Toward efficient calibration of higher-resolution Earth System Models

Christopher G. Fletcher[1], William McNally[1], and John G. Virgin[1]

[1]University of Waterloo

November 23, 2022

Abstract

Projections of future climate change to support decision-making require Earth system models (ESMs) running at high spatial resolution, but this is computationally prohibitive. A major challenge is the calibration (parameter tuning) during the development of ESMs, which requires running large numbers of simulations to identify the optimal values for parameters that are poorly constrained by observations. Here we train a convolutional neural network (CNN) on perturbed parameter ensembles from two lower-resolution (and thus much less expensive) versions of the same ESM, and a smaller number of higher-resolution simulations. Cross-validated results show that the CNN's skill exceeds that of a climatological baseline for most variables with as few as 5-10 examples of the higher-resolution ESM, and for all variables (including precipitation) with at least 20 examples. This proof-of-concept study offers the prospect of significantly more efficient calibration of ESMs, by reducing the required CPU time for calibration by 20-40 \%

Toward efficient calibration of higher-resolution Earth System Models

Christopher G. Fletcher[1]*, William McNally[2], and John G. Virgin[1]

[1]Department of Geography and Environmental Management, University of Waterloo, Canada.
[2]Department of Systems Design Engineering, University of Waterloo, Canada.

Key Points:

- Calibration of poorly-constrained parameters in higher-resolution Earth system models (ESMs) is computationally expensive
- A novel machine-learning technique from computer vision can replace the ESM during calibration, even for complex variables like precipitation
- The machine-learning calibration reduces the computational costs of calibration by 20-40%

*200 University Ave West, Waterloo, Ontario, Canada.

Corresponding author: Christopher G. Fletcher, `chris.fletcher@uwaterloo.ca`

Abstract

Projections of future climate change to support decision-making require Earth system models (ESMs) running at high spatial resolution, but this is computationally prohibitive. A major challenge is the calibration (parameter tuning) during the development of ESMs, which requires running large numbers of simulations to identify the optimal values for parameters that are poorly constrained by observations. Here we train a convolutional neural network (CNN) on perturbed parameter ensembles from two lower-resolution (and thus much less expensive) versions of the same ESM, and a smaller number of higher-resolution simulations. Cross-validated results show that the CNN's skill exceeds that of a climatological baseline for most variables with as few as 5-10 examples of the higher-resolution ESM, and for all variables (including precipitation) with at least 20 examples. This proof-of-concept study offers the prospect of significantly more efficient calibration of ESMs, by reducing the required CPU time for calibration by 20-40 %

Plain Language Summary

To determine how Earth's future climate will respond to greenhouse gas emissions requires building accurate computer models. Building these models requires a time-consuming calibration process to find optimal values for uncertain constants (parameters) in the model equations that represent small-scale processes. We took a method (called CNN) that is commonly used in image recognition applications and inverted it to replicate the calibration process of the climate model. The CNN reproduces all of the main features of the global simulation of the climate model, including for precipitation which varies a lot from place-to-place, in a fraction of the computational time. The CNN also makes use of information contained in outputs from simpler versions of the climate model, which are available at lower cost. Our results suggest that inserting an artificial intelligence method, like CNN, in the calibration process for a climate model can reduce the time required by 20-40%.

1 Introduction

Quantitative projections of future climate change, with a robust estimate of their uncertainty, are critical to inform policy and decision-making. Earth System Models (ESMs) forced by emissions scenarios are the primary tool used to provide these projections, and modern ESMs incorporate sophisticated representations of Earth system processes, are physically self-consistent, show high fidelity with observations, and are computationally efficient enough to run large ensembles (Kay et al., 2014; Danabasoglu et al., 2020; Gent et al., 2011). However, most ESMs participating in the sixth phase of the Couple Model Intercomparison Project (CMIP6) have spatial grid resolutions of $\mathcal{O}$(100 km) on a side, which is often much too coarse to provide useful information to stakeholders; for example, spatial resolution finer than 10 km is required to study hydrologic change at the scale of small watersheds (Erler et al., 2019). Limited-area versions of ESMs called Regional Climate Models (RCMs) are used to downscale ESM simulations to resolutions as fine as 1 km, but this approach creates other problems such as physical inconsistencies between the driving model and RCM, scale mismatches at the lateral boundaries, and computational inefficiencies limiting ensemble size (Racherla et al., 2012; Luca et al., 2016). Statistical downscaling can be effective, but omits small-scale feedbacks and implicitly assumes stationarity in the downscaling model (Lanzante et al., 2018). Machine-learning based downscaling methods may overcome some of these limitations (Beusch et al., 2020; Heinze-Deml et al., 2020).

It is clear that the optimal solution is to build global ESMs at resolutions of $\mathcal{O}$(10 km), but these models are computationally prohibitive to develop and cal-

ibrate (Schär et al., 2020). Emulation offers an efficient alternative, by using a simpler empirical model to learn the behaviours of a more complex dynamical model (Kennedy & O'Hagan, 2001). Modern statistical learning methods have enabled more sophisticated emulation of ESMs, however, most previous studies have focused on simplified outputs, either through spatial averaging (Fletcher et al., 2018; Lee et al., 2011), or by first applying dimension-reduction methods like PCA (Salter & Williamson, 2019). Several studies have built emulators that represent the spatial structure of the ESM response; however, these tend to emulate one output variable at a time (Salter et al., 2018; Regayre et al., 2018).

Here we present a novel application of a statistical learning technique popular in computer vision to emulate global output from a higher-resolution ESM as a function of a number of uncertain input parameters. We demonstrate that the emulator can be trained effectively using a combination of inexpensive lower-resolution examples from the same ESM, and a relatively small number of high-resolution examples. The fully-trained emulator is able to accurately predict the impact of the input parameters on full global maps of a suite of seven output variables from the ESM, including precipitation. This represents a potentially significant pathway to expediting the calibration process for future generations of higher-resolution ESMs.

2 Models and Methods

2.1 Earth System Model

The climate model used in this study is the National Center for Atmospheric Research (NCAR) Community Earth System Model (CESM) Version 1.0.4 (Gent et al., 2011). For computational efficiency, and to support our focus on the influence of atmospheric parameterization on model uncertainty, we conduct all simulations using the F-compset configuration of CESM, which includes interactive atmosphere and land surface models, and prescribed climatological ocean surface temperatures and sea ice representative of the pre-industrial period (1850). The atmospheric model component used here is the Community Atmosphere Model Version 4 (CAM4) fully documented in Collins et al. (2006). The details of CAM4 pertinent to this study include its representation of aerosol-radiation interactions, but not aerosol-cloud interactions, and its finite-volume dynamical core, which is run here at three horizontal resolutions (referred to as higher, medium and lower). The higher-resolution configuration is a 0.9° x 1.25° latitude-longtiude grid (henceforth f09), which was the same resolution used in the CESM simulations that were contributed to the CMIP5 project (Taylor et al., 2011). Two lower-resolution configurations of the same version of CESM are also employed here, a medium one at 1.9° x 2.5° (f19), and a lower-resolution one at 4° x 5° (f45). The physics time-step (1,800 s) and vertical resolution are identical in all three configurations.

