Learning to Correct Climate Projection Biases

Baoxiang Pan1, Gemma J. Anderson1, André Goncalves1, Donald D. Lucas1, Céline J. W. Bonfils1, Jiwoo Lee1, Yang Tian1, and Hsi-Yen Ma1

1Lawrence Livermore National Laboratory, Livermore, CA, USA

Abstract The fidelity of climate projections is often undermined by biases in climate models due to their simplification or misrepresentation of unresolved climate processes. While various bias correction methods have been developed to post-process model outputs to match observations, existing approaches usually focus on limited, low-order statistics, or break either the spatiotemporal consistency of the target variable, or its dependency upon model resolved dynamics. We develop a Regularized Adversarial Domain Adaptation (RADA) methodology to overcome these deficiencies, and enhance efficient identification and correction of climate model biases. Instead of pre-assuming the spatiotemporal characteristics of model data, we apply discriminative neural networks to distinguish historical climate simulation samples and observation samples. The evidences based on which the discriminative neural networks make distinctions are applied to train the domain adaptation neural networks to bias correct climate simulations. We regularize the domain adaptation neural networks using cycle-consistent statistical and dynamical constraints. An application to daily precipitation projection over the contiguous United States shows that our methodology can correct all the considered moments of daily precipitation at approximately 1° resolution, ensures spatiotemporal consistency and inter-field correlations, and can discriminate between different dynamical conditions. Our methodology offers a powerful tool for disentangling model parameterization biases from their interactions with the chaotic evolution of climate dynamics, opening a novel avenue toward big-data enhanced climate predictions.

Plain Language Summary Accurate climate prediction is crucial for understanding climate change and implementing effective climate adaptation strategies. However, climate models that are used to generate climate predictions have multifaceted biases that often need to be corrected before predictions can be considered usable. We develop a data-driven methodology that detects and corrects climate model biases using a game theory inspired machine learning technique. By applying physical and statistical constraints, our predictions are not only more accurate, but also more trustworthy as judged by our physical understandings.

1. Introduction

1.1. Two Approaches Toward Improving Climate Projection Accuracy

Understanding, predicting, and adapting to climate change is a problem of growing urgency. These endeavors rely upon climate projections computed by global atmosphere-ocean-land coupled general circulation models (GCMs). Unfortunately, the fidelity of climate projections is often undermined by GCMs’ biases due to their over-simplification, coarse grid cells resolution, or misrepresentation of the climate processes (Christensen et al., 2008; Flato et al., 2014; Maraun, 2016; Zelinka et al., 2020).

A solid avenue toward identifying and correcting GCM biases is to apply numerical weather prediction (NWP) techniques to isolate the bias and its propagation from interactions between GCM’s parameterized physics and resolved dynamics (H.-Y. Ma et al., 2014; Phillips et al., 2004; S. Xie et al., 2012). This process-oriented diagnostic technique mainly targets GCM biases associated with fast physical processes, such as precipitation, cloud, or convection on short time scales. However, if the bias impact requires longer time scales (1–2 weeks) to develop, this methodology alone may not be sufficient to disentangle the model biases and the unneglectable growth of initial state estimation errors (Smith, 2001).

A separate vein of research, which is more accessible for climate simulation end-users, is to downscale and correct GCM output biases at a post-processing step. The idea is to first establish a statistical connection between GCM’s historical simulations and climate observations, and then apply this statistical relationship to
reduce their mismatch. Despite several severe problems in their assumptions and implementations (Ehret et al., 2012; Maraun, 2013), these bias correction approaches offer an immediate “improvement” to climate simulations, and have been applied routinely in a broad range of applications, such as climate change impact analysis (H. Li et al., 2010; Wood et al., 2004), and climate-related decision making (L. Li et al., 2015; Teutschbein & Seibert, 2012).

The key difficulty for bias correction lies in that, due to the chaotic nature of geophysical fluid dynamics, the freely evolving historical climate simulations do not follow the evolution of historical climate variability. This prevents us from inferring GCM biases by contrasting simulation-observation time series, as is conventionally done in weather-to-seasonal forecast verification (F. Ma et al., 2016; Pan, Hsu, AghaKouchak, Sorooshian, & Higgins, 2019; Zhao et al., 2021). Correspondingly, we usually infer biases by comparing simulated and observed climate mean state or other simple statistics summarized over a long range period (i.e., multiple decades). On one hand, this conceals detailed bias information that is canceled out through error compensations. On the other hand, this leaves us with limited observational climatology references to develop and verify powerful bias identification and correction algorithms. Consequently, existing bias correction practices tend to conceal, rather than correct the GCM biases.

1.2. What Counts as a “Perfect” Climate Projection Bias Corrector?

In a critical discussion of bias correction applicability, Ehret et al. (2012) requires a “perfect” bias correction approach to be able to:

1. Correct at high spatiotemporal resolution all moments of the variable of interest,
2. Assure spatiotemporal consistency as well as inter-field correlations,
3. Discriminate between different weather situations, allow for the bias to be time-transient,
4. Include feedback effects.

To meet these requirements, they conclude that we will “inevitably arrive at a complexity of the bias correction method comparable to the GCM itself, but still lack the physical justification of the latter” (Ehret et al., 2012).

While we agree with this conclusion, here we use this statement as guidance to formulate the GCM bias correction task as a machine learning problem. That is, to identify and correct GCM biases, we need powerful black-box models that can tackle problems whose complexity is comparable to that of GCMs, and come up with suitable objective functions to train these black-box models toward the above listed requirements.

The high requirement on the capability of black-box models is no longer an obstacle. Recent years have witnessed the flourishing of deep learning that successfully solves complex, high-dimensional problems, including physical system simulations such as electromagnetic and fluid dynamics (Kutz, 2017; Mills et al., 2017), as well as climate applications such as parameterization (Pan, 2019; Rasp et al., 2018), downscaling (Miao et al., 2019; Pan, Hsu, AghaKouchak, & Sorooshian, 2019), analog forecasting (Pan, Anderson, Goncalves, et al., 2020; Weyn et al., 2019), inverse modeling (Zhang et al., 2020), and climate signal identification (Barnes et al., 2019).

While most of the applications above adopt a supervised learning problem setting, which maps an input to an output by learning from example input-output pairs, we do not have paired GCM simulations and observations to learn GCM biases, except in very limited retrospective forecast cases (Phillips et al., 2004; Rasp & Lerch, 2018). On one hand, it is difficult, if not impossible, to train powerful data-driven models with these limited paired data. On the other hand, we cannot afford hindcast based diagnosis for each GCM configuration. Is it possible to learn GCM biases from large amount unpaired historical climate simulation and observation data?

1.3. Correct Climate Projection Biases Using Adversarial Learning

We demonstrate that adversarial learning (Goodfellow et al., 2014) is a less-exploited, yet well-suited paradigm for identifying and correcting climate projection biases using unpaired simulation and observation data. Rather than pre-assuming the spatiotemporal characteristics of the model biases, we carefully design a discriminative function to identify arbitrary mismatches between the distribution of climate simulations
and observations. Feedback from this discriminative function is applied to learn a bias correction function to close these mismatches. We alternatively train the discriminative function and the bias correction function to improve the skills of both, until the corrected climate simulation is indistinguishable from the observation. This adversarial learning framework offers the potential for leveraging the big data of historical climate observations and simulations to disentangle the parameterization biases with model resolved dynamics, saving the huge effort in forcing a GCM to start from realistic initial conditions to expose its parameterization biases.

Although promising, adversarial learning can be highly unconstrained (Pan, Anderson, Lucas, et al., 2020; Wu et al., 2020; Yang et al., 2019). For instance, the bias correction function may map the GCM simulations to limited varieties of climate variability modes, or result in “corrections” that are inconsistent with GCM resolved dynamics. To address these issues, we introduce several statistical and dynamical constraints to regularize the data-driven bias corrector. We show that the Requirements 1–3 in Ehret et al. (2012) can potentially be well satisfied using this regularized adversarial domain adaptation (RADA) methodology. To include feedback effects (Requirement 4), we should either couple the bias corrector with dynamical simulations, or iteratively apply this methodology to correct the parameterization biases and alleviate their impacts on model resolved dynamics. We discuss the difficulties and opportunities here, and leave the tackling of the feedback problem (Requirement 4) for future work.

The rest of the paper is organized as follows. First, in Section 2, we introduce the notion of adversarial learning, the challenges of using an unregularized methodology, as well as the three regularization methods. A case study for correcting daily precipitation projection bias is presented in Section 3, and the results of this case study are described in Section 4. Section 5 discusses the individual and combined contributions of the three regularizers to improve the adversarial learning of GCM biases. Section 6 compares our approach with well-established univariate and multivariate bias correction methods. Section 7 discusses the perceptual performance and the training stopping criteria of the RADA methodology. We draw conclusions and layout the direction for future work in Section 8.

2. Methodology

2.1. Disentangle Parameterization and Dynamics Using GAN

We distinguish two sets of variables in climate simulation. The first set includes the variables that are directly resolved by discretizing the geophysical fluid dynamical equations, also informally referred to as the dynamics. The second set includes the remaining unresolved variables, also informally referred to as the physics. In climate simulation, the unresolved variables are estimated based on their statistical connections with the resolved variables. These parameterization processes contribute to the majority of climate simulation biases. Our objective is to identify and correct these parameterization biases. The challenge lies in disentangling the intricate parameterization biases from their interactions with the chaotic evolution of climate dynamics.

