JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001896

Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz '96 Model

David John GagneII[1], Hannah M. Christensen[1,2], Aneesh C. Subramanian[3], and Adam H. Monahan[4]

[1]National Center for Atmospheric Research, Boulder, CO, USA, [2]Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK, [3]Department of Atmospheric and Oceanic Sciences,University of Colorado, Boulder, CO, USA, [4]School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, Canada

Key Points:
- We have developed a generative adversarial network (GAN) stochastic parameterization of subgrid forcing for the Lorenz '96 dynamical model
- Online evaluation at weather and climate time scales reveals that some GAN configurations outperform a bespoke polynomial baseline model
- The GANs closely reproduce the spatiotemporal correlations and regimes of the Lorenz '96 system with only localized inputs

Correspondence to:
D. J. Gagne,
dgagne@ucar.edu

Citation:
Gagne, D. J., Christensen, H., Subramanian, A., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001896. https://doi.org/10.1029/2019MS001896

Received 16 SEP 2019
Accepted 14 FEB 2020
Accepted article online 18 FEB 2020

Abstract Stochastic parameterizations account for uncertainty in the representation of unresolved subgrid processes by sampling from the distribution of possible subgrid forcings. Some existing stochastic parameterizations utilize data-driven approaches to characterize uncertainty, but these approaches require significant structural assumptions that can limit their scalability. Machine learning models, including neural networks, are able to represent a wide range of distributions and build optimized mappings between a large number of inputs and subgrid forcings. Recent research on machine learning parameterizations has focused only on deterministic parameterizations. In this study, we develop a stochastic parameterization using the generative adversarial network (GAN) machine learning framework. The GAN stochastic parameterization is trained and evaluated on output from the Lorenz '96 model, which is a common baseline model for evaluating both parameterization and data assimilation techniques. We evaluate different ways of characterizing the input noise for the model and perform model runs with the GAN parameterization at weather and climate time scales. Some of the GAN configurations perform better than a baseline bespoke parameterization at both time scales, and the networks closely reproduce the spatiotemporal correlations and regimes of the Lorenz '96 system. We also find that, in general, those models which produce skillful forecasts are also associated with the best climate simulations.

Plain Language Summary Simulations of the atmosphere must approximate the effects of small-scale processes with simplified functions called parameterizations. Standard parameterizations only predict one output for a given input, but stochastic parameterizations can sample from all the possible outcomes that can occur under certain conditions. We have developed and evaluated a machine learning stochastic parameterization, which builds a mapping between large-scale current conditions and the range of small-scale outcomes from data about both. We test the machine learning stochastic parameterization in a simplified mathematical simulation that produces multiscale chaotic waves like the atmosphere. We find that some configurations of the machine learning stochastic parameterization perform slightly better than a simpler baseline stochastic parameterization over both weather- and climate-like time spans.

1. Introduction

A large source of weather and climate model uncertainty is the approximate representation of unresolved subgrid processes through parameterization schemes. Traditional, deterministic parameterization schemes represent the mean or most likely subgrid scale forcing for a given resolved-scale state. While model errors can be reduced to a certain degree through improvements to such parameterizations, they cannot be eliminated. Irreducible uncertainties result from a lack of scale separation between resolved and unresolved processes. Uncertainty in weather forecasts also arises because the chaotic nature of the atmosphere gives rise to sensitivity to uncertain initial conditions. Practically, uncertainty is represented in forecasts using ensembles of integrations of comprehensive weather and climate prediction models, first suggested by Leith (1975). To produce reliable probabilistic forecasts, the generation of the ensemble must include a representation of both model and initial condition uncertainty.

Initial condition uncertainty is addressed by perturbing the initial conditions of ensemble members, for example, by selecting directions of optimal perturbation growth using singular vectors (Buizza & Palmer,

1995) or by characterizing initial condition uncertainty during the data assimilation cycle (Isaksen et al., 2010). One approach for representing irreducible model uncertainty is stochastic parameterization of unresolved physical processes. A stochastic parameterization represents the probability distribution of possible subgrid scale tendencies conditioned on the large scale. Each ensemble member experiences one possible, equally likely realization of the subgrid-scale tendencies. A more detailed motivation for including stochastic parameterizations in weather and climate models is presented in Palmer (2012).

Stochastic approaches for numerical weather prediction (NWP) were originally proposed for use in the European Center for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (Buizza et al., 1999; Palmer et al., 1997). They were shown to substantially improve the quality of initialized ensemble forecasts and so became widely adopted by meteorological services around the world, where they are used to produce operational ensemble weather, subseasonal, and seasonal forecasts (Berner et al., 2010; Leutbecher et al., 2017; Palmer et al., 2009; Palmer, 2012; Reyes et al., 2009; Suselj et al., 2013; Sušelj et al., 2014; Stockdale et al., 2011; Sanchez et al., 2016; Teixeira & Reynolds, 2010; Weisheimer et al., 2014).

Recent work has assessed the impact of stochastic parameterization schemes in both idealized and state-of-the-art climate models for long-term integration (Ajayamohan et al., 2013; Christensen et al., 2017; Dawson & Palmer, 2015; Davini et al., 2017; Juricke & Jung, 2014; Strømmen et al., 2018; Williams, 2012; Wang et al., 2016). These studies demonstrate that including a stochastic representation of model uncertainty can go beyond improving initialized forecast reliability and can also lead to improvements in the model mean state (Berner et al., 2012; Palmer, 2001) and climate variability (Ajayamohan et al., 2013; Christensen et al., 2017; Dawson & Palmer, 2015) and change a model's climate sensitivity (Seiffert & von Storch, 2010). These impacts occur through nonlinear rectification, noise-enhanced variability, and noise-induced regime transitions (Berner et al., 2017). In this way, small-scale variability represented by stochastic physics can impact large spatiotemporal scales of climate variability.

Despite the historical disconnect between the weather and climate prediction communities, the boundaries between weather and climate prediction are somewhat artificial (Hurrell et al., 2009; Palmer et al., 2008; Shapiro et al., 2010). This disconnect is challenged by recent advances in prediction on time scales from weather to subseasonal-to-seasonal and decadal by operational weather forecasting centers around the world (Moncrieff et al., 2007; Vitart & Robertson, 2012) and the ability of global cloud-resolving models to both forecast the weather and simulate the long-term climate (Crueger et al., 2018; Satoh et al., 2019; Zangl et al., 2015). Nonlinearities in the climate system lead to an upscale transfer of energy (and therefore error) from smaller to larger scales (Lorenz, 1969; Palmer, 2001; Tribbia & Baumhefner, 2004). At the same time, slowly evolving modes of variability can produce predictable signals on shorter time scales (Hoskins, 2013; Vannitsem & Lucarini, 2016). Under the “seamless prediction” paradigm, the weather and climate communities should work together to develop Earth system models (Brunet et al., 2010; Christensen & Berner, 2019), as developments made in one community are expected to benefit the other. The development and use of stochastic parameterizations is a good example of this paradigm at work.

Recent years have seen substantial interest in the development of stochastic parameterization schemes in weather and climate models. Pragmatic approaches, such as the stochastically perturbed parameterization tendencies (SPPT) scheme (Buizza et al., 1999; Palmer et al., 2009), are widely used due to their ease of implementation and beneficial impacts on the model (Christensen et al., 2017; Leutbecher et al., 2017; Sanchez et al., 2016). Other schemes predict the statistics of model uncertainty using a theoretical understanding of the atmospheric processes involved, such as the statistics of convection (Bengtsson et al., 2019; Craig & Cohen, 2006; Khouider et al., 2010; Sakradzija & Klocke, 2018). A third approach is to make use of observations or high-resolution simulations to characterize variability that is unresolved in a low-resolution forecast model (Shutts & Palmer, 2007). This last approach can be used to develop data-driven stochastic schemes (Bessac et al., 2019; Dorrestijn et al., 2015) or to constrain tunable parameters in stochastic parameterizations (Christensen et al., 2015b; Christensen, 2019; Shutts & Pallares, 2014). A drawback of these data-driven approaches is that assumptions are made about the structure of the stochastic parameterization (e.g., the physical process to focus on, or the distribution of the stochastic term conditioned on the resolved state) in order to make the analysis tractable using conventional methods.

Machine learning models offer an approach to parameterize complex nonlinear subgrid processes in a potentially computationally efficient manner from data describing those processes. The family of machine learning models consist of mathematical models whose structure and parameters (often denoted weights)

optimize the predictive performance of a priori unknown relationships between input ("predictor") and output ("predictand") variables. Commonly used machine learning model frameworks range in complexity from simple linear regression to decision trees and neural networks (Hastie et al., 2009). More complex methods allow modeling of broader classes of predictor-predictand relationships but risk "overfitting" to spurious patterns and producing predictions with large variance over small changes to the inputs. Machine learning practitioners minimize the risk of overfitting through the use of regularization techniques that impose constraints on the predictions through the model structure and the optimization loss function. Regularization is critical for more complex machine learning models, so that they can converge to optimal and robust configurations in large parameter spaces. Machine learning for parameterizations has been considered since (Krasnopolsky et al., 2005), and recently, multiple groups have begun developing new parameterizations for a variety of processes (Bolton & Zanna, 2019; Gentine et al., 2018; Rasp et al., 2018; Schneider et al., 2017). The most common regularization in existing machine learning parameterization approaches has been multitask learning (Caruana, 1997), in which the machine learning model predicts multiple correlated values simultaneously and learns to preserve the correlations. However, these schemes have focused exclusively on deterministic parameterization approaches, but the need for stochastic perturbations is being recognized (Brenowitz & Bretherton, 2019).

One active area in current machine learning research is generative modeling, which focuses on models that create synthetic representative samples from distributions of arbitrary complexity. Generative adversarial networks, or GANs (Goodfellow et al., 2014), are a class of generative models that consist of two neural networks in mutual competition. The generator network produces synthetic samples similar to the original training data from a latent vector, and the critic, or discriminator, network examines samples from the generator and the training data in order to determine if a sample is real or synthetic. The discriminator acts as an adaptable loss function for the generator by learning features of the training data and teaching those features to the generator through back-propagation. The original GAN formulation used a latent vector of random numbers as the only input to the generator, but subsequent work on conditional GANs (Mirza & Osindero, 2014) introduced the ability to incorporate a combination of fixed and random inputs into the generator, enabling sampling from conditional distributions. Because the stochastic parameterization problem can be framed as sampling from the distribution of subgrid tendencies conditioned on the resolved state, conditional GANs have the potential to perform well on this task.

The purpose of this study is to evaluate how well GANs can parameterize the subgrid tendency component of an atmospheric model at weather and climate time scales. A key question is whether a GAN can learn uncertainty quantification within the parameterization framework, removing the need to retrospectively develop separate stochastic representations of structural model uncertainty. While the ultimate goal is to test these ideas in a full general circulation model (GCM: left for future work), as a proof of concept, we use the two-time scale model proposed in Lorenz (1996), henceforth the L96 system, as a test bed for assessing the use of GAN in atmospheric models. Simple chaotic dynamical systems such as L96 are useful for testing methods in atmospheric modeling due to their transparency and computational cheapness. The L96 system has been widely used as a test bed in studies including development of stochastic parameterization schemes (Arnold et al., 2013; Crommelin & Vanden-Eijnden, 2008; Chorin & Lu, 2015; Kwasniok, 2012; Wilks, 2005), data assimilation methodology (Fertig et al., 2007; Hatfield et al., 2018; Law et al., 2016), and using ML approaches to learn improved deterministic parameterization schemes (Dueben & Bauer, 2018; Schneider et al., 2017; Watson, 2019).