We conduct three perturbed parameter ensembles (PPEs) in total, one at each of the three spatial resolutions. In each PPE, 100 realizations are performed by perturbing nine atmospheric and aerosol parameters in CAM4 using values selected by Latin Hypercube Sampling (McKay et al., 1979). We emphasize that for the ith realization in each PPE the same set of parameter values are provided to CAM4. The nine parameters are the same ones that were perturbed in Fletcher et al. (2018), and they are listed in Table 1 but are not discussed here in detail because this study focuses on the application of a novel machine learning technique to a multi-resolution ensemble of PPEs. Several atmospheric parameters have different values in the default configuration for each resolution of CAM4 (these are the result of manual calibration by the model developers at NCAR), and we elect here to use each resolution's default values, rather than using identical parameter values across all simulations. Each realization is integrated for three years, and the outputs are

Table 1: List of parameters that are perturbed in this study, including for each parameter a description, the range of perturbed values, and the default value in CAM4 (where applicable).

parameter	description (CAM4 parameter name)	min	default	max	notes
x_1	Fraction of hygroscopic SO_4	0.0	0.0	1.0	Proxy for sulfate indirect effect (no units).
x_2	Spatial uniformity of BC (1 = globally uniform)	0.0	0.0	1.0	Proxy for BC aging and scavenging (no units).
x_3	Scaling factor for global BC mass	0.0	1.0	40.0	Proxy for uncertainty in BC emissions (no units)
x_4	Altitude for insertion of uniform BC layer	0.0	–	39.0	Proxy for vertical transport of BC (units km). Note: new parameter, no default.
x_5	RH threshold for low cloud formation (cldfrc_rhminl)	0.80	0.88	0.99	Value grid box RH must exceed before low cloud forms (no units)
x_6	Effective radius of liquid cloud droplets over ocean (cldopt_rliqocean)	8.4	14.0	19.6	(units microns)
x_7	Timescale for consumption rate of shallow CAPE (hkconv_cmftau)	900	1800	14440	(units seconds)
x_8	RH threshold for high cloud formation (cldfrc_rhminh)	0.50	0.50	0.85	Value grid box RH must exceed before high cloud forms (no units)
x_9	Timescale for consumption rate of deep CAPE (zmconv_tau)	1800	3600	28800	(units seconds)

time-averaged over all 36 months to reduce the influence of atmospheric internal variability.

The lower-resolution configurations of CESM use spatial grids of sizes 46×72 and 96×144, respectively. To ensure that the output from all three resolutions can be processed using the same machine learning architecture, the output data from the lower resolution ensembles are first upscaled using bilinear interpolation to match the largest (f09) grid size of 192×288. Nearest-neighbour and bi-cubic interpolation were also tested, to assess the sensitivity of our results to the method of regridding. Our findings (not shown) indicate that regardless of the method used, even for variables such as precipitation with large spatial variability, the conclusions of this work are unchanged.

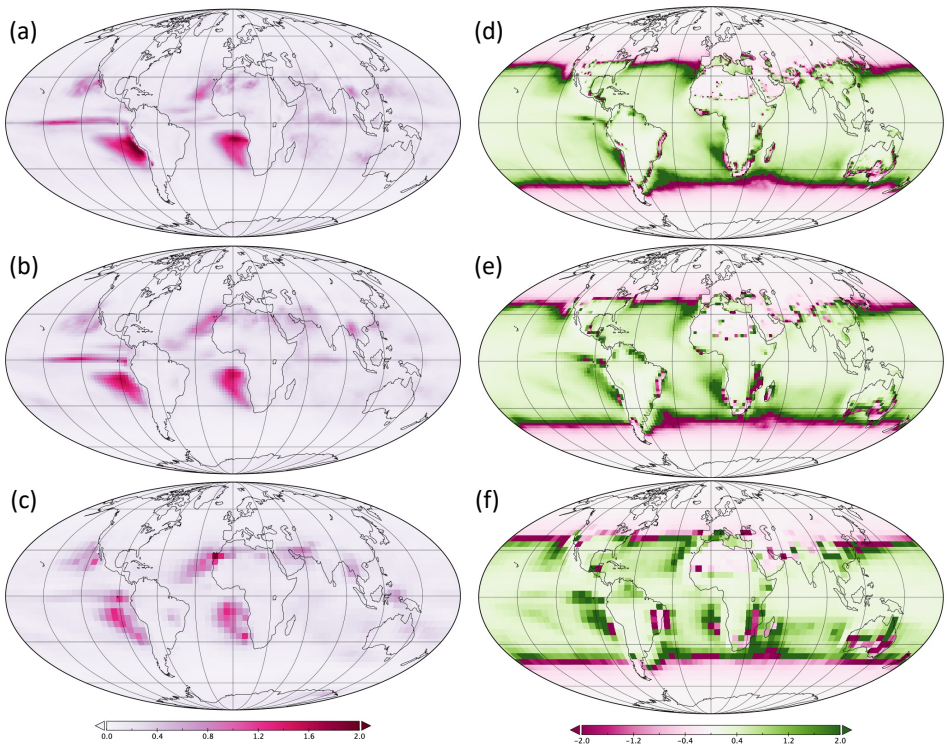


Figure 1: Coefficient of variation (calculated by dividing the ensemble mean by the ensemble standard deviation for each PPE) for the difference maps of annual mean (left) total precipitation (mm day^{-1}) and (right) net top-of-atmosphere radiative flux (W m^{-2}) over the 100 realizations of the perturbed parameter ensembles of CESM-CAM4 run at three horizontal resolutions: (a,d) f09, (b,e) f19 and (c,f) f45.

To quantify the impact on the atmospheric simulation from perturbations to the nine input parameters, Fig. 1 shows the coefficient of variation (i.e., ensemble spread) in total precipitation (PRECT) and net top-of-atmosphere radiative flux (FNET) at the three resolutions, where heavier shading represents greater variability within the ensemble. Total precipitation represents one of the most scientifically important, and most spatially variable, physical outputs, and therefore it neatly illustrates the relationship between the simulations at multiple resolutions (Fig. 1a-c). The regions of greatest variation are found in the (sub)tropical Pacific and Atlantic, where the parameter perturbations affect equatorial deep convection, and cloud formation in the subtropical dry zones off the western boundaries of Africa, North and South America. An important finding is that the magnitude of this variation increases at finer resolutions, suggesting that not all of the information about the influence of the parameters on precipitation is available at lower resolutions. In contrast, the impact of parameter perturbations on FNET is less sensitive to resolution because of the much lower spatial variability in that field (Fig. 1d-f). Even in the regions of greatest spread, roughly coinciding with western boundary currents and the edge of the global tropical belt, there is no evidence that variability changes as a function of resolution.

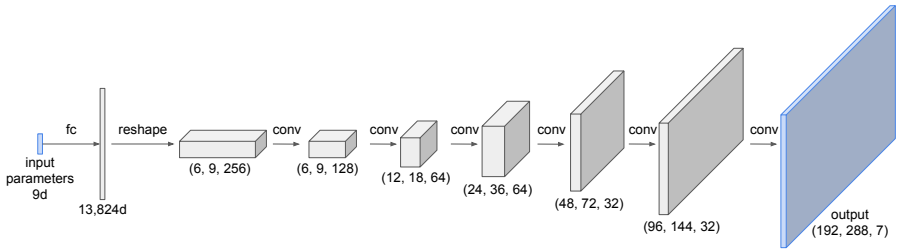


Figure 2: The architecture of the generative convolutional neural network used to predict seven spatially-resolved outputs of an ESM parameterized by nine aerosol forcing / atmospheric parameters. fc: fully-connected (dense) layer. conv: transpose convolution with a kernel size of 5x5.

