Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models with Real-Geography Boundary Conditions

Griffin Mooers[1,2], Michael Pritchard[1], Tom Beucler [1], Jordan Ott [2,1], Galen Yacalis [4], Pierre Baldi [2], Pierre Gentine [3]

[1]Department of Earth System Science, University of California at Irvine, CA, USA
[2]Department of Computer Science, University of California at Irvine, CA, USA
[3]Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA
[4]Department of Mathematics, University of California at Irvine, CA, USA

Key Points:

- Simple neural networks can emulate explicit convection including geographic complexity.
- 70 percent of zonal mean variance explained in the mid-to-upper troposphere; less in high variability regions.
- Deep convection is parameterizable in neural networks locally in space and time.

Corresponding author: Griffin Mooers, gmooers96@gmail.com

Abstract

We explore the potential of simple, feed-forward deep neural networks (DNNs) for emulating cloud superparameterization in realistic geography, using training data from the Super Parameterized Community Atmospheric Model (SPCAM5). Another novel aspect of our work is formal hyperparameter tuning in hundreds of trials to identify a network architecture of optimized skill. Our DNN explains over 70% of variance at the 15-minute sampling scale throughout the mid-to-upper troposphere in the zonal mean where the diurnal cycle drives the convective modes of variability. The spatial relationship between DNN skill and autocorrelation timescale confirms skill is weakest at the tropical, marine boundary layer (where regimes of highest temporal variability dominate). Spectral analysis in the time domain indicates no mode-specificity and good emulation of even the diurnal scale; a close look at the diurnal cycle reveals correct signals of land-sea contrast and vertical structure in the heating and moistening fields, but distortions with precipitation emulation. Sensitivity tests targeting precipitation skill reveal complementary effects of adding positive constraints vs. hyperparameter tuning, motivating the use of both in the future. A first attempt to couple a DNN emulated atmosphere to a land model produces convergent results further suggesting neural network viability. Skill is competitive with recent attempts by sophisticated Residual Neural Network architectures trained on added information, including memory of past states. Our results confirm the parameterizability of convection through machine learning and highlight advantages of casting this problem locally in space and time for accurate emulation and hopefully quick implementation of hybrid climate models.

Plain Language Summary

Machine learning methods have been used to replace parameterizations of atmospheric convection under very idealized scenarios (aquaplanets). The hope is that these machine learning emulators can help power a next generation of climate models with similar accuracy but at a fraction of the computational cost. But important questions are outstanding about how approachable more realistic convection (atmosphere over both land and ocean considered) is to neural networks. Recently, a first attempt at machine learning replicated convection was made under these realistic conditions. But it required a highly specialized neural network as well as memory of the previous behavior of the atmosphere. This design would make synthesising these machine learning emulators to climate models very difficult. Instead, we make a first attempt to simpler neural network to parameterize convection under realistic geography. Our success suggests that coupling machine learning emulators to climate models may be closer than previously thought.

1 Introduction

Although global atmospheric model simulations are increasingly high-resolution, even under optimistic scenarios of enhanced computing performance, physically resolving atmospheric turbulence controlling clouds will likely not be feasible for decades. Current climate model horizontal grid cells are typically 50-100 kilometers wide but the turbulent updrafts controlling cloud formation occur on scales of just tens to hundreds of meters and the microphysical processes regulating convection occur down at the micrometer scale (Schneider et al., 2017; Blossey et al., 2016; Morrison et al., 2020). This discrepancy creates large uncertainties about the precise details of deep convection on cloud feedbacks and climate change (Bony et al., 2015). Multi-scale methods such as embedding thousands of two-dimensional Cloud Resolving Models (CRMs) into General Circulation Model (GCM) grid cells (superpararmeterization) to directly resolve the spatial and temporal progression of moist convection, and more recently explicit kilometric-scale simulation of moist convection has improved the representation of deep convective clouds and the hydrological cycle (Randall et al., 2003; Schneider et al., 2017; G. Li &

Xie, 2012; Christensen et al., 2015; Daleu et al., 2015). These advancements allow models to simulate historically challenging modes of variability like the observed afternoon maxima of deep convection over continents and a more realistic probability distribution of precipitation that captures the tail (Kooperman et al., 2016; Z. Li et al., 2011). However, even the highest resolution global CRM simulations today require some assumption-prone parameterization for microphysics and sub-km turbulence, among other cloud processes Siebesma et al. (2007); Cheng and Xu (2011), although multi-scale algorithms still hold promise for making some of this explicitly tractable (Parishani et al., 2017; Jansson et al., 2019) .

Given these hurdles, using machine learning emulators to replace sub-grid convective physics in coarse-resolution climate models is an area of rapidly increasing interest. Following seminal works including V. Krasnopolsky et al. (2010); V. M. Krasnopolsky et al. (2008); Kisi et al. (2013a), recent breakthroughs from global aqua-planet simulations have provided proof of concept for these hybrid models. Gentine et al. (2018) showed 40-100M samples taken from a zonally symmetric aquaplanet simulation were sufficient to train a 5-10 layer DNN to emulate superparameterized convective heating and moistening in a hold-out validation, with R^2 greater than 0.7 in the midtroposphere. Building on these results, Rasp et al. (2018) demonstrated that a similar DNN could even be run in a prognostic setting, coupled to an advective scheme in the Community Atmosphere Model, thus generating accurate mean climate states and equatorial wave spectra at as low as five percent of the computational cost of actual superparameterization. Recently, O'Gorman and Dwyer (2018); Yuval and O'Gorman (2020a) showed using random forests that similar prognostic success can occur in idealized aquaplanets trained on coarse-grained three-dimensional output, using refined inputs and outputs tailored to the prognostic variables underlying SPCAM's embedded cloud resolving model. Whereas all of the above studies have focused on aquaplanets, skillfully replicating this convection emulation in more complex, realistic settings could help build a replacement for traditional subgrid parameterizations of deep convection in climate models. These subgrid parameterizations have been linked to inaccuracies in current simulations of fundamental climate components including the Inter-Tropical Convergence Zone, precipitation extremes, and atmospheric waves that are reduced by resolving convection explicitly (Voigt et al., 2017; Randall et al., 2003; Hohenegger & Stevens, 2016; Arnold & Randall, 2015).

Yet land will likely be a much more significant hurdle for neural network architectures may be required to achieve competitive emulation under realistic geography. For parameterizations of convective activity over land, the physical complexity of the emulation task is increased because traditional convection parameterizations developed based on the assumption of quasi-equilibrium (Arakawa & Schubert, 1974) no longer validate sufficiently well at faster than diurnal timescales (Donner & Phillips, 2003). At the time of this writing, a first pioneering attempt has been made to fit superparameterized convection in a realistic operational setting, with results suggesting that a sophisticated "residual" neural network architecture He et al. (2015) involving vertical convolution and skip-connections are critical to achieving reasonable fits (Han et al., 2020). In this case a sequential input vector involving state information from previous time steps was critical to emulate moist convection with real geography boundary conditions. This finding question the viability of parameterization that is cast instantaneously in time. If confirmed, this makes the full potential of physically interpretable NN emulators as a tool to advance scientific understanding beyond aquaplanets substantially harder to achieve – implementing a resnet with skip connections and sequential input is a daunting technical challenge, especially once coupling to a host climate model is considered.

In this context, we will explore whether simple, feed-forward DNNs are capable of emulating convection with real geography if enough hyperparameter tuning is deployed during the training process. This builds off the experience of Ott et al. (2020) for aquaplanets in which formal, expansive hyperparameter tuning was identified as essential for

the performance of a DNN to emulate convection. We focus on emulating superparameterization to avoid the ambiguities due to coarse-grained uniformly-resolved cloud resolving model output (Brenowitz & Bretherton, 2018). We readily acknowledge that other methods such as Random Forests (RF) (Yuval & O'Gorman, 2020b) also show success and promise in subgrid convection emulation, but treat this as beyond the scope of this study. Indeed, RFs have some advantages, including automatically respecting physical constraints that are linear in their outputs, such as energy conservation, and positive definite precipitation, which are not native to DNNs (Yuval & O'Gorman, 2020b). However, there are also ways to enforce such constraints in DNNs (Beucler et al., 2019; Kumar et al., 2020), and RFs also come with disadvantages. RFs with deep trees become computationally expensive in ways that make them harder to imagine scaling to especially large datasets and also require some large storage capacity which could be an issue to fully take advantage of Graphics Processing Unit (GPU) infrastructure (Yuval & O'Gorman, 2020b). RFs have also often had difficulty both generalizing out of sample and avoiding problems stemming from overfitting (Genuer et al., 2015; Herrera Cordova et al., 2019). Thus we are especially interested in mapping out the limits of simple DNN emulation.

Our goal is to understand what convective patterns, cycles, and modes of variation in a realistic setup of a superparameterized convection are learnable with a simple DNN. We additionally aim to establish a set of post-processing metrics to benchmark our own neural network's performance but also as a basis for transparent assessments for comparing different neural network emulators trained on similar data. Section 2 outlines the details of our training dataset below and introduce the design of a neural network, and automated hyperparameter tuning algorithm, capable of finding a reasonable fit. In Section 2.3, we lay out our validation benchmarks we use, some of which are common in the merging machine learning emulation/parameterization literature, and some which have not been discussed before. In Section 3 we go over the spatial and temporal details of our neural network predictions for parameterized convective tendencies. We also separately analyse the plausibility of the hydrological cycle as well as the potential to couple a DNN to a land model. Section 3.4 includes a summary of our work and potential directions for future research.

2 Methods

2.1 Climate Simulation Data

We leverage three different datasets to train, test, and benchmark DNNs emulating convection with real geography. The data are based on the Super Parameterized Community Atmospheric Model (SPCAM), a global climate model that nearly explicitly resolves atmospheric moist convection (Randall et al., 2003; Grabowski, 2001) by using idealized embedded CRMs. Each of its large-scale host global climate model grid cells embeds two-dimensional CRMs of optional horizontal resolution and physical extent, thus avoiding heuristic parameterization of subgrid moist convective processes (Randall et al., 2003; Benedict & Randall, 2009).

As a baseline, we first use outputs from SPCAM v.3 (SPCAM3) at T42 spectral truncation (i.e. 8,192 horizontal grid cells) driven with boundary conditions of a zonally symmetric aqua-planet; as in Rasp et al. (2018). We then build on previous aqua-planet emulation studies by generating a new dataset from a more modern version of SPCAM v.5 (SPCAM5) that includes higher horizontal resolution (1.9x2.5 degree finite volume dynamical core, i.e. 13,824 grid cells) and in which we incorporate realistic boundary conditions, including a land surface model, seasonality, and a zonally asymmetric annual climatology of sea surface temperatures (SSTs). The dataset is similar to one recently produced by Han et al. (2020) but with a few differences. The simulation itself is 10 years long, but selectively sub-sampled to every 10 days to avoid temporally au-

Simulation Datasets			
Details	CAM5	SPCAM3	SPCAM5
Spatio-temporal resolution	1.9°×2.5°×15min	2.8°×2.8°×30min	1.9°×2.5°×15min
Total Number of Days Simulated	93	93	3,650
Total Number of columns Simulated	123,420,672	36,569,088	4,843,929,600

DNN Setup				
Input	Size	Output	Scaling factor	size
Temperature (K)	30	Heating Tendency (K/s)	cp	30
Specific Humidity (kg/kg)	30	Moistening Tendency (kg/kg/s)	Ls	30
Sfc. Pressure (hPa)	1	TOA LW Flux (W/m^2)	-1e-3	1
Solar Insolation (W/m^2)	1	Sfc. LW Flux (W/m^2)	1e-3	1
Sensible Heat (W/m^2)	1	TOA SW Flux (W/m^2)	-1e-3	1
Latent Heat (W/m^2)	1	Sfc. SW Flux (W/m^2)	1e-3	1
		Precipitation (m/s)	1728000	1

Table 1. Details of the input and output vectors to the DNN. c_p refers to the specific heat capacity of air at a constant pressure and is assumed to be 1.00464e3 (J/kg/K) and L_s is the latent heat of sublimination of water in standard atmospheric conditions calculated by adding the latent heat of vaporization 2.501e6 (J/kg) and the latent heat of freezing 3.337e5 (J/kg). Precipitation is weighted by the same prefactor, 1728000, also used in Rasp et al. (2018) to ensure it is felt in the loss function of the DNN.

tocorrelated training samples. We also rely on a marginally shorter GCM timestep (15 minutes as opposed to 20) (Table 2.1). As in Gentine et al. (2018); Rasp et al. (2018), but unlike Han et al. (2020) we make the further simplifiation of using a reduced (32-km) CRM horizontal extent, i.e. CRMs with only 8-columns apiece, instead of the 128-km / 32-column CRM configuration. This decision is strategic since it simplifies the comparison to Gentine et al. (2018); Rasp et al. (2018), recognizing the findings of Pritchard et al. (2014) that (for deep convection) an unusually small CRM does not corrupt the representation of tropical dynamics in SPCAM; meanwhile there are reasons to think it may make the act of DNN emulation more tractable (Brenowitz et al., 2020; Ott et al., 2020). The codebase for running the "SPCAM5" simulations is the same employed by (Parishani et al., 2017), which is archived at `https://github.com/mspritch/UltraCAM-spcam2_0_cesm1_1_1`; this code was in turn forked from a development version of the CESM1.1.1 located on the NCAR central subversion repository under tag `spcam_cam5_2_00_forCESM1_1_1Rel_V09`, which dates to February 25, 2013. Finally, for reference, we analyze some output from the conventionally parameterized version of CAM5; this helps assess the emulation of SuperParameterization compared to conventional parameterization.