The evaluation consists of four primary questions. First, given inputs from the "true" L96 model run, how closely does the GAN approximate the true distribution of subgrid tendencies? Second, when an ensemble of L96 models with stochastic GAN parameterizations are integrated forward to medium-range weather prediction time scales, how quickly does the prediction error increase and how well does the ensemble spread capture the error growth? Third, when the L96 model with a stochastic GAN parameterization is integrated out to climate time scales, how well does the GAN simulation approximate the properties of the true climate? Fourth, how closely does the GAN represent both different regimes within the system and the probability of switching between them?

Details of the Lorenz '96 model and the GAN are presented in Section 2, and the results of the weather and climate analyses described above are presented in Section 3. Section 4 presents an overall discussion of

results. Conclusions follow in Section 5. A draft of this paper has been posted on the arXiv website (Gagne et al., 2019).

2. Methods

2.1. Lorenz '96 Model Configuration

The L96 system was designed as a "toy model" of the extratropical atmosphere, with simplified representations of advective nonlinearities and multiscale interactions (Lorenz, 1996). It consists of two scales of variables arranged around a latitude circle. The large-scale, low-frequency X variables are coupled to a larger number of small-scale high-frequency Y variables, with a two-way interaction between the Xs and Ys. It is the interaction between variables of different scales that makes the L96 system ideal for evaluating new ideas in parameterization development. The L96 system has proven useful in assessing new techniques that were later developed for use in GCMs (Crommelin & Vanden-Eijnden, 2008; Dorrestijn et al., 2013).

The X and Y variables evolve following:

$$\frac{\mathrm{d}X_k}{\mathrm{d}t} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j;\ k = 1, \ldots, K \tag{1a}$$

$$\frac{\mathrm{d}Y_j}{\mathrm{d}t} = -cbY_{j+1}(Y_{j+2} - Y_{j-1}) - cY_j + \frac{hc}{b}X_{(\lfloor (j-1)/J \rfloor + 1)};\ j = 1, \ldots, JK, \tag{1b}$$

where in the present study the number of X variables is $K = 8$ and the number of Y variables per X variable is $J = 32$. Further, we set the coupling constant to $h = 1$, the spatial-scale ratio to $b = 10$, and the temporal-scale ratio to $c = 10$. The forcing term $F = 20$ is set large enough to ensure chaotic behavior. The chosen parameter settings, which were used in Arnold et al. (2013), are such that one model time unit (MTU) is approximately equivalent to five atmospheric days, deduced by comparing error-doubling times in L96 and the atmosphere (Lorenz, 1996).

In this study, the full Lorenz '96 equations are treated as the "truth" which must be forecast or simulated. In the case of the atmosphere, the physical equations of motion of the system are known. However, due to limited computational resources, it is not possible to explicitly simulate the smallest scales, which are instead parameterized. Motivated by this requirement for weather and climate prediction, a forecast model for the L96 system is constructed by truncating the model equations and parameterizing the impact of the small Y scales on the resolved X scales:

$$\frac{\mathrm{d}X_k^*}{\mathrm{d}t} = -X_{k-1}^*(X_{k-2}^* - X_{k+1}^*) - X_k^* + F - \hat{\mathrm{U}}(X_k^*, t);\ k = 1, \ldots, K, \tag{2}$$

where $X_k^*(t)$ is the forecast estimate of $X_k(t)$ and $\hat{\mathrm{U}}(X_k^*, t)$ is the parameterized subgrid tendency. The parameterization $\hat{\mathrm{U}}$ approximates the true subgrid tendencies:

$$U(X, Y) = \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j, \tag{3}$$

which can be estimated from realizations of the "truth" time series as

$$U_k(t) = \left[-X_{k-1}(t)\left(X_{k-2}(t) - X_{k+1}(t)\right) - X_k(t) + F\right] - \left(\frac{X_k(t + dt_f) - X_k(t)}{dt_f}\right), \tag{4}$$

following Arnold et al. (2013). The time step dt_f equals the time step used in the forecast model for consistency.

A long "truth" run of the L96 model is performed to generate both training data for the machine learning models and a test period for both weather and climate evaluations. The "truth" run is integrated for 20,000 MTU using a fourth-order Runge-Kutta (RK4) time stepping scheme and a time step $dt = 0.001$ MTU. Output from the first 2,000 MTU are used for training, and the remaining 18,000 MTU are used for testing. A burn-in period of 2 MTU is discarded. All parameterized forecast models of the L96 use a forecast time step of $dt_f = 0.005$ MTU and a second-order Runge-Kutta (RK2) time stepping scheme. This RK2 scheme

is used instead of RK4 to represent the temporal discretization of the equations representing the resolved dynamics in an atmospheric forecasting model.

2.2. GAN Parameterizations

The GAN parameterization developed for the Lorenz '96 model in this study utilizes a conditional dense GAN to predict the subgrid tendency at the current time step given information about the state at the previous time step. We will investigate two classes of predictors of U_t: both X and U at the previous forecast time step and X alone. In the following discussion, we focus on GANs based on the first of these predictor sets; the construction of those associated with the second set is analogous to that of the first. Note that in this section, we move to a discrete time notation, with the time index indicated by the subscript, t, where adjacent indices are separated by the forecast time step, dt_f.

The GAN generator accepts $X_{t-1,k}$, $U_{t-1,k}$, and a latent Gaussian random vector $Z_{t-1,k}$ as input to estimate $\hat{U}_{t,k}$, or the predicted U at time t. The discriminator accepts $X_{t-1,k}$, $U_{t-1,k}$, and $V_{t,k}$ as inputs (where $V_{t,k}$ may be either $U_{t,k}$ if from the training data or $\hat{U}_{t,k}$ if from the generator) and outputs the probability that $V_{t,k}$ comes from the training data. All inputs and outputs are rescaled to have a mean of 0 and standard deviation of 1 based on the training data distributions. Note that we choose to develop local GANs, that is, GANs, which accept X and U for a given spatial index, k, and that predict $\hat{U}$ for that index, k, as opposed to GANs that accept vector X and U and thus include spatial information. This is to match the local nature of parameterization schemes in weather and climate models.

Each GAN we consider consists of the same neural network architecture with variations in the inputs and how noise is scaled and inserted into the network. A diagram of the architecture of the GAN networks is shown in Figure 1. Both the generator and discriminator networks contain two hidden layers with 16 neurons in each layer. The weights of the hidden layers are regularized with a L2, or Ridge, penalty (Hoerl & Kennard, 1970) with scale factor λ of 0.001, which was chosen after evaluating different λ values and selecting the one that minimized the Hellinger distance. Scaled exponential linear unit (SELU) activation functions (Klambauer et al., 2017) follow each hidden layer. SELU is a variation of the common rectified linear unit (ReLU) activation function with a scaled exponential transform for the negative values that helps ensure the output distribution retains a mean of 0 and standard deviation of 1. Larger numbers of neurons per hidden layer were evaluated and did not result in improved performance. Gaussian additive noise layers before each hidden layer and optionally the output layer inject noise into the latent representations of the input data. A batch normalization (Ioffe & Szegedy, 2015) output layer ensures that the output values retain a mean of 0 and and standard deviation of 1, which helps the generator converge to the true distribution faster.

The GAN training procedure iteratively updates the discriminator and generator networks until the networks reach an equilibrium in which the discriminator struggles to distinguish true from generated samples. The inputs, outputs, and connections between networks are shown in Figure 1. A batch B, or subset of samples drawn randomly without replacement from the training data, of truth run output is split in half. One subset is fed through the generator G and then into the discriminator D, and the other is sent directly to the discriminator. The discriminator weights are then updated based on the following loss function L_d,

$$L_d = \mathbb{E}_B[\log(D(X_{t-1,b}, U_{t-1,b}, U_{t,b})] + \mathbb{E}_B[\log(1 - D(G(X_{t-1,b}, U_{t-1,b}, Z_{t-1,b}))]. \quad (5)$$

$\mathbb{E}_b$ is the expected value over a single batch of data. Another batch of samples are drawn and sent through the generator and then the discriminator with frozen weights. The generator loss L_g is calculated as

$$L_g = \mathbb{E}_B[\log(D(G(X_{t-1,b}, U_{t-1,b}, Z_{t-1,b}))] \quad (6)$$

based on labeling the generated samples as originating from the truth run. This reversal forces the discriminator to teach the generator features that would worsen the discriminator's own performance.

Gaussian noise $Z_{t-1,k}$ is injected into the neural network at each iteration through both the input and hidden layers. We consider hidden layer Gaussian noise scaled to standard deviations σ_g of different orders of magnitude (Table 1) in order to evaluate how the magnitude of the noise affects the forecast spread and the representation of the model climate. In forecast mode, we test providing both white, or uncorrelated noise,

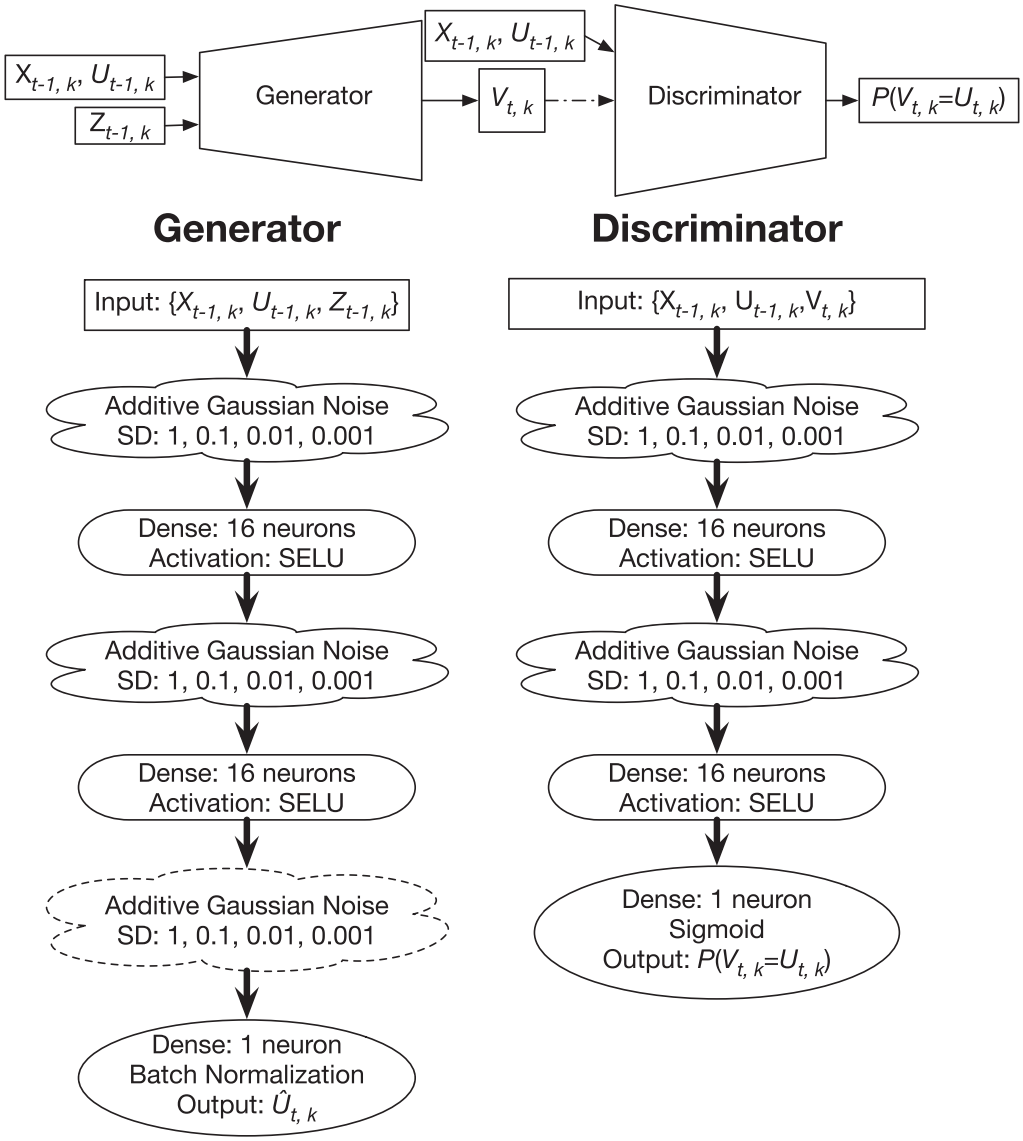

Figure 1. (Top) A diagram of how the GAN networks are connected for training. (Bottom) A diagram of the GAN network architectures used for the stochastic parameterization.

