On Metric Analysis for Deep Weather Generators

Maysa M. G. Macedo[1], Daniela Szwarcman[1], Jorge Guevara Diaz[1], Dario Augusto Borges Oliveira[1], Bianca Zadrozny[1]

IBM Research, Sao Paulo, Brazil [1]

Abstract

The definition of metrics for evaluating reconstruction image data from machine learning generative methods are well established for applications involving natural images. However, machine learning models applied to weather field precipitation data in the context of weather generators are still not sufficiently addressed. In this work, we discuss the use of various metrics for weather data generation and we propose the use of the Fréchet Inception Distance metric based on weights from a weather dataset.

Introduction

Synthetic data generation using Generative Adversarial Networks (GAN) (Goodfellow et al. 2014) and Variational autoencoders (VAE) (Kingma and Welling 2014) has gained a lot of attention in recent years, particularly for generating natural images (Karras et al. 2018; Brock, Donahue, and Simonyan 2019). Although there are many works that discuss the evaluation of these models, most of them are focused on the context of natural images. Theis, van den Oord, and Bethge (2016) alerted the scientific community about the issue of evaluating synthetic data according to their application and that a good performance with respect to one criterion might not necessarily imply good performance with respect to another criteria. Then, subsequent works on generative models came to expose the complex question of how to select metrics depending on the applications (Borji 2019; Abdella and Uysal 2020; Ding, Wang, and Zhao 2019; Mukherjee, Praveen, and Madumbu 2018). Borji (2019), for example, discusses pros and cons of GAN evaluation strategies, describing multiple qualitative and quantitative metrics for image synthesis.

Besides evaluation, another important reason to study metrics is to create better alternative training losses for different applications. Ding, Wang, and Zhao (2019) propose a new loss more suitable for generating trajectories in the context of autonomous cars, while Abdella and Uysal (2020) use the structural similarity index (SSIM) as part of a loss for a VAE in order to obtain better results for the well-know datasets MNIST and Fashion-MNIST. Mukherjee, Praveen, and Madumbu (2018) present a deep neural network to improve the visual quality of images captured under adverse weather conditions, like rain and fog. They use a perceptual loss to train the model and evaluate their results with SSIM and peak signal-to-noise ratio (PSNR) metrics.

Synthetic data has played an important role in the context of extreme weather impact studies. Impact models require long series of weather data representing different climate scenarios (Verdin et al. 2018), which are usually not available. Deep generative models have also been applied in the context of weather data generation: the authors in (Bhatia, Jain, and Hooi 2020) use a GAN to create precipitation scenarios, and Klemmer et al. (2021) propose a GAN to generate spatiotemporal patterns conditioned on extreme events. Oliveira et al. (2021), on the other hand, use a VAE to generate precipitation scenarios, claiming that it is possible to control the VAE generation without the need of auxiliary variables. However, the authors provide only a comparison of the distributions of generated samples, without a quantitative analysis on the samples' quality.

It is important to notice that the requirements that define a proper sample in the context of weather might be different than those of natural images. In this case, metrics widely used in computer vision tasks, such as the Frechet Inception Distance (FID) (Heusel et al. 2017), might not be suitable for assessing the quality of synthetic weather data. Ullrich et al. (2021) adopt some metrics to compare two statistical models that generate precipitation data. The authors emphasize the importance of maintaining some characteristics such as daily mean precipitation, daily extreme precipitation (99.9th percentile), wet-to-dry transition probabilities and wet day frequencies.

In this work, we propose an investigation of several metrics in the context of precipitation data synthesis with deep generative models. We use the VAE presented in (Oliveira et al. 2021) as the base model in our experiments and analyze the behavior of several metrics under different training conditions. We focus on the reconstruction quality of the VAE, that is, we compute the metrics comparing precipitation samples with their respective reconstructions. Additionally, we compare the FID metric using an Inception model trained on a weather dataset and another on the Imagenet dataset. The precipitation experiments cover seven regions of the world, addressing a great diversity of climate scenarios.

Method

The proposed methodology consists of analyzing the quality of precipitation samples generated with a VAE trained with different dataset sizes. More specifically, we investigate the reconstruction quality of the VAE when precipitation samples from the dataset are used as inputs. For this analysis, several metrics are used and we also propose an FID metric based on weather data.

We performed five different controlled transformations on the data to understand the response of each metric: blurring, shifting, additive bias, Gaussian noise, and swirl transform.

