Toward efficient calibration of higher-resolution Earth System Models

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

Abstract

Projections of future climate change to support decision-making require high spatial resolution, but this is computationally prohibitive with modern Earth system models (ESMs). A major challenge is the calibration (parameter tuning) process, which requires running large numbers of simulations to identify the optimal parameter values. Here we train a convolutional neural network (CNN) on simulations 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 %.

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 CMIP6 have spatial grid resolutions of $\mathcal{O}(100\text{ km})$ on a side, which is often much too course 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 $\mathcal{O}(10\text{ km})$, but these models are computationally prohibitive to develop and calibrate (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 relatively small number of input (calibration) 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 calibration 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.

[1] Department of Geography and Environmental Management, University of Waterloo, Canada. [2] Department of Systems Design Engineering, University of Waterloo, Canada.. Correspondence to: Christopher G. Fletcher, Department of Geography and Environmental Management, University of Waterloo, Canada. <chris.fletcher@uwaterloo.ca>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

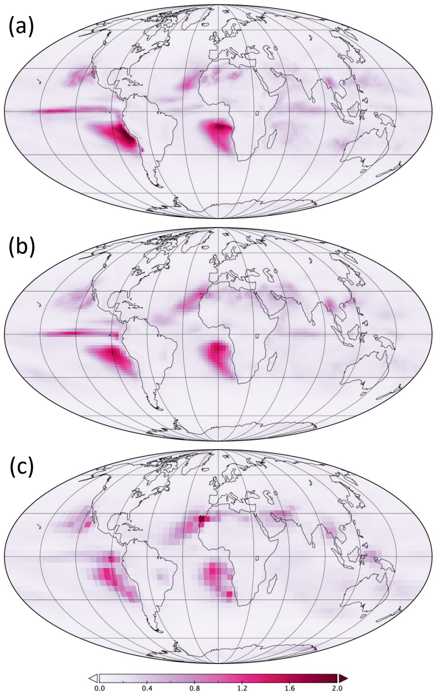


Figure 1. Coefficient of variation for the difference maps of annual mean total precipitation (mm/day) for the 100 realizations of the perturbed parameter ensembles of CESM-CAM4 run at three horizontal resolutions: (a) f09, (b) f19 and (c) f45.

2. Data and Methods

2.1. Earth System Model

The ESM studied here is the Community Earth System Model (CESM) Version 1.0.4 (Gent et al., 2011). All simulations include interactive atmosphere and land surface models, and prescribed climatological ocean surface temperatures and sea ice representative of the pre-industrial period. The atmospheric model (Community Atmosphere Model Version 4, CAM4) represents aerosol-radiation interactions, but not aerosol-cloud interactions, and employs a finite-volume dynamical core (Collins et al., 2006). CAM4 is run here at three horizontal resolutions: a higher-resolution 0.9° × 1.25° latitude-longitude grid (henceforth f09), which is the same resolution used in the CESM simulations that were contributed to the CMIP5 project (Taylor et al., 2011). Two less computationally expensive configurations are also employed here, a medium-resolution version at 1.9° × 2.5° (f19), and a lower-resolution version at 4° × 5° (f45).

To investigate the impact of resolution on model calibration, we conduct a 100-member perturbed parameter ensemble (PPE) at each of the three spatial resolutions. In each PPE the same set of nine uncertain parameters in CAM4 are perturbed using a set of values selected by Latin Hypercube Sampling. The nine parameters are identical to those perturbed by Fletcher et al. (2018) and they relate to the representation in CAM4 of the radiative forcing of anthropogenic aerosols, cloud amount, cloud optical properties, and convective precipitation. Each realization is integrated for three years, and the outputs are averaged over all 36 months to reduce the influence of atmospheric internal variability.

To quantify the impact on total precipitation (PRECT)—a scientifically important, and highly spatially variable, output in ESM simulations—from perturbations to the nine input parameters, Fig. 1 shows the coefficient of variation (i.e., ensemble spread) at the three resolutions, where heavier shading represents greater variability within the ensemble. 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 net top-of-atmosphere radiative flux (FNET) is largely insensitive to resolution because of the much lower spatial variability in that field (not shown).

The atmospheric response to perturbing the parameters occurs on a timescale of hours to days, which means that in this experimental configuration with prescribed ocean temperatures a spin-up period is not required; however, a spin-up period of multiple decades would likely be required if an interactive ocean model was coupled to CAM4, to allow the model's radiative balance to re-equilibrate. The lower-resolution configurations of CESM have spatial grid resolutions of 46×72 and 96×144, respectively. To ensure that the output from all three resolutions can be easily incorporated into the CNN, the lower resolution outputs are first upsampled using bilinear interpolation to match the f09 grid size of 192×288.