2.2 Convolutional Neural Network

We emulate spatially-resolved outputs from CESM as a function of the nine uncertain atmospheric parameters using a generative convolutional neural network (CNN), as depicted in Fig. 2. CNN models are very common in computer vision applications and are ideally-suited to spatially-resolved targets (I. J. Goodfellow et al., 2014). Given sufficient training examples, the CNN learns a statistical representation of the underlying physical equations that relate changes in the parameter values to the spatially-resolved outputs. The CNN architecture includes seven layers to map the 9d input parameter vector to global maps of seven output variables (192 $\times$ 288 $\times$ 7). With the exception of the final convolution (conv) layer, all depicted layers are followed by batch normalization (Ioffe & Szegedy, 2015) and a leaky rectified linear unit (Maas et al., 2013). The 9d input is first projected to a 13,824d feature space using a fully-connected (fc) layer. The size of the 13,824d feature space is selected to allow a simple reshaping to a volume 6$\times$9$\times$256, which facilitates a series of transpose convolutions—sometimes referred to as deconvolutions—using a kernel size of 5$\times$5. The first transpose convolution uses a stride of 1, and all following transpose convolutions use a stride of 2, which doubles the spatial dimensions of the feature space so that after five convolutions the spatial resolution of the feature space matches the higher-resolution CESM grid (192$\times$288). The final transpose convolution uses 7 output channels to match the desired number of output variables being predicted, which includes low cloud fraction (CLDL), shortwave cloud forcing (SWCF), net top-of-the-atmosphere radiative flux (FNET) and total precipitation (PRECT).

The CNN was implemented in TensorFlow 2.2 using the Keras API. The neural network contains a total of 1.36M trainable parameters and 288M multiply-accumulate operations (MACs). On a desktop computer housing an NVIDIA TITAN Xp graphical processing unit (GPU) and a 12-core Intel i7-8700K processor (3.70 GHz), the CNN can process approximately 165 samples per second using a batch size of 1, and approximately 3400 samples per second using a batch size of 256[1].

[1] Speed averaged over 1000 forward passes of the network.

2.3 Training and Validation

To evaluate the CNN's ability to emulate the ESM, the CNN was trained in cross-validation mode using 80 randomly selected high-resolution (f09) samples, and tested on the remaining 20 samples. This entire training-testing process was repeated 40 times to estimate the uncertainty in the model fit that arises due to sampling variability. In practice, we train the CNN to predict the difference between the temporally averaged outputs of the default version of ESM and a perturbed ESM with a non-default parameterization, which we refer to as a set of *difference maps*. This method of learning the residual can potentially lead to improved training performance (He et al., 2016), where the intuition is that, in the extreme case when the perturbed CESM equals the default CESM, it is easier for the network to learn a zero mapping than an identity mapping. A single training example comprises an input vector $\mathbf{x}$ representing the nine parameter values, and a target set of difference maps $\mathbf{Y}$. We denote the predicted set of difference maps as $\hat{\mathbf{Y}}$. Prior to training the CNN, the input and output data were normalized to a range of [0, 1] by subtracting the minimum value and dividing by the maximum value. This was performed on a per-channel basis (i.e., per parameter for the input vector, and per output variable in the set of difference maps) using all 100 samples.

Training a neural network involves minimizing a loss function representing the error between the predicted and target outputs by iteratively updating the network parameters using gradient descent and backpropagation (I. Goodfellow et al., 2016). Since the selection of an appropriate loss function is a subjective element of the CNN architecture for each application, two different loss functions are compared here. The first is the mean squared error (MSE), which is commonly used in computer vision applications (Ledig et al., 2017; McNally et al., 2020). For convenience, we define the per-channel mean squared error (MSE_k) as

$$MSE_k(\mathbf{Y}^k, \hat{\mathbf{Y}}^k) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (\mathbf{Y}^k_{i,j} - \hat{\mathbf{Y}}^k_{i,j})^2, \tag{1}$$

where M and N represent the number of latitudinal and longitudinal grid points, respectively, and k represents the channel index (i.e., $\mathbf{Y}^k$ is a slice of $\mathbf{Y}$ representing a single difference map for the kth output variable). The mean squared error loss (L_{MSE}) for a single training example can then be defined as

$$L_{MSE} = \frac{1}{K} \sum_{k=1}^{K} MSE_k(\mathbf{Y}^k, \hat{\mathbf{Y}}^k), \tag{2}$$

where K is the number of channels, or output variables, in the set of difference maps.

Additionally, we propose a new loss function inspired by the spatial skill score metric that is often used to quantify the accuracy of climate models (Pierce et al., 2009) and has been used in our previous work to quantify the fidelity of a perturbed model version to a reference case (Fletcher et al., 2018). The skill metric (henceforth SS) is numerically similar to other model validation metrics such as Kling-Gupta Efficiency (Gupta et al., 2009):

$$SS_X = r_{p,d}^2 - \left[r_{p,d} - (\sigma_p/\sigma_d)\right]^2 - \left[(\bar{p} - \bar{d})/\sigma_d\right]^2 \tag{3}$$

where for a global grid of particular output variable X (e.g., precipitation, low cloud amount, etc.), p denotes the test case, and d denotes the ground truth, $r_{p,d}$ is the anomaly (pattern) correlation between X in p and d, σ is the spatial standard deviation of X, and overbars denote the global mean of X. Six output variables are included in the calculation of SS: low cloud fraction (CLDL), total precipitation

(PRECT), net radiative flux at the top-of-atmosphere (FNET), shortwave cloud forcing (SWCF), longwave cloud forcing (LWCF), and vertically-integrated longwave heating rate (QRL). We calculate SS for each variable separately, and then average the SS values to obtain the final SS for each test case. The per-channel (i.e., per output variable) skill score (SS_k) and skill score loss (L_{SS}) are defined as

$$SS_k = \frac{MSE_k(\mathbf{Y}^k, \hat{\mathbf{Y}}^k)}{MSE_k(\overline{\mathbf{Y}^k}, \hat{\mathbf{Y}}^k)} \quad L_{SS} = \frac{1}{K}\sum_{k=1}^{K} SS_k \tag{4}$$

where

$$\overline{\mathbf{Y}^k} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbf{Y}^k_{i,j}. \tag{5}$$

A perfectly predicted difference map will have SS_k=0, whereas the maximum possible value of SS_k is 1. While we use this convention out of convenience for the minimization problem, in our results we report 1–SS_k to be consistent with the skill score metric used in the climate modeling literature (Pierce et al., 2009). The CNN was trained for 500 epochs using each loss function. The Adam optimizer (Kingma & Ba, 2014) was used with a batch size of 8 and a cosine decay learning rate schedule (Loshchilov & Hutter, 2016) with initial learning rate of 0.01. Training took approximately 150 seconds on the TITAN Xp GPU.

3 Results

3.1 Overall performance of the CNN

We begin by showing how well the CNN is able to predict the 20% of unseen high-resolution (f09) outputs of CESM when trained on the remaining 80% of f09 cases. To illustrate how the CNN-predicted outputs compare to the original simulation from CESM, Fig. 3 shows the actual CESM output and the CNN predictions for CLDL, FNET, PRECT and SWCF in a randomly selected test case[2]. Qualitatively, the CNN achieves a high degree of learning about the relationship between the input parameters and changes to the spatial outputs in CESM. This includes relatively complex features of the precipitation response to parameter changes; for example, enhanced monsoon circulations over east Asia, reduced precipitation in tropical South America, and the latitudinal separation within the ITCZ in the tropical eastern Pacific and Atlantic basins. We emphasize that the CNN is provided with only the nine parameter values as an input, and predicts all seven output fields in a single calculation (Fig. 2).