The metrics we evaluate are:

- **Statistical:** mean precipitation, variogram, Kling Gupta Efficiency (KGE), Mean Squared Error (MSE), Mean Absolute Error (MAE)
- **Connectivity:** Connectivity and Two-point probability function (Renard and Allard 2013; Torquato, Beasley, and Chiew 1988; Torquato and Haslach Jr 2002)
- **Geometrical:** fractal dimension (Bouda, Caplan, and Saiers 2016), Structural Similarity (Wang et al. 2004)
- **Spectral:** 95% of the maximum energy in a spectrum
- **Signal processing:** Peak signal-to-noise ratio (PSNR)
- **Distances:** Fréchet Inception Distance Climate dataset (FID) climate, Fréchet Inception Distance Imagenet dataset (FID) (Heusel et al. 2017; Fréchet 1906; Alt and Godau 1992)

Variational Autoencoder

VAEs (Kingma and Welling 2014) are an encoder decoder generative model that enables stochastic synthesis by regularizing the latent space to a known distribution. VAEs parameterize a posterior distribution $q(z|x)$ of discrete latent random variables z given the input data x, a prior distribution $p(z)$, and a decoder with a distribution $p(x|z)$ over input data.

The standard VAE loss has two components (Kingma and Welling 2014): a reconstruction term, accounting for sample quality, and a regularization term, that encourages the latent space to follow a known distribution.

Table 1 shows the architecture of the VAE used in our experiments. It is a 2D-structure that learn the latitude and longitude relation.

Frechet Inception Distance for climate data

Heusel et al. (2017) introduced the Frechet Inception Distance (FID) to evaluate the results of deep generative models such as GANs. The generated samples are fed into an Inception model (trained on ImageNet) and the output features are used to compute the distance metric. The FID considers a comparison between two ensembles of samples and it is suitable for natural images. More recently, Dai and Wipf (2019) proposed to compare samples from VAEs and GANs using the FID measure also based on ImageNet.

We understand that to use the FID metric in the context of weather data generation, the extracted features from the Inception network may not be adequate. One reason is that the data does not have three channels like color images. Also, the weather data do not have a reasonable representation in the Imagenet database. Thus, to calculate the FID we retrain, with 50 epochs, the Inception network with weights acquired from a 12-class classification task, where each class is a weather variable of the ERA5 and CHIRPS datasets. Details about the classes are in the Section . For this classification task we used 5000 samples of training/validation from January/2000 to December/2009 in 50 epochs, while the test was performed with 1000 samples from January/2015 to December/2019.

Table 1: Architecture of the 2-D VAE networks. In the *Processing* columns, we indicate the kernel size and the number of channels for the convolutional layers, and the number of units in the dense layers. The convolutional layers (*conv*) and the transposed convolutional layers (*convt*) have stride 2

Encoder		**Decoder**	
Input x (dim=64×64)		Input z (dim=30)	
Layer	Processing	Layer	Processing
conv 1-1	3×3, 128	dense	256
conv 1-2	3×3, 128	reshape	8×8
conv 1-3	3×3, 128	convt 2-1	3×3, 128
dense	500	convt 2-2	3×3, 128
dense-μ_x	30	convt 2-3	3×3, 128
dense-σ_x	30	convt 2-4	3×3, 1
sampling	-	-	

Experiments

Both experiments, the data reduction and the transformations involve 20 different samples for each task, where most metrics are related to pairwise comparisons, in which case their average values are calculated. In the case of FID metrics, the calculation is performed involving the 20 input samples and 20 simulated ones.

Data

We used daily precipitation from the CHIRPS dataset (Funk, Peterson, and Landsfeld 2015) to create our training and test datasets. CHIRPS is a quasi-global rainfall dataset with 0.05° of resolution, ranging from 1981 to near-present. For our experiments, we used samples from seven different places with a size of 3.2°×3.2° from January 1981 to December 2019. The data is around Belem-Brazil, Sydney-Australia, Parana-Brazil, Dhaka-Bangladesh, Alabama-United States, Barcelona-Spain, and Accra-Ghana.