To efficiently address this challenge, we introduce the generative adversarial neural network (GAN, Goodfellow et al., 2014), which serves as the backbone of our data-driven bias correction methodology. In GANs, a neural network parameterized function, named the generator (G, see Supporting Information S1 for notation clarification), learns a mapping from random noise to a target distribution, while an auxiliary neural network parameterized function, named the discriminator (D), guides the training of G by distinguishing candidates produced by G from the target distribution. The term adversarial learning refers to the idea that, G is not trained toward a pre-defined, fixed objective, but to fool D, which itself is dynamically updated to enhance its discrimination capability.

Here, instead of applying GANs to generate new samples, we make use of the adversarial learning idea to match samples from a source domain to corresponding samples in a target domain. This line of research is often named domain adaptation or domain translation in the machine learning literature (Liu et al., 2017; Tzeng et al., 2017; Wang et al., 2018; Wilson & Cook, 2020; Zhu et al., 2017). Specifically, as is illustrated in Figure 1, we apply G as a deterministic bias correction function that translates samples from a source domain X to their corresponding analog samples in a target domain Y, where X is GCM simulation of a target, parameterized variable (i.e., precipitation, Figure 1 bottom left), Y is its observation (Figure 1 top left). We
train \( D \) to distinguish whether a sample comes from \( Y \) or \( G(X) \) by minimizing a discrimination error (Figure 1 blue arrow, strict formulation provided latter). The evidences based on which \( D \) makes distinctions are applied to train \( G \) to make \( G(X) \) less distinguishable from \( Y \) (Figure 1 red arrow, strict formulation provided latter), hence closing the mismatch between the distributions of \( X \) and \( Y \).

How can we guarantee \( D \) distinguishes \( Y \) and \( G(X) \) relying purely upon the GCM parameterization bias information, not the uncontrollable dynamical state variance information? We explain two critical configurations that make \( D \) work here. First, we suppress the irrelevant dynamical variance signal: we train \( D \) using \( Y \) and \( G(X) \) samples with random dynamical states (Figure 1 wave arrows). Since the dynamical state differences do not contribute to the discrimination objective, \( D \) will learn representations that suppress the impact of samples’ dynamical state variances, and turn to the parameterization bias information to distinguish \( Y \) and \( G(X) \). Second, we enhance the parameterization bias signal: we consider large spatial patterns (i.e., continental scale in Figure 1) of the target variable, rather than pixel-by-pixel values, to distinguish the parameterization bias signal. This is inspired by Barnes et al. (2019) and Sippel et al. (2020), which tell that, intricate climate signals, when viewed from large spatial scale, are more detectable.

In practice, we use the Wasserstein GAN (Arjovsky et al., 2017, see Supporting Information S1 for details), giving its training stability. Specifically, we consider \( D \) with continuous output expressing how likely a sample is from \( Y \) rather than \( G(X) \). We alternatively train \( G \) to increase the false positive rate of \( G(X) \) being taken as \( Y \), and train \( D \) to enhance the distinction between \( G(X) \) and \( Y \). These are achieved by applying stochastic gradient descent/ascent toward the following min/max objective:

\[
\min_G \max_D \mathbb{E}_{Y \sim P_Y} [D(Y)] - \mathbb{E}_{X \sim P_X} [D(G(X))]
\]  

(1)

Here \( \mathbb{E}_{Y \sim P_Y} \) and \( \mathbb{E}_{X \sim P_X} \) are expectations taken over probability distribution of \( Y \) and \( X \). Competition between \( G \) and \( D \) driven by stochastic gradient descent/ascent improves the skills of both, until the distribution of \( G(X) \) is indistinguishable from \( Y \) as viewed by \( D \). This adversarial learning framework automatically identifies multifaceted mismatches between domain distributions, and simultaneously reduces the distribution mismatches.
2.2. Regularizing the Adversarial Learning Process

2.2.1. Necessity for Regularization

The adversarial loss in Equation 1 serves as a general, high-level objective for correcting GCM biases. Various solutions can meet this adversarial learning objective. However, many of them may appear authentic upon visual inspection, but are physically unreasonable. Specifically, we may face the following three challenges:

1. **Mode collapse.** \( G \) may map the GCM simulations to limited varieties of climate variability, lacking diversity as compared to the uncorrected simulations. An extreme example is that, \( G \) maps all simulations to a same, observation-alike output. This situation satisfies the adversarial learning objective in Equation 1 well, but yields biased, less-diverse distribution estimation.

2. **Dynamical inconsistency.** This means that, even if \( G(x) \) is statistically indistinguishable from samples of \( Y \), the resulting \( G(x) \) may not be physically consistent with the GCM’s resolved dynamics. Enforcing dynamical consistency implicitly alleviates mode collapse.

3. **Dynamical invariance.** For two same GCM simulation samples, if their embedding dynamical conditions are different, they may exhibit distinct mismatches with observations. It is difficult to discriminate dynamics-dependent biases, if the bias corrector is not conditioned on the resolved dynamics.

These three challenges roughly correspond to the Requirement 1–3 in Ehret et al. (2012). To overcome these challenges, we introduce three corresponding statistical and dynamical regularizers to constrain \( G \) in adversarial learning for correcting GCM biases. Here regularization refers to adding extra information to exclude physically unreasonable solutions in an ill-posed optimization problem. Part of the introduced regularizers here are well-established tools developed by the machine learning community (Section 2.2.2), while the rest are uniquely designed for the climate projection bias correction problem (Sections 2.2.3 and 2.2.4). We demonstrate the benefits of these regularizers with ablation analysis in the latter introduced case study (Section 5).

2.2.2. Cycle Consistency

We alleviate the mode collapse problem using the cycle consistency regularization, which has quickly become the de facto regularization method for learning GAN-based domain adaptation (Kim et al., 2017; Zhu et al., 2017). The idea is to enforce one-to-one correspondence between domains by coupling an adversarial domain adaptation model \([G, D]\) with its inverse model \([G^{-1}, D^{-1}]\). In this inverse model, \( G^{-1} \) maps \( Y \) back to \( X \), and \( D^{-1} \) guides the learning of \( G^{-1} \) by distinguishing samples from \( X \) and \( G^{-1}(Y) \). To couple the two opposite-directed adversarial learning processes, we minimize the following cycle consistency loss \( \mathcal{L}_c \) besides the adversarial losses (Zhu et al., 2017):

\[
\mathcal{L}_c = \mathbb{E}_{x \sim p_X} \left[ \| G^{-1}\left( G(x) \right) - x \|_1 \right] + \mathbb{E}_{y \sim p_Y} \left[ \| G\left( G^{-1}(y) \right) - y \|_1 \right]
\]

(2)

Here, \( \| \|_1 \) is the \( L^1 \)-norm, which has shown to work better than \( L^2 \)-norm, as it tends to produce sharper results (Zhu et al., 2017). To minimize \( \mathcal{L}_c \) pushes \( G^{-1}(G(X)) \approx X \) and \( G(G^{-1}(Y)) \approx Y \), which encourages learning invertible mappings between \( X \) and \( Y \).

The physical explanation for enforcing invertible mapping is that, provided with a same resolvable, underlying dynamical state \( s \), the GCM simulation of a target variable is preferred to be identical to its observation (we neglect stochastic parameterization noise here, and revisit this issue in Section 8.2.1.) Given a same set of plausible dynamical states \( S \) subjected to certain external forcing, an optimal domain adaptation function \( G \) should be invertible as bridged by \( S \). Therefore, if the GCM historical simulation can generate a same range of climate dynamical state as the observed climate system, we can translate \( X \) to \( G(X) \) that is distributed nearly identically to \( Y \).

Since this neural network based methodology can easily scale up to high dimensional massive data (Bottou, 2010), by matching the distribution of \( G(X) \) and \( Y \), we can automatically correct at high spatiotemporal resolution all moments of the variable of interest (Requirement 1 in Ehret et al., 2012).

2.2.3. Dynamical Consistency

Although cycle consistency helps to restrict the learning to a smaller feasible solution space of invertible functions, we may still end up finding many invertible mappings between the two considered domains (see...
Cohen et al., 2018 for a failure case example). To guarantee that an individual climate simulation sample \( x \) is meaningfully translated to its corresponding analog sample in \( Y \), we should have \( x \) and \( G(x) \) correspond to a same dynamical state \( s_x \). We enforce this by maintaining the inter-field consistency between the target variable and its underlying dynamical state during training.

To achieve dynamical consistency, we first build statistical models that estimate the target variable (\( X \) and \( Y \)) using its underlying dynamical state information (\( S_X \) and \( S_Y \), see Supporting Information S1 for details):

\[
F^*_X = \arg \min_{\tilde{p}_X} \mathbb{E}_{x \sim p_X} \left( \| F^*_X(s_x) - x \|^2 \right)
\]

\[
F^*_Y = \arg \min_{\tilde{p}_Y} \mathbb{E}_{y \sim p_Y} \left( \| F^*_Y(s_y) - y \|^2 \right)
\]

(3)

Here \( F^*_X \) and \( F^*_Y \) are arbitrary form (i.e., neural network, see Supporting Information S1) statistical models developed for the GCM simulated climate system and the observed climate system; \( s_x \) and \( s_y \) are dynamical state representations corresponding to \( x \) and \( y \), which are typically resolvable variables not directly impaired by GCM biases. \( F^*_X \) and \( F^*_Y \) are often known as statistical downscaling models, as they derive statistical relationships between small-scale, parameterized variables and large-scale, resolved dynamics. The purpose for developing and applying these models here is not to enhance resolution, but to ensure the consistency between \( G(X) \) and \( S_X \), \( G^{-1}(Y) \) and \( S_Y \), which is described below.