2.2 Neural Network Design

We design a DNN takes the same inputs as standard convective parameterizations in CAM to predict subgrid-scale tendencies at each vertical level and across each timestep across the earth. More specifically, inputs are atmospheric thermodynamics components

in the nine year SPCAM5 data training simulation including: both temperature (K) and specific humidity (kg/kg) for each of the 30 vertical levels spanning the column, as well as surface latent heat flux (W/m^2), surface sensible heat flux (W/m^2), top of atmosphere (TOA) solar insolation (W/m^2), and surface pressure (hPa). Concatenating these input variables creates an input vector to the neural network of length 64. Each of the input variables was pre-normalized before exposure to the neural network by subtracting its mean and dividing by its range, with these statistics computed and applied separately for each vertical level, in the case of the vertically-resolved temperature and humidity profiles(Table 1). Some previous aqua-planet experiments also used the meridional wind vertical profile as part of the input vector to the neural network, but it was omitted in this case as offline tests indicate it had an insignificant effect on the skill of the trained network while increasing the input vector length by 30 and thus substantially increasing training time (Rasp et al., 2018; Gentine et al., 2018); we note that (Han et al., 2020) also deem this an avoidable input.

Our DNN ultimately predicts the subgrid scale time tendency of temperature (K/s), which includes the subgrid advection of temperature by convection and fine-scale turbulence, as well as grid average radiative heating throughout the column, or heating tendency for short. It also predicts the subgrid scale time-tendency of specific humidity throughout the column (kg/kg/s) - or moistening tendency for short. A scalar is predicted for precipitation (m/s) converted into (mm/day). Additional scalars for the long and shortwave net radiative fluxes (W/m^2) at both the surface and the top of the atmosphere are output by the neural network. This fully concatenated output vector is length 65 (Table 1). The state variables that comprise the output vector are based from differing units, making the ultimate Mean Squared Error (MSE) of the neural network devoid of physically meaningfully units. To ensure that all of the predicted variables are felt in the optimizer we apply multiplicative prefactors as in (Rasp et al., 2018) such that they are all of comparable magnitude, recognizing that other choices can also be made such as additionally weighting by the mass of each pressure level (Beucler et al., 2019), although this approach is not taken here.

2.3 Performance Analysis

How well a neural network emulator appears to perform is contingent upon statistical validation choices. Multiple conventions have been used and the degree of spatial averaging before applying error statistics has not been sufficiently reported to do inter-study comparison confidently, though spatial averaging is common practice (Han et al., 2020; Rasp et al., 2018; Brenowitz & Bretherton, 2018). In some cases snapshots of non-averaged data (Rasp et al., 2018) have helped reveal issues at the finest resolved scales and zonally averaged temporal standard deviations (Han et al., 2020) have helped reveal second order issues in NN fits at the finest resolved scales. Precipitation time series and Probability Density Functions (PDFs) (Rasp et al., 2018; Han et al., 2020; O'Gorman & Dwyer, 2018) have also been used to assess neural network performance.

In our case, to examine the magnitude of the error between the neural network prediction and the SPCAM target data, which we treat as truth, a sum of squared errors (SSE) will be calculated separately for each longitude and latitude and, in the case of 3D variables, vertical level:

$$\mathrm{SSE} = \sum_{\mathrm{j}=1}^{\mathrm{N_t}} (\mathrm{y_j} - \hat{\mathrm{y}}_\mathrm{j})^2 \tag{1}$$

where $\mathrm{N_t}$ is the length of the time series, y is the target SPCAM5 data, and $\hat{\mathrm{y}}$ is the corresponding neural network predicted value based on coarse-grained variables. In this case, we examine the performance of the neural network predicting heating and moistening tendencies. The primary metric for determining the performance of the DNN in emu-

lating the expected climate is determined by the coefficient of determination R^2 defined as:

$$R^2 = 1 - \frac{\text{SSE}}{\sum_{j=1}^{N_t}(y_j - \bar{y})^2}, \tag{2}$$

where $\bar{y}$ is the temporally-averaged heating or moistening tendency at a given latitude, longitude, and level.

Latitude-longitude cross sections at specific vertical levels will reveal this error structure at its native 15 minute sampling interval (Figure 3). In other portions of the analysis some spatial averaging is done prior to of the error calculation, as in Han et al. (2020). For instance, in pressure-latitude cross-sections(Figures 4, 1), we first average the predictions and targets in longitude, prior to computing R^2 over the time dimension. Likewise, in pressure-longitude cross-sections, we compute the meridional mean before computing R^2.

The coefficient of determination will thus be our prime analysis metric for evaluating heating and moistening tendencies and their spatial error structure, applied to data ranging from completely unaveraged, to data averaged over narrow latitude intervals, to data averaged around complete latitude circles. We will also examine R^2 at its native time resolution (15 minute sampling; Figures 1, 2, 3, 4, 6) – as the strong diurnal cycle over land regions could bias the analysis between land and ocean regions – vs. prior averaging at daily-mean timescale (Figure 4).

In order to elucidate if there is any detectable "mode-specific" performance, i.e. certain temporal patterns that are especially predictable such as the diurnal cycle over continents, in our moist convection emulation, we also examine a single month (April) of neural network predictions in the frequency domain by calculating the Power Spectral Density (PSD) Φ:

$$\Phi_k = \sum_{j=0}^{N_t-1} y_j e^{\frac{-2\pi i j k}{N_t}} \tag{3}$$

$$\text{PSD}_k = \frac{2\Delta t}{N_t}|\Phi_k|^2 \tag{4}$$

where y is a time series of values of convective heating or moistening tendency at a given location, Δt is the sampling interval, and $i^2 = -1$ (Cooley & Tukey, 1965). The power spectral densities of heating and moistening tendencies are examined both regionally and globally as follows: we mass weight each vertical level and then a PSD value is calculated for each frequency bin at each latitude, longitude and pressure grid cell, focusing on timescales up to a month to examine synoptic-to-subdiurnal timescales. Then, the spectral coefficients are combined into a single, average power spectrum, formed separately for the globe, ocean-only, and land-only areas.

Finally, to hone in on a regime in the lower tropical atmosphere that our R^2 analysis suggests is especially difficult for our DNN to emulate, we will we will analyze the temporal autocorrelations of the subgrid-scale tendencies. Our goal is to understand the regions where the DNN emulation of superparameterized convection is significantly worse than the global average performance. As a proxy for the "stochasticity" of atmospheric motions, we leverage our analysis of the frequency domain to take a power spectrum and then reconstruct an auto-correlation function, or ACF - a measure of the self similarity between a given signal in frequency space and a phase shifted (delayed) version of itself.

$$\mathrm{ACF_j} = \frac{1}{\mathrm{N_t}} \sum_{\mathrm{k=0}}^{\mathrm{N_t}-1} \left(|\Phi_\mathrm{k}|^2 \, \mathrm{e}^{\frac{2\pi \mathrm{i j k}}{\mathrm{N_t}}} \right) \tag{5}$$

Fast signals that decorrelate quickly are more likely to be of stochastic nature. We thus compare the time to e-folding decay in ACF with R^2 skill score across the planetary boundary layer to test for correlations between DNN skill and the timescale of dominant atmospheric signals (diurnal cycle, Rossby waves) visible in vertically resolved heating and moistening tendencies 4. We also the inverse of the e-folding decay timescale, which we will refer to as the "auto-correlation frequency" e.g. 1/ACF, to examine the patterns between R^2 coefficient of determination globally and the stochasticity of the present atmospheric signals. This comparison offers a possible explanation for much of the performance of our DNN in the planetary boundary layer.

To better quantify the differences between true and reproduced PSDs, we rely on the Log-Spectral Distance (D) (Wang et al., 2003):

$$\mathrm{D} = \sqrt{\frac{1}{\mathrm{N_t}} \sum_{\mathrm{k=1}}^{\mathrm{N_t}-1} \left[\ln \left(\frac{\Phi_{\mathrm{real_k}}}{\Phi_{\mathrm{pred_k}}} \right) \right]^2} \tag{6}$$

where the summation is done in frequency space. We examine the same spectral statistics for precipitation and calculate the global error between DNN predicted precipitation and SPCAM5 target data. We also determine the diurnal timing of maximum precipitation globally, but only at locations where the amplitude of precipitation is least two times greater than standard deviation of the diurnal cycle of precipitation. This avoids examining noise in high latitudes .

2.4 Formal Hyperparameter Tuning Process

In several previous aquaplanet studies, small volumes of training data (as low as three months) and manual hyperparameter tuning were sufficient to achieve acceptable Machine Learning (ML) emulator performance (Rasp et al., 2018; Gentine et al., 2018). However, the same was not true for simple feed-forward DNNs applied to twelve months of real-geography training SPCAM data in the study of (Han et al., 2020).

Whether this difficulty is due to convection itself being unparameterizable in the limits of realistic geography, at least as a function that is local in time, as well as reliance on simple DNNs, is unclear. It is equally possible that data of this complexity requires augmentation of neural networks used – i.e. considerably more training data and a formal hyperparameter tuning process (Ott et al., 2020). A main goal of this paper is to test the latter hypothesis.

We thus exploit the potential of our 9-year training simulation. As a first step we subsampled by a factor of 10 after sensitivity tests (not shown) indicated little difference in the fit skill from manual tuning attempts, likely due to redundant information from temporally autocorrelated state data. This reduced the training sample volume to 108 months, which could be managed on single-GPU. Our initial baseline architecture was inspired by previous literature, i.e. composed of five fully connected layers with 256 nodes each. However baseline skill from this manual configuration was unsatisfying (Figure 2), and other manual attempts to explore alternate choices of hyperparameters and learning rate variations were likewise unsuccessful (not shown).

We attained much better results after adopting a formal hyperparameter tuning. Automated NN architecture searches have just begun to prove their value in climate modeling - both for optimizing offline fits (Beucler et al., 2019) and even prognostic online

Name	Range	Parameter Type	Best Model
Batch Normalization	yes, no	Choice	yes
Dropout	[0., 0.25]	Continuous	0.01
LeakyReLU	[0., 0.4]	Continuous	0.15
Learning Rate	[0.00001, 0.01]	Continuous (log)	0.000227
Learning Rate Decay	[0.5, 1.]	Continuous	0.91
Number of Layers	[3, 12]	Discrete	7
Number of Nodes	128, 256, 512	Choice	512
Optimizer	Adam, SGD, RMSProp	Choice	Adam

Table 2. Hyperparameter Space. The resulting best model configuration is shown in the rightmost column.

coupled performance (Ott et al., 2020). Using similar approaches, we thus implemented a resource-intensive automated DNN training process, conducting a formal search over the following hyperparameters: batch normalization, dropout, leaky ReLU coefficient (Maas et al., 2013), learning rate, learning rate decay, number of layers, nodes per layer, and the optimizer (Kingma & Ba, 2014). All parameters and their corresponding ranges for the search are shown in Table 2.

This hyperparameter search took place in two stages, using "Sherpa" (Hertel et al., 2020), a Python library for hyperparameter tuning. First, a large suite (over 200) candidate DNN models were fit using a random search algorithm. The random search has the advantage of making no assumptions about the network architecture or the task of interest. In this stage all hyperparameters, except the learning rate and learning rate decay, are modified. Excluding learning rate parameters in the first stage is strategic to ensure that any increases in performance are due to more skilled architectures rather than gradient update schedules.

Following the initial search, a second search was conducted on the best performing model uncovered by the first stage. This secondary investigation, which tested another fifty models, focused exclusively on the learning rate and learning rate decay settings. This procedure allows the network to be trained with the best possible learning schedule so as to maximize the networks performance for best-performing architecture uncovered in the first stage.

Thus in total more than 250 network architectures were tested, we noticed a dramatic improvement in performance from the hyperparameter search quantified by the difference between the baseline model's MSE and the MSE of the stage 2 model. We also observed the benefit of tuning the learning rate and the learning rate decay in stage two. The validation loss of the stage 2 model descends smoothly and consistently compared to the baseline or stage 1 model. The final result of the hyper-parameter search is shown in Table 2. We discuss below in the results section the extent to which hyperparameter tuning improves the benchmarks discussed above.

While it would be interesting to know whether skillful models could have been obtained with less data volume, this is impossible to precisely quantify without performing Sherpa hyperparameter tuning on a smaller dataset – something we opted not to do due to the heavy GPU requirements of applying Sherpa. We strategically only utilized it on our richest training dataset to conserve resources. For context on the resource requirement, each candidate NN architecture required roughly twenty-four hours to train (about one hour per epoch). Eight models could be run in parallel on a single GPU and thus, using four GPUs, we could train about thirty two models per day. In total, with four GPUs (12 GB memory each), it took eight days to train all 250 models.