Quantitative metrics averaged over all test cases and all seven output variables confirm the high fidelity of the CNN predictions, with very low average mean-squared error (MSE=4.07e^{-4}), and a high average skill score (SS=0.817). The likeness of the CNN predictions to the CESM output is very high for CLDL, FNET and SWCF, as measured by the skill score metric (SS>0.9). Precipitation represents the most challenging target for the CNN because of its very high spatial variability, and its mean skill score over all cases is somewhat lower than the multi-variable mean (SS=0.727). However, as shown by the single case in Fig. 3i, the CNN still represents the spatial variations in precipitation from the original CESM simulation with high fidelity. The finer-scale details that are *not* predicted by the CNN must, by definition, be explained by processes other than the parameters, for example atmospheric internal variability. That the CNN does not capture these details would

[2] Very similar features are found in all cases, and we avoid showing the mean here because we wanted to illustrate the CNN's ability to predict the finer-scale details of the response.

likely not be of critical importance to most model developers, since calibration is most often concerned with the representation of larger-scale features.

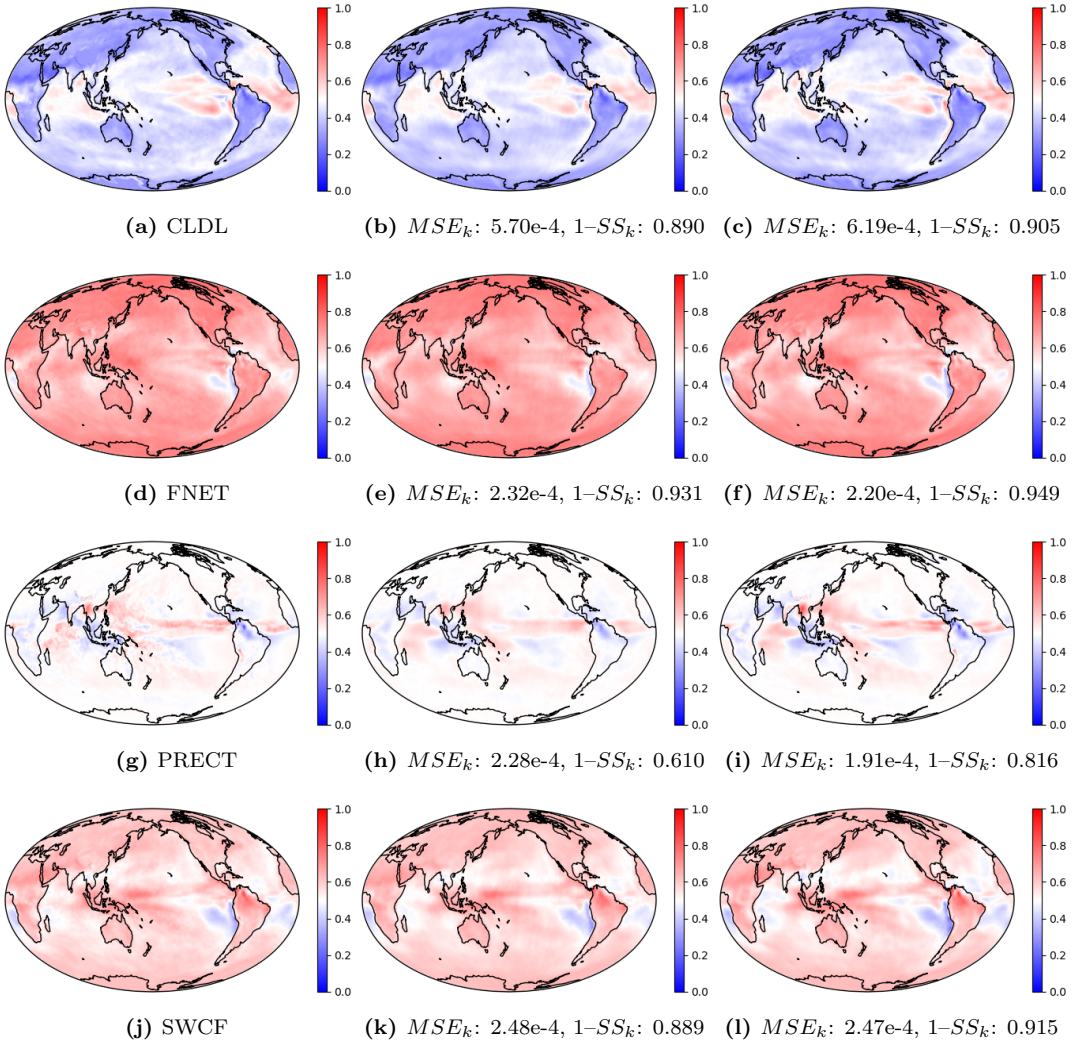
Comparing the middle and righthand columns of Fig. 3, the choice of the loss function used to train the CNN has an impact on prediction skill. For the single case shown, predictions trained on L_{SS} capture more sharply the features in the original simulation (left column), and this results in a higher skill score ($1–SS_k$) and lower MSE for most variables compared to predictions trained on L_{MSE}. Evaluating the effect of loss function across all cases, training on L_{SS} increases the average skill across all seven output variables by 15%, and for precipitation the skill is increased by 25% (not shown). The strong suggestion is that L_{SS} enables the CNN to better learn the spatial details of the output fields because it incorporates information about the spatial correlation and variability of these physical quantities in the training process.

3.2 Predicting high-resolution cases using low-resolution data

Having established that the CNN is able to produce high-fidelity predictions of CESM outputs taking only the parameter values as input, we next describe a practical application of the CNN-based emulator that follows the approach of Anderson and Lucas (2018) to extract information about a higher-resolution ESM from simulations with less expensive lower-resolution versions of the same ESM. We use the same CNN architecture as above (Fig. 2), but this time the training data includes all 200 lower resolution cases (f19 and f45), in addition to a number of high-resolution cases (n_{hr}) that is sequentially increased in our experiments from zero to 80. The goal is to determine how many higher-resolution examples the CNN requires before it can adequately learn the behaviour of the higher-resolution version of CESM. At each value of n_{hr}, 40 random trials are conducted and a separate CNN is trained in each random trial. This multi-resolution CNN is validated against predictions of the difference maps from 20 randomly selected f09 test samples that are excluded from the training data.

The input vector to the network was modified to include a tenth parameter representing the spatial resolution of the ESM from which the sample originated. Because the spatial grid areas vary by a factor of 4 between resolutions, the resolution values were set to 1, $\frac{1}{4}$, and $\frac{1}{16}$ for the low (f45), medium (f19), and high (f09) resolution cases, respectively (Anderson & Lucas, 2018). The training runs were configured in the same way as described in Section 2.3, except that the number of epochs was reduced to 200, and the batch size was increased to 16. These changes were made to reduce the overall number of training iterations required over many trials, and we verified that they did not have a material effect on the accuracy of the trained models (not shown).

The primary result is to illustrate how the skill metric ($1–SS$) varies as n_{hr} is progressively increased in the training data (Fig. 4). The mean skill score of the CNN averaged over all seven output variables is around 0.6 when the CNN is trained on *only* the lower-resolution cases (i.e., when n_{hr}=0; Fig. 4a). The skill increases approximately linearly to around 0.8 as more higher-resolution cases are included in the training data, but it plateaus for n_{hr} >40. This demonstrates that, when averaged over all variables, the lower-resolution versions of CESM alone provide the CNN with around 75 % of the information required to predict higher-resolution outputs. Increased prediction skill is achieved by introducing the higher-resolution training cases, and around 40 higher-resolution examples is optimal. Above n_{hr}=40 the returns diminish considerably, and so the benefit of running additional costly higher-resolution cases appears small.