We use the resulting \( F^*_X \) and \( F^*_Y \) to constrain \( G \) and \( G^{-1} \) by minimizing the following dynamical consistency loss \( \mathcal{L}_D \):

\[
\mathcal{L}_D = \mathbb{E}_{x \sim p_X} \left[ \max \left( \| F^*_X(s_x) - G(x) \|^2, \delta_X \right) - \delta_Y \right] + \mathbb{E}_{y \sim p_Y} \left[ \max \left( \| F^*_Y(s_y) - G^{-1}(y) \|^2, \delta_X \right) - \delta_X \right]
\]

(4)

Here \( \delta_X = \mathbb{E}_{x \sim p_X} \left( \| F^*_X(s_x) - x \|^2 \right) \), \( \delta_Y = \mathbb{E}_{y \sim p_Y} \left( \| F^*_Y(s_y) - y \|^2 \right) \), which are the \( L^2 \) regression losses of the statistical models \( F^*_X \) and \( F^*_Y \). By minimizing \( \mathcal{L}_D \), we impose the inter-field correlation of \( \{X, S_X\} \) and \( \{Y, S_Y\} \) to \( \{G^{-1}(Y), S_Y\} \) and \( \{G(X), S_X\} \), penalizing errors beyond the statistical downscaling constraints quantified by \( \delta_X \) and \( \delta_Y \). Here, we tolerate errors smaller than \( \delta_X \) and \( \delta_Y \); if \( \| F^*_X(s_x) - G(x) \|^2 \) and \( \| F^*_Y(s_y) - G^{-1}(y) \|^2 \) are smaller than \( \delta_X \) and \( \delta_Y \), \( \mathcal{L}_D \) diminishes to 0. This setting admits the fact that \( X \) and \( Y \) are not uniquely determined by \( S_X \) and \( S_Y \), and avoids the bias corrector from producing the mean of all possible outcomes, leaving space for optimizing the adversarial loss (Equation 1) within the dynamical consistency constraint.

To minimize \( \mathcal{L}_D \) loosely guarantees that \( G(X) \) and \( G^{-1}(Y) \) are consistent with their underlying dynamical states \( S_X \) and \( S_Y \). Also, this constraint implicitly encourages spatiotemporal consistency of the target variable, and further alleviates mode collapse. Provided that \( F^*_X \) and \( F^*_Y \) impose strong constraints on the variability of \( X \) and \( Y \), which has been intensively testified by statistical downscaling studies (Miao et al., 2019; Pan, Hsu, AghaKouchak, & Sorooshian, 2019; Sun & Tang, 2020), we could guarantee the inter-field correlation as well as the spatiotemporal consistency of the bias correction. The Requirement 2 in Ehret et al. (2012) can thus be fulfilled.

2.2.4. Dynamical Dependency

Biases in GCM simulations in part depend on the underlying dynamics. For instance, consider a common GCM bias of overestimating the number of wet days (Piani et al., 2010), the knowledge of the dynamical state helps identifying if a simulated drizzle event is a case of false alarm or true positive. This suggests that the bias correction function \( G \) should take into consideration of both the target variable \( X \) and the dynamical state \( S_X \). Similarly, \( G^{-1} \) should take into input of both \( Y \) and \( S_Y \). As we make \( G \) and \( G^{-1} \) functions of \( \{S_X, X\} \) and \( \{S_Y, Y\} \), we can potentially discriminate between different weather situations, allowing the bias corrector to be time-transient (Requirement 3 in Ehret et al., 2012).

2.3. Regularized Adversarial Domain Adaptation

We summarize the final bias correction methodology here: as is illustrated in Figure 2a, we apply two discriminative neural networks \( D \) and \( D^{-1} \) to learn arbitrary mismatch between the distribution of historical climate simulation \( X \) and observation \( Y \). Meanwhile, we adopt dynamics-dependent domain adaptation neural networks \( G(X, S_X) \) and \( G^{-1}(Y, S_Y) \) to close the mismatch under cycle consistency constraint and
dynamical consistency constraint. The final objective function of this regularized adversarial domain adaptation (RADA) bias corrector takes the following form:

\[
\mathcal{L}_{\text{RADA}} = \mathbb{E}_{y \sim p_Y} \left[ D(y) - \mathbb{E}_{x \sim p_X} \left[ D \left( G(x,s_x) \right) \right] \right] + \mathbb{E}_{x \sim p_X} \left[ D^{-1}(x) - \mathbb{E}_{y \sim p_Y} \left[ D^{-1} \left( G^{-1}(y,s_y) \right) \right] \right]
\]

\[
+ \lambda_C \mathbb{E}_{x \sim p_X} \left[ \left\| G^{-1} \left( G(x,s_x),s_x \right) - x \right\| \right] + \mathbb{E}_{y \sim p_Y} \left[ \left\| G \left( G^{-1}(y,s_y),s_y \right) - y \right\| \right]
\]

\[
+ \lambda_D \left[ \mathbb{E}_{x \sim p_X} \left[ \max \left( \left\| F_X \left( s_x \right) - G(x,s_x) \right\|, \delta_X \right) - \delta_X \right] + \mathbb{E}_{y \sim p_Y} \left[ \max \left( \left\| F_Y \left( s_y \right) - G^{-1}(y,s_y) \right\|, \delta_Y \right) - \delta_Y \right] \right]
\]

(5)

\[\mathcal{L}_{\text{RADA}}\] includes the following four components: (a) the adversarial loss for mapping \(X\) to \(Y\), (b) the adversarial loss for the inverse mapping from \(Y\) to \(X\), (c) the cycle consistency loss, and (d) the dynamical consistency loss. We weight these losses using hyperparameters \(\lambda_C\) and \(\lambda_D\).

To train the RADA model, we first train \(F_X\) and \(F_Y\) to minimize the statistical downscaling loss, then, we fix the trained models \(F_X^*\) and \(F_Y^*\), and alternatively train \(\{G, G^{-1}\}\) to minimize (a)–(d), and train \(\{D, D^{-1}\}\) to maximize (a) and (b). Unlike supervised learning, where we stop training as the supervision loss reaches a minimum for the validation set, there is no consensus about when the adversarial training obtains optimal

Figure 2. Structure of the Regularized Adversarial Domain Adaptation model. (a) Overall structure of the model. The model applies a conditional generative adversarial network \(\{G, D\}\) (red arrows) to obtain \(G(X,S_X)\) that is identically distributed with \(Y\) as viewed by \(D\). The adversarial learning is regularized by (i) coupling with an inverse model \(\{G^{-1}, D^{-1}\}\) (blue arrows), and (ii) maintaining the coherency between resolved dynamics and the target variable (brown arrows). (b) U-net (Ronneberger et al., 2015) architecture of the submodules \(F_X\) and \(F_Y\). (c and d) Res-net (He et al., 2016) architecture of the submodules \(G, G^{-1}\) and \(D, D^{-1}\).
performance, and the training should be terminated. A common strategy is to visually inspect the generated results and stop the training if there is no visually perceived improvement. However, as we will show in Section 7, this strategy does not work in our problem, as perceptual judgments can be easily fooled, and yield poor quantitative performance during test. Here, we compare some crucial precipitation indices (first to fourth order moments, see Section 3.4 for details) between corrected simulations and observations for the validation set at the end of each training epoch, and terminate the training when these statistics match best.

3. Experimental Design

3.1. General Settings

To test the proposed methodology, we consider a case study of correcting the Community Earth System Model version 2 (CESM2, Danabasoglu et al., 2020) daily precipitation projection over the contiguous United States (CONUS). We consider precipitation because it is among the most important yet least well-simulated variables in GCMs (Marvel & Bonfils, 2013; Tapiador et al., 2019). Later studies may explore the possibility of scaling up this methodology from univariate fields to multivariate fields, from continental scale to global scale, and from moderate spatiotemporal resolution to high resolution, by leveraging the scalability of deep neural networks.

We quantify the precipitation estimation accuracy before and after correction using a suite of commonly applied skill indices computed for a holdout test set. The applied data, data processing methods, model configuration, and evaluation approaches are described below.

3.2. Data

3.2.1. GCM Simulation

CESM2 is an open-source, fully coupled, global climate model that provides state-of-the-art computer simulations of the Earth’s past, present, and future climate states (Danabasoglu et al., 2020). We use CESM2 historical simulation contributing to the Coupled Model Intercomparison Project phase 6 (CMIP6, Eyring et al., 2016). Besides the daily precipitation products, we use the simulated daily averaged sea level pressure (SLP), geopotential height and specific humidity at 500 hPa (500hPaEz and 500hPaEq) as dynamical constraints to support model training (Pan, Hsu, AghaKouchak, & Sorooshian, 2019). All the simulation data are of a common resolution of approximately 1°.

3.2.2. Observation

We use the U.S. National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) unified gauge-based analysis of daily precipitation over CONUS (0.25°, P. Xie et al., 2010) as observational reference for bias correction. The daily averaged SLP, 500hPaEz and 500hPaEq records from the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the twentieth century (ERA-20C, 1.25°, Poli et al., 2016) are used as dynamical constraints for the observations.