Label	SPCAM Data	Region	Variable	timestep	25th	50th	75th
a	aqua	Ocean	Heating	15 min.	0.06	0.27	0.55
b	aqua	Ocean	Moistening	15 min.	0.00	0.00	0.21
c	real-geog.	Ocean	Heating	15 min.	0.33	0.58	0.80
d	real-geog.	Land	Heating	15 min.	0.44	0.67	0.84
e	real-geog. (No Sherpa)	Ocean	Heating	15 min.	0.14	0.40	0.64
f	real-geog. (No Sherpa)	Land	Heating	15 min.	0.07	0.38	0.65
g	real-geog.(No Sherpa)	Ocean	Moistening	15 min.	0.00	0.02	0.32
h	real-geog. (No Sherpa)	Land	Moistening	15 min.	0.00	0.00	0.35
i	real-geog.	Ocean	Moistening	15 min.	0.00	0.18	0.52
j	real-geog.	Land	Moistening	15 min.	0.00	0.04	0.55
j	real-geog.	Ocean	Heating	Daily	0.44	0.73	0.88
l	real-geog.	Land	Heating	Daily	0.49	0.75	0.87
m	real-geog.	Ocean	Moistening	Daily	0.00	0.47	0.82
n	real-geog.	Land	Moistening	Daily	0.00	0.08	0.77

Table 3. Statistical breakdown of skill score showing percentiles summarizing skill variability in 3D space, i.e. from a flattened vector of R^2 values that were calculated across just the time dimension separately for each longitude, latitude, and pressure level, using raw data at the 15-minute sampling scale. For SPCAM5 DNNs, assume Sherpa is used for automated tuning unless explicitly mentioned otherwise.

3 Results

In this section, we use the diagnostics outlined in Section 2.3 to benchmark the performance of our DNNs. We quantify the overall performance of our DNN in emulating atmospheric heating and moistening tendencies in Sections 3.1, 3.2, and analyze the emulation of precipitation in Section 3.3. Note that since we used the first nine years to train the network, we benchmark the performance of our DNN on the remaining tenth year that we held out for validation.

3.1 Spatial Structures

When assessed on the skill of its raw predictions with no averaging, our manually tuned DNN emulates the (temporal) mean vertically resolved heating and moistening tendency profiles at least as skillfully as manually tuned aqua-planet NNs (Table 3, a-b vs. e-h). The heating rate skills in real-geography mode are comparable over land (median R^2 of 0.38) and ocean (median R^2 of 0.40) and these even outperform baseline values from manually-tuned aquaplanet simulations (median R^2 of 0.27). Effects of hyperparameter tuning are dramatic, boosting the median and 75th percentile values for heating and moistening R^2 to around than 0.6 and 0.8 respectively, for both heating and moistening (Table 3 c,d vs. e,f). The fact that one quarter of the domain is emulated with $R^2 > 0.8$, even prior to any averaging in space or time indicates our final DNN setup (full training data volume and hyperparameter tuning) generates a good fit. The variation between percentiles in Table 3 indicates spatial variations in the skill, suggesting certain regions are more accurately emulated than others.

A first look at spatial structures in the skill affirms a generally appropriate fit with familiar structures relative to aquaplanet expectations but also some interesting differences. Figure 1 presents the skill of zonally averaged DNN predictions. Here we compare our prototype manually tuned DNN's skill on aquaplanet target data (a,b; representative of what was used in Rasp et al. (2018) against our optimized real-geography

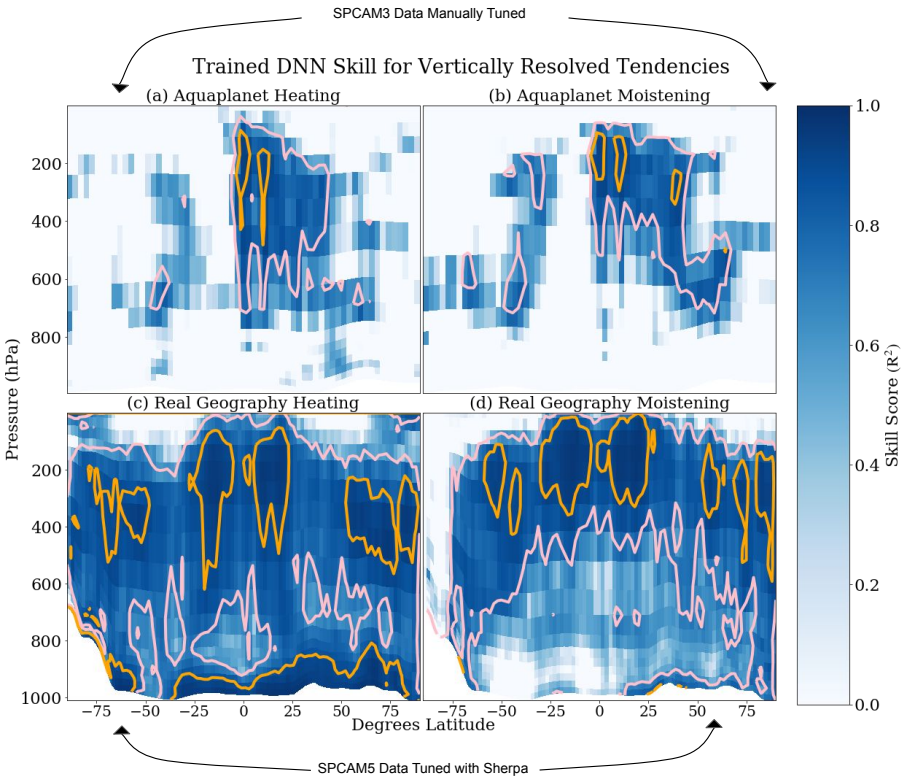

Figure 1. The R^2 coefficient of determination for zonally averaged DNN predictions. The performance of a manually tuned neural network emulating aquaplanet data (a and b) is compared to our optimal DNN emulation of full complexity real-geography data after formal tuning (c and d). Skill shown separately for heating tendency in K/s (a and c) and moistening tendency in kg/kg/s (b and d). Both neural networks capture deep convection in the tropics well, but overall the Sherpa tuning and more plentiful training data provided to the SPCAM5 data allows the neural network to perform better in spite of the associated higher complexity native to the real-geography data. Areas where R^2 is greater than 0.7 agreement between are contoured in pink and areas greater than 0.9 in orange.

DNN's skill. Achieving realistic performance on zonal means is an easier target due to the averaging between land and marine atmosphere and the smoothing of the sharpest temporal variations in convection. However, this zonal mean perspective still provides a useful composite view of emulation of the real geography atmosphere. Here the R^2 for zonal mean net diabatic heating and moistening is greater than 0.7 throughout the free troposphere, agnostic to latitude (Figure 1c and d). This widespread skill in the upper troposphere is further amplified by cores of R^2 greater than 0.9 around mid-latitude storm tracks and locations of deep tropical convection above the southern and northern bounds of the ITCZ. The results confirm the appropriateness of simple DNNs architectures for the "parameterizability" of convection. Our simple DNN can skillfully emulate heating and moistening tendencies of convection across latitude and vertical level in line with SPCAM5 target data (Figure 2). It is also reassuring that the best skill (R^2 over 0.9 for zonal mean predictions) occurs in important regions of the troposphere (Inter-tropical Convergence Zone, mid-latitude storm tracks) where mean diabatic heating couples to the general circulation (Figures 1, 2). This was true also in Rasp et al. (2018) and can be seen in our own aquaplanet benchmark (Figure 1a,b). However, unlike the aquaplanet,

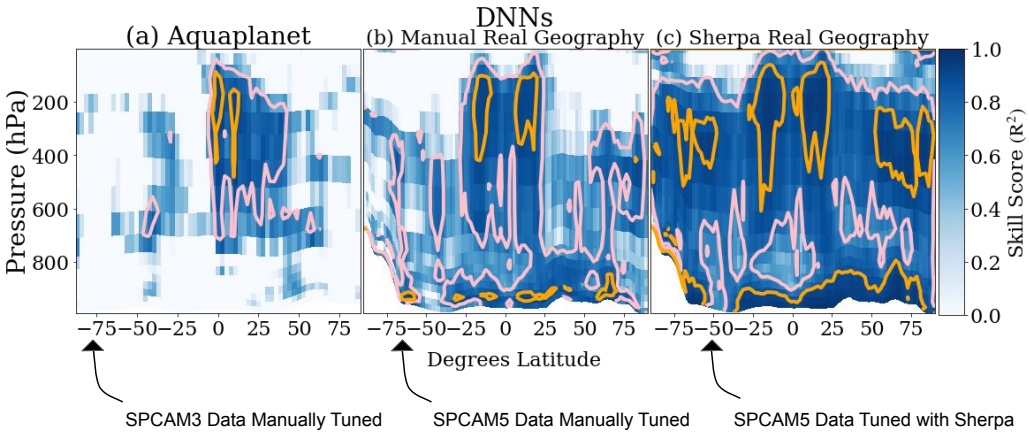

Figure 2. The effects of upgrading from aquaplanet (SPCAM3 - panel A) to real geography data (SPCAM5 - panels b, c) and then applying automated Sherpa hyperparameter tuning (panel c) can be seen on DNN skill. Both real geography trained networks above received an identical eight years of training data, while the aquaplanet network is given 6 months in line with Rasp et al. (2018) where it is shown that skill plateaus above 3 months of training data. The difference in skill in emulating vertically resolved heating tendency in K/s is shown. Longitude is again averaged to get a global cross-section. The Sherpa based hyper-parameter tuning substantially boosts the performance of the neural network, particularly outside of tropical regions. Areas where R^2 is greater than 0.7 agreement between are contoured in pink and areas greater than 0.9 in orange.

there is a new region of high skill in the real-geography emulator with R^2 greater than 0.9 in the planetary boundary layer for heating tendency emulation. This appears to be a continental signal, evidenced by both higher skill in the northern hemisphere (Figure 1c) and comparatively lower near-surface zonal mean skill at the latitudes of the Southern Ocean. Though less skillful overall, there is a similar pattern in the convective moistening tendency at the boundary layer, where the highest DNN performance at the surface (R^2 over 0.7) is confined the the land heavy northern hemisphere (Figure 1d).

We confirm some existence of land-sea spatial structures in emulation skill by examining maps of predictions before any spatial averaging. At the lowest model level, our DNN predictions achieve R^2 greater than 0.7 (greater than 0.9 in continental interiors; Figure 3a). However, this spatial pattern is inverted when examining skill on a model layer in the midtropopshere, near 500 hPa. At this altitude, our DNN now makes the most accurate predictions on the 15-minute timescale over the extratropical marine atmosphere but struggles over continents and deep convecting regions of the tropics (Figure 3b). We speculate that a strong deterministically predictable component of diurnal variability in surface heating and moistening associated with large surface flux diurnal variations over land could allow the low-level heating skill to be enhanced there; diurnal signals will be examined in detail in Section 3.3. Meanwhile, in the upper troposphere we expect to see skill deficits in regions of tropical and continental convection (on this 15-min timescale).

The aquaplanet vs. real-geography comparison in Figure 1 combined two factors that could have influenced skill. The first is the new existence of real-geography data, in which the realistic continent setup will add more deterministic components to convection in the atmosphere, boosting odds of our DNN finding good a good fit. The effects of the hyperparameter tuning are also paramount to consider. To separate their in-

fluence, Figure 2 shows the effect of adding real geography data (while maintaining manual tuning). This confirms an increase in skill over the continental boundary layer independent of hyperparameter tuning, consistent with the existence of new deterministic signals in real-geography settings likely associated with the strong, predictable diurnal cycle over land. The dramatic nature of the skill improvement from migrating from a manual tuning environment to automated hyperparameter tuning underscores the crucial role that Sherpa played in identifying the optimal DNN that will be primarily discussed going forward.

We now focus on the spatial structures where even the optimized neural network still struggles. For this we return to assessing zonal mean predictions, since these are less exposed to details of stochasticity but particularly important to get correctly when using NNs prognostically. Figure 4a and b present the greatest emulation challenge for our optimal DNN ($R^2 < 0.3$ for mean temporal variance) which occurs throughout the lower troposphere, excluding the continental boundary layer, and especially for convective moistening. This is consistent with previous aqua-planet simulations (Rasp et al., 2018; Gentine et al., 2018), and the study of Han et al. (2020). Boundary layer moistening is an especially challenging target for machine learning emulation, particularly when focusing on the 15-minute sampling interval. This is also verified when looking at the temporal standard deviations of the heating and moistening tendencies, where much of the spatial field is emulated well, but under-predictions occur for moistening tendency in the lower troposphere (Figure A1).

We conclude with an animation demonstrating unfiltered, non-composited views of the convective tendency emulation of over a two week period in July at the following link `https://tinyurl.com/y4selg5g`. The animation shows the evolution of total diabatic heating and moistening on a model level near 600 hPa, in the lower-to-mid troposphere. In the right-hand panels, the diurnal cycle of peak nocturnal radiative cooling can be seen propagating from east to west tracking the earth's rotation. It is punctuated by local features of positive diabatic heating from latent heating within slow moving weather systems, as well as the stationary lagged diurnal convective response to the passage of the sun over Central Eurasia and America. No geographic distortions of synoptic disturbances are detectable – even heating tendencies from tropical convective clusters and the Atlantic convergence zone and Pacific Inter-tropical Convergence Zone are all well emulated from this view. On the left hand panels the moistening associated with this convection provides an especially clear view of the secondary issues with the emulated convection. The main distortion compared to truth is the lack of stochasticity, which manifests as geographic static in the benchmark training data (top panels) but is absent in the DNN tendencies, as expected.