and red, or correlated noise $Z_{r,t,k}$ to the GAN. The red noise is generated using an AR(1) process with a lag-1 temporal autocorrelation ϕ_g equal to the lag-1 autocorrelation of the deterministic residuals of the GAN,

$$\phi_g = \frac{\mathbb{E}[(U_t - \hat{U}_t^d)(U_{t-1} - \hat{U}_{t-1}^d)]}{\sigma_{\hat{U}}^2}; \sigma_g = (1 - \phi_g^2)^{1/2} \tag{7a}$$

$$\epsilon_g \sim \mathcal{N}(0, \sigma_g) \tag{7b}$$

$$Z_{r,t,k} = \phi_g Z_{t-1,k} + \epsilon_g \tag{7c}$$

The color of the noise is not relevant during the training process because the GAN only uses inputs from the previous time step like parameterizations in full-complexity weather and climate models. Both white- and red-noise representations are trained in the same way. The noise values are kept constant through the inte-

Table 1
Summary of the GAN Configurations Tested

Short name	Input variables	Noise magnitude	Noise correlation	Output layer noise?
XU-lrg-w	$X_{t-1,k}$, $U_{t-1,k}$	1	white	yes
XU-med-w	$X_{t-1,k}$, $U_{t-1,k}$	0.1	white	yes
XU-sml-w	$X_{t-1,k}$, $U_{t-1,k}$	0.01	white	yes
XU-tny-w	$X_{t-1,k}$, $U_{t-1,k}$	0.001	white	yes
X-med-w	$X_{t-1,k}$	0.1	white	yes
X-sml-w	$X_{t-1,k}$	0.01	white	yes
X-tny-w	$X_{t-1,k}$	0.001	white	yes
XU-lrg-r	$X_{t-1,k}$, $U_{t-1,k}$	1	red	yes
XU-med-r	$X_{t-1,k}$, $U_{t-1,k}$	0.1	red	yes
XU-sml-r	$X_{t-1,k}$, $U_{t-1,k}$	0.01	red	yes
XU-tny-r	$X_{t-1,k}$, $U_{t-1,k}$	0.001	red	yes
X-med-w	$X_{t-1,k}$	0.1	red	yes
X-sml-r	$X_{t-1,k}$	0.01	red	yes
X-tny-r	$X_{t-1,k}$	0.001	red	yes
XU-lrg-w*	$X_{t-1,k}$, $U_{t-1,k}$	1	white	no
XU-med-w*	$X_{t-1,k}$, $U_{t-1,k}$	0.1	white	no
XU-sml-w*	$X_{t-1,k}$, $U_{t-1,k}$	0.01	white	no
XU-tny-w*	$X_{t-1,k}$, $U_{t-1,k}$	0.001	white	no
X-sml-w*	$X_{t-1,k}$	0.01	white	no
X-tny-w*	$X_{t-1,k}$	0.001	white	no

gration of a single time step. The difference between the white- and red-noise representations only manifests when they are incorporated as parameterizations in the full model (Equation (1)).

The GANs are all trained with a consistent set of optimization parameters. The GANs are updated through stochastic gradient descent with a batch size (number of examples randomly drawn without replacement from the training data) of 1,024 and a learning rate of 0.0001 with the Adam optimizer (Kingma & Ba, 2015). The GANs are trained on 639,936 samples and are validated on 28,797,096 samples from different portions of the truth run. The GANs are trained for 30 epochs, or passes through the training data. The model weights are saved for analysis every epoch for the first 20 epochs and then every 2 epochs between epochs 20 and 30. The GANs are developed with the Keras v2.2 machine learning API coupled with Tensorflow v1.13.

The GAN configurations considered in this study are summarized in Table 1. A short name of the format "predictors–noise magnitude–noise correlation" is introduced to simplify identification of different GANs. For example, "XU-med-r" refers to the GAN that takes X and U as predictors and uses medium (med) magnitude red (r) noise. While most GANs tested include the optional additive noise layer before the output layer, the sensitivity to this choice was also considered. GANs that do not include this final noise layer follow the naming convention above, but are indicated by an asterisk.

2.3. Polynomial Regression Parameterization

The GAN stochastic parameterization is evaluated against a cubic polynomial regression parameterization, $\hat{U}_{t,k}$, similar to the model used in Arnold et al. (2013).

$$\begin{aligned} \hat{U}_{t,k} &= U^d_{t,k} + \epsilon_{t,k} \\ U^d_{t,k} &= aX^3_{t-1,k} + bX^2_{t-1,k} + cX_{t-1,k} + d \end{aligned} \tag{8}$$

The parameters $[a, b, c, d]$ are determined by a least squares fit to the (X, U) data from the L96 "truth" training run. It is known that the simple cubic polynomial deterministic parameterization $U^d_{t,k}$ is a poor forecast model for the L96 system (Arnold et al., 2013; Christensen et al., 2015a; Wilks, 2005), as X does not uniquely determine U. The variability in the (X, U) relationship is accounted for using a temporally correlated additive

noise term:

$$\epsilon_{t,k} = \phi\epsilon_{t-1,k} + \sigma_\epsilon(1-\phi^2)^{1/2}z_{t,k}, \tag{9}$$

where $z \sim \mathcal{N}(0,1)$, the first-order autoregressive parameters (ϕ, σ_ϵ) are fit from the residual time series $r_t = U_t - U_t^d$, and the ϵ_k processes are independent for different X variables.

The polynomial parameterization has been specifically designed to represent the impact of the Y variables in this version of the L96 model, just as traditional parameterization schemes are designed to represent a given process in a GCM. Previous studies have demonstrated that the polynomial parameterization with additive noise performs very well (Arnold et al., 2013; Christensen et al., 2015a; Wilks, 2005). Although the multiplicative noise polynomial parameterization does perform slightly better in Arnold et al. (2013), we compare the GAN with the additive noise polynomial to ensure consistency in noise inclusion process. This "bespoke" parameterization is therefore a stringent benchmark against which to test GAN parameterizations.

3. Results

3.1. Metrics

The accuracy of ensemble weather forecasts can be summarized by evaluating the root mean square error (RMSE) of the ensemble mean. The lower the RMSE, the more accurate the forecast. The RMSE at lead time τ is defined as

$$RMSE(\tau) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(X_o(t) - X_m(t_{\text{init}}\quad,\tau)\right)^2}, \tag{10}$$

where N is the number of forecast-observation pairs, $X_o(t)$ is the observed state at time t, and $X_m(t_{\text{init}}, \tau)$ is the ensemble mean forecast at lead time τ, initialized at $t_{\text{init}}\quad$, such that $t = t_{\text{init}} + \tau$.

If an ensemble forecast correctly accounts for all sources of uncertainty such that the forecast of the spread of the ensemble and measured probabilities of an event are statistically consistent, the forecast is said to be *reliable* (Wilks, 2011). In this study, we assess the reliability of the ensemble using the spread-error relationship (Leutbecher & Palmer, 2008; Leutbecher, 2010). This states that, for an unbiased and reliable forecasting system, the root mean square error in the ensemble mean is related to the average ensemble variance:

$$\frac{M}{M-1}\,\overline{\text{estimate ensemble variance}} = \frac{M}{M+1}\,\overline{\text{mean square error}}, \tag{11}$$

where M is the number of ensemble members and the variance and mean error have been estimated by averaging over many forecast-verification pairs. For the large ensemble size used in this study, $M = 40$, we can consider the correction factor to be close to 1. A skillful probabilistic forecast will have as small an RMSE as possible, while also demonstrating a statistical match between RMSE and average ensemble spread.

The simplest definition of the "climate" of the L96 system is the probability density function (PDF) of the individual $X_{t,k}$ values. The climatological skill can therefore be summarized by quantifying the difference between the true and forecast PDF. The Hellinger distance H is calculated for each forecast model:

$$H(p,q) = \frac{1}{2}\int\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x, \tag{12}$$

where $p(x)$ is the PDF of forecast $X_{t,k}$ values and $q(x)$ is the PDF of truth $X_{t,k}$ values (Pollard, 2002). The smaller H, the closer the forecast model pdf is to the truth pdf. We also considered the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), motivated by information theory, but found it provided no additional information over the Hellinger distance, so results for the KL are not shown for brevity.

Evidence that the L96 model displays distinct dynamical regimes of behavior for the parameter set considered was presented in Christensen et al. (2015a), in which regime affiliations were determined using a metric based on the temporally-localized spatial covariance. Christensen et al. (2015a) found that during the more common regime (regime frequency ~80%), the eight X variables exhibit wave-2 like behavior around the ring, while in the rarer regime, the X variables exhibit wave-1 type behavior. Another approach to characterizing regime structure that makes use of both the instantaneous state and the recent past of the system

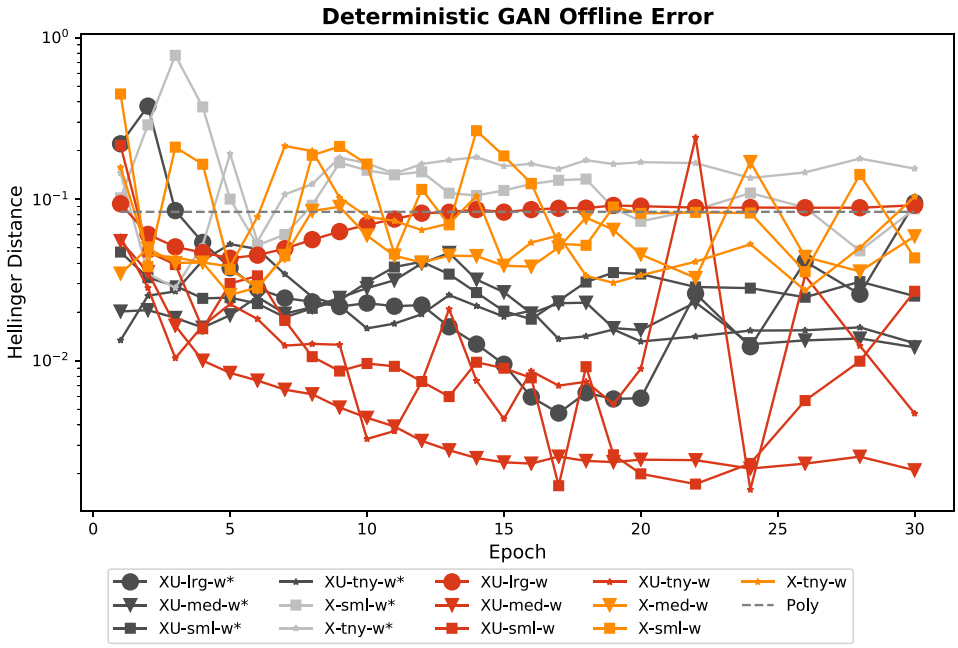

Figure 2. Off-line assessment of GAN performance. Hellinger distances between the GAN subgrid tendency distributions given input X and U values from the truth run and the truth run subgrid forcing distribution as a function of training epoch.

is hidden Markov model (HMM) analysis (Franzke et al., 2008; Monahan et al., 2015; Rabiner, 1989). In an HMM analysis, it is assumed that underlying the observed state variables is an unobserved Markov chain taking discrete values. The HMM algorithm provides maximum likelihood estimates of the probability distributions of the state variables conditioned on the instantaneous hidden state values, the stochastic matrix Q of transition probabilities for each time step, and an optimal estimate of the hidden state sequence.