3.2.3. Data Processing

The precipitation field data covering CONUS (25°N-50°N, 65°W-125°W) land areas are used. We apply the dynamical field data with an extra 5° spatial extension (20°N–55°N, 60°W–130°W) to constrain the precipitation field. The observed precipitation, SLP, 500hPaEz and 500hPaEq reanalyzes are regridded to the same resolution as CESM2 simulation using bi-linear spatial averaging (for CPC precipitation observation) or interpolation (for ERA-20C). To accommodate the long-tail distribution nature of precipitation, we regularize the precipitation data using a natural logarithm transformation. We normalize the dynamical variables by subtracting the grid mean, and dividing by the grid standard deviation. It should be noted that, different data processing approaches might influence the model performance. We leave a detailed examination on the impact of data processing for future work.

We split the data into training, validation, and test sets. The training and validation sets cover the period of 1950–1994 (16,435 days). To sample different climate periods for model training, for each decade from 1950
to 1994, we put data from the first 8 years into the training set, and the remaining 2 years into the validation set. Data from 1995 to 2004 (3,652 days) are held out in the test set to offer unbiased evaluation of the bias corrector.

3.3. Model Configuration

The RADA model architecture is illustrated in Figure 2. We apply convolutional neural networks (CNNs, Krizhevsky et al., 2012; LeCun & Bengio, 1995) as building blocks for the model, given their capability to exploit the multi-scale spatial coherency of geodata (Baño-Medina et al., 2020; Pan, Hsu, AghaKouchak, & Sorooshian, 2019; Sadeghi et al., 2020).

For the downscaling submodules of $F^x$ and $F^y$, we apply a U-net form of CNN (Figure 2h, Ronneberger et al., 2015). The U-net applies a convolution based contracting path to capture the resolved dynamical field information, and a symmetric transposed convolution based expanding path that gradually refines precipitation field estimation. Skip connections between symmetrical convolution and transposed convolution blocks ease the training by forcing the deeper layers to learn meaningful representations that are not well captured by shallower layers (He et al., 2016). Similarly, we apply CNNs with skip connections for $[G, G^{-1}]$ (Figure 2c) and $[D, D^{-1}]$ (Figure 2d).

We use the Adam optimizer (Kingma & Ba, 2014) with batch size of 32, learning rate of $10^{-4}$, and train the model for 300 epochs using 36 GPUs on the Wolfram Mathematica 12.1 deep learning platform (Wolfram Research, Inc., n.d.). The training does not require the climate simulation and observation data to be synchronized. Considering the impact of seasonal cycle, internal climate variability, and climate change, in each iteration of a training epoch, we take random mini-batch of daily precipitation observations, and sample climate simulations from the same months and ±5 years. This strategy greatly increases the training sample diversity, while excludes the impact of seasonality and climate change. For each training epoch, we shuffle the observational data in the training set, split the data into random mini-batches. For each mini-batch, we sample the corresponding simulation data (same months and ±5 years), obtain the gradient of $\mathcal{L}_{\text{RADA}}$ with respect to all the model parameters using this mini-batch of data, and update the model parameters accordingly. We repeat this process for all the mini-batches, completing one epoch of training.

We take an ensemble of trained neural networks, each representing a condition where no individual index measuring the precipitation estimation accuracy can be improved without worsening other indices (Pareto front). The considered indices are the spatial average root mean square error between observation’s and simulation’s first to fourth order moment of precipitation distribution (see Section 3.4). The average of the four neural network outputs are applied as the final bias correction result.

3.4. Performance Evaluation

The discriminator $D$ can in principle detect any mismatch between the distribution of observations and simulations, given that neural networks are universal function approximators (Leshno et al., 1993). Therefore, we could potentially evaluate the bias correction performance based on any statistics. Here we focus on the commonly applied measure of precipitation distribution characteristics, including the moments, quantile, frequency, intensity, and seasonality characteristics at grid scale, as well as spatial coherency and dynamical consistency at field scale. The considered indices are listed in Table 1.

For grid scale assessment, we consider the first to fourth order moments (mean, standard deviation, skewness, and kurtosis), the three tertiles (33%, 66%, 99% quantile), the probability of precipitation, the average precipitation intensity, the 1-day, 3-day, and 5-day maximum precipitation, ratio of winter/summer (December–February/June–August) precipitation. We assess the significance of bias correction improvements using the following bootstrap test: We first generate 100 bootstrap samples of precipitation observation and simulation time series before and after correction. Thereafter, we calculate the ratio that the target index after bias correction better matches that of observation. The confidence interval is thereafter determined using a percentile method. All tests are one-tailed and use a confidence level of 99%. We further apply the spatial correlation ($r$) and spatial average root mean square error (RMSE) between observation and simulation for these indices to roughly summarize the climate simulation performance before and after correction.
Table 1

Indices for Assessing the Bias Correction Performance

| Scale | Scope   | Statistics                                      | Calculation                                                                 |
|-------|---------|------------------------------------------------|-----------------------------------------------------------------------------|
| Grid  | Moment  | Mean ($\mu$)                                    | $\mathbb{E}_{y \sim p(y)}, \mathbb{E}_{x \sim f(x)}, \mathbb{E}_{x \sim f(G(x))}$ |
|       |         | Standard deviation ($\sigma$)                    | $\sqrt{\mathbb{E}_{y \sim p(y)}(y - \mu_y)^2}, \sqrt{\mathbb{E}_{x \sim f(x)}(x - \mu_x)^2},$ |
|       |         | Skewness (Skew)                                 | $\mathbb{E}_{y \sim p(y)} \cdot \frac{y - \mu_y}{\sigma_Y} \cdot \frac{X - \mu_X}{\sigma_X}$, $\mathbb{E}_{x \sim f(x)} \cdot \frac{X - \mu_X}{\sigma_X}$ |
|       |         | Kurtosis (Kurt)                                 | $\mathbb{E}_{y \sim p(y)} \cdot \frac{y - \mu_y}{\sigma_Y} \cdot \frac{X - \mu_X}{\sigma_X}^3$, $\mathbb{E}_{x \sim f(x)} \cdot \frac{X - \mu_X}{\sigma_X}^3$ |
| Quantile | 1-3 tertile ($Q_{33.3\%}$, 66.7%, 99%)       | 33%, 66%, 99% quantile                              | Total precipitation days                                                    |

| Field | Intra-field | First and second empirical orthogonal function of precipitation field ($\text{EOF}_{i,j}$) | Leading orthonormal eigenvectors of precipitation correlation matrix |
|       | Inter-field | Applicability of $F^x_i$ and $F^y_i$ for mapping $S_k$ to $Y$, $S_k$ to $X$ or $G(X)$ | $r(F^x_i(S_k), Y), r(F^x_i(S_k), X), r(F^x_i(S_k), G(X))$ |

$$
r = \frac{\sum_{i,j}(\zeta^Y_{i,j} - \bar{\zeta}^Y)(\zeta^X_{i,j} - \bar{\zeta}^X)}{\sqrt{\sum_{i,j}(\zeta^X_{i,j} - \bar{\zeta}^X)^2(\zeta^Y_{i,j} - \bar{\zeta}^Y)^2}} \quad (6)
$$

$$
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i,j}(\zeta^X_{i,j} - \zeta^Y_{i,j})^2} \quad (7)
$$

Here $\zeta^X_{i,j}$ and $\zeta^Y_{i,j}$ are the considered indices for simulation and observation at geogrid $(i,j)$, $\bar{\zeta}^X$ and $\bar{\zeta}^Y$ are the spatial average of the statistics, $N$ is the total number of geogrids.

For field scale assessment, we consider the linear spatial structure of the precipitation field revealed by its leading two empirical orthogonal function (EOF$_{1,2}$). Also, we consider the inter-field coherency between precipitation and resolved dynamics by evaluating how the simulation-based and observation-based statistical downscaling model ($F^x_i$ and $F^y_i$) works for the observation and simulation data. We illustrate the rationale for these evaluations in the following section.
4. Bias Correction Results

4.1. General Performance

We evaluate the RADA methodology by examining if the distribution of the precipitation indices better match observations after bias correction. Evaluation results for the test set (1995–2004) are presented in Figures 3–8. Each sub-figure here shows the spatial maps of a considered precipitation index for observations (left), CESM2 historical simulations before (middle) and after (right) bias correction.

In general, the RADA methodology demonstrates favorable performance by faithfully representing the spatial variability of precipitation distribution, improving the spatial correlation, and reducing the spatial average RMSE between observations and simulations for all the considered grid-scale indices (Figures 3–6). We achieve significant skill improvements for a broad range of regions (stippled grids on the bias corrected maps), while the limited regions show significant skill drop (stippled grids on the uncorrected maps) for some low-precipitation related indices are mostly from mountain areas, suggesting the need for dedicated effort in modeling orographic effect in future development of our approach. Regarding the field-scale
assessments, our method better simulates the linear spatial structure of the precipitation field (Figure 7), and maintains the coherency between precipitation and the resolved dynamics (Figure 8).