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. Given sufficient training examples, the CNN learns a statistical representation of the underlying physical equations that relate changes in the parameters to the outputs. The CNN architecture includes seven layers to map the 9d input feature vector to global maps of seven output variables (192 × 288 × 7). With the exception of the final convolution (conv) layer, all depicted layers are followed

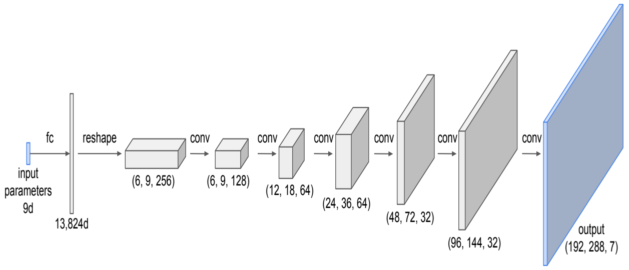


Figure 2. The architecture of the generative convolutional neural network used to predict seven spatially-resolved outputs of an ESM controlled by nine atmospheric parameters. fc: fully-connected (dense) layer. conv: transpose convolution with a kernel size of 5×5.

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 6×9×256, which facilitates a series of transpose convolutions—sometimes referred to as deconvolutions—using a kernel size of 5×5. 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 (192×288). 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). In this paper, we focus exclusively on FNET and PRECT, which are representative of the general features of the full results. The CNN was implemented in TensorFlow 2.2 using the Keras API.

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. Since we are interested in calibrating CESM's response to the nine input parameters, we train the CNN to predict the *difference* between the outputs of each perturbed model and the reference configuration with all parameters set to their defaults, which we refer to as 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, $\mathbf{x}$ and $\mathbf{Y}$ are normalized to the range [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.

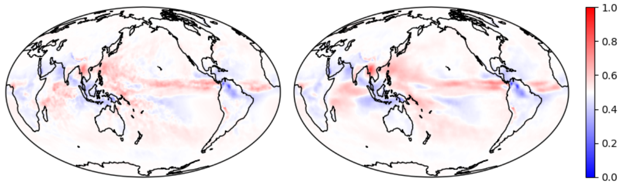
Training a neural network involves minimizing a loss function representing the error between the predicted and target outputs by iteratively updating the network parameters using gradient descent and backpropagation (Goodfellow et al., 2016). Since the selection of an appropriate loss function is a subjective element of the CNN architecture for each application, two different loss functions are compared here. The first uses the mean squared error (L_{MSE}), which is commonly used in computer vision applications (Ledig et al., 2017; McNally et al., 2020). The second is a new loss function (L_{SS}) inspired by a spatial skill score metric (SS) that is often used to quantify the accuracy of climate models at reproducing the spatial pattern, and amplitude, of a reference field (Pierce et al., 2009). The accuracy of the predictions is evaluated using the MSE and SS between the difference maps simulated by CESM and the difference maps predicted by the CNN.

3. Results

3.1. Overall performance of the CNN

We begin by showing how well the CNN, trained using the L_{SS} loss function, is able to predict the total precipitation field for 20% of unseen high-resolution (f09) outputs of CESM when trained on the remaining 80%. Looking at a randomly-selected difference map, Figs. 3a,b show that 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).

This qualitatively high performance for a single case and a single variable is reinforced by the quantitative metrics averaged over all test cases and all seven output variables: the CNN produces low average MSE (4.07e-4), and a high average skill score (0.817). Precipitation represents the most challenging target for the CNN because of its very high



(a) PRECT

(b) MSE_k: 1.91e-4, 1–SS_k: 0.817

Figure 3. Global mean annual mean total precipitation (normalized, dimensionless) for a randomly sampled test case. The left panel shows the original simulation from CESM, and the right panel shows the prediction from the CNN trained with the (L_{SS}) loss function. The values below the right panel show the mean-squared error (MSE_k), and the skill metric (1–SS_k), for this individual case compared to the original simulation in the left column.

spatial variability, and its mean skill score over all cases is somewhat lower than the multi-variable mean (0.727). However, as shown in the single case in Fig. 2, the CNN still represents the spatial variations from the original CESM simulation with high fidelity.

