

# Adversarial Rendering: A Novel Deep Learning Framework for Unsupervised Visual Content Synthesis

Harsh Hemantkumar Patel

\*\*\*

**Abstract**—Recent advances in generative adversarial networks (GANs) have unlocked new frontiers in unsupervised visual content synthesis. In this work, we introduce "Adversarial Rendering," a groundbreaking framework that integrates multi-scale feature extraction with adaptive loss functions to generate high-fidelity, semantically coherent images across diverse visual domains. By incorporating novel architectural innovations and self-supervised training strategies, our approach effectively captures intricate textures and complex structures without relying on paired datasets. Extensive evaluations on benchmark image synthesis and style transfer tasks demonstrate that our method surpasses state-of-the-art models in both visual realism and quantitative performance metrics. Beyond image generation, we illustrate the framework's versatility in applications such as data augmentation and cross-domain translation, highlighting its potential to drive advancements in multimedia content creation and computer vision research.

## I. INTRODUCTION

Accurate and high-resolution climate data at local scales is crucial for supporting society's adaptation strategies to rapidly changing climate conditions. While Earth System Models (ESMs) provide some of the most advanced climate projections on a global scale, their spatial resolution is often too coarse for regional or local decision-making needs. This creates a demand for downscaling techniques, which transform the large-scale, low-resolution outputs of global climate models into finer-scale climate information that is more relevant for local applications. High-resolution climate data across both space and time is essential for a wide range of practical uses, including detailed modeling of extreme weather events like floods, wildfires, and storms, as well as for guiding infrastructure development and ecological management decisions such as selecting appropriate tree species in changing environments. However, creating accurate downscaled datasets is often challenged by limited computational resources, which constrain the complexity and resolution achievable in practice. To be most useful, downscaling methods should balance computational efficiency, adaptability across diverse climatic regions, and the ability to represent extreme events effectively. This last point emphasizes the importance of generating multiple downscaled climate realizations or ensembles that capture the full variability of weather and climate scenarios.

Downscaling methods for converting coarse-resolution climate model outputs to finer regional scales typically fall into two broad categories: dynamical downscaling and statistical downscaling. Dynamical downscaling involves running a high-resolution regional climate model that physically simulates atmospheric processes over a limited area, using boundary conditions derived from the larger-scale, low-resolution Earth System Model output. This approach explicitly resolves local meteorological dynamics, allowing it to capture fine-scale spatial features and physical processes that global models cannot represent due to their coarse grids. In contrast, statistical downscaling relies on identifying and applying empirical relationships between large-scale climate variables and local climate features. These relationships can be derived using various statistical tools, including regression models, interpolation techniques, cluster analyses, lapse rate adjustments, and geostatistical methods such as Gaussian processes or kriging. Each approach has distinct advantages and limitations related to accuracy, computational cost, and flexibility.

Among these two, dynamical downscaling is often regarded as more physically robust because it directly simulates atmospheric physics and can better represent complex spatial patterns such as mountain-induced rain shadows or elevation-dependent temperature changes. Regional climate models (RCMs), which have been developed and refined over several decades, typically operate at resolutions between 20 to 40 kilometers and are frequently employed to downscale ESM outputs or reanalysis data to this intermediate spatial scale. For many practical applications particularly those requiring detailed local climate information it is desirable to move to even higher resolutions. Convection-permitting models (CPMs), which operate at resolutions between 1 to 4 kilometers, offer the ability to explicitly simulate convective weather phenomena like thunderstorms, providing much more realistic weather representations. However, the very high computational demands of CPMs limit their use in extensive climate change projections or operational forecasting, especially

when multiple model runs are required to assess uncertainty. Statistical downscaling is generally much more computationally efficient than dynamical downscaling, but can have limited capacity to downscale variables with complex dependence structures. Also, traditional statistical downscaling usually requires observations to calibrate models properly. Over the past few years, deep learning has been introduced as a new statistical downscaling strategy which can potentially capitalize on the benefits of both traditional statistical and dynamical downscaling. Specifically, deep neural networks can be trained on HR output from dynamic downscaling, and the network learns to “emulate” the dynamic downscaling, given a suite of LR climate information. This paper focuses on stochastic ensemble downscaling, where the deep learning model attempts to learn the conditional distribution of the HR fields, and then sample realisations from those. In this framework, the LR climate input are then the conditioning fields, since they condition the distribution that the model attempts to learn. So far, deep-learning methods have demonstrated potential for creating down scalings of similar quality to dynamic methods, but with the efficiency of standard statistical methods. The deeplearning downscaling we present in this paper attempts to emulate the results of convection permitting models, downscaling to high spatial resolution at hourly time steps.