To train the Inception network for our proposed FID based on weather data, we used ERA5 (Di Napoli et al. 2021) and CHIRPS. The ERA5 dataset has global coverage in the resolution of 0.25° with hourly information, but here, we used its daily averages. We selected 11 variables from ERA5 (along with the precipitation from CHIRPS) to create a classification dataset where each variable defines a class. The ERA5 variables are:

- *Surface variables*: 2 metre temperature, Volumetric soil water layer 1,10 metre U wind component, 10 metre V wind component, Mean sea level pressure, Sea surface temperature

- *Pressure level variables*: Geopotential at 500mb, Geopotential at 150mb, Temperature at 10mb, Component of wind U at 200mb, Component of wind V at 200mb.

Gradual data transformations

We consider the same test data used in the VAE experiments for the transformations. For the blurring transform, we vary sigma values from 0 to 9. For shifting transform, we apply a shift of pixels ranging from 0 to 9 lines on the test samples. In the Gaussian noise case, we gradually increase sigma from 0.001 to 0.01, and for the swirl transform, we use strengths from 0 to 9. Finally, for additive bias, we increased the intensity values of the samples from 0.0% to 2.19%.

VAE gradual reduction data training

Given the total number of samples for the training data, we created five smaller datasets, with reductions of 25%, 50%, 75%, 90%, and 95%, respectively. We train the VAE with each dataset and compare the results of the trained models. The idea is to capture the degradation of the reconstruction quality caused by the decreasing number of training samples. As we randomly selected examples to compose the smaller datasets, we performed this five times to analyze the variability of the results. We trained a different VAE for each region considered.

Results

The target metrics of this study were calculated for all transformations and VAE generation varying the number of training samples. Figure 1 depicts the five VAE reconstructions for each reduction of data and Figure 2 shows the metrics computed for these reconstructions for all 7 places including the proposed climate FID metric.

Figure 3 shows all transformations and the respective level of distortion and Figure 4 depicts the computed metrics to twenty samples against the twenty simulated one along the increasing of transformation level intensity.

The results of the reconstruction of weather scenarios using VAE, depicted in Figure 1 show a level of blurring that increases with the gradual reduction of samples in the training phase. When we gather the results of all the regions covered, presented in Figure 2, it is noticeable that each VAE model presents very different results and that the metrics do not always indicate the same trend along the reduction of data reduction for all locations. Regarding the comparison between FID climate, our proposal, and FID Imagenet, the results for FID climate are stable throughout the process of degradation of the reconstructions. However the results of FID Imagenet show a slight tendency to decrease the distance between the collections of samples of input and simulated, which does not reflect reality.

To better understand this behavior, we can analyze the results of the metrics for each type of transformation. Figure 4 shows that the FID climate has a greater capacity to represent blurring degradation, which we understand to be the main effect in the VAE reconstructions. While the FID Imagenet no longer represents the blurring increase from the value $\sigma = 2$. The same happens with the swirl transformation in which the FID Imagenet can not represent the gradual change of the samples.

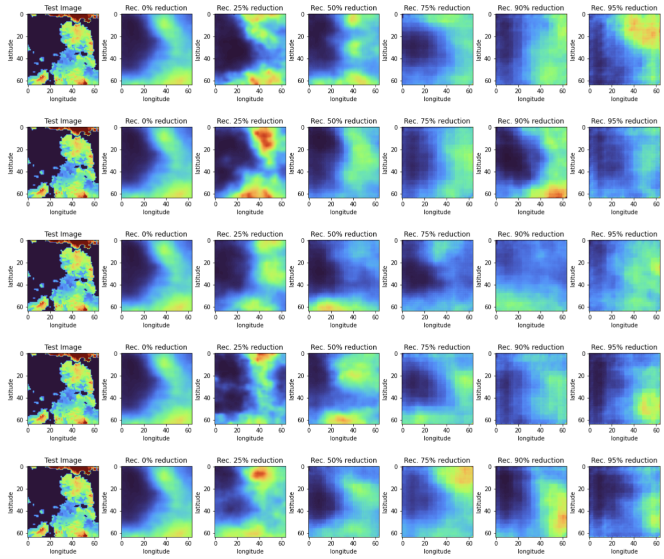

Figure 1: Sequences of VAE reconstruction results of Alabama-United States for each dataset reduction (5 times) for each percentage of reduction.

Even though it is obvious that FID climate is better than FID imagenet in blurring transformation, we believe it is necessary to elaborate a more complex classification task to be solved by Inception network so that we can perform more tests with VAE.

Still regarding the VAE training data reduction experiment, it is important to analyze some other metrics that were noteworthy, such as structural similarity, PSNR and KGE, which show a good ability to represent the degradation of results for the 7 locations.