3.2. Off-line Assessment of GAN Performance

The GAN parameterizations are first evaluated on how closely their output subgrid forcing distributions match those of the truth run when the GANs are supplied with input X and U values from the truth run. This is summarized by the Hellinger distance in Figure 2. Most of the GANs show a trend of decreasing Hellinger distance for the first few epochs followed by mostly stable oscillations. GANs with both $X_{t-1,k}$ and $U_{t-1,k}$ as input tend to perform better in the off-line analysis than those with only $X_{t-1,k}$. Larger input noise standard deviations seem to reduce the amount of fluctuation in the Hellinger distance between epochs, but there does not appear to be a consistent correlation with noise standard deviation and Hellinger distance. Note that the weights as fitted at the end of epoch 30 are used in the forecast networks, regardless of whether the GAN at this epoch shows the minimum off-line Hellinger distance, because we wanted to give each network a consistent amount of training time.

3.3. GAN Simulation of Subgrid-Scale Tendency Distribution

The joint distributions of X_{t-1} and U_t from the different model runs reveal how the noise standard deviation affects the model climate (Figure 3). Larger noise standard deviations increase the range of X values appearing in the run but do not appear to change the range of U values output by the GAN. The X-only GANs did the best job in capturing the shape of the truth distribution although some, such as X-sml-r, have a translational bias in their joint distribution that the Hellinger distance penalizes. While the XU-w* and X-w* GANs capture the bulk of the distribution well, there are spurious points outside the bounds of the truth distribution for all of these models. The XU-*-r GANs produce either an overly diffuse but unbiased distribution or fail to capture the true distribution entirely. The polynomial model captures the conditional mean and shape of the distribution very well but produced an overly smooth representation of the truth and did not capture the bifurcation in the positive X_{t-1} and positive U_t quadrant.

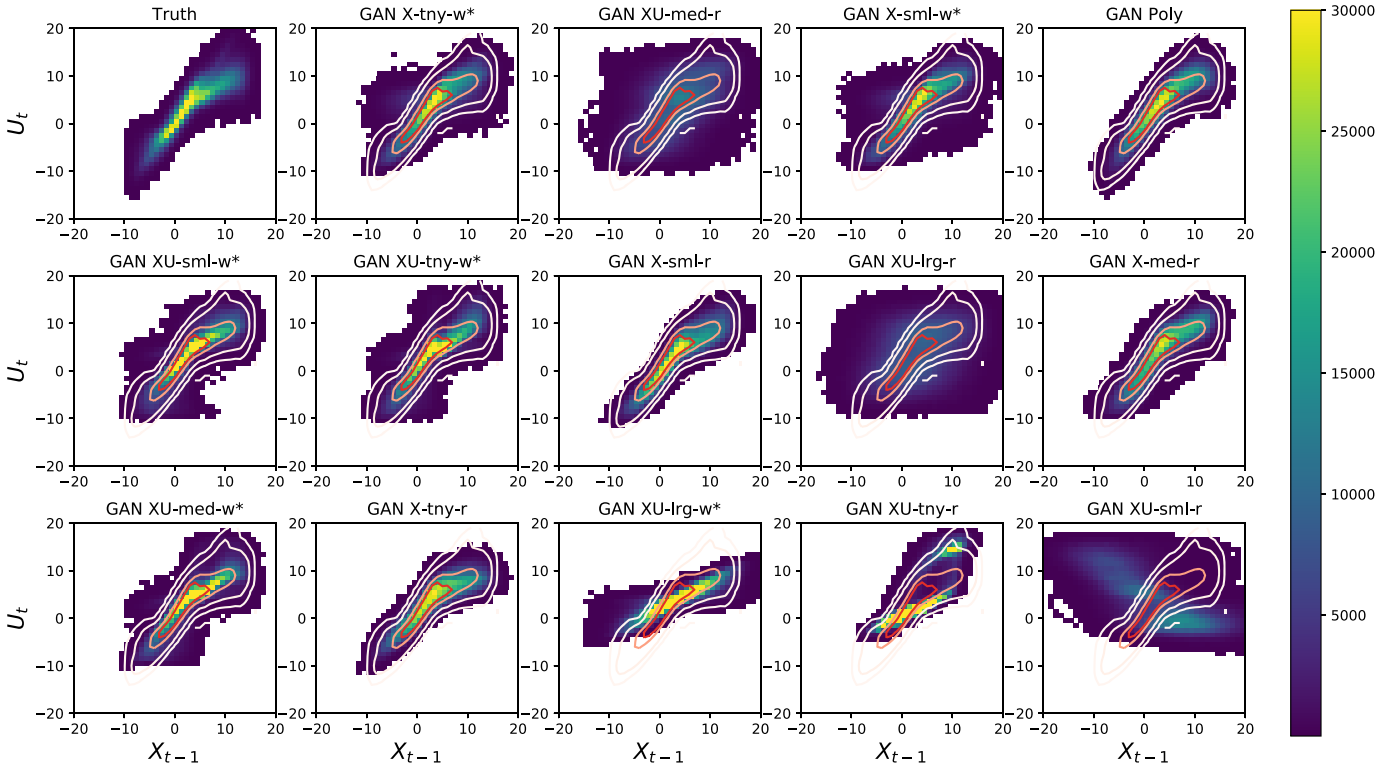

Figure 3. Joint distributions (2-D histograms) of X_{t-1} and U_t for each GAN configuration. The truth joint distribution is overlaid in red contours on each forecast model distribution. The distributions are ordered from left to right descending in terms of their relative total marginal Hellinger distances.

3.4. Weather Evaluation

The parameterized models for the Lorenz '96 system are evaluated in a weather forecast framework. An extensive set of reforecast experiments were produced for 751 initial conditions selected from the attractor. An ensemble of 40 forecasts was produced from each initial condition (i.e., no initial condition perturbations are used). Different random seeds are used for each ensemble member to generate the stochastic perturbations used in the GAN or polynomial parameterizations.

Figure 4 shows the RMSE and spread for all weather experiments at 1 MTU. A reduction in RMSE indicates an ensemble forecast that more closely tracks the observations. A good match between RMSE and ensemble spread indicates a reliable forecast. The best performing GANs in terms of RMSE are X-tny-r, X-tny-w, and X-tny-w*. All of these models performed slightly better than the polynomial regression, which was competitive with most GANs in terms of both RMSE and the ratio of RMSE to spread. The spread of the white noise GANs is generally underdispersive. Most of the red noise GANs, on the other hand, are somewhat overdispersive with X-med-r having the spread/error ratio closest to 1. Red noise perturbs the model in a similar direction for a longer period, which results in greater ensemble spread compared with white noise. Figure 5 shows the RMSE and ensemble spread for a subset of the ensemble forecasts of the X variables performed using the GAN parameterizations or the bespoke polynomial parameterization. X-sml-w* demonstrates both low RMSE and similar spread to RMSE throughout the forecast period. X-sml-r and X-sml-w feature similar RMSE through 1 MTU, but X-sml-r has smaller RMSE after that point and a better spread-to-error ratio throughout the period. The XU GANs have higher RMSEs than their X counterparts because the XU models may have overfit to the strong correlation between $U_{t-1,k}$ and $U_{t,k}$ in the training data. Inspection of the input weights revealed that XU GANs generally weigh $U_{t-1,k}$ more highly than $X_{t-1,k}$. At maxima and minima in the waves, the XU models may be biased toward extending the current growth forward, which can be a source of error in forecast runs.

3.5. Climate Evaluation

The GAN parameterizations are also tested on their ability to characterize the climate of the Lorenz '96 system. First, the ability to reproduce the pdf of the X variables is evaluated. Each forecast model and the full

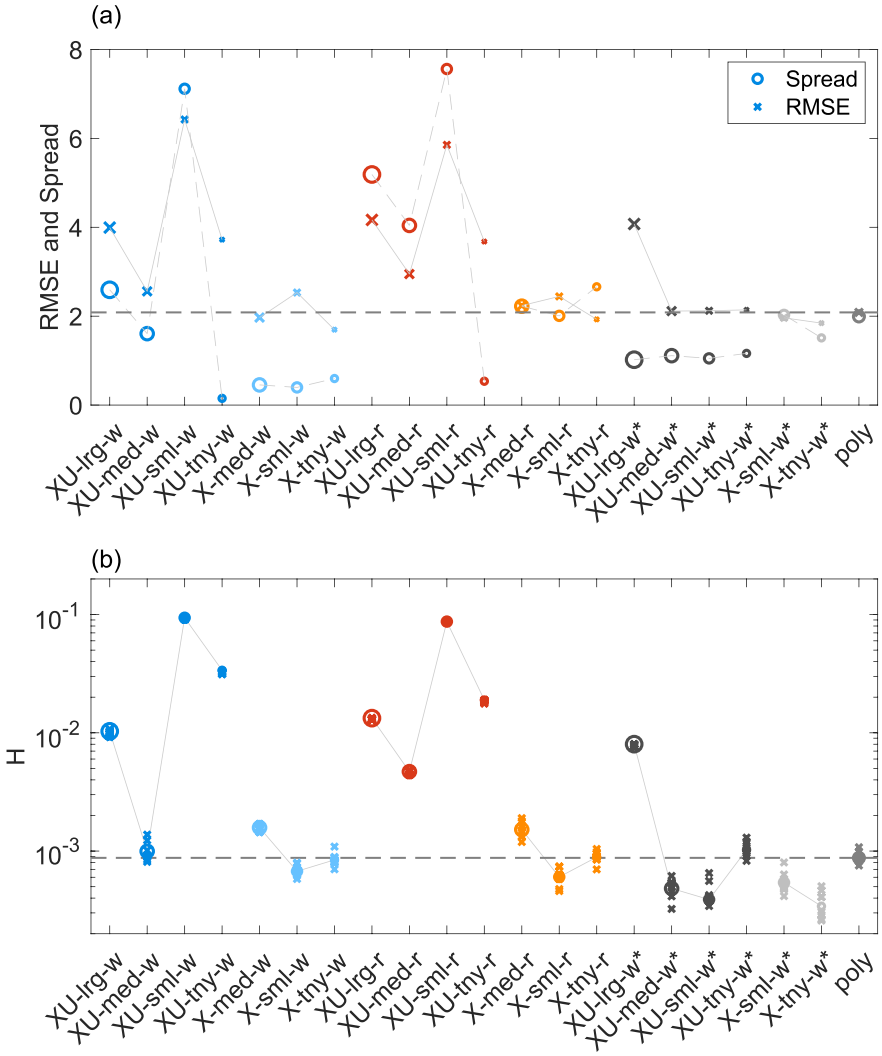

Figure 4. Summary of performance of different parameterized models (x-axis). (a) Weather forecast performance. Ensemble spread (circles) and RMSE (crosses) for experiments with white and red noise in GANs at time step 201. The horizontal dashed line indicates the RMSE for the polynomial forecast model. Ideally, a forecast model will produce forecasts with small RMSE while maintaining the match between spread and RMSE. (b) Climate performance. The Hellinger distance between each forecast pdf and the "true" pdf. The metric in question is calculated for the best estimate of the climatological pdf, averaging across all X variables (circles), as well as for each X variable in turn, for example, comparing true and forecast X_1 pdfs, etc (crosses). The latter gives an indication of the sampling uncertainty. The horizontal dashed line indicates the mean value of H for the polynomial model.