3.2 Temporal Variability

Why do we see such considerable variations in the skill of our DNN as a function of geographic location and altitude? One hypothesis is that the DNN learns "mode-specific" fluctuations. A first test of this is re-examining spatial skill structures after averaging predictions from their native timescale of 15 minutes to the daily mean timescale instead. Figure 4 shows the corresponding skill for daily-mean zonal-mean predictions. From this view, the vast majority of the atmosphere can be emulated, in both heat and moisture, with R^2 greater than 0.9. Meanwhile comparing to the faster timescale, the skill deficit in the lower troposphere for convective moistening is not nearly as dramatic. The fact that spatial structures are to some extent sensitive to temporal averaging is consistent with the hypothesis that the DNN may exhibit skill that is mode-specific.

This motivates our spectral analysis (Figure 5) on the SPCAM5 target data and DNN predictions. Switching to frequency space is a clean way to determine if specific modes of variation such as the diurnal heating cycle and synoptic storm propagation are

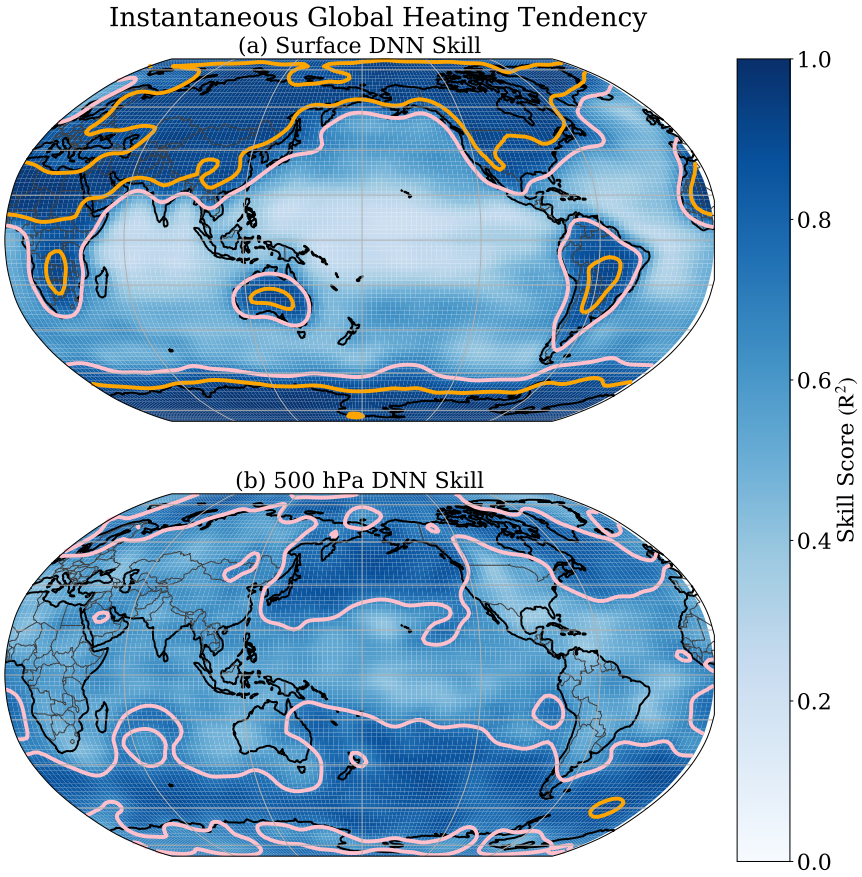

Figure 3. The neural network skill in emulating the (a) the lowest model level showing planetary boundary layer heating tendency and (b) The model level closest to 500 hPa heating tendency (K/s), both at the native 15 minute timestep interval. The neural network learns to represent locations over continents and the mid-latitudes best down at the surface, while locations of mid latitude storm tracks are the neural networks strongest performance higher in the atmosphere. The tropics, in particular tropical locations over oceans give the neural network the greatest challenge in emulating sub-grid heating tendency. Areas where R^2 is greater than 0.7 agreement between are contoured in pink and areas greater than 0.9 in orange. The cross-section was smoothed by a series of 1D convolutional filters for interpretability.

driving preferential modes of DNN performance. The PSD is calculated separately at each unique latitude, longitude and vertical level from the CAM5 data, SPCAM5 data, and the corresponding DNN output, and appropriately weighted by layer mass to achieve the contribution to the column integral in W/m^2. These location-specific PSD are then averaged together horizontally and vertically to arrive at a globally representative spectrum.

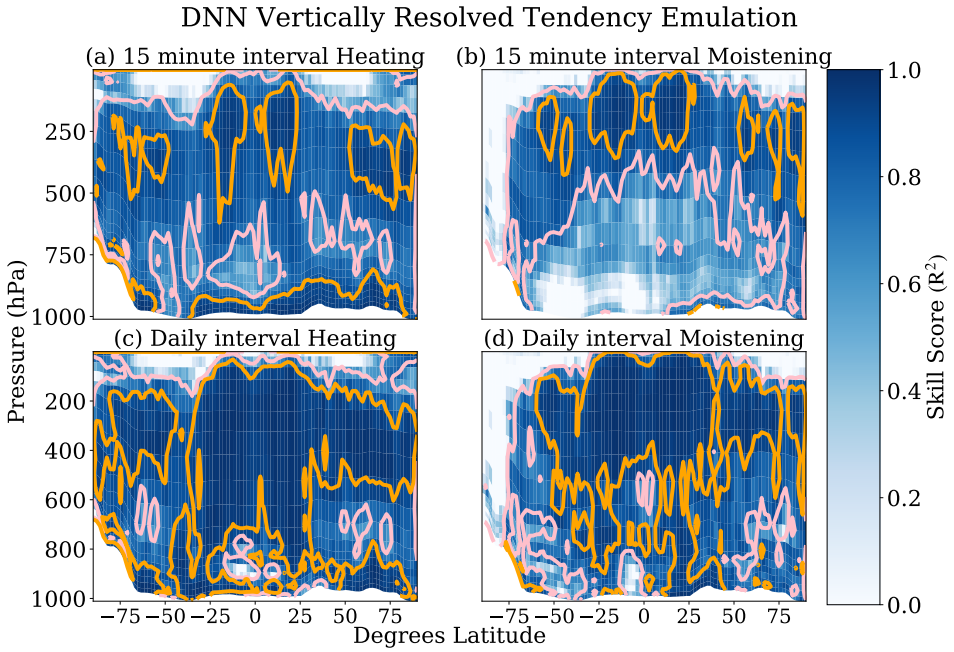

Figure 4. Neural network performance at time step interval (a and b) is contrasted with performance at the diurnal scale (c and d). Representation of heating tendency in K/s (a and c) and moistening tendency in kg/kg/s (b and d) are both examined. Longitude is again averaged out. In both vertically resolved heating and moistening, there is an across the board gain in skill at higher timescales. Areas where R^2 is greater than 0.7 agreement between are contoured in pink and areas greater than 0.9 in orange.

In contrast to our hypothesis, the results do not show any major mode-specificity to the DNN skill on the hourly-to-weekly timescale. All of the most important spectral features in the target data exhibit comparable variance in the DNN predictions, including the main variance source from disturbances slower than one day, but also the discrete variance from diurnal, semi-diurnal, and other harmonics of the daily cycle. While there is an expected underprediction of total variance for all of these modes, the DNN skill is not obviously preferential to any modes as the diurnal timescale or longer. Interestingly, there is very little land-sea contrast in vertically integrated predictions from the view of frequency space (Figure 5c,d vs. e,f). Quantitatively, the total log spectral distance between the SPCAM5 benchmark data and the neural network prediction is 2.40 for globally averaged mass-weighted heating tendencies and 2.68 for globally averaged mass-weighted moistening tendencies from Equation 8.

The very weak "mode" of fast variability on timescales less than 2-6 hours is where our DNN performs poorest based on Figure 5. Evidently there is something native to high frequency heating and moistening convective tendencies that challenges our DNN. This suggests an alternate hypotheses that geographic structures in our DNNs skill might be an artifact of variance sorting (i.e. the fact that the fastest signals are also the weakest potentially downweights their contribution to the loss function) or of stochasticity (since fast variations can often be stochastic in origin). The conventionally parameterized CAM's relative strength at these higher frequencies may be a result of noisy parameterizations.

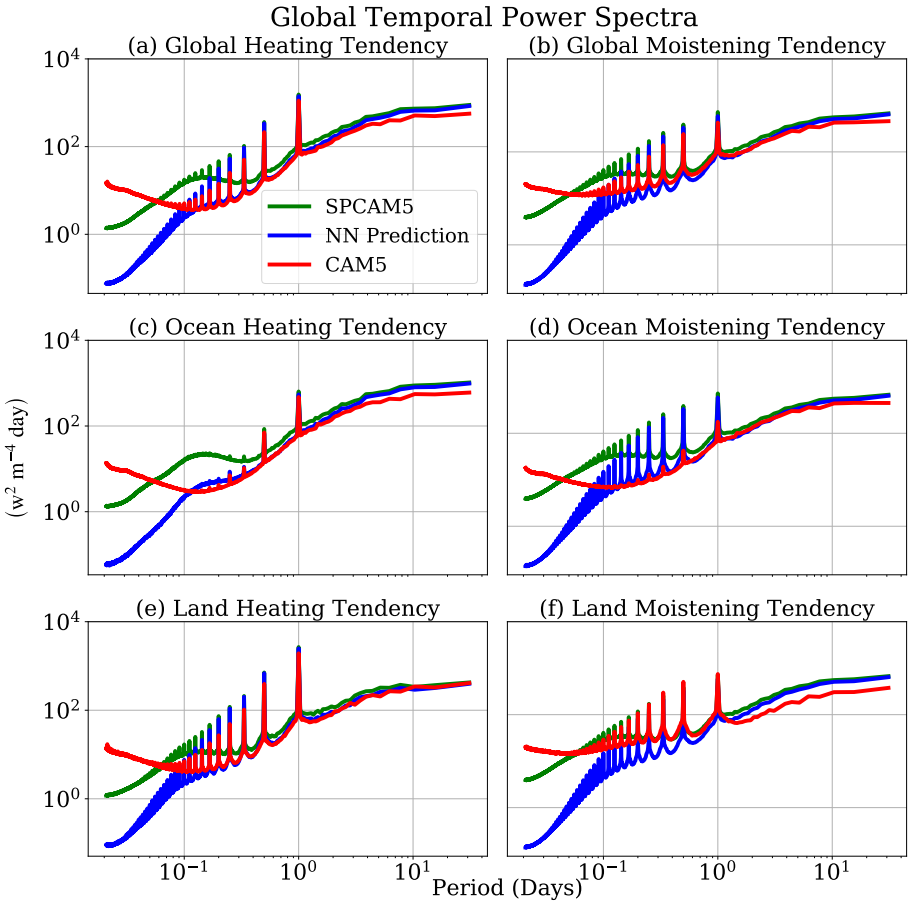

Figure 5. The temporal power spectrum for vertically resolved heating tendency in $w^2/m^4 day^{-1}$ (a) and vertically resolved moistening tendency in $w^2/m^4 day^{-1}$ (b) are calculated at each latitude, longitude and elevation across the globe. These spectra are then averaged together to see how much variance the neural network captures globally compared to its SPCAM5 training data and a CAM5 baseline. Further tests are done exclusively over marine locations (c and d) and over continental ones (e and f). The peaks correspond to the solar radiation driving the diurnal cycle, though this is stronger on land (a) than in marine locations (b). Multi-taper spectra were also calculated for both tendencies but showed no qualitative difference with the results above calculated through the numpy fft package.

To further test these hypotheses, we construct a proxy of stochasticity, and compare its geographic structure to the geographic structure of DNN R^2 skill. The proxy is the e-folding time at which signals of atmospheric variance decouple into noise based on the auto-correlation, i.e. the decorrelation timescale. Qualitatively, the patterns of R^2 heating skill and the de-correlation timescale are very similar with Pearson correlation coefficient of 0.57 based on the Pearson correlation coefficient (Lehmann, 1992), indicating at least moderate pattern correlation between these two variables. The fact that the regions of least DNN skill are also the regions of fastest signal decorrelation (pur-

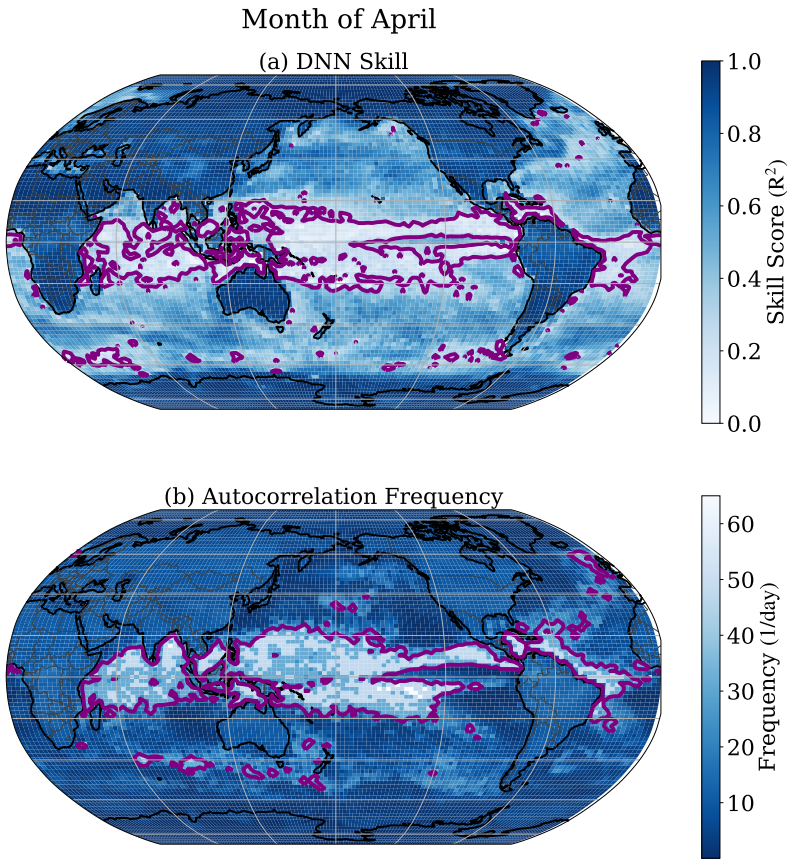

Figure 6. A comparison between the neural network R^2 skill in emulating the vertically resolved heating tendency in K/s (a) and the auto-correlation frequency of the SPCAM5 heating tendencies (b). Both cross sections are taken at the lowest pressure level in the model. Qualitatively the patterns closely match. The areas of lowest skill score (bottom tenth percentile) and highest auto-correlation frequency (90th percentile) are both contoured in purple.

ple contours in Figure 7) supports the view that imperfect emulation of fast, stochastic signals is mainly responsible for sculpting the spatial structures in the DNN's skill score.