On the other hand, some metrics were established as not suitable for this data reduction task as binary fractal dimension, energy and connectivity.

Conclusion

The generation of climate scenarios is important to mainly simulate events that happen less frequently. Thus, we need to advance in the correct evaluation of these reconstructions, finding domain metrics that better represent this data. In this work we identify metrics that can be used for these precipitation data in addition to the FID metric with specific weights based on a climate database with 12 modalities. The results presented in some transformations indicate that exploration in this direction may be promising for the advancement of the area of deep weather generators.

References

Abdella, A.; and Uysal, I. 2020. A Statistical Comparative Study on Image Reconstruction and Clustering With Novel

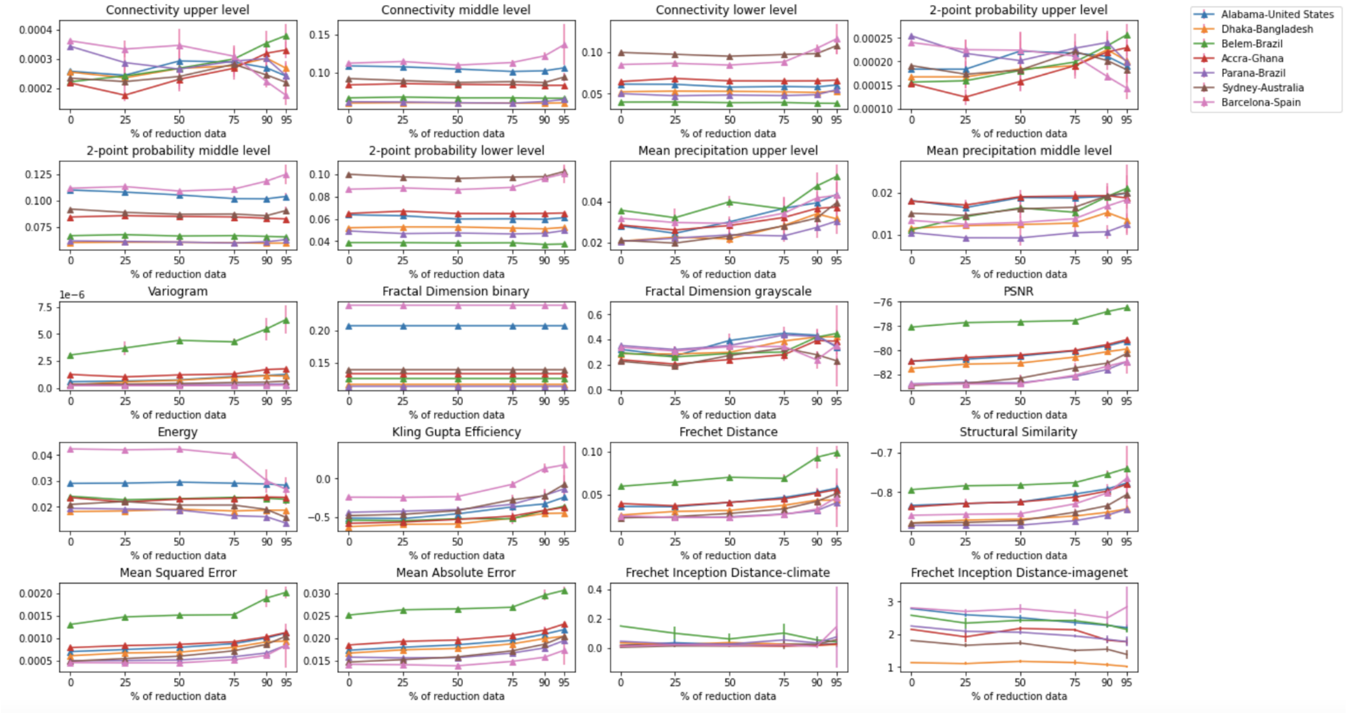

Figure 2: Metrics computed for reconstruction VAE evaluation with twenty samples for each place on gradual reduction training data.