Most of the recent research in generative deep learning based downscaling has employed conditional Generative Adversarial Networks (GANs). GANs, first introduced by [1] and adapted into a conditional version by [2], contain two deep convolutional networks, the Generator and the Critic. During training, these networks compete; the Generator tries to fool the Critic by producing output similar to the training data, and the Critic tries to distinguish between real and generated samples. Theoretically, the GAN will learn to sample from the conditional distribution of the HR variables, conditioned on the LR fields. [3] Developed a GAN framework which produced accurate downscaling of wind components. Recent work by [4] adapted this framework to be fully stochastic, allowing the GAN to sample multiple realizations from the learned conditional distribution. Their study showed that when applied to downscaling wind components, the stochastic GAN was well calibrated and was better at predicting extremes. This current work uses the basic network architecture developed by [3] and then adapted to be fully stochastic in [4].

While substantial research has investigated GAN downscaling, there multiple questions that should still be addressed prior to use in an operational setting. First, most studies to date have focused on downscaling single variables. Many studies [5]–[7] have focused on precipitation, and while [3] and [4] employed multivariate GANs, they focused solely on downscaling wind components. However, operational down- scaling is often required to provide a suite of variables. Temperature, humidity, precipitation, and wind are essential variables for a broad range of applications including fire weather, hydrology, ecology, and urban planning. While the stochastic GAN developed in [4] produced accurate down- scaling of wind components, its ability to extend to other variables has not previously been assessed. Multiple studies [5], [6] have shown that GANs can struggle to capture extreme precipitation events, a task which is crucial for adaptation planning. [4] found that their stochastic GAN was better at capturing extremes in wind components than an equivalent deterministic model. Since many traditional statistical downscaling methods succeed for means, but struggle to capture extremes, it will be important to ascertain the extent to which GAN downscaling can capture important extremes.

Downscaling of multiple climate variables invites the possibility of using fully multivariate GANs, where multiple variables are predicted from one model. Such an approach could improve dependence structures between variables, especially at small scales. While some dependence between variables will be inherited from the LR conditioning fields, multivariate models may be able to create correct dependence of fine-scale generated features, leading to improved consistency. However, most studies so far have only used univariate GANs [5]– [7], and while [3] and [4] showed success with multivariate downscaling of wind components, it is uncertain what the costs and benefits of multivariate prediction might be when extended to more variables. Wind components are very closely physically linked with similar distributions and dependence structures, are are not a very challenging multivariate down- scaling problem. From an operational perspective, multivariate GAN downscaling is desirable, as it decreases the number of models requiring training.

Commonly, GANs used in downscaling research input all the conditioning information (i.e., covariates) at LR. While this is necessary for variables coming from LR models, there is often pertinent surface information (such as topography) available at high resolution. Intuitively, providing surface information at a higher resolution should improve the performance of the model. However this hypothesis has not been systematically tested, and it requires adjusting the architecture of the Generator network from that using only LR covariates.

[5] included HR topography information by first convolving it down to LR and then concatenating with the climate variables essentially trying to fit more information into the LR architecture. [4] Used a different approach with two parallel Generator streams for covariates of different resolution. It is important to assess how these architectural changes impact the results.

Most studies investigating GAN downscaling use small domains for computational reasons, and the ability of GAN frameworks to generalize over large spatial reasons has yet to be assessed. Applying GAN downscaling over large contiguous areas poses many challenges. Computational constraints aside, it is unclear whether a single GAN framework can optimally downscale the idiosyncratic weather of disparate regions in a large study area, as most research to-date has developed and tested GAN frameworks in a single region. Showing that a GAN framework can generalize over space is a first step to developing models capable of downscaling large regions.