The choice of the loss function used to train the CNN has a major impact on performance. Training on L_{MSE} reduces the average skill across all seven output variables by 15%, and for precipitation the skill is reduced by 25%. 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 lower-resolution data

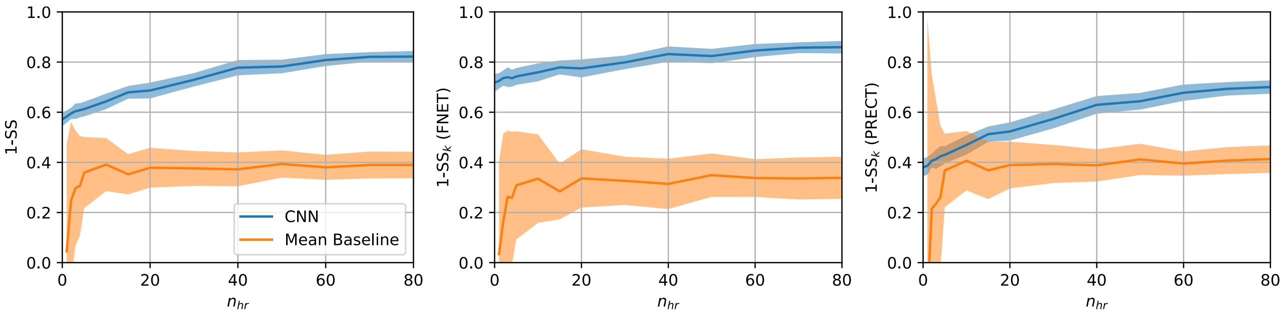
We next describe a practical application of the CNN-based emulator that follows the approach of Anderson & Lucas (2018) to extract information about a higher-resolution ESM from simulations from less expensive lower-resolution versions of the same ESM. We use the same CNN architecture as above (Fig. 2), but this time the training data includes all 200 of the lower resolution (f19 and f45) cases, in addition to a number of high-resolution cases (n_{hr}) that is sequentially increased from 0 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 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 our results show clearly that around 40 higher-resolution examples represents about as many as are required. Above $n_{hr} = 40$ the returns diminish considerably, and so the benefit of running additional costly higher-resolution cases appears small.

To evaluate the benefit of using the CNN over a simpler approach, a baseline skill value is obtained by assuming the CNN predicts spatial maps that equal the climatological mean of each variable for all values of the input parameters. The orange line in Fig. 4a shows that the baseline model achieves a mean skill score of just under 0.4 when more than 10 higher-resolution examples are included in the calculation of the climatological mean. Importantly, the fully-trained CNN outperforms the baseline for all values of n_{hr}. Net top-of-atmosphere radiative flux shows systematically higher skill, and precipitation shows systematically lower skill, suggesting that skill decreases with increasing spatial complexity in the target variable. For all variables, the skill of the CNN-based emulator plateaus at $n_{hr} < 80$, suggesting that underfitting due to too few training cases is *not* limiting the CNN's skill. Our conclusion is that fine-scale details of the output (e.g., Fig. 3a for precipitation) are related more to internal atmospheric variability than to parameter uncertainty, and are thus not being captured by the CNN.

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 ESM output. While computationally-efficient and ideally 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 are likely to contain less noise than observation-based data. Alternative approaches to constructing a multi-resolution emulator are conceivable; for example, as an image-to-image translation, where the lower-resolution data are the inputs to the CNN, and the higher-resolution version is the target. However, since the aim 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 CNN architecture. For this reason we



(a) Mean **(b)** FNET **(c)** PRECT

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. (left) the mean skill over all seven outputs, (middle) skill for FNET only, and (right) skill for PRECT only. 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.

believe that the architecture shown in Fig. 2 represents a simple and efficient mapping of the vector of parameter values onto an array of spatially-resolved ESM outputs.

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. 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 $\mathcal{O}(10\text{ km})$. An interesting 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 a different 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. 3b 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, rather than against the default version of the ESM (Hourdin et al., 2016). 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).

Acknowledgements

We thank the Microsoft AI for Earth program via the Waterloo AI Institute for funding, Alberto Costa Nogueira Junior, and three anonymous reviewers for their helpful comments.

References

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

Beusch, L., Gudmundsson, L., and Seneviratne, S. I. 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, February 2020. ISSN 2190-4979. doi: 10.5194/esd-11-139-2020. URL https://esd.copernicus.org/articles/11/139/2020/. Publisher: Copernicus GmbH.

Boyle, J. S., Klein, S. A., Lucas, D. D., Ma, H.-Y., Tannahill, J., and Xie, S. The parametric sensitivity of CAM5's MJO. *Journal of Geophysical Research: Atmospheres*, 120(4):2014JD022507, February 2015. ISSN 2169-8996. doi: 10.1002/2014JD022507. URL http://onlinelibrary.wiley.com/doi/10.1002/2014JD022507/abstract.

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., Briegleb, B. P., Bitz, C. M., Lin, S.-J., and Zhang, M. The Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). *Journal of Climate*, 19(11): 2144–2161, June 2006. ISSN 0894-8755. doi: 10.1175/JCLI3760.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI3760.1.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., Kampenhout, L. v., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G. The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. ISSN 1942-2466. doi: https://doi.org/10.1029/2019MS001916. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001916.