(a) CLDL **(b)** MSE_k: 5.70e-4, 1–SS_k: 0.890 **(c)** MSE_k: 6.19e-4, 1–SS_k: 0.905

(d) FNET **(e)** MSE_k: 2.32e-4, 1–SS_k: 0.931 **(f)** MSE_k: 2.20e-4, 1–SS_k: 0.949

(g) PRECT **(h)** MSE_k: 2.28e-4, 1–SS_k: 0.610 **(i)** MSE_k: 1.91e-4, 1–SS_k: 0.816

(j) SWCF **(k)** MSE_k: 2.48e-4, 1–SS_k: 0.889 **(l)** MSE_k: 2.47e-4, 1–SS_k: 0.915

Figure 3: Global annual mean outputs for a randomly sampled test case in four of the seven predicted output variables: (top row) low cloud fraction, (second row) net top-of-atmosphere radiative flux, (third row) total precipitation, (bottom row) shortwave cloud forcing. All quantities have been normalized and so units are dimensionless. Left column shows the original simulations from CESM. The middle (right) column shows the predictions from the CNN trained with the L_{MSE} (L_{SS}) loss function (see Section 2.3 for details). The values below each panel in the middle and right columns show the mean-squared error (MSE_k), and the skill metric (1–SS_k), compared to the original simulation in the left column.

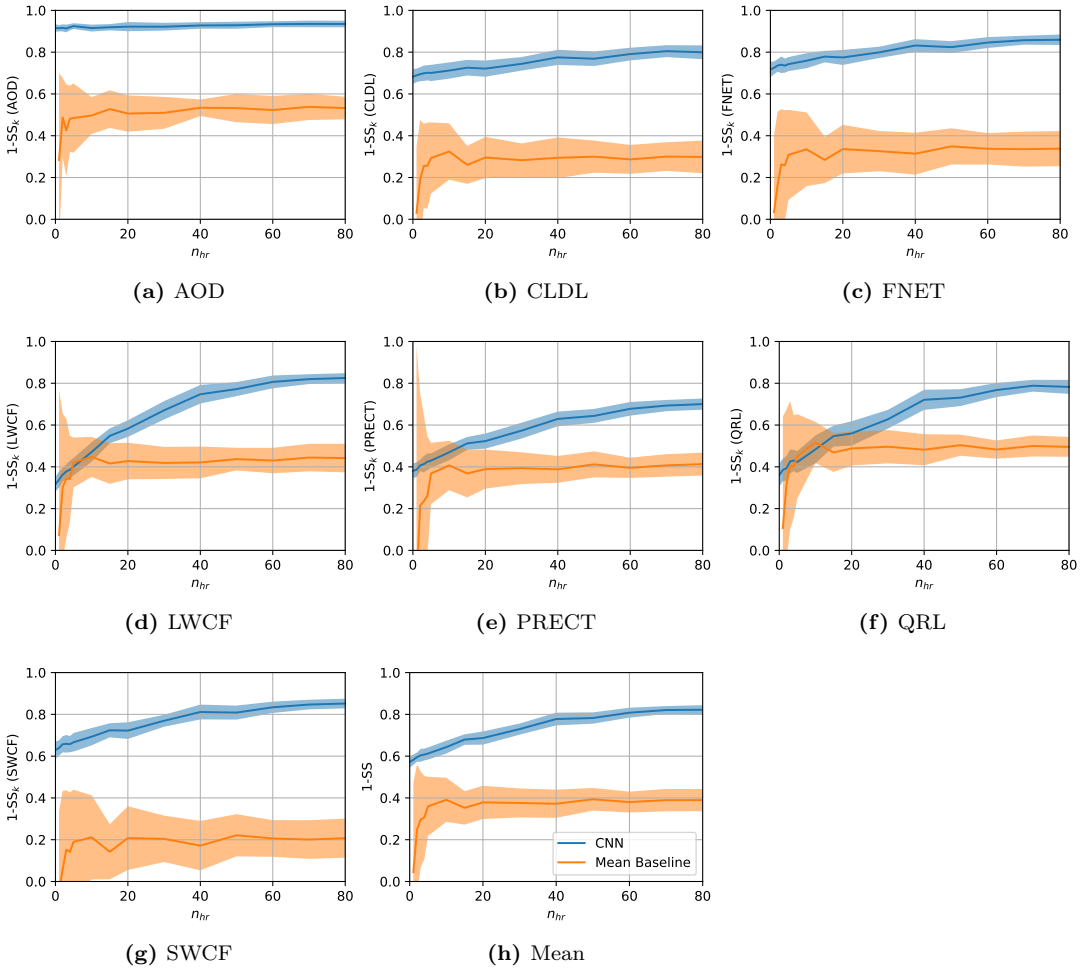
Since our CNN-based emulator is relatively complex, and has not been applied to ESMs before, it is important to benchmark the benefit of using this approach versus a simpler one. In studies that apply machine learning methods for climate modelling, a commonly-used benchmark is a general linear model (GLM; e.g., Fletcher et al. (2018)). However, since here the CNN emulates the entire set of spatially-resolved output arrays as a single function of the nine input parameters, it was not possible to provide a simple comparison using a GLM. Instead, we compare the skill derived from the fully-trained CNN to a baseline skill value obtained from a null model that assumes the CNN simply predicts the climatological mean output at each grid cell for each combination of input parameters. The orange line shows that the baseline null model achieves a mean skill score of around 0.4 when at least 10 higher-resolution examples are included in the calculation of the climatological mean. Importantly, the fully-trained CNN model outperforms the baseline skill for all values of n_{hr}.

The mean skill score peaks at 0.8 because the CNN produces different levels of skill for different output variables. Aerosol optical depth (AOD) shows uniformly high skill, even with n_{hr}=0 (Fig. 4b), because it is highly constrained by the prescribed aerosol mass concentrations employed by CAM4, and was shown to be a strong linear function of a single parameter (x_1) (Fletcher et al., 2018). In contrast, other variables like total precipitation exhibit systematically lower skill values (0.4 at n_{hr}=0, peaking at 0.7 with n_{hr}=80, Fig. 4f). This mirrors the result seen in Fig. 3 for a single case, where lower skill is found for variables like precipitation whose spatial output contains finer-scale spatial variability. This is a result very familiar to climate modellers, who have reported for decades that the spatial distribution of precipitation is the most challenging target for coarse-resolution ESMs (Luca et al., 2016). Here, the result serves as a reminder that the CNN does not represent a panacea: the skill of its predictions depends on the spatial complexity of the output variable being emulated. It also suggests that the skill of the CNN-based emulator has an upper-limit that does *not* appear to be caused by underfitting due to too few training cases. For all variables, including precipitation, the skill score has effectively plateaued at n_{hr}=80, suggesting that it is unlikely to improve substantially further even if a much larger training sample was available. Our conclusion is that fine-scale details of the output (e.g., Fig. 3g for precipitation) are related more to internal atmospheric variability than to parameter uncertainty, and thus cannot be captured by the CNN, by definition.