4.2. Moments

For moment based evaluations (Figure 3), the improvements are reflected in the following four aspects. First, we obtain considerable improvements for the western mountain regions, especially for the Sierra Nevada, where the orographic lifting and shadowing impacts from the narrow mountain strips pose notorious difficulty for GCM simulations (Gershunov et al., 2019; Kapnick et al., 2018). Second, localized high precipitation variability incubated by peculiar geographical conditions can be well reproduced, such as for Southern California Bight. Third, the bias correction results better represent the intense and highly variate coastal precipitation along the Gulf of Mexico, and alleviate the overestimation along the eastern coast. Finally, the bias correction results better match observations regarding the inland precipitation distribution.

4.3. Quantile, Frequency, and Intensity

For quantile, frequency, and intensity based evaluations (Figures 4 and 5), we significantly alleviate the over estimation of low precipitation events for the Northern Great Plains, the Midwest, and the Southeast (Figure 4a). Also, we significantly draw down the precipitation frequency estimation to better match observations for most regions (Figure 4d). Exceptions are for mountain regions: our model tends to exacerbate
Figure 5. Similar to Figure 3 but for correcting the average intensity (a), 1-day, 3-day, and 5-day maximum precipitation (b–d).

Figure 6. Similar to Figure 3 but for the winter (a, December to January) and summer (b, June to August) precipitation ratio.
the overestimation of 33% quantile for part of the Appalachian Mountain Range (Figure 4a), while excessively eliminate certain amount of precipitation events for part of the Rocky Mountain range (Figure 4a). These deficiencies may result from sharing feature-learning convolution kernels with regions of distinct precipitation orographical impacts.

Another favorable aspect of the RADA approach shown here is that, we obtain considerable improvement in simulating precipitation extremes, as reflected in the 99% quantile (Figure 4c), 1-day maximum (Figure 5b), 3-day maximum (Figure 5c), and 5-day maximum (Figure 5d) estimation maps. The improvements are particularly obvious for the west coast and the east north central regions. We can largely attribute this improvement to the following two reasons: first, although extreme precipitations are rare events locally, they are ubiquitous as viewed from a broader, that is, continental, spatial scope. By modeling the continental precipitation field using CNNs, we obtain large sample diversity of extreme events. Second, while most existing climate-related machine learning methods apply maximum likelihood objective, which work

Figure 7. Spatial loading of the first (a) and second (b) principal components (PCs) of the precipitation field. The variance explained by PC₁ and PC₂ are labeled on top of each map. The spatial correlation coefficient and spatial average root mean square error between simulation-revealed EOFs and observation-revealed EOFs are calculated and denoted.

Figure 8. Applicability (correlation skill) of the statistical downscaling model \( F_X \) and \( F_Y \) for the observed and simulated precipitation field, where \( F_X \) and \( F_Y \) are U-net based statistical downscaling models trained using the CESM2 historical simulation data and observational data. The spatial correlation coefficient and spatial average root mean square error (RMSE) of the downscaling correlation skills are calculated and denoted.
poorly for modeling the tail of the distribution for the target variable (Pan, 2019; Van Horn & Perona, 2017), we apply the adversarial learning objective, encouraging the model to faithfully generate extremes as revealed by the observations.

### 4.4. Seasonality

In terms of seasonality (Figure 6), winter precipitation variability in CONUS is largely dominated by extratropical cyclone related storm tracks. The cyclone-related precipitation can be enhanced or suppressed as the atmospheric flows encounter the orography. The bias correction results better represent the spatial variability of winter precipitation ratio for the orographical complicated regions, such as the western mountain regions (Figure 6a). Also, we improve the representation of the dry winter climate for the Northern Rockies and plains (Figure 6a). For summer precipitation (Figure 6b), improvements are reflected in better representing the spatial variability of precipitation for the North American Monsoon in the southwest, coastal landing of tropical cyclones for the Gulf of Mexico and southeast coast. Furthermore, we also enhance the summer precipitation ratio to better match observations for the great plain.

### 4.5. Spatial Consistency

Regarding the spatial consistency of the precipitation estimation, we draw the first and second leading EOFs of the observed, simulated, and bias-corrected daily precipitation fields (Figure 7). The EOFs represent the spatial loading of the leading principal components (PCs) that account for most precipitation variability. Therefore, these EOFs could roughly tell the linear spatial structure of the precipitation field. For EOF1 (Figure 7a), the bias correction result agrees better with observations, with a spatial correlation coefficient improved from 0.85 to 0.93, and the spatial RMSE dropped from $2.1 \times 10^{-2}$ to $1.3 \times 10^{-2}$, suggesting that the leading mode of precipitation spatial variability can be better revealed using our approach. Besides, we lowered the ratio of explained variance of PC1 from 2.22% to 1.51%, which better matches observation (1.29%). For EOF2 (Figure 7b), while there is a considerable discrepancy among the maps, in particular over the Western U.S., the bias correction results still agree better with the observation, especially for the Southeast. Bias corrected results achieve a spatial correlation of 0.57 regarding PC1 loading, as compared to 0.15 in the original CESM2 simulation. These results suggest that the RADA methodology can maintain, and even improve, the spatial consistency of the target variable.

### 4.6. Inter-Field Correlation

Considering the inter-field correlation between precipitation and the resolved dynamics, we evaluate how the simulation-based and observation-based statistical models $F_X^\star$ and $F_Y^\star$ work for the observation and simulation data by calculating the $r$ skill of these two models (Figure 8). Conventionally, we consider GCM biases at a single variable level, the assessments here examine GCM biases at a functional relationship level. For the simulation-based downscaling results (Figure 8a), $F_X^\star$, which is trained using CESM2 historical simulation data, works particularly well for estimating CESM2 simulated precipitation using the simulated dynamical field information, achieving a spatial average $r$ skill score of 0.89 during test (Figure 8a middle, $r\left(F_X^\star\left(S_X, X\right)\right)$). However, this GCM revealed inter-field relationship works far worse when applied to observations, with the average $r$ skill score dropped to 0.42 (Figure 8a left, $r\left(F_X^\star\left(S_X, Y\right)\right)$). This drop of skill can result from the following two reasons, first, the GCM imposes an overly stringent constraint on simulated precipitation variability; second, the downscaling relationship revealed the GCM has non-linear biases.

We dissect the impacts from both sides by examining the observation-based downscaling results (Figure 8b). The impact of the first aspect (over stringency) is reflected in the fact that, $F_Y^\star$, trained using historical observational data, cannot pose as stringent a constraint on observed precipitation (Figure 8b left, $r\left(F_Y^\star\left(S_Y, Y\right)\right)$) as $F_X^\star$ when applied to simulation data (Figure 8a middle, $r\left(F_X^\star\left(S_X, X\right)\right)$). In other words, the GCM underestimates the conditional variability of the parameterized variable given the resolved dynamics. The impact of the second aspect (non-linear bias) is reflected in the fact that, $F_Y^\star$ achieves an overall higher $r$ skill score for
estimating the observed precipitation using the dynamical field information (Figure 8b left, \( r\left(F_Y(S_X), Y\right) \)), as compared to \( F_X \) when applied to observations (Figure 8a left, \( r\left(F_X(S_Y), Y\right) \)).

How are these two aspects of deficiencies reflected in the RADA bias correction results? First, as we apply \( F_X \) to estimate the bias corrected precipitation (Figure 8a right, \( r\left(F_X(S_X), G(X)\right) \)), we obtain lower \( r \) skill compared to the estimation of the uncorrected precipitation (Figure 8a middle, \( r\left(F_X(S_X), X\right) \)), which is more consistent with the observation-based result (Figure 8a left, \( r\left(F_X(S_Y), Y\right) \)). This suggests that we may alleviate the deficiency in CESM2 precipitation estimation resulting from either over-stringency or non-linear bias. Second, \( F_Y \) obtains better applicability for the bias correction result (Figure 8b right, \( r\left(F_Y(S_X), G(X)\right) \)), compared to its applicability for the original CESM2 precipitation estimation (Figure 8b middle, \( r\left(F_Y(S_X), X\right) \)). On one hand, this confirm that we alleviate the non-linear bias in estimating precipitation using the RADA methodology; on the other hand, since we still achieve considerable higher \( r \) skill for using \( F_Y \) to estimate the bias correction precipitation (Figure 8b right, \( r\left(F_Y(S_X), G(X)\right) \)), compared to estimating the observed precipitation (Figure 8b left, \( r\left(F_Y(S_Y), Y\right) \)), the over stringency problem is not fully resolved by this deterministic bias correction approach. An implication here is that we may need to inject random noise into the bias corrector for probabilistic estimation of the unresolved precipitation processes, as is explored in stochastic parameterization practices (Berner et al., 2017). We revisit this issue in Section 8.2.1.

5. Ablation Analysis

We have introduced three regularizers to enhance adversarial learning of GCM biases, namely, cycle consistency (Section 2.2.2), dynamical consistency (Section 2.2.3), and dynamical dependency (Section 2.2.4). Here, we quantify the contribution from each regularizer or their combinations using an ablation analysis method: we monitor the change of bias correction performance by systematically removing the regularizers in the RADA network architecture. The combinations of these three regularizers consist seven GAN-based bias corrector baselines.

We show the bias correction results for each of the baseline models in Figures S1–S7 in Supporting Information S1. Each figure there draws the spatial maps of all the considered precipitation indices for observations and simulations before and after bias correction. To compare models’ performances, Figure 9 draws models’ relative spatial correlation skills: for each model and each considered precipitation distribution characteristics, we first calculate the spatial correlation between the observation-revealed precipitation indices and simulation-revealed precipitation indices, thereafter, we re-scale the models’ correlation skills to the range of \([0, 1]\) by subtracting the minimum correlation skill achieved by the nine considered precipitation simulations (CESM2 original simulation, RADA bias correction, and seven GAN-based bias correction), and dividing the result by the difference between the maximum and minimum correlation skill achieved by these models. The skills of CESM2 original simulation (gray dashed) and RADA corrected simulation (black solid) are plotted as benchmarks.