Erler, A. R., Frey, S. K., Khader, O., d'Orgeville, M., Park, Y.-J., Hwang, H.-T., Lapen, D. R., Peltier, W. R., and Sudicky, E. A. Simulating Climate Change Impacts on Surface Water Resources Within a Lake-Affected Region Using Regional Climate Projections. *Water Resources Research*, 55(1):130–155, 2019. ISSN 1944-7973. doi: https://doi.org/10.1029/2018WR024381. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024381. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR024381.

Fletcher, C. G., Kravitz, B., and Badawy, B. Quantifying uncertainty from aerosol and atmospheric parameters and their impact on climate sensitivity. *Atmospheric Chemistry and Physics*, 18(23):17529–17543, December 2018. ISSN 1680-7324. doi: 10.5194/acp-18-17529-2018. URL https://www.atmos-chem-phys.net/18/17529/2018/.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z.-L., and Zhang, M. The Community Climate System Model Version 4. *Journal of Climate*, 24(19):4973–4991, October 2011. ISSN 0894-8755, 1520-0442. doi: 10.1175/2011JCLI4083.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2011JCLI4083.1.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

Heinze-Deml, C., Sippel, S., Pendergrass, A. G., Lehner, F., and Meinshausen, N. Latent Linear Adjustment Autoencoders v1.0: A novel method for estimating and emulating dynamic precipitation at high resolution. *Geoscientific Model Development Discussions*, pp. 1–39, October 2020. ISSN 1991-959X. doi: 10.5194/gmd-2020-275. URL https://gmd.copernicus.org/preprints/gmd-2020-275/. Publisher: Copernicus GmbH.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, July 2016. ISSN 0003-0007. doi: 10.1175/BAMS-D-15-00135. URL http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-15-00135.1.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., and Volodina, V. 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, 2021. ISSN 1942-2466. doi: 10.1029/2020MS002225. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002225. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002225.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 2015.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M. 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, November 2014. ISSN 0003-0007. doi: 10.1175/BAMS-D-13-00255.1. URL http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-13-00255.1.

Kennedy, M. C. and O'Hagan, A. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, January 2001. ISSN 1467-9868. doi: 10.1111/1467-9868.00294. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00294/abstract.

Lanzante, J. R., Dixon, K. W., Nath, M. J., Whitlock, C. E., and Adams-Smith, D. Some Pitfalls in Statistical Downscaling of Future Climate. *Bulletin of the American Meteorological Society*, 99(4): 791–803, April 2018. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-17-0046.1. URL 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.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W., and Spracklen, D. V. Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. *Atmos. Chem. Phys.*, 11(23): 12253–12273, December 2011. ISSN 1680-7324. doi: 10.5194/acp-11-12253-2011. URL https://www.atmos-chem-phys.net/11/12253/2011/.

Luca, A. D., Argüeso, D., Evans, J. P., Elía, R. d., and Laprise, R. Quantifying the overall added value of dynamical downscaling and the contribution from different spatial scales. *Journal of Geophysical Research: Atmospheres*, 121(4):1575–1590, 2016. ISSN 2169-8996. doi: https://doi.org/10.1002/2015JD024009. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD024009. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD024009.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Marchine Learning*, 2013.

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

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

Racherla, P. N., Shindell, D. T., and Faluvegi, G. S. 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), 2012. ISSN 2156-2202. doi: 10.1029/2012JD018091. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD018091.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S. Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF. *Atmospheric Chemistry and Physics*, 18(13):9975–10006, July 2018. ISSN 1680-7316. doi: https://doi.org/10.5194/acp-18-9975-2018. URL https://www.atmos-chem-phys.net/18/9975/2018/.

Salter, J. M. and Williamson, D. B. Efficient calibration for high-dimensional computer model output using basis methods. *arXiv:1906.05758 [stat]*, June 2019. URL http://arxiv.org/abs/1906.05758. arXiv: 1906.05758.

Salter, J. M., Williamson, D. B., Scinocca, J., and Kharin, V. Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases. *Journal of the American Statistical Association*, pp. 1–24, September 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2018.1514306. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1514306.

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Girolamo, S. D., Hentgen, L., Hoefler, T., Lapillonne, X., Leutwyler, D., Osterried, K., Panosetti, D., Rüdisühli, S., Schlemmer, L., Schulthess, T. C., Sprenger, M., Ubbiali, S., and Wernli, H. Kilometer-Scale Climate Models: Prospects and Challenges. *Bulletin of the American Meteorological Society*, 101(5): E567–E587, May 2020. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-18-0167.1. URL 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.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4):485–498, October 2011. ISSN 0003-0007. doi: 10.1175/BAMS-D-11-00094.1. URL http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1.

Watson, P. A. G. 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, 2019. ISSN 1942-2466. doi: https://doi.org/10.1029/2018MS001597. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001597. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001597.

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