L96 system were used to produce a long simulation of length 10,000 MTU. Figure 6a shows kernel density estimates of the marginal pdfs of $X_{t,k}$ from the full L96 system and from a sample of parameterized models. The pdf of the true L96 system is markedly non-Gaussian, with enhanced density forming a "hump" at around $X = 8$. Compared to the true distribution, the XU-sml-w and XU-sml-r models both perform poorly, producing very similar pdfs with too large a standard deviation and that are too symmetric. However, the other displayed models skillfully reproduce the true pdf. Figure 6b shows the difference between each forecast pdf and the true pdf. Several GANs perform as well if not better than the benchmark bespoke polynomial parameterization.

Figure 4b shows the Hellinger distance H evaluated for each parameterized model. The filled circles indicate the value of the metric when the pdf is evaluated across all X variables in both parameterized and truth time series. The crosses give an indication of sampling variability and indicate the metrics comparing pairs of X variables, that is, $X_{t,j}$ and $X_{t,k}$ for $j \neq k$. Quantifying the parameterized model performance in this

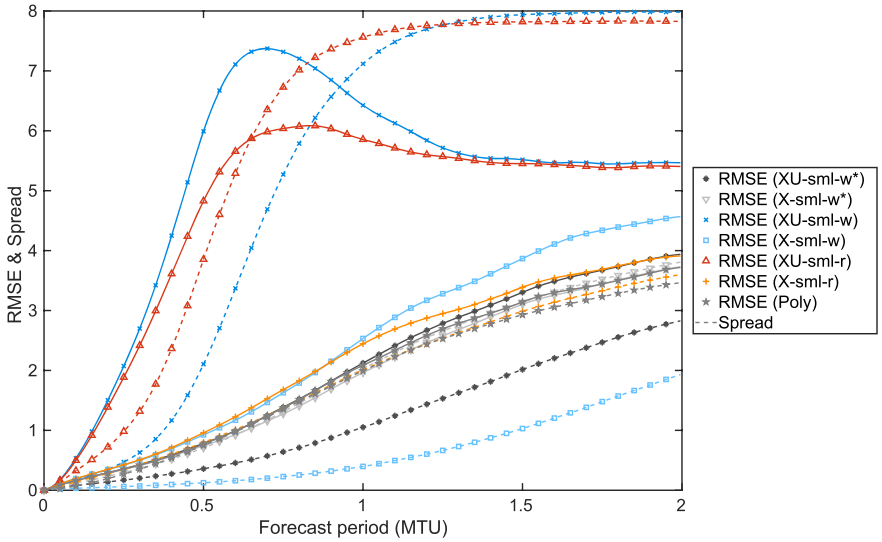

Figure 5. RMSE (lines with dots) and ensemble spread (dashed lines) for a subset of experiments with white and red noise in GANs. Note that 400 forecast time steps corresponds to 2 MTU, or 10 "atmospheric days."

way allows for easy ranking of the different models. While the AR(1) stochastic polynomial parameterized forecast model is very skillful (Arnold et al., 2013), several GANs outperform the polynomial model.

In addition to correctly capturing the distribution of the X variables, it is desirable that a parameterized model will capture the spatiotemporal behavior of the system. This is assessed by considering the spatial and temporal correlations in both the true system and parameterized models. The diagnostic is shown for a subset of the tested parameterized models in Figure 7. It is evident that the parameterized models that skillfully capture the pdf of X also skillfully represent the spatiotemporal characteristics of the system. The X-sml-w* scheme performs particularly well, improving over the stochastic polynomial approach and other GANs by having a temporal correlation pattern in sync with the truth but with lower magnitude peaks.

Following the regime results presented in Christensen et al. (2015a), we use HMM analysis to classify into two clusters the instantaneous states in the four-dimensional space spanned by the norms of the projections of X on wavenumbers 1 through 4 (denoted $Z_j, j = 1, \ldots, 4$). Because of the spatial homogeneity of the $X_{t,k}$ statistics, these wavenumber projections correspond to the empirical orthogonal functions.

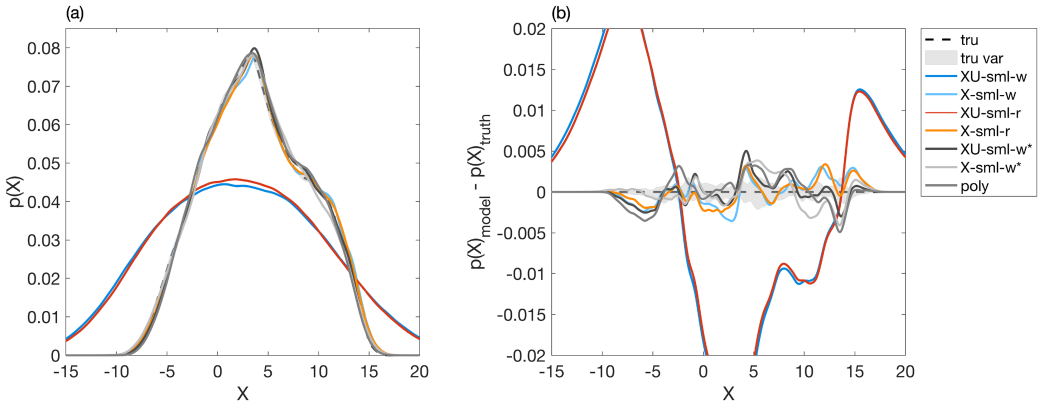

Figure 6. The skill of the forecast models at reproducing the climate of the Lorenz '96 system defined as the pdf of the X variables. (a) The pdf of the Lorenz '96 system (black dashed) compared to a subset of the forecast models. (b) The difference between the forecast and true pdfs shown in (a). Sampling variability is indicated by shading the range in metrics for each of the 8 X variables of the full Lorenz '96 system.

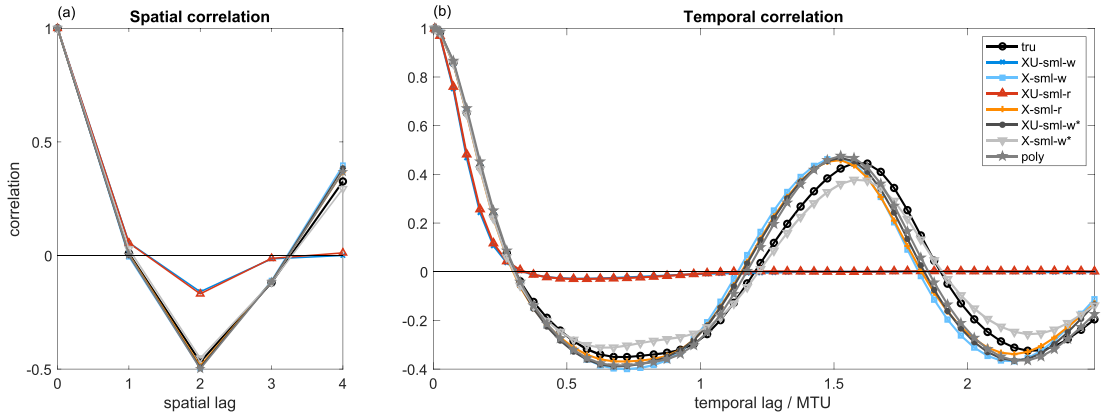

Figure 7. The skill of the parameterized models (colors) at reproducing the "true" (a) spatial correlation and (b) temporal correlation of the X variables in the Lorenz '96 system (black), calculated from the climatological simulation. The sampling variability in these metrics, as indicated by the variability between the metric calculated for different X variables, is narrower than the plotted line width.

Kernel density estimates of the joint pdfs of the projections of X on wavenumbers 1 and 2 are presented in Figure 8, along with estimates of the joint distributions conditioned on the HMM regime sequence. The conditional distributions have been scaled by the probability of each state so that the full joint pdf is the sum of the conditional pdfs. For reference, the unconditional joint pdf for truth is shown in gray contours in each panel. The stochastic matrix Q shows the probability of remaining in each regime (diagonal values) or transitioning from one regime to the other (off-diagonal values). As in Figures 6 and 7, only a subset of results are displayed.

The clear separation of the truth simulation into two distinct regimes is modeled well by the polynomial parameterization and most of the GAN parameterizations. With the exception of XU-sml-w and XU-sml-r, the regime spatial structures and stochastic matrices are captured well. The GANs are slightly more likely to transition between regimes than the truth run, while the polynomial run is slightly more likely to stay in the same regime. Consistent with the other climate performance results presented above, the joint distribution of $|Z_i|$ produced by XU-sml-w and XU-sml-r are strongly biased, with no evidence of a meaningful separation into two distinct regimes of behavior.

To quantify the forecast model skill at reproducing the L96 behavior in wavenumber space, Figure 9 shows H calculated over the marginal pdfs of $|Z_1|$ and $|Z_2|$ as upward and downward triangles, respectively. Several

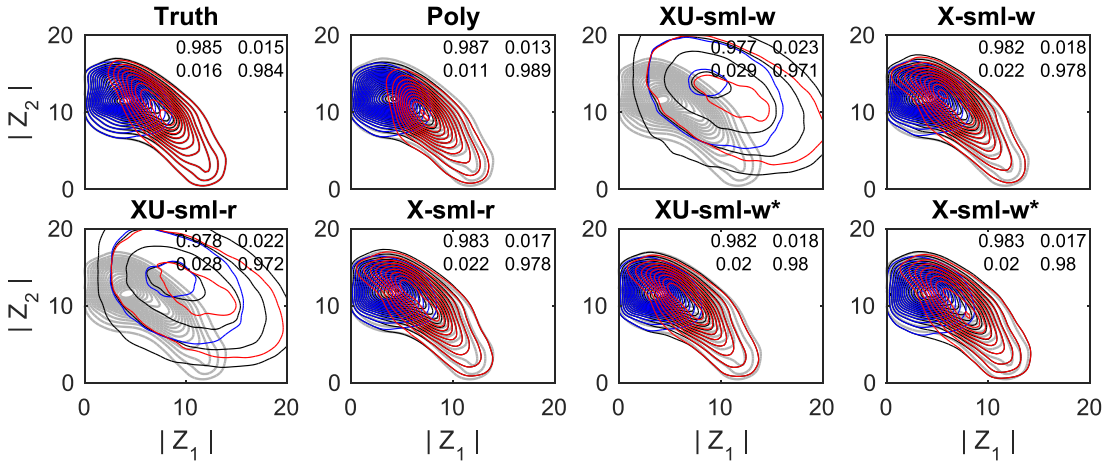

Figure 8. Joint distributions of the projections of X on wavenumbers 1 and 2 ($|Z_1|$ and $|Z_2|$, respectively) conditioned on HMM regime occupation (red and blue contours) for the truth simulation subset of parameterized simulations. In all but the upper left panel, thin black lines display the unconditional joint distribution. In all panels, the grey curves denote the full joint distribution (without regime occupation conditioning) from the truth simulation. The conditional distributions have been scaled by the relative probabilities of each state. Inset: HMM stochastic matrix Q.