To illustrate this further and within multiple geographic regimes, Figure 7 examines temporal auto-correlations in different DNN skill regimes as follows. First, all grid cells having poorest skill are identified, and the temporal auto-correlation of the benchmark time series data is calculated from time lags 0 to +0.4 days. This is done separately for each grid cell and then composited into a single auto-correlation plot (red line). Repeating the procedure for those horizontal grid cells where the DNN fit has highest skill (blue line) reveals the characteristic difference in de-correlation timescale in the high-skill vs. low-skill spatial regions. Repeating this procedure for the midlatitudes (Figure 7) confirms the same timescale-selectively of skill exists across multiple geographic regimes,

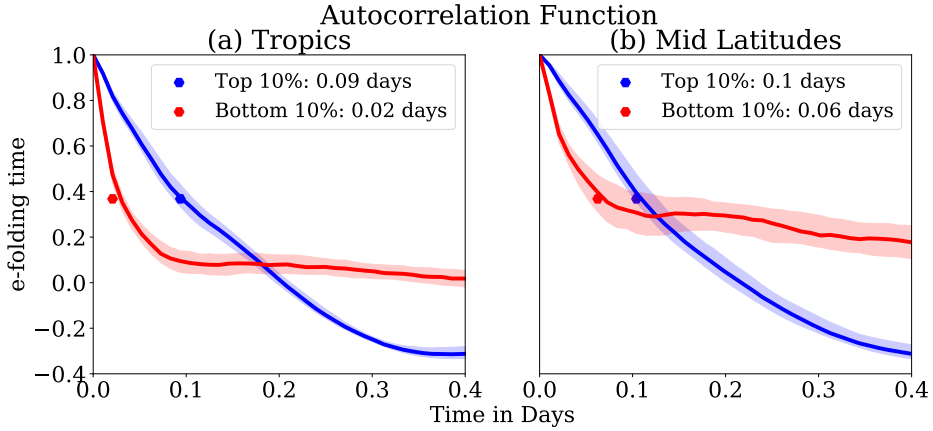

Figure 7. The time to e-folding decay from the autocorrelation function for the median at every surface location of the top 10 percent of R^2 skill score locations and the bottom 10 percent of R^2 skill score of heating tendency in K/s at locations in the tropics (15°S to 15°N) (a) and the mid latitudes (30°N to 60°N) (b). The corresponding inter-quartile regions are shaded in as a marker for statistical significance. The month chosen to test in the validation data is April.

even though this was not obvious in Figure 6. The relationship is robust and clear – locations where signals tend to decorrelate faster are locations where DNN skill is lower. This is consistent when examining both the planetary boundary layer of the oceans and the continents each in isolation as well as globally (not shown).

3.3 Bayesian Methods vs. Physical Constraints for Emulating the Diurnal Cycle

So far we have shown comprehensive skill for our optimized DNN except for some fast-varying signals that de-correlate quickly. On the one hand it is not obviously a problem to have skill deficits for the stochastic component of superparameterized convection, since it does not appear to be critical to most emergent behaviors of SPCAM – though some have suggested that a close representation of the stochastic component is still necessary to properly model large scale weather phenomena (Jones et al., 2019; Neelin et al., 2008). On the other hand some fast-varying signals are also critical to regional climate simulation and should be deterministically predictable – the diurnal cycle provides a perfect example. To what extent can the DNN emulate the details of the diurnal cycle?

A first look at the composite height vs. time-of-day structure of convective moistening (Figure 8) is reassuring – the DNN captures the coherent temporal transition of shallow to deep convection in the afternoon over land (moistening above cooling growing after sunrise into the mid-troposphere;Figure 8a vs. b). Over the ocean, the opposite phase of peak moistening-above-drying happens during the night between 8pm and 6a; this land-sea contrast is also successfully emulated (Figure 8b vs. d). Thus our first impression is that the DNN correctly recreates the land sea contrast of diurnal convection present in the SPCAM5 target data (Figure 8). It is also worth noting this emulated composite in Figure 8 shows that compared to our SPCAM5 target data, the diurnal cycle produced by the DNN is less stochastic.

This motivates a closer look at the full geographic structure, focusing now on the diurnal cycle of *precipitation*, which reveals some interesting surprises. The benchmark SPCAM5 target data (Figure 9b) resembles observations – over land regions, a strong,

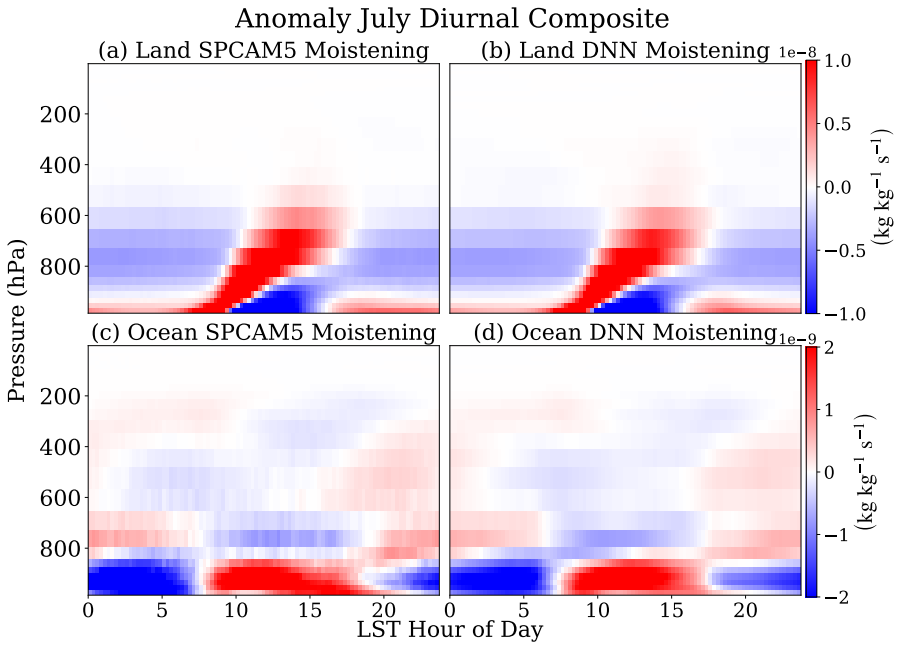

Figure 8. A comparison between the moistening tendency of SPCAM5 target data (a and c) and DNN predictions (b and d) in (kg/kg/s) over continental (a and b) and marine (c and d) locations respectively. The composite is taken over the month of July and the anomaly in which the mean of the diurnal cycle is subtracted out is shown above.

predictable diurnal precipitation cycle over tropical rainforests and summer continents, with lagged afternoon maximum precipitation (Local Solar Time between 13:00 and 18:00). In contrast, a weak diurnal cycle of precipitation over the oceans that peaks in the end of night and early morning, and is especially detectable in subtropical stratocumulus regions. Note the benefits of superparameterization relative to conventional parameterization (Figure 9b vs. a) have been previously reported as a reduced detectability of the diurnal mode except where it is supposed to be strong such as over tropical rainforests or where it can be especially detectable such as over marine subtropical drizzle regimes. Here we see a surprise in the optimized DNN's precipitation predictions (Figure 9c): Although the stratocumulus marine drizzle cycle appears to be well emulated, consistent with the diurnal moistening composites seen in Figure 8d, over land there is incorrect timing and spatial extent of maximum precipitation (Figure 9c). This is paradoxically at odds with Figure 8a and b which indicated excellent emulation of the diurnal cycle of convective moistening over land regions. DNN detection of a cycle of precipitation over desert regions is physically unrealistic and the timing of the onset of deep convection and heavy precipitation on land is several hours premature, much like CAM5 data (Figure 9c, b, and a).

Why is precipitation emulated less skillfully? Our working hypothesis is that in hindsight our DNN architecture did not respect an important physical constraint that distinguishes this variable. Unlike moistening and heating tendencies, precipitation should be positive definite. To test this hypothesis we introduce two additional neural networks in Figure 9 to elucidate contradictions between emulation of precipitation and convective tendencies. We compare our DNN that ignored positive definite nature of precipitation but included substantial hyperparameter tuning (called "baseline DNN" for the

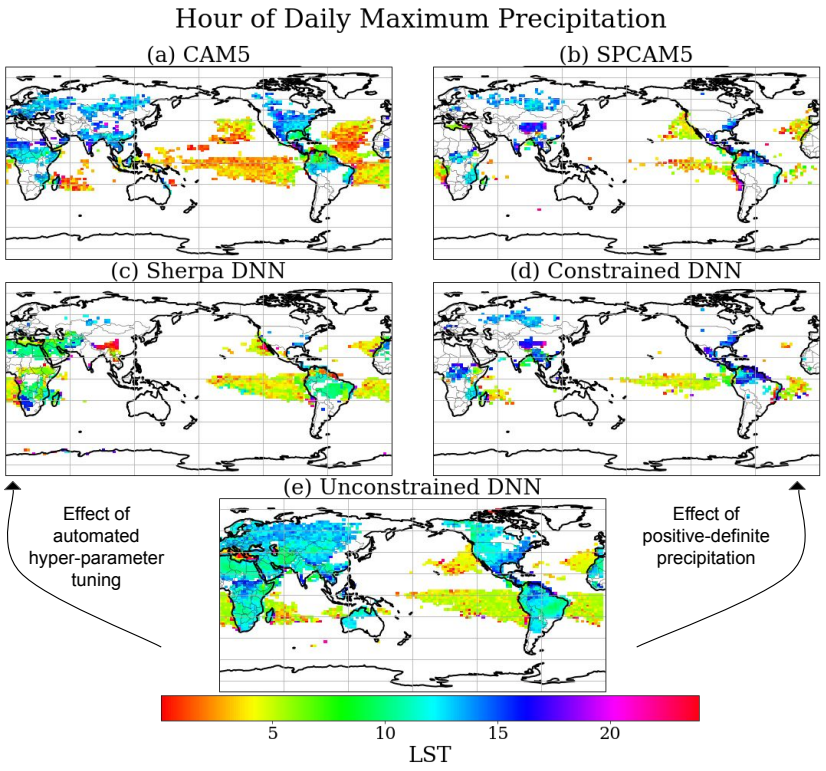

Figure 9. A comparison between CAM5 (a), SPCAM5 (b), and our neural network with automated hyperparameter tuning (previously referred to as just DNN – now called Sherpa) (c), a neural network with positive definite precipitation through ReLU (Constrained) (d), and an archaic version of our DNN without automated hyperparameter tuning or physical constraints (Unconstrained) (e). The figures show the hour of maximum precipitation in mm/day during the boreal summer (months of June, July, and August). The time of maximum precipitation is colored in only over areas with a significant diurnal amplitude in precipitation rate as defined in Section 2.3. There is close agreement in the both the timing of the onset of maximum precipitation and the spatial extent of the diurnal cycle between SPCAM5 data and the Constrained neural network predictions over land and good agreement between the Sherpa DNN and target data over the ocean. The constrained DNN and CAM5 baseline both have a noon-centric bias for maximum precipitation timing and detects a diurnal cycle of precipitation over many locations where observational data would suggest there is not.

remainder of this section for clarity) (Figure 9c) against a physically constrained DNN (Constrained) in which a ReLU activation is applied to the last precipitation node to ensure rainfall predictions are positive definite, but which is not formally re-tuned (Figure 9d) as well as a DNN (Unconstrained) which has neither physical constraints nor automated hyperparameter tuning (Figure 9e).

