org/10.1016/j.atmosres.2019.06.023