3.3 Sources of uncertainty in the emulator

The width of the shaded blue envelope in Fig. 4 displays the variance in the CNN predictions due to sampling variability across the 40 realizations of the training-testing process, while the orange shading displays the variance in the climatological mean computed over n_{hr} samples. For all values of n_{hr} the impact of sampling is smaller for the trained CNN model than for the null model, showing that there is greater variability in skill when using the climatological mean to predict each output variable than using the CNN. At n_{hr} <20 the uncertainty in the null model is very large, which relates to the instability of the calculation of the climatological average from the small sample of higher-resolution cases available to train the CNN. At n_{hr} >50 the uncertainty in both models has stabilized; however, it remains larger for the null model than the fully-trained CNN. We do not have a precise explanation for this behaviour, other than to say that it strongly suggests that the trained CNN model is able to extract a larger deterministic signature from the training data, which causes the performance of the CNN to be less variable from realization to realization. We also note that the variance in the climatological mean is larger for variables like FNET and SWCF, and smaller for PRECT and QRL, and

the climatological mean variance appears unrelated to the overall skill or uncertainty of the CNN model.



(a) AOD **(b)** CLDL **(c)** FNET

(d) LWCF **(e)** PRECT **(f)** QRL

(g) SWCF **(h)** Mean

Figure 4: The blue line shows the skill of the CNN in predicting high-resolution difference maps after being trained on the full lower- and medium-resolution ensembles, plus an increasing number (n_{hr}) of high-resolution samples. Panels (a-g) show the skill for the individual seven outputs, and panel (h) shows the mean skill. The orange line shows the skill from using the climatological mean of the n_{hr} high resolution samples included in the training set. The shading indicates the cross-validated uncertainty from 40 randomized trials.

Since the CNN is a multivariate model, it is instructive to examine which of the input parameters is most important for predicting the seven output variables. Other machine learning methods like randomForest provide feature importance by default (Anderson & Lucas, 2018). However, to obtain feature importance for the CNN we first permute the parameter values by randomly shuffling them among the 20 held-back high-resolution cases being predicted in each of the 40 resampling iterations (see Section 2.3), and then calculate the average reduction in prediction skill between the default and permuted realizations. The results show that parameter x_5—which directly controls the amount of low cloud in the model—is the most

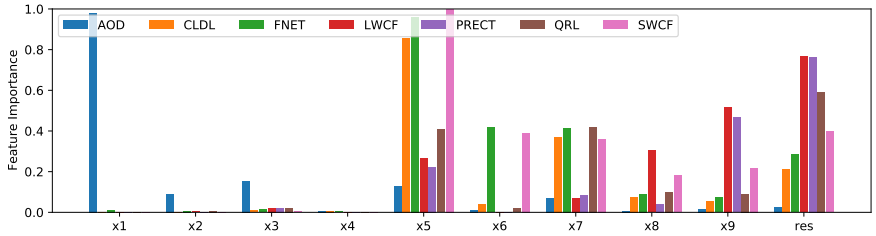


Figure 5: Normalized parameter importance for all nine atmospheric input parameters (x_1–x_9) and the resolution parameter (res) from the multi-resolution emulator with $n_{hr} = 40$. The height of each colored bar shows the importance of a given parameter to that output variable, relative to the highest importance (SWCF for x_5).

important for multiple outputs, notably low cloud fraction, net radiation and shortwave cloud forcing (Fig. 5). In agreement with Anderson and Lucas (2018), spatial resolution (res) is most important for precipitation and longwave cloud forcing, with both tending to be large in regions of tropical deep convection. Parameter x_9, which controls the timescale for the consumption of CAPE in deep convection, is also moderately important for precipitation, indicating that the coupling between precipitation and resolution occurs primarily through the Zhang-McFarlane deep convection parameterization in CAM4 (Neale et al., 2013).

Operationally, the degree to which the emulated global maps can be used to support calibration depends on the accuracy of the CNN predictions. Using our framework we can explicitly validate this accuracy through the mean prediction error for the spatially-resolved predictions against the original CESM simulation. The prediction errors are generally small, but vary depending on which output variable is being predicted. Very small errors are found everywhere for low cloud fraction (CLDL, Fig. 6 top row), whereas other outputs like net radiative flux, precipitation and shortwave cloud forcing (SWCF) show locally larger amplitude errors (Fig. 6). The global mean error for all quantities *except* SWCF is roughly 1-2 % of the magnitude of the climatological mean, whereas for SWCF it is almost 10 % because of a substantial positive bias across most regions of the globe that is particularly strong over the west Pacific and Amazon basin. In all variables the error decreases approximately linearly with increasing n_{hr} (left to right in Fig. 6), further demonstrating the value of including some high-resolution information in the training data for the CNN. Interestingly, the spatial *pattern* of the errors remains very similar for different n_{hr}, indicating that the CNN predictions are systematically better/worse in some regions than others, for reasons that we do not yet fully understand. There is a clear difference in the spatial pattern of errors between CLDL and SWCF, despite SWCF being controlled primarily by the spatial distribution of low clouds (Zelinka et al., 2020). One possible explanation is that the regions of largest SWCF error in the Pacific tend to coincide in CESM-CAM4 with regions of vertically deep convection and widespread *high* cloud layers, which are associated with significant obscuration of the low cloud radiative effect (Virgin et al., 2021).

4 Discussion and Conclusions

The convolutional neural network (CNN) approach employed here is popular in the field of computer vision but has not, to our knowledge, been used previously to emulate an Earth system model. While computationally-efficient and ideally

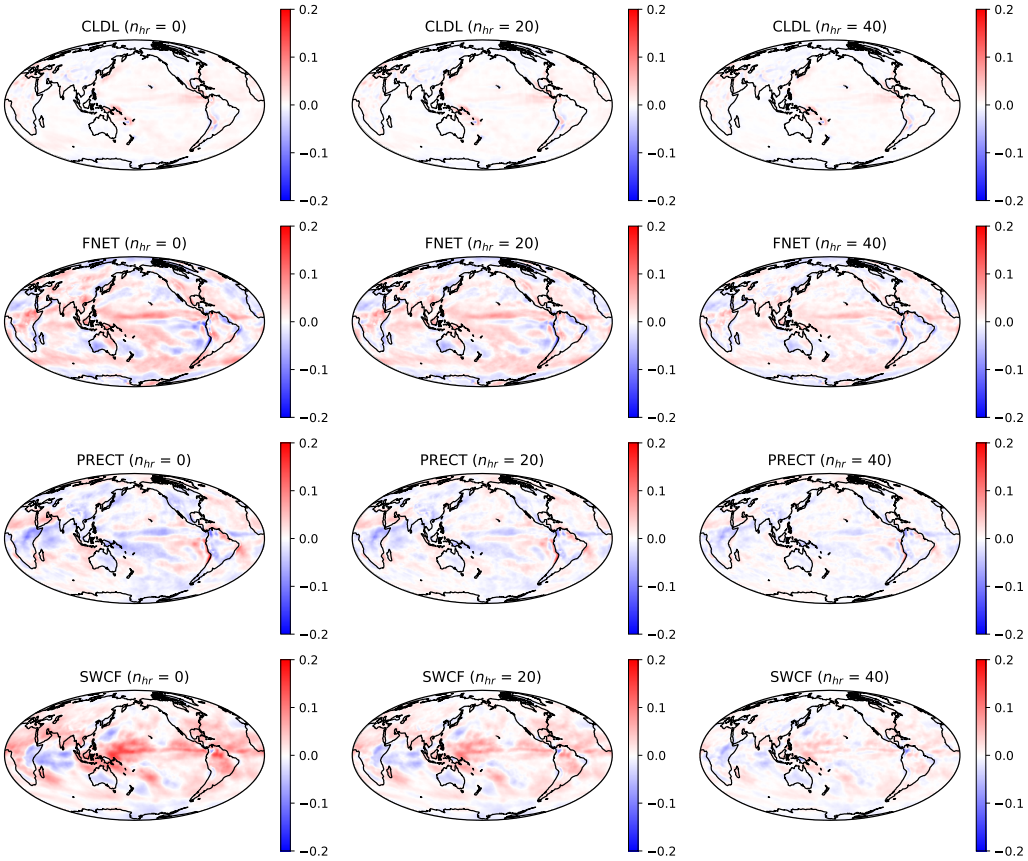


Figure 6: Global annual mean prediction errors for the CNN in four of the seven predicted output variables: (top row) low cloud fraction, (second row) net top-of-atmosphere radiative flux, (third row) total precipitation, (bottom row) shortwave cloud forcing. All quantities have been normalized and so units are dimensionless. The left column is for predictions that include zero high-resolution examples ($n_{hr} = 0$) in the training data for the CNN, the middle column is for $n_{hr} = 20$, and the right column is for $n_{hr} = 40$.