The results show the importance of adding a positive constraint on precipitation rivals the importance of doing hyperparameter tuning yet these techniques induce improvements in different geographic regimes. Unsurprisingly, the Unconstrained DNN that includes neither benefits of physical constraints or hyperparameter tuning struggles with emulating diurnal precipitation everywhere. The timing of maximum daily precipitation is wrong over land (noon-centric) and the diurnal cycle far too detectable there (Figure 9b

vs. e). Adding the positive-definite constraint alone produces dramatic improvements over land – the diurnal cycle becomes detectable only where it should be, over tropical rainforest and continental Eurasia, and with the appropriate timing (Figure 9d vs. e). This appears to solve the problem that our baseline DNN suffered (Figure 9c). However the positive constraint alone is unable to produce successful results over the ocean. Only the tuned DNN (Figure 9c) emulates the subtle marine stratocumulus diurnal cycle well.

The limitations of hyperparameter tuning alone become apparent. We have highlighted the power of automated hyperparameter tuning on convective tendencies in Figure 2, but revealed that our baseline DNN did not take into account physical laws governing its simultaneous prediction of precipitation, instead corrupting it. Difficulty emulating details of precipitation cycles are certainly not unique to this study but do point to larger growing pains in the machine learning-physical science nexus. Similar problems with capturing the physics behind precipitation through neural networks have been discussed in Yen et al. (2019); Brenowitz and Bretherton (2018) where neural networks created non-trivial negative precipitation as well. As in Kumar et al. (2020), we show that augmenting our DNN with a positive constraint mitigates many of the errors in precipitation emulation (Figure 9b vs. d). For our Constrained DNN, even though it was not re-tuned, the land-sea contrast in the timing of peak precipitation is captured accurately: nocturnal over oceans, late-afternoon over the hottest and moistest continental regions. This is a quintessential example of the importance of prior understanding of the details of emulated variables and the success of physically informed neural networks in STEM fields more broadly (Toms et al., n.d.). In hindsight it would be logical to complement the benefits of hyperparameter tuning with such constraints – an important topic for future work.

To further assess these trade-offs we now look beyond just the diurnal cycle to examine the full PDF of precipitation across our sensitivity tests. The baseline DNN does best in capturing the global amount distribution (Figure 10 blue vs. green). Consistent with the diurnal cycle analysis the baseline DNN performs especially well over the ocean (Figure 10d blue vs. green). However its issues over land are even more striking from the viewpoint of the full PDF where the 3 DNNs have radically different values for the amount mode – i.e. rainfall rate delivering maximal precipitation. Whereas the diurnal cycle analysis had suggested a positive definite constraint alone brought continental precipitation into focus, we can see from the amount distribution that beyond diurnal timing its statistics are incorrect (blue dashed vs. green), with far too intense extreme rainfall events. Meanwhile the baseline DNN has a pronounced problem of producing too much drizzle over land which is also a problem seen in precipitation from standard parameterization.

Taken as a whole our precipitation results suggest that this is an area where further refinement of our baseline DNN is needed. Over the oceans, the DNN captures much of the pdf of precipitation (Figure 10), including moderate to heavy regimes at the tail that challenge many climate models, as well as the diurnal cycle of precipitation over the oceans (Figure 9b vs. c). But there is significant corruption of the emulated signal over continental locations, particular with regards to the timing of onset of heaviest precipitation and the intensity of rain delivering most surface accumulation. For an even more information-rich view, we have attached as supplemental information an animation showing two weeks of July precipitation from baseline CAM5, target SPCAM5 data, and DNN emulation which can be found at `https://tinyurl.com/y4adqsrs`.

In comparing sensitivity to adding a constraints vs. leveraging bayesian optimization methods (Toms et al., n.d.), both methods provide unique, disparate performance enhancements with the Constrained DNN performing better over land and the baseline DNN doing better over the oceans (Figures 10,9). But synthesizing both tools may ultimately be necessary since neither a physically constrained neural network architecture nor automated hyperparameter tuned network on its own could capture the full com-

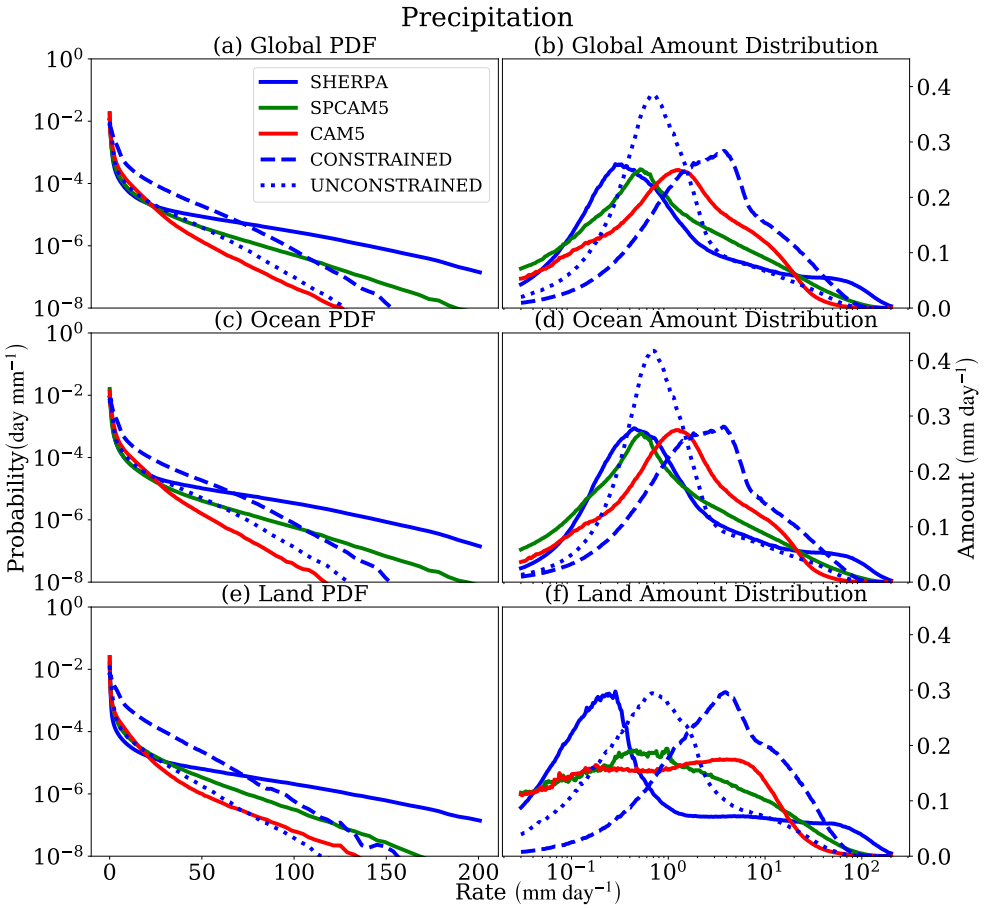

Figure 10. The Probability Density Function across the range of simulated precipitation rates (a) and the amount distribution (b) of precipitation in which the probability density function is multiplied by the bin-averaged values of precipitation. We design the histograms based on the methods outlined in (Watterson & Dix, 2003), which have been widely adopted in literature including in formative works such as Pendergrass and Hartmann (2014). We implement logarithmically distributed rain-rate bins - in our case each bin width grows by 3 percent to ensure the entirety of the precipitation PDF is reflected. For more detail, we include an archaic version of our DNN without an automated hyperparameter tuning or physical constraints (Unconstrained), a DNN with positive definite precipitation through ReLU (Constrained), and our Sherpa DNN discussed previously in the methods section (Sherpa). Comparisons are also made exclusively over marine areas (c and e) and continental ones (d and f)

plexity and timing of the diurnal cycle of precipitation over both land and ocean. We recommend an integration of both these tools for future attempts of this work. Meanwhile it is worth recalling that these corruptions are less obvious in the diurnal cycle of heating and moistening which is better emulated, perhaps because it dominates the loss function, or perhaps because – unlike for precipitation – there is no internal inconsistency

with the values of these variables and the assumed DNN architecture. But since precipitation is a critical input to land surface models, resolving the issues revealed in this section will be an important next step towards realizing that dream. Other issues at the frontier of coupling emulators to land surface models are discussed next.

3.4 Towards Interactive Land Coupling

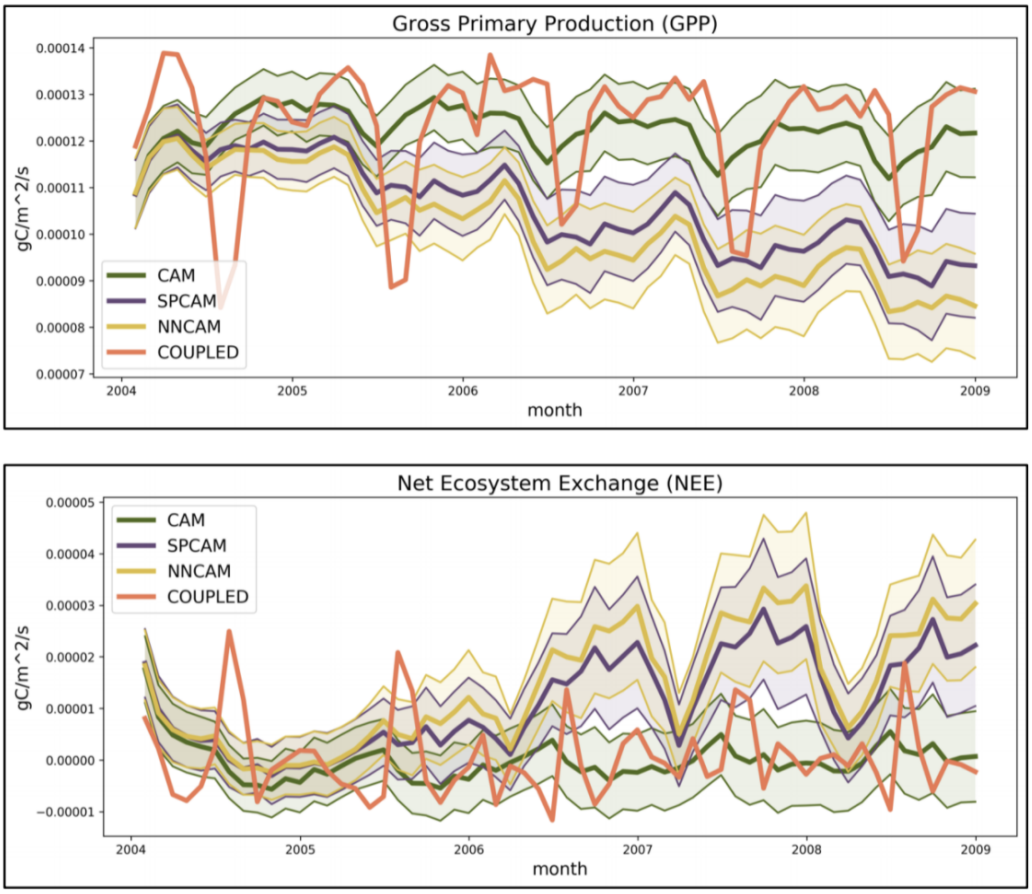

Figure 11. The Gross Primary Production (GPP) and Net Ecosystem Exchange (NEE) monthly based on CAM data are contrasted against SPCAM and a a neural network, the results of which are derived from one way land coupling.

Taken together, most of the above results look good enough that it is natural to wonder if prognostic tests using an emulator like this might produce stable simulations as was shown for an aquaplanet by Rasp et al. (2018), but in a real-geography setting. This would be exciting to test but our view is that as yet it is premature to try. For instance, beyond its corruptions of continental precipitation, the DNN we have described does not predict everything that would be needed to drive an interactive land surface model in practice. It is even unknown whether the imperfections in the near-surface state of the DNN's predictions would even be compatible with land surface modeling.

As a first credibility test on the latter front we thus report some results from "offline" standalone land surface model simulations driven with actual (vs. emulated) surface state data from an aquaplanet training dataset (Yacalis, 2018). Such simulations are easier to perform than fully interactive land-atmosphere coupled simulations and pro-

vide a quick test of the null hypothesis that corruptions of the surface state by the DNN might be incompatible with land modeling in general.

The results in Figure 11 reaffirm the potential for prognostic tests. The Figure shows the carbon cycle flux responses from over a hundred Coupled Land Model (CLM) simulations, each with Amazon-like conditions, driven by high frequency forcing data taken from different tropical grid cells of actual vs. emulated SPCAM aquaplanet data. Relative to CLM's conventional coupled behavior (orange lines) these integrations drift to an unusual attractor, which can be understood by the unusual aquaplanet surface state (e.g. high wind speeds from a frictionless surface). Despite this idealization, the key point is that the CLM drifts to the same new attractor regardless of whether the emulated surface inputs or the actual surface inputs are used to drive it, including details of multiple nonlinear cycles that we have traced to threshold physics associated with wildfire and carbon cycle feedbacks interior to CLM's biogeochemistry modules. These similar trajectories, despite the nonlinearities inherent to CLM physics, are strong evidence against the null hypothesis. This adds support to the idea of trying DNN convection emulators like this in fully interactive real-geography simulations, if they can be adapted to produce all necessary output fields. Further details are provided in Yacalis (2018).