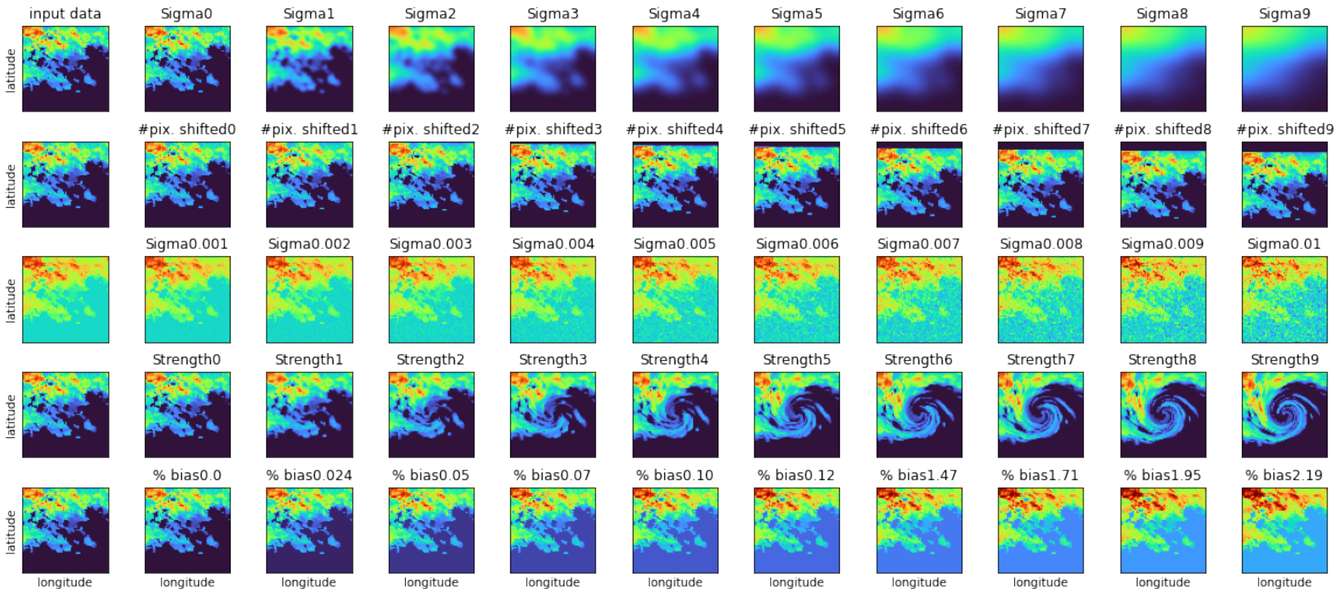

Figure 3: Lines sequences of transformations: blurring, shifting, Gaussian noise, swirl transform and additive bias.