suited for predicting multivariate spatially-resolved outputs, CNN models typically require large ($\mathcal{O}(10^4)$) training sets to produce accurate predictions. In this study, we obtained useful predictive skill with around 240 training samples of differing spatial resolutions, and this is likely because the training and validation data are both computer-generated by the same ESM, meaning they contain less noise than observation-based data. Alternative approaches to constructing a multi-resolution emulator are conceivable; for example, using an image-to-image translation, where the lower-resolution data are the inputs to the CNN, and the higher-resolution version is the target (downscaling). However, since the motivation here is the calibration of uncertain parameters, one would also have to consider how the perturbed parameter values that correspond to each training case would be incorporated into the downscaling process. For this reason we believe that the architecture shown in Fig. 2 represents a simple and efficient way for model developers to assess the impact of parameter values on an array of spatially-resolved ESM outputs.

The deterministic nature of the CNN predictions constitutes a limitation of the approach, because the CNN does not include an estimate of its own uncertainty that would be available with probabilistic methods such as Gaussian Process regression (Lee et al., 2011). We attempt to account for prediction uncertainty by repeating the entire CNN training and testing procedure 40 times using random sampling, but this obviously is limited to sampling only the variability present within the ensembles of model output. An additional source of uncertainty not fully accounted for here is internal variability in the ESM simulation itself; the 36-month climatology may not be sufficient to properly characterize all timescales of atmospheric internal variability (Milinski et al., 2020). The input parameter values are also uncertain (Table 1), but we did not demonstrate here how they would actually be "calibrated", in the sense of using the CNN to identify their optimal values. In practice, modeling centres could employ 'history-matching' using the CNN, by comparing the predicted maps against a set of reference observations to identify regions of parameter space that produce plausible climates (McNeall et al., 2016).

Our results show that a highly accurate emulator can be trained using relatively few iterations of the higher-resolution ESM, thus offering the potential for significantly improved efficiency in the calibration process. To illustrate the time and resource saving associated with our approach, the CPU time required to run CESM-CAM4 at f09 resolution is a factor of 16 higher than at f45 resolution. As a result, the total CPU time required to complete the two 100-member ensembles at lower resolution, plus n_{hr}=20 (n_{hr}=40) higher-resolution simulations, is reduced by 40 % (20 %) compared to producing only a 100-member ensemble of the higher-resolution model. Assuming that similar statistical relationships extend to grid resolutions finer than f09—which are more relevant for decision-makers—one could theoretically expect even greater efficiency gains for an ESM with resolution $\mathcal{O}$(10 km). An interesting follow-on question is whether the CNN in this study, trained on output from CESM, could be used to emulate other ESMs. In principle, useful predictive information on the relationship between aerosol, cloud and precipitation parameters in CESM *could* be applied to help calibrate other models, but one important limitation is that different ESMs employ different physical parameterization schemes. This means that some/all of the parameters being calibrated in CESM are unlikely to exist in other ESMs; in fact, many of the parameters being calibrated in this study, described in detail in Fletcher et al. (2018), have been replaced or superseded in more recent versions of CESM-CAM (Boyle et al., 2015; Danabasoglu et al., 2020). It seems likely, therefore, that a unique CNN would need to be trained for each ESM, unless they shared parameterization schemes.

The choice of n_{hr} is somewhat subjective, and depends on what constitutes sufficiently high skill of the emulator to enable calibration. With this CNN the skill

score for precipitation only reaches 0.7 at n_{hr}=40, yet model developers may consider the predicted pattern of precipitation in Fig. 3i to be adequate. If the target field is more spatially homogeneous, like FNET, then only n_{hr}=20 may be required, and these decisions will likely differ for individual modeling centers. The outcome may also be sensitive to the region, and/or season, of interest. We consider only parametric uncertainty here, and emulation could feasibly be used to examine structural uncertainty in ESMs (Watson-Parris, 2020; Watson, 2019). Future work will also evaluate the CNN-based emulator in an operational-like setting, where the calibration of parameters is typically performed by minimizing the difference between the ESM and observational data (Hourdin et al., 2016), rather than against the default version of the ESM as here. The computational efficiency of the emulator means that using it to replace the ESM in the calibration process allows for a much larger sample of parameter combinations to be evaluated, with the implication that the final calibrated model will provide a better representation of the observed climate (Hourdin et al., 2021). Finally, even greater computational efficiency gains could be made by using the CNN-based emulator to calibrate higher-resolution configurations of fully-coupled ESMs with an interactive ocean model, including training the CNN to predict temporally-resolved outputs from transient climate simulations (for example, with time-evolving greenhouse gas forcing).

Acknowledgments

We thank the Microsoft AI for Earth program via the Waterloo AI Institute for funding, and three anonymous reviewers at the ICML 2021 conference for their helpful comments. The model simulation output data used for training and testing the CNN emulator in the study, and all code required to perform the analysis and plotting, are available in our research group GitHub repository via `https://github.com/Fletcher-Climate-Group` under a creative commons license.

References

Anderson, G. J., & Lucas, D. D. (2018). Machine learning predictions of a multiresolution climate model ensemble. *Geophysical Research Letters*, *45*(9), 4273–4280.

Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020, February). Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*, *11*(1), 139–159. Retrieved 2021-06-21, from `https://esd.copernicus.org/articles/11/139/2020/` (Publisher: Copernicus GmbH) doi: 10.5194/esd-11-139-2020

Boyle, J. S., Klein, S. A., Lucas, D. D., Ma, H.-Y., Tannahill, J., & Xie, S. (2015, February). The parametric sensitivity of CAM5's MJO. *Journal of Geophysical Research: Atmospheres*, *120*(4), 2014JD022507. Retrieved 2017-11-20, from `http://onlinelibrary.wiley.com/doi/10.1002/2014JD022507/abstract` doi: 10.1002/2014JD022507

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., ... Zhang, M. (2006, June). The Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). *Journal of Climate*, *19*(11), 2144–2161. Retrieved 2017-06-29, from `http://journals.ametsoc.org/doi/abs/10.1175/JCLI3760.1` doi: 10.1175/JCLI3760.1

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., ... Strand, W. G. (2020). The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. Retrieved 2021-01-20, from `https://agupubs`

.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001916) doi: https://doi.org/10.1029/2019MS001916

Erler, A. R., Frey, S. K., Khader, O., d'Orgeville, M., Park, Y.-J., Hwang, H.-T., ... Sudicky, E. A. (2019). Simulating Climate Change Impacts on Surface Water Resources Within a Lake-Affected Region Using Regional Climate Projections. *Water Resources Research*, *55*(1), 130–155. Retrieved 2021-06-01, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024381 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR024381) doi: https://doi.org/10.1029/2018WR024381