4 Conclusion

We find that a simple deep neural network architecture can skillfully emulate the deterministic part of subgrid-scale diabatic heating and moistening tendencies from global superparameterization. For the zonal mean, emulated tendencies explain over 70% of the actual variance at the 15-minute sampling scale and over 90% of the actual variance at the daily-mean sampling scale throughout most of the mid-to upper troposphere. On regional scales, heating skill is best at low altitudes over land, and at mid-levels over extratropical oceans – both regions where we expect convection to be deterministically set by the large-scale thermodynamic state. On diurnal timescales, convective responses to solar heating are emulated correctly, including land-sea contrast and vertical structure. Full temporal spectral analysis reveals no obvious mode-specifity to what is vs. isn't emulated other than imperfections in the goodness of fit on short timescales (model time step interval of 15 minutes). Pearson correlation coefficient of 0.57 between DNN skill and autocorrelation statistics confirm these errors are highest in stochastic regions where the deterministic component of diabatic tendencies is weaker such as the tropical, marine boundary layer, and the mid-to-upper troposphere over convective land regions. But on longer timescales, particularly where there are distinct, deterministic patterns of atmospheric variation like the diurnal heating of the continents or baroclinic Rossby wave disturbances along mid latitude storm tracks, our DNN effectively emulates superparameterized diabatic processes. We find the highest R^2 coefficient of determination values (typically greater than 0.9) for daily- and zonal-mean predictions especially compelling (Figure 4c and d). Despite issues in precipitation emulation, our DNN captures much of the marine PDF of precipitation, though it has an unexpected drizzle bias over land that can be partially reduced via a positive constraint on precipitation 9. Despite these imperfections, precipitation statistics produced by the DNN are superior to conventional parameterization.

The accuracy achieved by our neural network suggests that simple feed-forward DNNs may still be the best way to create next generation, hybrid climate emulators. Skip-connections in conjunction with convolutions would seem to have possible advantages in allowing multiscale structures to be simultaneously prioritized in the loss function (Han et al., 2020). But our DNN achieves similar (Figures A1) to superior (Figure A2b vs. e), skill compared to the Resnet trained recently on similar data in Han et al. (2020). This would suggest that model architecture choices like skip connections and convolutional layers are less important to achieving a good fit for a neural network in the emulation of convection.

More broadly, these results also speak to an ongoing question of what sets the "parameterizability" of deep convection as reverse engineered by machine learning methods trained on superparameterized simulations, recognizing that despite their constraints SP includes nontraditional degrees of freedom like convective memory and organization. Our results suggest that convective memory may be much less essential than previously believed (Han et al., 2020; Colin et al., 2018), at least for simple feed-forward DNNs. Our results indicate that simple DNNs may not require memory from previous timesteps in learning convective tendencies – predictions independent of space and time may be the better way forward to achieve successful moist convection emulation. We find that the atmospheric modes of variation where convective memory would be most helpful (diurnal cycle, onset of afternoon deep convection/heavy precipitation, synoptic storms) are patterns our DNN, with no memory used in training, learns the best. Our issues in DNN emulation are greatest over regions where the controlling signals happen at the shortest temporal scales – especially around the tropical, marine boundary layer. These are exactly the places convective memory would be least helpful.

Looking ahead, we suggest a simple deep neural network, powered by automated hyperparameter tuning as well as physical constraints, may be the most realistic way for ML to emulate superparameterized moist convection in a realistic atmosphere with real-geography boundary conditions. This is also a more direct way to achieve two way coupling with a host climate model since simple DNNs can be rapidly deployed today as prognostic Fortran hybrid models thanks to new automated software (Ott et al., 2020). More broadly, for general deep learning applications in STEM, we believe our experience sheds light on the importance of incorporating a synthesis of physical science knowledge and exploiting machine learning methods when designing appropriate neural networks to tackle problems such as moist convection emulation. We have found like many others (Toms et al., n.d.) that while each of these design choices show notable improvements to emulation performance on their own, both are likely needed in conjuncture to utilize our DNNs to their fullest potential.

A note on emulating stochastic effects is appropriate, even if we have not attempted it here. To replace CRMs in a convection simulation, deep neural networks will likely eventually need to learn not just deterministic but also stochastic parameterizations, which are crucial to error and bias reduction (Kisi et al., 2013b; Palmer, 2019). Even in superparameterized climate simulations such stochastic effects, while not critical to mean climate, have been linked to some important regional precipitation extremes (Jones et al., 2019). Our results indicate there are high variance modes of moist convection that will be difficult for any simple DNN to represent perfectly without a faster time step interval in training data or a stochastic modeling component may prove to be necessary. Incorporating a generative component in future neural network simulations could enable accurate emulation of moist convection globally beyond locations with large scale, more deterministic modes of variance controlling circulation, convection, and precipitation. However, given the counter-effect of sea surface temperature gradients in the tropics on the details of convective organization and aggregation, replication of CRM subgrid-scale convective details in the boundary layer may be less paramount than in other portions of the atmosphere where our DNN already performs skillfully (Coppin & Bony, 2018), which would work to our advantage, and the strengths of our DNN are centered around deterministic components of convection, in creating a viable hybrid climate model from this trained DNN.

Meanwhile, a next step for the specific case of emulating superparameterization should be an online test of the performance of our trained DNN in prognostic mode to determine if the neural network is skilled enough to produce physically plausible outputs from the coupled run. In this limit, the secondary effects of stochasticity noted by (Jones et al., 2019) argue deterministic DNNs are appropriate. Although coupling emulated atmospheres to prognostic land models is mostly an unexplored frontier, we are optimistic

based on our first pilot tests that it is readily approachable; imperfections of the fit do not break standalone land model simulations. But carrying this forward into fully prognostic coupled tests will require significant work, such as expanding this prototype DNN's output vector to include additional variables needed to allow fully interactive land model coupling, and associated tuning. Even if this next step proves successful, simple DNNs should not be thought of as a panacea for all flaws in climate models – they cannot in their present application resolve biases induced by imperfect microphysics parameterization and the resulting errors in associated turbulence and cloud-radiative effects produced by superparameterized models. However, neural networks do still have broad use for sidestepping the computational bottlenecks that currently limit the global modeling community's ability to approach eddy-resolving scales. We remain excited about that potential, especially given our findings here that such approaches can be made remarkably skillful beyond aquaplanets.

Acknowledgments

GM acknowledges support from a NSF graduate fellowship under grant 1633631 and thanks the MAPS Program. GM, MP and TB acknowledge additional funding from NSF grants OAC-1835863 and AGS-1734164. The work of JO and PB in part supported by NSF NRT grant 1633631 to PB. The authors thank Stephan Rasp for the initial Github repository cloned for this work and as well as his guidance. We also are grateful to Stephan Mandt and Ruihan Yang for helpful conversations and ideas that advanced this work. The authors thank Chris Terai and Liran Peng for assistance with the simulation of SPCAM and CAM data. Computational resources were provided by the Extreme Science and Engineering Discovery Environment supported by NSF grant number ACI-1548562 (charge number TG-ATM190002). Code for preprocessing, training, and post-processing figure generation can be found at `https://github.com/gmooers96/Real_Geography_Manuscript`. All data for training, validation and post-processing can be taken from `https://tinyurl.com/y27n7zjs`.

References

Arakawa, A., & Schubert, W. H. (1974, April). Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *Journal of Atmospheric Sciences*, *31*(3), 674-701. doi: 10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2

Arnold, N. P., & Randall, D. A. (2015). Global-scale convective aggregation: Implications for the madden-julian oscillation. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1499-1518. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015MS000498` doi: 10.1002/2015MS000498

Benedict, J., & Randall, D. (2009, 11). Structure of the madden-julian oscillation in the superparameterized cam. *Journal of The Atmospheric Sciences - J ATMOS SCI*, *66*. doi: 10.1175/2009JAS3030.1

Beucler, T., Pritchard, M., Rasp, S., Gentine, P., Ott, J., & Baldi, P. (2019). Enforcing analytic constraints in neural-networks emulating physical systems. *arXiv preprint arXiv:1909.00912*.

Blossey, P. N., Bretherton, C. S., Cheng, A., Endo, S., Heus, T., Lock, A. P., & van der Dussen, J. J. (2016). Cgils phase 2 les intercomparison of response of subtropical marine low cloud regimes to co2 quadrupling and a cmip3 composite forcing change. *Journal of Advances in Modeling Earth Systems*, *8*(4), 1714-1726. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000765` doi: 10.1002/2016MS000765

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., ... Webb, M. J. (2015). Clouds, circulation and climate sensitivity. *Nature Geo-*

science, *8*(4), 261–268. Retrieved from https://doi.org/10.1038/ngeo2398 doi: 10.1038/ngeo2398

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). *Interpreting and stabilizing machine-learning parametrizations of convection.*

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289-6298. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL078510 doi: 10.1029/2018GL078510

Cheng, A., & Xu, K.-M. (2011, 07). Improved low-cloud simulation from a multiscale modeling framework with a third-order turbulence closure in its cloud-resolving model component. *Journal of Geophysical Research*, *116*. doi: 10.1029/2010JD015362

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, *44*(7), 2195–2214. Retrieved from https://doi.org/10.1007/s00382-014-2239-9 doi: 10.1007/s00382-014-2239-9

Colin, M., Sherwood, S., Geoffroy, O., Bony, S., & Fuchs, D. (2018, 12). Identifying the sources of convective memory in cloud-resolving simulations. *Journal of the Atmospheric Sciences*, *76*. doi: 10.1175/JAS-D-18-0036.1

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, *19*, 297–301. (URL: http://cr.yp.to/bib/entries.html#1965/cooley)

Coppin, D., & Bony, S. (2018). On the interplay between convective aggregation, surface temperature gradients, and climate sensitivity. *Journal of Advances in Modeling Earth Systems*, *10*(12), 3123-3138. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001406 doi: 10.1029/2018MS001406

Daleu, C. L., Plant, R. S., Woolnough, S. J., Sessions, S., Herman, M. J., Sobel, A., ... van Ulft, L. (2015). Intercomparison of methods of coupling between convection and large-scale circulation: 1. comparison over uniform surface conditions. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1576-1601. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015MS000468 doi: 10.1002/2015MS000468

Donner, L. J., & Phillips, V. T. (2003). Boundary layer control on convective available potential energy: Implications for cumulus parameterization. *Journal of Geophysical Research: Atmospheres*, *108*(D22). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003JD003773 doi: 10.1029/2003JD003773

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL078202 doi: 10.1029/2018GL078202

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2015, 11). Random forests for big data. *Big Data Research*. doi: 10.1016/j.bdr.2017.07.003

Grabowski, W. W. (2001, 05). Coupling cloud processes with the large-scale dynamics using the cloud-resolving convection parameterization (crcp). *J. Atmos. Sci.*, *58*, 978-997. doi: 10.1175/1520-0469(2001)058<0978:CCPWTL>2.0.CO;2

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002076 (e2020MS002076 2020MS002076) doi: 10.1029/2020MS002076

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.*

Herrera Cordova, V., Khoshgoftaar, T., Villanustre, F., & Furht, B. (2019, 12). Random forest implementation and optimization for big data analytics on lexisnexis's high performance computing cluster platform. *Journal of Big Data*, *6*. doi: 10.1186/s40537-019-0232-1

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*. (In press. Also: arXiv:2005.04048. Software available at: https://github.com/sherpa-ai/sherpa)

Hohenegger, C., & Stevens, B. (2016). Coupled radiative convective equilibrium simulations with explicit and parameterized convection. *Journal of Advances in Modeling Earth Systems*, *8*(3), 1468-1482. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000666` doi: 10.1002/2016MS000666

Jansson, F., van¬†den¬†Oord, G., Pelupessy, I., Gr–nqvist, J. H., Siebesma, A. P., & Crommelin, D. (2019). Regional superparameterization in a global circulation model using large eddy simulations. *Journal of Advances in Modeling Earth Systems*, *11*(9), 2958-2979. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001600` doi: 10.1029/2018MS001600

Jones, T. R., Randall, D. A., & Branson, M. D. (2019). Multiple-instance superparameterization: 2. the effects of stochastic convection on the simulated climate. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3521-3544. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001611` doi: 10.1029/2019MS001611

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

Kisi, O., Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013a). 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. Retrieved from `https://doi.org/10.1155/2013/485913` doi: 10.1155/2013/485913

Kisi, O., Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013b). 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. Retrieved from `https://doi.org/10.1155/2013/485913` doi: 10.1155/2013/485913

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016). Impacts of cloud superparameterization on projected daily rainfall intensity climate changes in multiple versions of the community earth system model. *Journal of Advances in Modeling Earth Systems*, *8*(4), 1727-1750. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000715` doi: 10.1002/2016MS000715

Krasnopolsky, V., Fox-Rabinovitz, M., Hou, Y., Lord, S., & Belochitski, A. (2010, 05). Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: Climate simulations and seasonal predictions. *Monthly Weather Review - MON WEATHER REV*, *138*, 1822-1842. doi: 10.1175/2009MWR3149.1

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Tolman, H. L., & Belochitski, A. A. (2008). Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors. *Neural Networks*, *21*(2), 535 - 543. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0893608007002663` (Advances in Neural Networks Research: IJCNN '07)

doi: https://doi.org/10.1016/j.neunet.2007.12.019

Kumar, A., Islam, T., Sekimoto, Y., Mattmann, C., & Wilson, B. (2020, 03). Convcast: An embedded convolutional lstm based architecture for precipitation nowcasting using satellite data. *PLOS ONE*, *15*(3), 1-18. Retrieved from https://doi.org/10.1371/journal.pone.0230114 doi: 10.1371/journal.pone.0230114