The key findings are summarized here. Applying unconstrained GAN for bias correction yields poor performance regarding most of the considered precipitation distribution characteristics (Figure 9a). This is because, GANs are trained to produce individual trust-worthy samples, not accurate probability distribution estimations (Goodfellow, 2016). The overall performance is generally improved as more regularizers are included to constrain the bias corrector (Figures 9b–9g), although the inclusion of a new regularizer might hamper the performance for certain aspects. The RADA bias corrector containing all three regularizers achieves overall the best performance compared to the rest GAN-based models. These evidences empirically confirm the arguments made in Section 2.2.1, and justify the contribution of the proposed regularizers.
6. Comparison Study

In this section, we compare the RADA methodology with widely applied univariate and multivariate bias correction approaches. The first approach is quantile mapping, which is a univariate correction technique that maps quantiles of historical climate simulations to quantiles of corresponding climate observations (Mamalakis et al., 2017; Michelangeli et al., 2009):

\[ x_c = \text{CDF}_Y^{-1}(\text{CDF}_X(x)) \quad (8) \]

Here \( x \) and \( x_c \) are the grid-scale target variable simulation before and after bias correction, \( \text{CDF}_Y^{-1} \) is the inverse of the empirical cumulative distribution function of the observation, \( \text{CDF}_X \) is the empirical cumulative distribution function of the original simulation. Both \( \text{CDF}_Y^{-1} \) and \( \text{CDF}_X \) are built on the training set data as defined in Section 3.2.3. Model performance is evaluated on the test set. The implementation details are described in the Supporting Information S1.

The quantile mapping approach neglects the spatial consistency of the target variable in its correction. Regarding this deficiency, many multivariate bias correction methods have been developed, although the robustness and benefits of these multivariate methods are frequently questioned (Maraun & Widmann, 2018; Räty et al., 2018), and remain to be rigorously tested (François et al., 2020; Maraun, 2016). Here we consider the Multivariate Bias Correction with N-dimensional probability density function transform (MBCn, Cannon, 2018; Pitie et al., 2005) approach to correct precipitation projection biases while maintaining the spatial consistency. MBCn is adopted for the following three reasons: First, MBCn is claimed to achieve competitive performance compared to univariate or other multivariate bias correction approaches (François et al., 2020; Cannon, 2018). Second, the computation cost for scaling up MBCn to high dimensional data...
is relatively low. Third, we do not need to pre-assume the parametric distribution form of the considered climate variable.

MBCn maps a multivariate source distribution to a same-dimensional target distribution by iteratively applying the following three steps. First, rotate the multivariate source distribution and target distribution with a same uniformly distributed random orthogonal matrix. Second, apply quantile mapping to match the source distribution with the target distribution for each of the rotated dimensions. Third, apply the inverse rotation matrix to yield source values for the next iteration. By the end of each iteration, an energy distance function (Rizzo & Székely, 2016) is applied to measure the distance between corrected source distribution and target distribution. We consider two strategies for stopping the iterative bias correction process: the first strategy, following the setting in Cannon (2018), is to stop the iteration as the energy distance function evaluated for the training set converges; the second strategy, as advocated in the discussion section of Cannon (2018), is to use early stopping (Prechelt, 1998): we stop the iteration as the energy distance function evaluated for the validation set stops decreasing. We evaluate the model performance using the test set data. The implementations can be found in the Supporting Information S1. More details can be found in Cannon (2018) and François et al. (2020).

The comparison results are presented in a similar manner as Section 5. We show the bias correction results for quantile mapping and MBCn in Figure S8 in Supporting Information S1 (quantile mapping) and Figures S9 and S10 in Supporting Information S1 (MBCn without/with early stopping). Each figure draws the spatial maps of all the considered precipitation indices for observations and simulations before and after bias correction. Figure 10 draws models’ relative spatial correlation skills (see Section 5 for computation details).

The quantile mapping approach (Figure 10a) obtains good performance for correcting grid-scale precipitation distribution characteristics, slightly surpassing the RADA approach regarding some of the considered indices ($\mu$, $\sigma$, $Q_{33\%}$, $Q_{99\%}$, and Int). These are reasonable results as we empirically align the simulated precipitation distribution to match observed precipitation distribution for each grid cell. Meanwhile, the RADA approach demonstrates advantage for better matching observations regarding the higher-order moments (Skew and Kurt), $Q_{33\%}$, the probability of precipitation, the daily maximum precipitation, the seasonal precipitation ratio, and all the field-scale precipitation characteristics. For grid-scale assessments, the disadvantage of quantile mapping here may due to the fact that these statistics are sensitive to sampling variance, requiring more samples to build empirical CDFs that better characterize these statistics, while the advantage of RADA is achieved by generalizing from limited available data using deep neural networks. For field-scale assessments, the disadvantage of quantile mapping is due to the fact that it does not explicitly take into consideration of spatiotemporal consistency or inter-field correlation, while the RADA is designed to overcome these deficiencies.

Figure 10. Relative spatial correlation skills for (a) the quantile mapping approach, (b) the MBCn approach without early stopping, and (c) the MBCn approach with early stopping. For each model and each considered precipitation distribution characteristics, we calculate the spatial correlation between the observation-revealed precipitation indices and simulation-revealed precipitation indices. We re-scale the models’ correlation skills to the range of [0, 1] using min-max normalization based on correlation skills achieved by the three baseline models, the original CESM2 simulation, and the Regularized Adversarial Domain Adaptation (RADA) model. The skills of CESM2 original simulation (gray dashed) and RADA corrected simulation (black) are plotted as benchmarks.
The MBCn approach (Figures 10b and 10c, red) obtains relatively lower skill compared to either the quantile mapping approach or the RADA approach. Even worse, many of the statistical indices, after MBCn bias correction, match less well with observations. This degradation of skill is particularly obvious if the early stopping strategy is not adopted (Figure 10b). We provide the following explanations for these phenomena. First, most of the reported implementations of MBCn apply cross-validation evaluation (Cannon, 2018; François et al., 2020), which tend to yield misleadingly promising performance, see Maraun et al. (2017) and Maraun & Widmann (2018) for theoretical and empirical proofs. As we evaluate the MBCn approach using a hold-out test set, we find the iterative bias correction strategy in MBCn generalizes poorly, as compared to either the straightforward, one-step quantile mapping approach, or the sophisticated, but well regularized RADA approach. Second, in each iteration, MBCn uses a random orthogonal matrix to partially de-correlate the precipitation field data, and carry out bias correction on the basis of the orthogonal matrix. This linear assumption may not work well for the highly variate precipitation field data. Overall, the results here suggest the difficulty for extending univariate bias correction to multivariate cases, and highlight the contribution of the RADA methodology.

Finally, it is worth noting that, many recent works have tried to correct climate model biases conditioned on model resolved dynamics (Addor et al., 2016; Bellprat et al., 2013; Wetterhall et al., 2012), which echoes the idea of applying resolved dynamics to regularize the bias correction in the RADA methodology. However, these existing works often overly simplify the dynamical states conditioned on which to correct model biases. For instance, Manzanas and Gutiérrez (2019) showed that, conditioning the bias corrector on large scale circulation patterns, such as El Niño/Southern Oscillation (Manzanas & Gutiérrez, 2019), enhances model’s seasonal forecasting capability. The RADA methodology suggests a potential direction for correcting climate model biases conditioned on more detailed dynamical state information.

7. Perceptual Evaluation

The RADA methodology, derived from the idea of adversarial learning, builds an evolving game between the neural network parameterized bias detector and bias corrector under cycle-consistent statistical and dynamical constraints. The overall loss function ($\mathcal{L}_{\text{RADA}}$, Equation 5) does not directly inform the model performance or when to stop the training. In many GAN applications, the training is stopped when there is no perceptual improvement of the generated samples (Salimans et al., 2016). Here, we evaluate the perceptual performance of the RADA methodology. We empirically prove that, compared to the judgment made by the neural network parameterized bias detector, the perceptual judgment made by human beings can be easily fooled in our problem setting, yielding less optimal bias correction performance.

We consider a binary distinction of climate observation samples and simulation samples, judged by a human being test taker and a separately trained deep neural network. This problem formulation is similar to the climatic Turing test proposed by Palmer (2016). The deep neural network applied here has same architecture as the RADA discriminative neural network, but unlike the setting in Wasserstein GAN, the weights are not clipped to fully exploit its discrimination power. Both the human being and the deep neural network are trained to distinguish historical climate observation samples and simulation samples, and are tested to make distinctions for a holdout test set. We compare the human being and the deep neural network performance for the distinction task before and after RADA bias correction. Results are shown in Figure 11.

The top of Figure 11 shows randomly selected samples of precipitation observation maps (first row) and the CESM2 precipitation simulation maps before (second row) and after (third row) bias correction. The simulations do not match the observations, as the freely evolving historical climate simulations do not follow the evolution of historical climate variability. Although unpaired, we can still easily distinguish the observations (first row) and the original simulations (second row): the observations are of higher spatial variability, and are featured by well-pronounced orographical impact; the simulations are of lower spatial variability, and are featured by well-expanded drizzles. The RADA bias correction results (third row) largely overcome the perceptual deficiencies, while still matching the spatial distribution pattern of the original simulations.