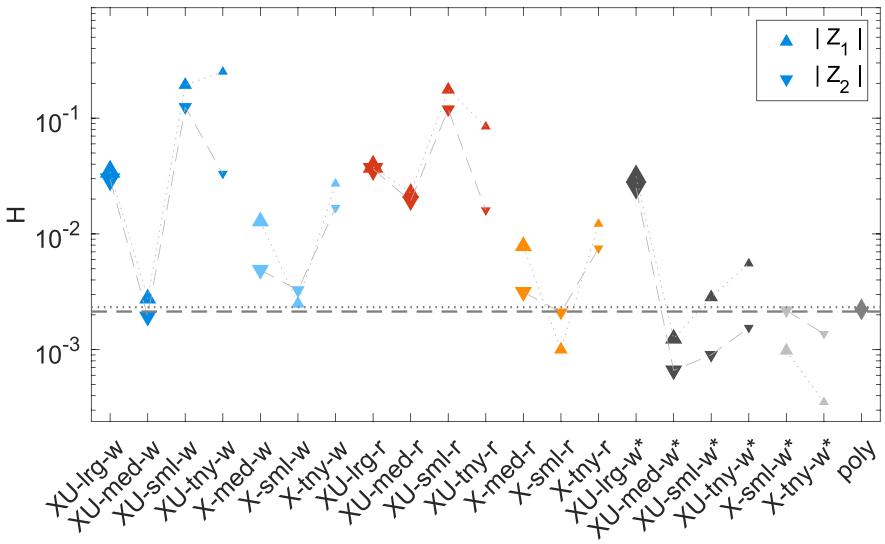

Figure 9. The Hellinger distance between each forecast pdf and the "true" pdf, considering the projection of the X variables onto (upward triangles) wavenumber 1 and (upward triangles) wavenumber 2.

of the GANs evaluated are competitive with, or improve upon, the polynomial parameterization scheme. The best performing GANs also performed the best in terms of other climate metrics (e.g., see Figure 4).

We note that in particular, models which accurately capture the regime behavior will also show good correlation statistics when averaged over a long time series. The regime analysis can help diagnose why a model shows poor correlation statistics. For example, X-sml-w* accurately captures the marginal distributions of the two regimes, but the frequency of occurrence of the red regime, dominated by wavenumber-1 behavior, is slightly too high (at 54% compared to 51% in the "truth" run). This reduces the magnitude of the negative correlation at a lag of 0.75 MTU and the positive correlation at a lag of 1.5 MTU observed in Figure 11.

3.6. Wavelet Analysis

To further investigate why the white noise and red noise GANs differ in online performance, a wavelet analysis is performed on time series of outputs from the climate runs. A continuous wavelet transform with the Ricker wavelet decomposes the time series into contributions from different periods. We choose a continuous wavelet transform to have more control over the spectral sampling. The total energy E for a given period is represented as

$$E = \frac{1}{T}\sum_{t=1}^{T} w_t^2 \tag{13}$$

where w is the wavelet magnitude at a given time step t. The total power for each period from each model is shown in Figure 10. All sml GANs except the XU-sml-r and XU-sml-w follow the truth power curve closely. The polynomial regression follows the truth closely although it tends to underestimate the power slightly for each period. The GANs peak in power at the same period when the temporal correlation in Figure 7 crosses 0. The GANs with poor Hellinger distances also contained more energy for longer periods.

A clearer evaluation of the wavelet differences can be found by calculating the mean absolute percentage difference from the truth run at different wavelengths. The absolute difference between the truth and parameterized runs increases with increasing wavelengths, so the percentage difference is employed to control for this trend:

$$MAPD = \frac{1}{T}\sum_{t=1}^{T} \frac{|E_{g,t} - E_{u,t}|}{E_{u,t}}. \tag{14}$$

The MAPD scores with wavelength in Figure 10 shows that none of the GANs consistently perform better at all periods, but some do provide closer matches to the truth spectrum for the longer periods. In the peak

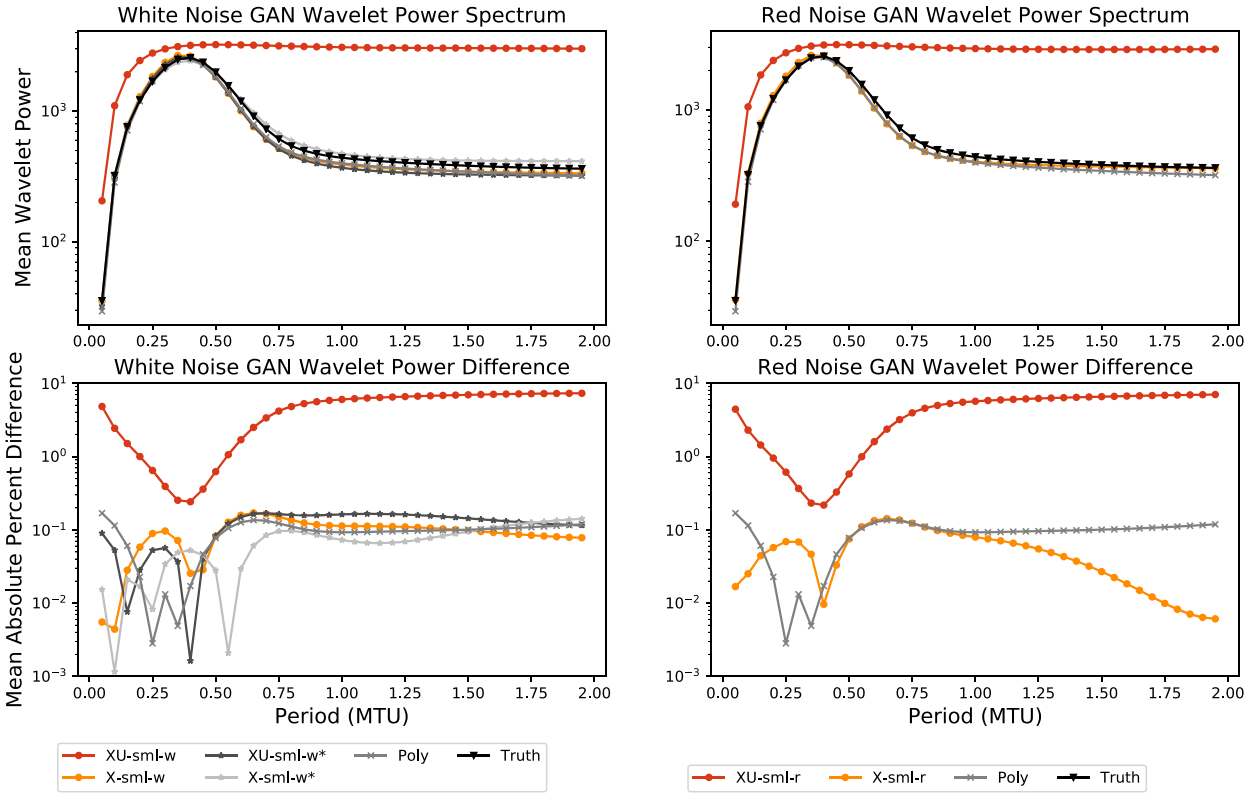

Figure 10. (Top) Wavelet power spectra for the white and red noise GAN climate runs as well as the polynomial and truth runs. (Bottom) Mean absolute percent differences between the truth wavelet power spectrum and each model.

energy period, the different GANs have minimum error for slightly different periods before increasing in error again. The X-sml-r GAN uniquely shows decreasing MAPD with increasing period, while the white noise GANs generally have similar differences across the range of evaluated periods.

4. Discussion

In this study, we choose to focus on GANs for stochastic parameterization primarily because the framework offers a way to embed stochasticity directly into the model and training process instead of adding it a posteriori to a deterministic parameterization. The GAN's use of the discriminator as an adaptive loss function is particularly attractive for weather and climate applications because it reduces the need for developing a handcrafted loss and can be scaled to higher-dimensional and more complex outputs, including spatial fields.

Several of the GANs tested show a weather and climate skill that is competitive with a bespoke polynomial parameterization scheme. For climatological skill, several different metrics were considered including the distribution of the X variable, spatiotemporal correlation statistics, and regime behavior. We found that forecast models that performed well according to one online metric also performed consistently well for all online metrics. The good performance of the GAN is encouraging, demonstrating that GANs can indeed be used as explicit stochastic parameterizations of uncertain subgrid processes directly from data. Furthermore, a small number of GANs improve upon the bespoke polynomial approach, indicating the potential for such machine-learned approaches to improve on our current hand-designed parameterization schemes, given suitable training data. While this has been proposed and demonstrated for deterministic parameterizations (Rasp et al., 2018; Schneider et al., 2017), this is a first demonstration for the case of an explicitly stochastic parameterization.

Comparison of Figures 4a and 4b indicates a relationship between forecast models that perform well on weather and climate time scales. To quantify this further, Figure 11 shows the correlation between weather

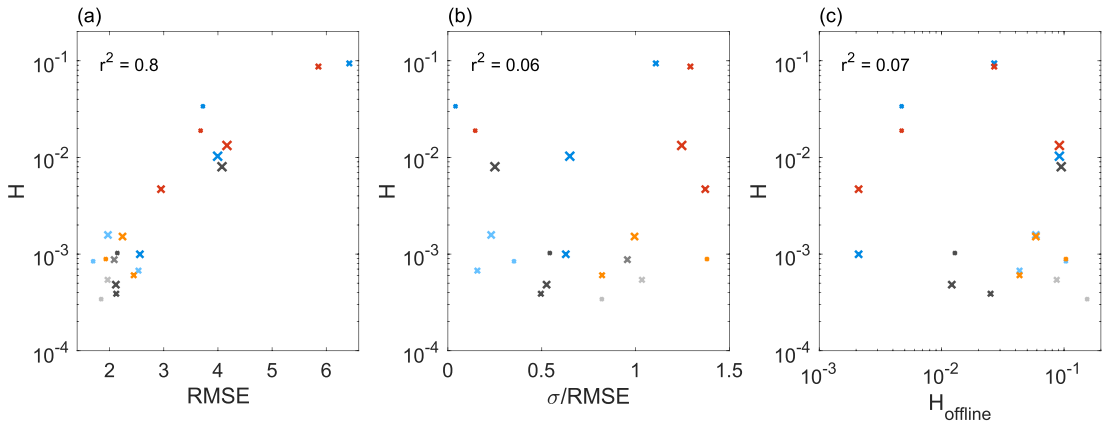

Figure 11. Correlation between weather forecast skill and climate performance. (a) Weather forecast RMSE versus climate Hellinger distance. (b) Weather forecast spread-error ratio versus climate Hellinger distance. (c) Off-line versus online Hellinger distance. Colors indicate forecast model, as in Figure 4, and marker sizes correspond to the noise standard deviation.

forecast skill and climate performance for each model considered. Models which produce weather forecasts with a lower RMSE also show good statistics on climate time scales. In contrast, the *reliability* of weather forecasts, that is, the statistical match between spread and error, is a poor predictor of climate performance. This is reflected by the competitive performance of the white noise GAN for producing a realistic climate, whereas on weather time scales, red noise increases the spread and can thereby substantially improve the forecast reliability (e.g., consider the X-med, X-sml, and X-tny GAN for white noise and red noise, respectively; Figure 4). This relationship between initialized and climatological performance has been discussed in the context of global models (Williams et al., 2013). It provides further evidence that parameterizations can first be tested in weather forecasts before being used in climate models, as promoted by the "seamless prediction" framework (Brunet et al., 2010).