Lehmann, E. L. (1992). Introduction to student (1908) the probable error of a mean. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 29–32). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4612-4380-9_3 doi: 10.1007/978-1-4612-4380-9_3

Li, G., & Xie, S.-P. (2012). Origins of tropical-wide sst biases in cmip multi-model ensembles. *Geophysical Research Letters*, *39*(22). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL053777 doi: 10.1029/2012GL053777

Li, Z., Niu, F., Fan, J., Liu, Y., Rosenfeld, D., & Ding, Y. (2011). Long-term impacts of aerosols on the vertical development of clouds and precipitation. *Nature Geoscience*, *4*(12), 888–894. Retrieved from https://doi.org/10.1038/ngeo1313 doi: 10.1038/ngeo1313

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *in icml workshop on deep learning for audio, speech and language processing.*

Morrison, H., van Lier-Walqui, M., Fridlind, A. M., Grabowski, W. W., Harrington, J. Y., Hoose, C., ... Xue, L. (2020). Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS001689. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001689 (e2019MS001689 2019MS001689) doi: 10.1029/2019MS001689

Neelin, J. D., Peters, O., Lin, J. W.-B., Hales, K., & Holloway, C. E. (2008). Rethinking convective quasi-equilibrium: observational constraints for stochastic convective schemes in climate models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1875), 2579-2602. Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2008.0056 doi: 10.1098/rsta.2008.0056

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. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001351 doi: 10.1029/2018MS001351

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A fortran-keras deep learning bridge for scientific computing. *Scientific Programming*. (Article ID 8888811. https://doi.org/10.1155/2020/8888811. Also: arXiv:2005.04048)

Palmer, T. N. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), 463–471. Retrieved from https://doi.org/10.1038/s42254-019-0062-2 doi: 10.1038/s42254-019-0062-2

Parishani, H., Pritchard, M. S., Bretherton, C. S., Wyant, M. C., & Khairoutdinov, M. (2017). Toward low-cloud-permitting cloud superparameterization with explicit boundary layer turbulence. *Journal of Advances in Modeling Earth Systems*, *9*(3), 1542-1571. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS000968 doi: 10.1002/2017MS000968

Pendergrass, A., & Hartmann, D. (2014, 11). Two modes of change of the distribution of rain*. *Journal of Climate*, *27*, 8357-8371. doi: 10.1175/JCLI-D-14-00182.1

Pritchard, M. S., Bretherton, C. S., & DeMott, C. A. (2014). Restricting 32,Äì128 km horizontal scales hardly affects the mjo in the superparameterized community atmosphere model v.3.0 but the number of cloud-resolving grid columns constrains vertical mixing. *Journal of Advances in Modeling Earth Systems*, *6*(3), 723-739. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014MS000340 doi: 10.1002/2014MS000340

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003, 11). Breaking the Cloud Parameterization Deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547-1564. Retrieved from https://doi.org/10.1175/BAMS-84-11-1547 doi: 10.1175/BAMS-84-11-1547

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. Retrieved from https://www.pnas.org/content/115/39/9684 doi: 10.1073/pnas.1810286115

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3–5. Retrieved from https://doi.org/10.1038/nclimate3190 doi: 10.1038/nclimate3190

Siebesma, A., Soares, P., & Teixeira, J. (2007, 04). A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of The Atmospheric Sciences - J ATMOS SCI*, *64*. doi: 10.1175/JAS3888.1

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (n.d.). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *n/a*(n/a), e2019MS002002. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002 (e2019MS002002 2019MS002002) doi: 10.1029/2019MS002002

Voigt, A., Pincus, R., Stevens, B., Bony, S., Boucher, O., Bellouin, N., ... Zhang, H. (2017). Fast and slow shifts of the zonal-mean intertropical convergence zone in response to an idealized anthropogenic aerosol. *Journal of Advances in Modeling Earth Systems*, *9*(2), 870-892. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000902 doi: 10.1002/2016MS000902

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems computers, 2003* (Vol. 2, p. 1398-1402 Vol.2).

Watterson, I. G., & Dix, M. R. (2003). Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution. *Journal of Geophysical Research: Atmospheres*, *108*(D13). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002928 doi: 10.1029/2002JD002928

Yacalis, G. (2018). Artificial neural network impact on cloud parameterization and land-atmosphere interactions..

Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E., & Chen, C.-C. (2019). Application of the deep learning for the prediction of rainfall in southern taiwan. *Scientific Reports*, *9*(1), 12774. Retrieved from https://doi.org/10.1038/s41598-019-49242-6 doi: 10.1038/s41598-019-49242-6

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

Yuval, J., & O'Gorman, P. A. (2020b). *Use of machine learning to improve simulations of climate.*

Appendix A Performance Comparison with Previous Work

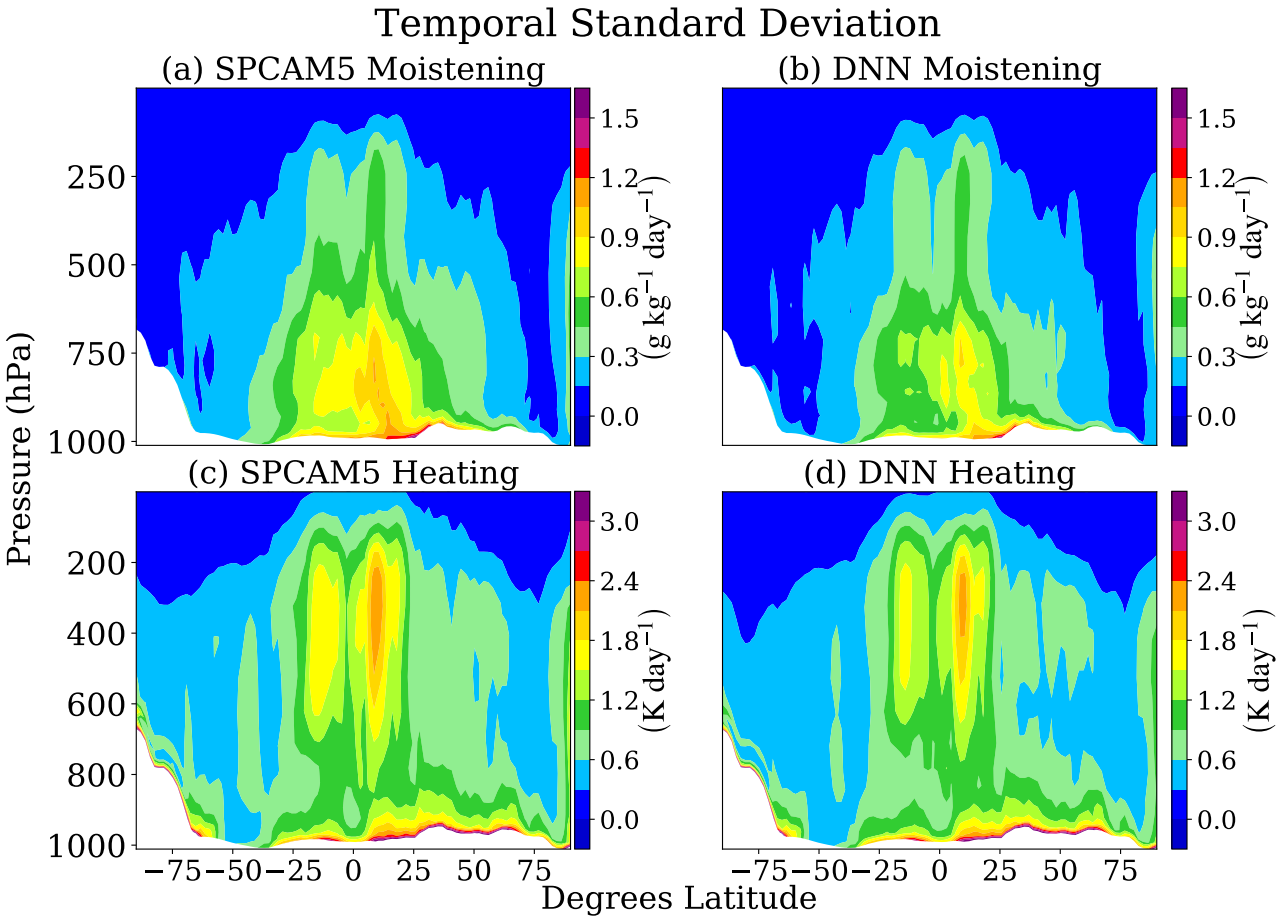

Figure A1. The temporal standard deviation of annual heating and moistening tendencies. Units converted to (K/day) and (g/kg/day) respectively to contrast with the performance of a Resnet (Han et al., 2020).

Figure A1 shows the extent to which the variations of the atmosphere, particularly those driven by deep convection and latent heating can be captured by a DNN with minimal under-prediction. The spatio-temporal patterns are replicated over the annual data.;

Overall, when looking at the annual mean, our DNN preforms well globally. These are some emulation troubles with intense tropical precipitation, but the heating and moistening tendency emulation is very close to the target data. In particular, figure A2 shows modes of variation can be captured at the surface by our DNN. Overall, the DNN seems to perform at least as well, if not better at locations like the planetary boundary layer when compared to a Resnet.

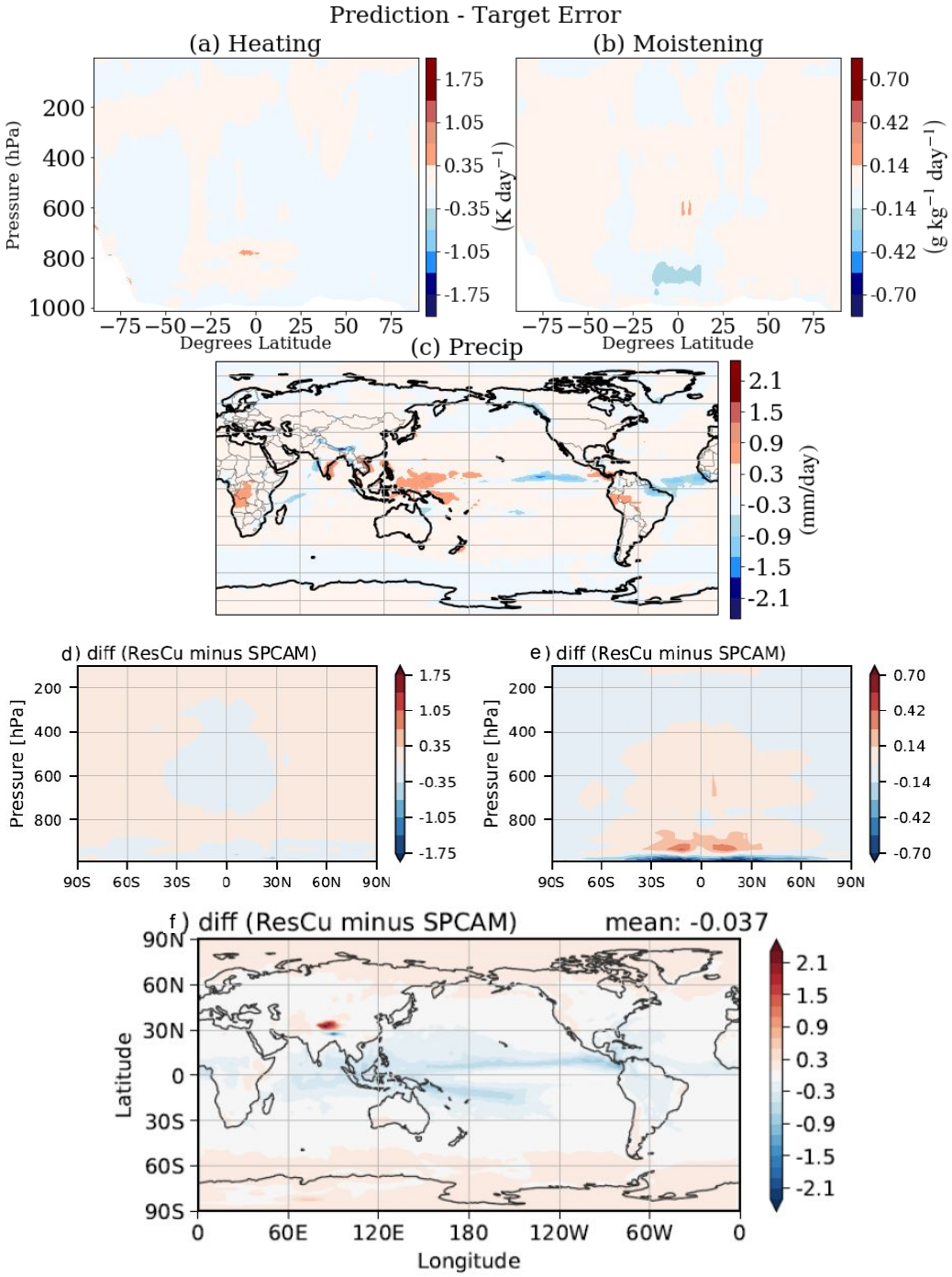

Figure A2. The difference between annual target SPCAM5 data and thee DNN predictions for heating tendency (K/day), moistening tendency (g/kg/day) and precipitation (mm/day). The 3 panels on the bottom have been taken from (Han et al., 2020) to provide direct comparisons between the performance of our DNN and their Resnet on simulation data with realistic, fully complexity real-geography boundary conditions