The bottom of Figure 11 summarizes the performance of the human being test taker (top) and the deep neural network (bottom) for distinguishing observation samples and simulation samples before (left) and after (right) bias correction. Here, TP, FP, TN, and FN are true positive rate, false positive rate, true negative
rate, and false negative rate, respectively. Before bias correction, both the human being test taker and the deep neural network can well distinguish the observation samples and the simulation samples. While the human being test taker misclassifies 5% of the samples, the accuracy of the deep neural network is nearly perfect. After bias correction, the human being test taker assigns 67.92% of simulation samples to observations, while the deep neural network shows a large advantage in the distinction task. These results suggest that, first, the perceptual judgment made by human beings is not as accurate as the judgment made by the discriminative deep neural network. In the current problem setting and many other physical applications, we should not rely purely upon perceptual judgment to evaluate the model performance, or to determine the training stopping criteria. Second, since a separately trained deep neural network can still distinguish the observation samples and the simulation samples with moderate accuracy, the current bias correction model may not have fully exploited the bias information in the data. We discuss the limitations of the model, and suggest potential directions for improvements in Section 8.2. Last, while the climatic Turing test (Palmer, 2016) was originally proposed as a subjective evaluation of climate models, our methodology shows that, this Turing test can serve as a powerful tool for not only model diagnosis, but also model improvement. Eventually, this methodology may potentially allow us to build a digital twin (Bauer et al., 2021) of the climate system.

8. Conclusion and Future Work

8.1. Contributions

The accuracy of climate projection is continuously marred by biases of climate models, due to their simplification or misrepresentation of unresolved climate processes. The community has reached a consensus
that, to significantly improve climate projection accuracy in the next-generation climate models, we should leverage the information hidden in multi-resolution simulations and multi-source observations, using the powerful tools of machine learning (Gentine et al., 2018; Schneider et al., 2017). Current endeavors along this direction mainly focus on learning data-driven parameterization schemes from high-resolution simulations, either deterministically (Rasp et al., 2018; Yuval & O’Gorman, 2020) or probabilistically (Gagne et al., 2020; Schneider et al., 2020).

Despite these efforts, climate models equipped with enhanced parameterization schemes (either conceptual based, or machine learning based) may still disagree with observations in many aspects, reflecting the gap between parameterization development and application, and highlighting the sophisticated structure of climate model biases. These biases cannot be readily revealed by contrasting historical climate simulation and observation time series, as the freely evolving historical climate simulations do not follow the evolution of actually observed historical climate variability. To identify and alleviate these biases, we may either carry out costly process-oriented diagnosis, or adjust model output distribution toward observations regardless of physical justifications. It remains challenging to efficiently apply the simulation-observation “disagreement” information to diagnose and improve climate models.

In this study, we develop the regularized adversarial domain adaptation (RADA) methodology that learns to identify and correct climate projection biases using unpaired climate simulation and observation data. We are inspired by the fact that, climate experts can easily distinguish maps of observations and model simulations, even though they represent different climate states (Palmer, 2016). As we apply a deep neural network for this observation-simulation distinction task, the evidences based on which this network derives its judgment roughly tell where the model bias lies, which can be inversely applied to train a domain adaptation network to have simulations less distinguishable from observations. Eventually, this may lead us to the ideal of digital twin (Bauer et al., 2021) for developing climate models. The adversarial learning process can be highly unconstrained, yielding superficially authentic yet physically unreasonable results. We turn the three requirements that Ehret et al. (2012) conceived for a “perfect” bias corrector into regularizers to support the adversarial learning of climate model biases. Specifically, we include cycle consistency regularization to encourage model to produce identical simulations as observations under same resolved dynamical states; we include dynamical consistency regularization to ensure spatiotemporal consistency as well as inter-field correlations in bias correction; we make the bias corrector dynamical dependent to discriminate between different dynamical conditions.

We apply this RADA methodology for correcting the Community Earth System Model version 2 daily precipitation projection over the Contiguous United States. Results show that our methodology can correct all the considered moments of daily precipitation at approximately 1° resolution, ensure spatiotemporal consistency and inter-field correlations, and discriminate between different dynamical states. We apply an ablation study to justify the contribution of the introduced regularizers. Compared to existing univariate and multivariate bias correction approach, our methodology can well correct all the considered statistics at grid scale, and perform better in maintaining intra-field and inter-field consistency.

The key message we want to deliver is that, we should learn from big data of climate observations and simulations to criticize and improve the ever-complicated climate models. Logically, sophisticated climate models should come with equally sophisticated diagnosis tools (Hofstadter, 1979), which could be built upon our methodology. We summarize three of our key contributions as follows:

1. We develop a powerful and efficient paradigm for learning climate projection biases using unpaired climate simulation and observation data.
2. We introduce useful regularization approaches to support physically consistent adversarial learning of climate projection biases. New regularizers can be more easily included following the strategies developed here.
3. We describe comprehensive training and evaluation details for implementing and assessing the proposed data-driven bias correction methodology, demonstrating the advantage of our approach over existing univariate or multivariate bias correction methods.
8.2. Limitations and Future Work

We have developed an adversarial learning based data-driven climate model bias corrector (RADA), equipped with three regularizers targeted toward the three requirements that Ehret et al. (2012) conceived for a “perfect” bias correction algorithm. Here, we highlight four limitations of the current RADA methodology, and discuss our future work directions.

8.2.1. Scalability

The current RADA methodology is applied to correct daily precipitation projection at 1° resolution. In order to make this methodology applicable for GCMs with increasingly high resolutions, we should address the following two problems: first, how to scale the learning algorithm to high spatiotemporal resolution cases; second, how to include stochastic component to account for the inherent uncertainty in parameterizing the unresolved processes (Berner et al., 2017). Below, we discuss the challenges and possible solutions of these two problems.

Since the development of the first GAN model (Goodfellow et al., 2014), many works have explored the possibility of scaling adversarial learning to large models and datasets, achieving consistent progresses regarding training stability, computation efficiency, and output quality (Karras et al., 2017, 2019; Kurakin et al., 2016; Radford et al., 2015). In particular, Karras et al. (2017) showed that, starting from a low resolution, as we progressively add new layers to refine the model output details, we can achieve robust, high-resolution results in adversarial learning. In our future work, we plan to apply this idea to extend the RADA methodology from moderate resolution to high resolution.

While it is relatively straightforward to follow existing works to scale the learning machine to high resolution cases, we should notice the physical limitation here: as the GCM resolution is increased, there is no clear scale separation between resolved dynamics and parameterized physics (Gerard, 2007). In this case, the resolved dynamics cannot support a strict one-to-one mapping between observations and simulations. This partially explains why the cycle consistency regularization (Section 2.2.2) does not significantly enhance bias correction performance (Figure 9b). To address this problem, we should include stochastic component to account for the inherent uncertainty in parameterizing the unresolved processes (Berner et al., 2017). Similar issue has been studied in the machine learning literature (Almahairi et al., 2018; Park et al., 2020). Instead of learning strict one-to-one mapping between domains, these methods are designed to learn probabilistic, many-to-many mappings. In particular, by adding auxiliary latent variables to the cycle-consistency GAN model, Almahairi et al. (2018) showed that we can learn mappings that capture diversity of the outputs in domain adaptation. In our future work, we plan to explore similar idea to learn stochastic bias correction for high dimensional cases.

8.2.2. Explainability

A unique feature of the RADA methodology is that, we do not pre-assume where the GCM biases are, and apply universal function approximators (deep neural nets, Leshno et al., 1993) to identify (via $D$) and correct (via $G$) these biases. Although this setting may potentially allow us to identify and correct arbitrary GCM biases, the spatiotemporal characteristics of the GCM biases and their dependency on the GCM resolved dynamics remain vague. To apply this modeling framework for climate studies, the neural network should provide human-understandable justifications for its outputs, and help model developers to diagnose and improve GCMs. Several neural network interpretability techniques, including saliency maps (Simonyan et al., 2013), occlusion sensitivity analysis (Zeiler & Fergus, 2014), backward optimization (Olah et al., 2017) and layerwise relevance propagation (Bach et al., 2015; Toms et al., 2020), could be applied to verify if the RADA methodology identifies meaningful model biases, enhance our understanding of model deficiencies, and support model improvement.

8.2.3. Applicability in a Changing Climate

We have trained the RADA model using historical climate simulation and observation data from 1950 to 1994, and test the bias correction performance using a hold-out test set, covering 1995–2004. The results suggest that the RADA methodology applies well for a “future,” changing scenario. Meanwhile, the evaluation here does not guarantee that the bias corrector trained using historical data can extrapolate well in a severely changed climate, such as the climate projections forced by high greenhouse gas emission scenarios.
On the one hand, we should strictly exam the bias correction performance for different climate periods; on the other hand, we should apply the detailed bias information offered by the RADA methodology to diagnose and improve the GCMs, translating the black-box model diagnosed bias information into robust knowledge encoded in climate models. The following section offers a potential direction to achieve this goal.

### 8.2.4. Feedback Effects

The three regularizers in the RADA methodology target toward the three requirements that Ehret et al. (2012) conceived for a “perfect” bias correction algorithm. So far, we have not explicitly considered the feedback effect associated with GCM parameterization biases, which consists the last requirement from Ehret et al. (2012), as well as the most difficult challenge in correcting climate projection biases.