The disparity between the off-line verification statistics and those from the climate and weather runs (Figure 4c) highlights the challenges in training GANs for parameterization tasks. Neither the values of the generator loss or the off-line evaluation of the GAN samples correlated with their performance in the forecast model integrations. The generator and discriminator networks optimize until they reach an equilibrium, but there is no guarantee that the equilibrium is stable or optimal. Some of the differences in the results may be due to particular GANs converging on a poor equilibrium state as opposed to other factors being tested. GANs and other machine learning parameterization models are trained under the assumption that the data are independent and identically distributed, but in practice are applied to spatially and temporally correlated fields sequentially, potentially introducing nonlinear feedback effects. GANs are more complex to train than other machine learning methods because they require two neural networks and do not output an informative loss function. Larger magnitude additive noise appears to help prevent runaway feedbacks from model biases at the expense of increasing weather prediction errors. The inclusion of the batch normalization output layer appeared to assist both training and prediction by limiting the possible extremes reached during integration.

Other generative neural network frameworks should also be investigated to determine if they can provide similar or better performance with a less sensitive training process. One stochastic machine learning approach is Monte Carlo dropout (Gal & Ghahramani, 2016), in which dropout layers are activated during inference mode to produce an ensemble of stochastic samples for the same input. Monte Carlo dropout can be applied to any neural network architecture and only requires tuning the dropout rate but may produce artifacts in the predicted distribution, especially with large dropout rates. Conditional variational autoencoders (Kingma & Welling, 2014), a Bayesian generative model that regularizes the latent space as a vector of Gaussian distributions, are another potentially viable approach. Variational autoencoders have a more constrained latent space than GANs but often produce overly smooth samples and have their own training challenges.

Wavelet analysis helped uncover differences in model performance across different time scales. While the other evaluation metrics focused on distributional or error metrics in the time domain, the wavelet power

spectrum separated the time series into different periods and enabled comparisons of the energy embedded in different scales. In particular, the wavelet analysis revealed that some of the GANs added energy to the system either at long periods or across all periods in some cases.

The standard deviation of the noise does impact both the training of the GANs and the resulting weather and climate model runs. Using too large a standard deviation limits the ability of the GAN to discover the structure of the true distribution of the data. Standard deviations that are too small may result in either the generator or discriminator becoming overly good at their tasks during training, which results in the GAN equilibrium being broken. During simulations, noise standard deviations that are too small can result in the system becoming trapped within one regime and never escaping. We tested optimizing the noise standard deviation during the training process but did not achieve consistent convergence to the same noise value in repeated experiments.

In addition to the GAN configurations evaluated in the paper, other GAN settings were tested and were found to have similar or worse performance. Given the relative simplicity of the Lorenz '96 system, adding neurons in each hidden layer did not improve performance. Using the SELU activation function generally resulted in faster equilibrium convergence than the ReLU. Varying the scaling factor for the L2 regularization on each hidden layer did affect model performance. Using a larger value greatly reduced the variance of the predictions, but using a smaller value resulted in peaks of the final distributions being too far apart, especially when both $X_{t-1,k}$ and $U_{t-1,k}$ were used as inputs. We also tested deriving U from a 1-D convolutional GAN that reproduced the set of Y values associated with each X. That approach did produce somewhat realistic Y values but contained "checkerboard" artifacts from the convolutions and upscaling, especially at the edges. The sum of the Y values was also not equal to U derived from Equation (9).

The L96 system is commonly used as a test bed for new ideas in parameterization, and ideas tested using the system can be readily developed further for use in higher complexity Earth system models. However, the L96 system has many fewer dimensions than an Earth system model and a relatively simple target distribution. The relative simplicity of the L96 system may have also led to the more complex GAN overfitting to the data compared with the simpler polynomial parameterization. For more complex, higher dimensional systems, the extra representational capacity of the GAN may provide more benefit than can be realized in L96. The computational simplicity of L96 also allows for the production of extremely large training data sets with little compute resources. Producing higher complexity Earth system model datasets requires trading off among spatial resolution, temporal output frequency, and temporal coverage based on the amount of computational and storage resources available.

5. Conclusions

In this study, we have developed an explicitly stochastic GAN framework for the parameterization of sub-grid processes in the Lorenz '96 dynamical system. After testing a wide range of GANs with different noise characteristics, we identified a subset of models that outperform a benchmark bespoke cubic polynomial parameterization. Returning to the questions posed in Section 1, we found that this subset of GANs approximates well the joint distribution of the resolved state and the subgrid-scale tendency. This model subset also produces the most accurate weather forecasts (i.e., lowest RMSE in the ensemble mean). Some GANs with red noise produce reliable weather forecasts (Figure 4), in which the ensemble spread is a good indicator of the error in the ensemble mean. Based on the comparison of the same white and red noise GANs, the correlation structure of the noise is most critical for producing reliable forecasts. However, these are not necessarily the GANs that produce the most accurate forecasts. The subset of models with the most accurate weather forecasts produce the most accurate climate simulations, as characterized by probability distributions, space and time correlations, and regime dynamics. However, we note that the GANs which produce skillful weather and climate forecasts were different to those which performed well in "off-line" mode (Figure 11).

Although the GANs required an iterative development process to maximize model performance and were very sensitive to the noise magnitude and other hyperparameter settings, they do show promise as an approach for stochastic parameterization of physical processes in more complex weather and climate models. Applying other recently developed GAN losses and regularizers (Kurach et al., 2019) could help reduce the chance for the GAN to experience a failure mode.

The experiments presented here demonstrate that GANs are a promising machine learning approach for developing stochastic parameterization in complex GCMs. Key lessons learned and unanswered questions include the following:

- While including the tendency from the previous time step provides a natural approach for building temporal dependence into the parameterization, it can lead to accumulation of error in the forecast, such that local-in-time parameterization should also be considered.
- Autocorrelated noise is important for a skillful weather forecast but appears less important for capturing the climatological distribution.
- It is possible that spatial correlations are also important in a higher complexity Earth system model, which could not be assessed here due to the simplicity of the Lorenz '96 system.
- It is possible that the noise characteristics could also be learned by the GAN framework to automate the tuning of the stochasticity.

Future work will use these lessons to guide further investigations in the Lorenz '96 framework and to develop machine-learned stochastic parameterization schemes for use in higher complexity Earth system models. Noise can also be incorporated into the truth run training data to evaluate the effect of simulated observational error on training performance. Other noise inclusion processes, such as multiplicative noise, should also be evaluated in GANs to see if they result in further improvements over the current additive noise approach. GANs of a similar level of complexity to those used for L96 could emulate local effects, such as some warm rain formation processes.

Acknowledgments
This research started in a working group supported by the Statistical and Applied Mathematical Sciences Institute (SAMSI). D. J. G. and H. M. C. were funded by National Center for Atmospheric Research Advanced Study Program Postdoctoral Fellowships and by the National Science Foundation through NCAR's Cooperative Agreement AGS-1852977. H. M. C. was funded by Natural Environment Research Council grant number NE/P018238/1. A. H. M. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and thanks SAMSI for hosting him in the autumn of 2017. The NCAR Cheyenne supercomputer was used to generate the model runs analyzed in this paper. Sue Ellen Haupt and Joseph Tribbia provided helpful reviews of the paper prior to submission. The source code and model configuration files necessary for replicating the results of this paper can be accessed at https://doi.org/10.5281/zenodo.3663121.

References

Ajayamohan, R. S., Khouider, B., & Majda, A. J. (2013). Realistic initiation and dynamics of the Madden-Julian oscillation in a coarse resolution aquaplanet GCM. *Geophysical Research Letters*, *40*, 6252–6257. https://doi.org/10.1002/2013GL058187

Arnold, H. M., Moroz, I. M., & Palmer, T. N. (2013). Stochastic parameterizations and model uncertainty in the Lorenz '96 system. *Philosophical Transactions of the Royal Society A*, *371*, 20110479. https://doi.org/10.1098/rsta.2011.0479

Bengtsson, L., Bao, J., Pegion, P., Penland, C., Michelson, S., & Whitaker, J. (2019). A model framework for stochastic representation of uncertainties associated with physical processes in NOAA's Next Generation Global Prediction System (NGGPS). *Monthly Weather Review*, *147*, 893–911. https://doi.org/10.1175/MWR-D-18-0238.1

Berner, J., Achatz, U., Batte, L., De La Cámara, A., Christensen, H., Colangeli, M., & Yano, Y. I. (2017). Stochastic parameterization: Towards a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*, 565–588.

Berner, J., Jung, T., & Palmer, T. N. (2012). Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *Journal of Climate*, *25*(14), 4946–4962.

Berner, J., Shutts, G. J., Leutbecher, M., & Palmer, T. N. (2010). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, *66*(3), 603–626.

Bessac, J., Monahan, A. H., Christensen, H. M., & Weitzel, N. (2019). Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. *Monthly Weather Review*, *147*, 1447–1469.

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*, 376–399. https://doi.org/10.1029/2018MS001472

Brenowitz, N., & Bretherton, C. (2019). Spatially extended tests of a neural network parameterization trained by coarse-graining. arXiv preprint, https://arxiv.org/abs/1904.03327

Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G. N., & Shukla, J. (2010). Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, *91*(10), 1397–1406. https://doi.org/10.1175/2010BAMS3013.1

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*, 2887–2908. https://doi.org/10.1002/qj.49712556006

Buizza, R., & Palmer, T. N. (1995). The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Sciences*, *52*(9), 1434–1456.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75. https://doi.org/10.1023/A:1007379606734

Chorin, A. J., & Lu, F. (2015). Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 9804–9809. https://doi.org/10.1073/pnas.1512080112

Christensen, H. M. (2019). Constraining stochastic parametrisation schemes using high-resolution simulations. arXiv Preprint, https://arxiv.org/abs/1904.04503

Christensen, H. M., & Berner, J. (2019). From reliable weather forecasts to skilful climate response: A dynamical systems approach. *Quarterly Journal of the Royal Meteorological Society*, *145*, 1052–1069. https://doi.org/10.1002/qj.3476

Christensen, H. M., Berner, J., Coleman, D., & Palmer, T. N. (2017). Stochastic parametrisation and the El Niño-Southern oscillation. *Journal of Climate*, *30*(1), 17–38.

Christensen, H., Moroz, I., & Palmer, T. (2015a). Simulating weather regimes: Impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, *44*(7-8), 2195–2214.

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015b). Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *Journal of the Atmospheric Sciences*, *72*(6), 2525–2544.

Craig, G. C., & Cohen, B. G. (2006). Fluctuations in an equilibrium convective ensemble. Part I: Theoretical formulation. *Journal of the Atmospheric Sciences*, *63*(8), 1996–2004.