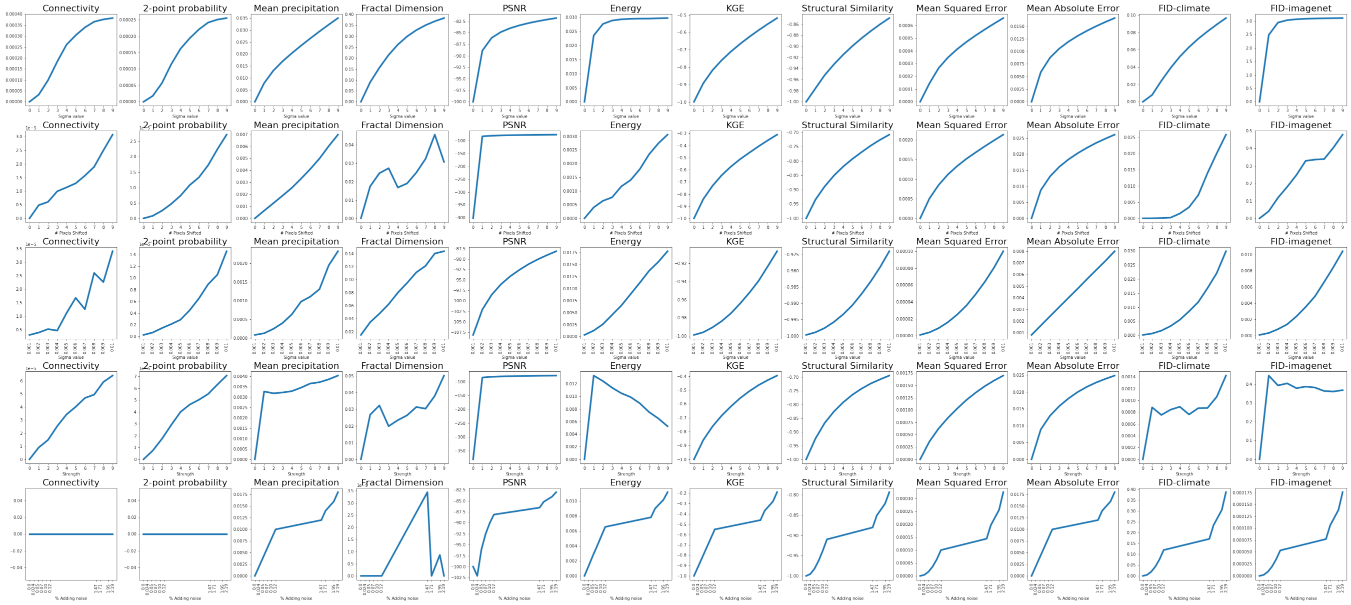

Figure 4: Some of the metrics computed for five different transformations, Blurring ,Shifting ,Gaussian noise, Swirl, Additive bias in respective lines from top to bottom. The metrics FID climate and FID Imagenet are located at the last two lines.

VAE Cost Function. *IEEE Access*, 8: 25626–25637.

Alt, H.; and Godau, M. 1992. Measuring the Resemblance of Polygonal Curves. In *Proceedings of the Eighth Annual Symposium on Computational Geometry*, SCG ’92, 102–109. New York, NY, USA: Association for Computing Machinery. ISBN 0897915178.

Bhatia, S.; Jain, A.; and Hooi, B. 2020. ExGAN: Adversarial Generation of Extreme Samples. arXiv:2009.08454.

Borji, A. 2019. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179: 41–65.

Bouda, M.; Caplan, J. S.; and Saiers, J. E. 2016. Box-Counting Dimension Revisited: Presenting an Efficient Method of Minimizing Quantization Error and an Assessment of the Self-Similarity of Structural Root Systems. *Frontiers in Plant Science*, 7.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Dai, B.; and Wipf, D. 2019. Diagnosing and Enhancing VAE Models. In *International Conference on Learning Representations*.

Di Napoli, C.; Barnard, C.; Prudhomme, C.; Cloke, H. L.; and Pappenberger, F. 2021. ERA5-HEAT: A global gridded historical dataset of human thermal comfort indices from climate reanalysis. *Geoscience Data Journal*, 8(1): 2–10.

Ding, W.; Wang, W.; and Zhao, D. 2019. A Multi-Vehicle Trajectories Generator to Simulate Vehicle-to-Vehicle Encountering Scenarios. In *2019 International Conference on Robotics and Automation (ICRA)*, 4255–4261.

Fréchet, M. M. 1906. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1): 1–72.

Funk, C.; Peterson, P.; and Landsfeld, M. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 150066.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, 6629–6640. Curran Associates Inc. ISBN 9781510860964.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Klemmer, K.; Saha, S.; Kahl, M.; Xu, T.; and Zhu, X. X. 2021. Generative modeling of spatio-temporal weather patterns with extreme event conditioning. *CoRR*, abs/2104.12469.

Mukherjee, J.; Praveen, K.; and Madumbu, V. 2018. Visual Quality Enhancement Of Images Under Adverse Weather

Conditions. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, 3059–3066.

Oliveira, D. A. B.; Guevara, J.; Zadrozny, B.; and Watson, C. D. 2021. Controlling Weather Field Synthesis Using Variational Autoencoders. arXiv:2108.00048.

Renard, P.; and Allard, D. 2013. Connectivity metrics for subsurface flow and transport. *Advances in Water Resources*, 51: 168–196.

Theis, L.; van den Oord, A.; and Bethge, M. 2016. A note on the evaluation of generative models. In *International Conference on Learning Representations*.

Torquato, S.; Beasley, J.; and Chiew, Y. 1988. Two-point cluster function for continuum percolation. *The Journal of chemical physics*, 88(10): 6540–6547.

Torquato, S.; and Haslach Jr, H. 2002. Random heterogeneous materials: microstructure and macroscopic properties. *Appl. Mech. Rev.*, 55(4): B62–B63.

Ullrich, S. L.; Hegnauer, M.; Nguyen, D. V.; Merz, B.; Kwadijk, J.; and Vorogushyn, S. 2021. Comparative evaluation of two types of stochastic weather generators for synthetic precipitation in the Rhine basin. *Journal of Hydrology*, 601: 126544.

Verdin, A.; Rajagopalan, B.; Kleiber, W.; Podestá, G.; and Bert, F. 2018. A conditional stochastic weather generator for seasonal to multi-decadal simulations. *Journal of Hydrology*, 556: 835 – 846.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.