There are potentially two directions toward tackling the feedback effects from a bias correction perspective, one is to couple the bias corrector with dynamical simulations, the other is to iteratively correct the parameterization biases and alleviate their impacts on model resolved dynamics.

The latter direction is more feasible as it poses lower requirement on the spatiotemporal resolution of the bias corrector, and lighter burden on model development. Provided that the bias corrector can pinpoint detailed spatiotemporal bias information without costly initialization and long-term running of climate models, we could obtain sufficient guidance for calibrating and improving climate models (Anderson & Lucas, 2018; Gong et al., 2016; Hourdin et al., 2017; Rasp et al., 2018; Yuval & O’Gorman, 2020). By iteratively training and applying this bias corrector to overcome the parameterization biases, we may gradually alleviate their impacts on model dynamics, therefore resolve the feedback impacts of parameterization biases.

To support the blueprint drawn above, we should test if the bias corrector can offer detailed spatiotemporal bias information. A perfect test bed is weather forecast postprocessing. Historically, postprocessing in weather forecast and bias correction in climate projection are considered as two disciplines (Maraun, 2016), although they result from similar model biases.

The barrier that prevents us from carrying out a weather forecast postprocessing verification of our methodology is that, retrospective weather forecast experiments usually apply different GCM settings compared with climate projections, regarding the resolution, numerical solver, and parameterization choices. The community has realized the necessity for applying an integrated forecasting system across weather to seasonal scale. Here we advocate that the unification should extend to climate scale for better identification and correcting of forecasting model biases. Before this is carried out, our future works plan to test if the learned bias corrector can enhance weather to seasonal forecast in a semi-supervised learning manner, alleviating the sample limitation problem in learning forecasting biases.

### Data Availability Statement

The Community Earth System Model version 2 (CESM2) historical simulation data are available through https://esgf-node.llnl.gov/projects/cmip6/. The CPC unified gauge-based analysis of daily precipitation data are available through https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html. The ECMWF atmospheric reanalysis of the twentieth century (ERA-20C) data are available through https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-20c. The code is available through https://github.com/panbaoxiang/RADA.

### References

Addor, N., Rohrer, M., Furrer, R., & Selbert, J. (2016). Propagation of biases in climate models from the synoptic to the regional scale: Implications for bias adjustment. *Journal of Geophysical Research: Atmospheres*, 121(5), 2073–2089. https://doi.org/10.1002/2015jd024040

Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., & Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International conference on machine learning* (pp. 195–204).

Anderson, G. J., & Lucas, D. D. (2018). Machine learning predictions of a multiresolution climate model ensemble. *Geophysical Research Letters*, 45(9), 4273–4280. https://doi.org/10.1002/2018gl077049

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. -R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Baño-Medina, J. L., García Manzanas, R., Gutiérrez Lorente, J. M., et al. (2020). Configuration and intercomparison of deep learning neural models for statistical downsampling.
Ma, F., Ye, A., Deng, X., Zhou, Z., Liu, X., Duan, Q., et al. (2016). Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental china. International Journal of Climatology, 36(1), 132–144. https://doi.org/10.1002/joc.4333

Ma, H.-Y., Xie, S., Klein, S., Williams, K., Boyle, J., Bony, S., et al. (2014). On the correspondence between mean forecast errors and climate errors in CMIP5 models. Journal of Climate, 27(4), 1781–1798. https://doi.org/10.1175/jcli-d-13-00474.1

Mamalakis, A., Langouinis, A., Deidda, R., & Marroco, M. (2017). A parametric approach for simultaneous bias correction and high-resolution downsampling of climate model rainfall. Water Resources Research, 53(3), 2149–2179. https://doi.org/10.1002/2016wr198578

Manzanas, R., & Gutiérrez, J. M. (2019). Process-conditioned bias correction for seasonal forecasting: A case-study with ENSO in Peru. Climate Dynamics, 52(3), 1673–1683. https://doi.org/10.1007/s00382-018-4226-z

Maraun, D. (2013). Bias correction, quantile mapping, and downsampling: Revisiting the inflation issue. Journal of Climate, 26(6), 2137–2143. https://doi.org/10.1175/jcli-d-12-00821.1

Maraun, D. (2016). Bias correcting climate change simulations—A critical review. Current Climate Change Reports, 2(4), 211–220. https://doi.org/10.1007/s40641-016-0050-x

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., et al. (2017). Towards process-informed bias correction of climate change simulations. Nature Climate Change, 7(11), 764–773. https://doi.org/10.1038/nclimate3418

Maraun, D., & Widmann, M. (2018). Cross-validation of bias-corrected climate simulations is misleading. Hydrology and Earth System Sciences, 22(9), 4867–4873. https://doi.org/10.5194/hess-22-4867-2018

Marvel, K., & Bonfils, C. (2013). Identifying external influences on global precipitation. Proceedings of the National Academy of Sciences, 110(48), 19301–19306. https://doi.org/10.1073/pnas.1314382110

Miao, Q., Pan, B., Wang, H., Hsu, K., & Sorooshian, S. (2019). Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water, 11(5), 977. https://doi.org/10.3390/w11050977

Michelangeli, P. -A., Vrac, M., & Loukos, H. (2009). Probabilistic downscaling approaches: Application to wind cumulative distribution functions. Geophysical Research Letters, 36(11). https://doi.org/10.1029/2009gl038401

Mills, K., Spanner, M., & Tamblin, I. (2017). Deep learning and the schrödinger equation. Physical Review A, 96(4), 042113. https://doi.org/10.1103/physreva.96.042113

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. Distill, (2)(11), e7. https://doi.org/10.23915/distill.00007

Palmer, T. (2016). A personal perspective on modelling the climate system. Geophysical Research Letters, 43(21), 9205–9216. https://doi.org/10.1002/2016gl069948

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., & Higgins, W. (2019). Improving precipitation estimation using convolutional neural network. Water Resources Research, 55(3), 2301–2321. https://doi.org/10.1029/2018wr024090

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., & Higgins, W. (2019). Precipitation prediction skill for the west coast United States: From short to extended range. Journal of Climate, 32(1), 161–182. https://doi.org/10.1175/jcli-d-18-0355.1

Park, T., Efros, A. A., Zhang, R., & Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In Advances in Neural Information Processing Systems 33 (pp. 319–345). NeurIPS 2020.

Pan, R., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., & Lee, J. (2020). Improving seasonal forecast using probabilistic deep learning.

Pan, R., Anderson, G. J., Lucas, D. D., Goncalves, A., Bonfils, C., Lee, J., & Tian, Y. (2020). Identifying and correcting climate projection biases using artificial intelligence. AGU Fall Meeting.

Park, T., Efros, A. A., Zhang, R., & Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In European conference on computer vision (pp. 339–345). https://doi.org/10.1007/978-3-030-58545-7_19

Phillips, T. J., Potter, G. L., Williamson, D. L., Cederwall, R. T., Boyle, J. S., Fiorino, M., et al. (2004). Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. Bulletin of the American Meteorological Society, 85(12), 1903–1916. https://doi.org/10.1175/bams-85-12-1903

Piani, C., Weedon, G., Best, M., Gomes, S., Viterbo, P., Hagemann, S., & Haerter, J. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. Journal of Hydrology, 393(3–4), 199–215. https://doi.org/10.1016/j.jhydrol.2010.10.024

Pitie, F., Kokaram, A. C., & Dahyot, R. (2005). N-dimensional probability density function transfer and its application to color transfer. Tenth IEEE International Conference on Computer Vision (iccv’05), 1, 1343–1349. https://doi.org/10.1109/iccv.2005.166

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., et al. (2016). Era-20c: An atmospheric reanalysis of the twentieth century. Journal of Climate, 29(11), 4083–4097. https://doi.org/10.1175/jcli-d-15-0556.1

Prechtl, L. (1998). Automatic early stopping using cross validation: Quantifying the criteria. Neural Networks, 11(4), 761–767. https://doi.org/10.1016/s0893-6080(98)00010-0

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11), 3885–3900. https://doi.org/10.1175/mwr-d-18-0187.1

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. Proceedings of the National Academy of Sciences, 115(39), 9684–9689. https://doi.org/10.1073/pnas.1802861115

Rätty, O., Räisänen, J., Bossard, T., & Donnelly, C. (2018). Intercomparison of univariate and joint bias correction methods in changing climate from a hydrological perspective. Climate, 6(2), 33. https://doi.org/10.3390/cli6020033

Rizzo, M. L., & Székely, G. J. (2016). Energy distance. wiley Interdisciplinary Reviews: Computational Statistics, 8(1), 27–38. https://doi.org/10.1002/wics.1375

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International conference on medical image computing and computer-assisted intervention (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Sadeghi, M., Nguyen, F., Hsu, K., & Sorooshian, S. (2020). Improving near real-time precipitation estimation using a u-net convolutional neural network and geographical information. Environmental Modelling & Software, 134, 104856. https://doi.org/10.1016/j.envsoft.2020.104856

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANS. Advances in Neural Information Processing Systems, 29, 2234–2242.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. Geophysical Research Letters, 44(24), 12–396. https://doi.org/10.1002/2017gl076101

Schneider, T., Stuart, A. M., & Wu, J.-L. (2020). Learning stochastic closures using ensemble kalman inversion.

Simonoyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps.