Crommelin, D., & Vanden-Eijnden, E. (2008). Subgrid-scale parametrisation with conditional Markov chains. *Journal of the Atmospheric Sciences*, *65*(8), 2661–2675.

Crueger, T., Giorgetta, M. A., Brokopf, R., Esch, M., Fiedler, S., Hohenegger, C., & Stevens, B. (2018). ICON-A, The atmosphere component of the ICON Earth system model: II. Model evaluation. *Journal of Advances in Modeling Earth Systems*, *10*, 1638–1662. https://doi.org/10.1029/2017MS001233

Davini, P., von Hardenberg, J., Corti, S., Christensen, H. M., Juricke, S., Subramanian, A., & Palmer, T. N. (2017). Climate SPHINX: Evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model. *Geoscientific Model Development*, *10*(3), 1383–1402.

Dawson, A., & Palmer, T. N. (2015). Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dynamics*, *44*, 2177–2193. https://doi.org/10.1007/s00382-014-2238-x

Dorrestijn, J., Crommelin, D. T., Biello, J. A., & Böing, S. J. (2013). A data-driven multi-cloud model for stochastic parametrization of deep convection. *Philosophical Transactions of the Royal Society A*, *371*, 1991.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., Jonker, H. J. J., & Jakob, C. (2015). Stochastic parameterization of convective area fractions with a multicloud model inferred from observational data. *Journal of the Atmospheric Sciences*, *72*, 854–869.

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*, 3999–4009. https://doi.org/10.5194/gmd-2018-148

Fertig, E. J., Harlim, J., & Hunt, B. R. (2007). A comparative study of 4D-VAR and a 4D ensemble Kalman filter: Perfect model simulations with Lorenz-96. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *59*(1), 96–100.

Franzke, C., Crommelin, D., Fischer, A., & Majda, A. J. (2008). A hidden Markov model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, *21*, 1740–1757.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2019). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz '96 model. arXiv Preprint, https://arxiv.org/abs/1909.04711

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). New York, New York, USA: PMLR. http://proceedings.mlr.press/v48/gal16.html

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*, 5742–5751. https://doi.org/10.1029/2018GL078202

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial networks. Advances in neural information processing systems 27 (pp. 2672–2680), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. https://web.stanford.edu/hastie/ElemStatLearn/

Hatfield, S., Subramanian, A., Palmer, T., & Düben, P. (2018). Improving weather forecast skill through reduced precision data assimilation. *Monthly Weather Review*, *146*, 49–62. https://doi.org/10.1175/MWR-D-17-0132.1

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Techometrics*, *12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoskins, B. (2013). The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, *139*(672), 573–584. https://doi.org/10.1002/qj.1991

Hurrell, J., Meehl, G. A., Bader, D., Delworth, T. L., Kirtman, B., & Wielicki, B. (2009). A unified modeling approach to climate system prediction. *Bulletin of the American Meteorological Society*, *90*, 1819–1832. https://doi.org/10.1175/2009BAMS2752.1

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv Preprint, https://arxiv.org/abs/1502.03167

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., & Raynaud, L. (2010). *Ensemble of data assimilations at ECMWF*. Shinfield park, Reading: European Centre for Medium-Range Weather Forecasts. 636.

Juricke, S., & Jung, T. (2014). Influence of stochastic sea ice parametrization on climate and the role of atmosphere–sea ice–ocean interaction. *Philosophical Transactions of the Royal Society A*, *372*(2018). https://doi.org/10.1098/rsta.2013.0283

Khouider, B., Biello, J., & Majda, A. J. (2010). A stochastic multicloud model for tropical convection. *Communications in Mathematical Sciences*, *8*(1), 187–216.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, San Diego, CA. https://arxiv.org/abs/1412.6980

Kingma, D., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the international conference on learning representations 2014*. https://arxiv.org/abs/1312.6114

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. Advances in Neural Information Processing Systems (NIPS), https://arxiv.org/abs/1706.02515

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370–1383. https://doi.org/10.1175/MWR2923.1

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Kurach, K., Lučić, M., Zhai, X., Michalski, M., & Gelly, S. (2019). A large-scale study on regularization and normalization in GANs. In *International Conference on Machine Learning*. https://arxiv.org/abs/1807.04720

Kwasniok, F. (2012). Data-based stochastic subgrid-scale parametrisation: An approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society A*, *370*(1962), 1061–1086.

Law, K. J., Sanz-Alonso, D., Shukla, A., & Stuart, A. M. (2016). Filter accuracy for the Lorenz 96 model: Fixed versus adaptive observation operators. *Physica D: Nonlinear Phenomena*, *325*, 1–13.

Leith, C. E. (1975). Numerical weather prediction. *Reviews of Geophysics*, *13*(3), 681.

Leutbecher, M. (2010). Diagnosis of ensemble forecasting systems, *Seminar on Diagnosis of Forecasting and Data Assimilation Systems, 7 - 10 September 2009* (pp. 235–266). Shinfield Park, Reading: ECMWF.

Leutbecher, M., Lock, S. J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., & Weisheimer, A. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, *143*(707), 2315–2339.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*, 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307.

Lorenz, E. N. (1996). Predictability—A problem partly solved, *Proceedings, Seminar on Predictability, 4-8 September 1995* (Vol. 1, pp. 1–18). Shinfield Park, Reading: ECMWF.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint, https://arxiv.org/abs/1411.1784

Monahan, A. H., Rees, T., He, Y., & McFarlane, N. (2015). Multiple regimes of wind, stratification, and turbulence in the stable boundary layer. *Journal of the Atmospheric Sciences*, *72*, 3178–3198.

Moncrieff, M. W., Shapiro, M. A., Slingo, J. M., & Molteni, F. (2007). Collaborative research at the intersection of weather and climate. *WMO Bulletin*, *56*, 204–211.

Palmer, T. N. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, *127*(572), 279–304.

Palmer, T. N. (2012). Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *138*(665), 841–861.

Palmer, T. N., Barkmeijer, J., Buizza, R., & Petroliagis, T. (1997). The ECMWF ensemble prediction system. *Meteorological Applications*, *4*(04), 301–304.

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., & Weisheimer, A. (2009). Stochastic parametrization and model uncertainty.

Palmer, T. N., Doblas-Reyes, F. J., & Weisheimer, A. (2009). Toward seamless prediction: Calibration of climate change projections using seasonal forecasts reply.

Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., & Rodwell, M. J. (2008). Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society*, *89*(4), 459–470.

Pollard, D. (2002). A user's guide to measure theoretic probability.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286.

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.1810286115

Reyes, F. D., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J. M., & Palmer, T. N. (2009). Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *135*, 1538–1559.

Sakradzija, M., & Klocke, D. (2018). Physically constrained stochastic shallow convection in realistic kilometer-scale simulations. *Journal of Advances in Modeling Earth Systems*, *10*, 2755–2776. https://doi.org/10.1029/2018MS001358

Sanchez, C., Williams, K. D., & Collins, M. (2016). Improved stochastic physics schemes for global weather and climate models. *Quarterly Journal of the Royal Meteorological Society*, *142*, 147–159. https://doi.org/10.1002/qj.2640

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S. J., Putman, W. M., & Düben, P. (2019). Global cloud-resolving models. *Current Climate Change Reports*, *5*, 172–184. https://doi.org/10.1007/s40641-019-00131-0

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), 12,396–12,417. https://doi.org/10.1002/2017GL076101

Seiffert, R., & von Storch, J. S. (2010). A stochastic analysis of the impact of small-scale fluctuations on the tropospheric temperature response to CO_2 doubling. *Journal of Climate*, *23*, 2307–2319. https://doi.org/10.1175/2009JCLI3043.1

Shapiro, M., Shukla, J., Brunet, G., Nobre, C., Béland, M., Dole, R., & Wallace, J. M. (2010). An Earth-system prediction initiative for the twenty-first century. *Bulletin of the American Meteorological Society*, *91*(10), 1377–1388. https://doi.org/10.1175/2010BAMS2944.1

Shutts, G. J., & Pallares, A. C. (2014). Assessing parametrization uncertainty associated with horizontal resolution in numerical weather prediction models. *Philosophical Transactions of the Royal Society A*, *372*(2018), 20130284.

Shutts, G. J., & Palmer, T. N. (2007). Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *Journal of Climate*, *20*(2), 187–202.

Stockdale, T. N., Anderson, D. L. T., Balmaseda, M. A., Doblas-Reyes, F., Ferranti, L., Mogensen, K., & Vitart, F. (2011). ECMWF seasonal forecast system 3 and its prediction of sea surface temperature. *Climate Dynamics*, *37*(3-4), 455–471.

Strømmen, K., Christensen, H. M., Berner, J., & Palmer, T. N. (2018). The impact of stochastic parametrisations on the representation of the Asian summer monsoon. *Journal of Climate*, *50*(5–6), 2269–2282.

Sušelj, K., Hogan, T. F., & Teixeira, J. (2014). Implementation of a stochastic eddy-diffusivity/mass-flux parameterization into the Navy Global Environmental Model. *Weather and forecasting*, *29*(6), 1374–1390.

Suselj, K., Teixeira, J., & Chung, D. (2013). A unified model for moist convective boundary layers based on a stochastic eddy-diffusivity/mass-flux parameterization. *Journal of Atmospheric Sciences*, *70*(7), 1929–1953.

Teixeira, J., & Reynolds, C. A. (2010). Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Monthly Weather Review*, *136*(2), 483–496.

Tribbia, J. J., & Baumhefner, D. P. (2004). Scale interactions and atmospheric predictability: An updated perspective. *Monthly Weather Review*, *132*(3), 703–713.

Vannitsem, S. S., & Lucarini, V. (2016). Statistical and dynamical properties of covariant Lyapunov vectors in a coupled atmosphere-ocean model—Multiscale effects, geometric degeneracy, and error dynamics. *Journal of Physics A: Mathematical and Theoretical*, *49*(22), 224001.

Vitart, F., & Robertson, A. W. (2012). Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *World Meteorological Organisation Bulletin*, *61*(2), 23.

Wang, Y., Zhang, G. J., & Craig, G. C. (2016). Stochastic convective parameterization improving the simulation of tropical precipitation variability in the NCAR CAM5. *Geophysical Research Letters*, *43*, 6612–6619. https://doi.org/10.1002/2016GL069818

Watson, P. A. G. (2019). Applying machine learning to improve simulations of dynamical systems using empirical error correction. *Journal of Advances in Modeling Earth Systems*, *11*, 1402–1417. https://doi.org/10.1029/2018MS001597

Weisheimer, A., Corti, S., & Palmer, T. (2014). Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. *Philosophical Transactions of the Royal Society A*, *372*, 20130290.

Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, *131*(606), 389–407.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd). San Diego, CA: Academic Press.

Williams, P. D. (2012). Climatic impacts of stochastic fluctuations in air-sea fluxes. *Geophysical Research Letters*, *39*, L10705. https://doi.org/10.1029/2012GL051813

Williams, K. D., Bodas-Salcedo, A., Déqué, M., Fermepin, S., Medeiros, B., Watanabe, M., & Williamson, D. L. (2013). The transpose-AMIP II experiment and its application to the understanding of southern ocean cloud biases in climate models. *Journal of Climate*, *26*(10), 3258–3274. https://doi.org/10.1175/JCLI-D-12-00429.1

Zangl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2015). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 563–579. https://doi.org/10.1002/qj.2378