Fletcher, C. G., Kravitz, B., & Badawy, B. (2018, December). Quantifying uncertainty from aerosol and atmospheric parameters and their impact on climate sensitivity. *Atmospheric Chemistry and Physics*, *18*(23), 17529–17543. Retrieved 2019-03-06, from https://www.atmos-chem-phys.net/18/17529/2018/ doi: 10.5194/acp-18-17529-2018

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., ... Zhang, M. (2011, October). The Community Climate System Model Version 4. *Journal of Climate*, *24*(19), 4973–4991. Retrieved 2012-01-16, from http://journals.ametsoc.org/doi/abs/10.1175/2011JCLI4083.1 doi: 10.1175/2011JCLI4083.1

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009, October). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. Retrieved 2018-01-15, from http://www.sciencedirect.com/science/article/pii/S0022169409004843 doi: 10.1016/j.jhydrol.2009.08.003

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Heinze-Deml, C., Sippel, S., Pendergrass, A. G., Lehner, F., & Meinshausen, N. (2020, October). Latent Linear Adjustment Autoencoders v1.0: A novel method for estimating and emulating dynamic precipitation at high resolution. *Geoscientific Model Development Discussions*, 1–39. Retrieved 2021-06-21, from https://gmd.copernicus.org/preprints/gmd-2020-275/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-2020-275

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... Williamson, D. (2016, July). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*. Retrieved 2016-12-02, from http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-15-00135.1 doi: 10.1175/BAMS-D-15-00135.1

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., ... Volodina, V. (2021). Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2020MS002225. Retrieved 2021-07-06, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002225 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002225) doi: 10.1029/2020MS002225

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the international*

conference on machine learning.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2014, November). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. Retrieved 2017-04-10, from http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-13-00255.1 doi: 10.1175/BAMS-D-13-00255.1

Kennedy, M. C., & O'Hagan, A. (2001, January). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464. Retrieved 2018-03-09, from http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00294/abstract doi: 10.1111/1467-9868.00294

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

Lanzante, J. R., Dixon, K. W., Nath, M. J., Whitlock, C. E., & Adams-Smith, D. (2018, April). Some Pitfalls in Statistical Downscaling of Future Climate. *Bulletin of the American Meteorological Society*, *99*(4), 791–803. Retrieved 2021-05-31, from https://journals.ametsoc.org/view/journals/bams/99/4/bams-d-17-0046.1.xml (Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society) doi: 10.1175/BAMS-D-17-0046.1

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... others (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W., & Spracklen, D. V. (2011, December). Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. *Atmos. Chem. Phys.*, *11*(23), 12253–12273. Retrieved 2018-03-19, from https://www.atmos-chem-phys.net/11/12253/2011/ doi: 10.5194/acp-11-12253-2011

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Luca, A. D., Argüeso, D., Evans, J. P., Elía, R. d., & Laprise, R. (2016). Quantifying the overall added value of dynamical downscaling and the contribution from different spatial scales. *Journal of Geophysical Research: Atmospheres*, *121*(4), 1575–1590. Retrieved 2021-06-01, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD024009 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD024009) doi: https://doi.org/10.1002/2015JD024009

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the international conference on marchine learning.*

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979, May). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, *21*(2), 239. Retrieved 2018-11-06, from https://www.jstor.org/stable/1268522?origin=crossref doi: 10.2307/1268522

McNally, W., Vats, K., Wong, A., & McPhee, J. (2020). Evopose2d: Pushing the boundaries of 2d human pose estimation using neuroevolution. *arXiv preprint arXiv:2011.08446*.

McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A., & Sexton, D. (2016, November). The impact of structural error on parameter constraint in a climate model. *Earth Syst. Dynam.*, *7*(4), 917–935. Retrieved 2018-03-09, from https://www.earth-syst-dynam.net/7/917/2016/ doi:

10.5194/esd-7-917-2016

Milinski, S., Maher, N., & Olonscheck, D. (2020, October). How large does a large ensemble need to be? *Earth System Dynamics*, *11*(4), 885–901. Retrieved 2021-08-24, from https://esd.copernicus.org/articles/11/885/2020/ (Publisher: Copernicus GmbH) doi: 10.5194/esd-11-885-2020

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013, January). The Mean Climate of the Community Atmosphere Model (CAM4) in Forced SST and Fully Coupled Experiments. *Journal of Climate*, *26*(14), 5150–5168. Retrieved 2016-11-26, from http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-12-00236.1 doi: 10.1175/JCLI-D-12-00236.1

Pierce, D. W., Barnett, T. P., Santer, B. D., & Gleckler, P. J. (2009). Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, *106*(21), 8441–8446.

Racherla, P. N., Shindell, D. T., & Faluvegi, G. S. (2012). The added value to global model projections of climate change by dynamical downscaling: A case study over the continental U.S. using the GISS-ModelE2 and WRF models. *Journal of Geophysical Research: Atmospheres*, *117*(D20). Retrieved 2019-10-29, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD018091 doi: 10.1029/2012JD018091

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., ... Carslaw, K. S. (2018, July). Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF. *Atmospheric Chemistry and Physics*, *18*(13), 9975–10006. Retrieved 2019-10-29, from https://www.atmos-chem-phys.net/18/9975/2018/ doi: https://doi.org/10.5194/acp-18-9975-2018

Salter, J. M., & Williamson, D. B. (2019, June). Efficient calibration for high-dimensional computer model output using basis methods. *arXiv:1906.05758 [stat]*. Retrieved 2020-06-05, from http://arxiv.org/abs/1906.05758 (arXiv: 1906.05758)

Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2018, September). Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases. *Journal of the American Statistical Association*, 1–24. Retrieved 2019-10-11, from https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1514306 doi: 10.1080/01621459.2018.1514306

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Girolamo, S. D., ... Wernli, H. (2020, May). Kilometer-Scale Climate Models: Prospects and Challenges. *Bulletin of the American Meteorological Society*, *101*(5), E567–E587. Retrieved 2021-06-01, from https://journals.ametsoc.org/view/journals/bams/101/5/bams-d-18-0167.1.xml (Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society) doi: 10.1175/BAMS-D-18-0167.1

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2011, October). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. Retrieved 2016-06-05, from http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1 doi: 10.1175/BAMS-D-11-00094.1

Virgin, J. G., Fletcher, C. G., Cole, J. N. S., von Salzen, K., & Mitovski, T. (2021, August). Cloud Feedbacks from CanESM2 to CanESM5.0 and their influence on climate sensitivity. *Geoscientific Model Development*, *14*(9), 5355–5372. Retrieved 2021-09-20, from https://gmd.copernicus.org/articles/14/5355/2021/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-14-5355-2021

Watson, P. A. G. (2019). Applying Machine Learning to Improve Simulations of a Chaotic Dynamical System Using Empirical Error Correction. *Journal of Advances in Modeling Earth Systems*,

11(5), 1402–1417. Retrieved 2021-06-01, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001597 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001597) doi: https://doi.org/10.1029/2018MS001597

Watson-Parris, D. (2020, October). Machine learning for weather and climate are worlds apart. *arXiv:2008.10679 [physics, stat]*. Retrieved 2020-11-01, from http://arxiv.org/abs/2008.10679 (arXiv: 2008.10679)

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1), e2019GL085782. Retrieved 2020-01-17, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782 doi: 10.1029/2019GL085782