This current study aims to address these some of these questions and investigate the applicability of stochastic GANs to future operational downscaling. Specifically, we apply the stochastic GAN framework to temperature, humidity, precipitation, as well as both wind components. Initial analyses are conducted in the region of complex topography (include Vancouver Island, the Coast Mountains, and the Interior Plateau on the Pacific Northwest of North America) considered by [4]. We then investigate the advantages and disadvantages of univariate versus multivariate prediction. The paper then assesses the utility of providing HR topography to the GAN. Finally, we test the GAN framework on all five variables in a second region in Northeastern British Columbia and Alberta (a region with relatively flat topography) and assess the spatial generalizability of the GAN framework.

## II. METHODS

All Generative Adversarial Networks (GANs) presented in this study share a consistent foundational architecture. Our training process utilizes paired datasets consisting of low- resolution (LR) conditioning inputs (covariates), static high resolution (HR) surface characteristics, and corresponding high-resolution target fields. The GAN framework is designed to learn a functional mapping from the LR input covariates and surface features to the detailed HR output fields. To maintain uniformity across experiments, we standardize the spatial dimensions and resolutions of all data inputs and outputs: the high-resolution fields are represented as 128 by 128 pixel grids, which approximately correspond to an area of 270 by 450 kilometers (or  $4^\circ \times 4^\circ$  in latitude and longitude), while the low-resolution inputs are 16 by 16 pixels, reflecting an eightfold downscaling factor.

### A. Data

A natural use of GANs in a downscaling setting involves training on paired HR regional weather model output and LR ESM or reanalysis data. We follow this approach, using ERA5 reanalysis variables as the LR predictors, and a Western Canada Weather Research and Forecasting (WRF) model output [8] as the paired HR training data. The WRF model is a state of the art numerical weather model, designed for convection-permitting scale (3-4 km) forecasts. This specific WRF run was driven by ERA-interim, covers all of British Columbia, and has a 4 km grid-size resolution. We used ERA5 as the paired LR data since it provided all the covariates we planned to use, whereas ERA-interim only provided a subset of them. Although there will be some differences between ERA5 and ERA-Interim, they represent the same realization of the climate system so it is reasonable to use ERA5 as the paired LR dataset. However, it is important to note that in this downscaling scenario, where HR and LR data each come from separate models, there will be large-scale biases between the WRF output and the ERA5 conditioning fields. The WRF model generates internal variability, which will also cause differences between models on common scales. Thus, a successful GAN must learn to bias correct, downscale, and accommodate internal variability.

Unless otherwise noted, the GANs we consider use a suite of seven LR covariates: temperature (2m), specific humidity (2m), precipitation, wind components (10m), convective available potential energy (CAPE) and surface pressure. We also include HR topography as an invariant field. For the majority of our analyses, we consider a rectangular region in Southwestern BC, Canada ( $49^\circ$  to  $53^\circ$  N,  $122^\circ$  to  $126^\circ$  W; henceforth called the Southwest region), as its high degree of topographic complexity represents a realistically challenging downscaling scenario (figure 1). [4] tested stochastic GANs for downscaling wind components in this region; we extend the analysis to downscale temperature, specific humidity, and precipitation. As is common in deep learning we standardize all variables to mean zero and unit standard deviation prior to training.

To investigate the applicability of this GAN framework to different regions, we also investigate a second region in the northeast of BC and the northwest of Alberta (the Northeast region), which in contrast to the Southwest region has flat topography and is influenced by different weather processes. We also investigate using a land use index, which includes water bodies and forest types, as a second HR invariant field.

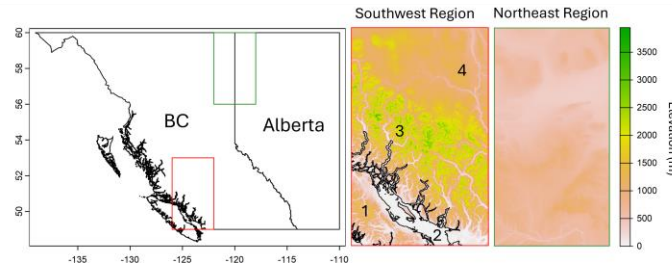


Fig. 1. Maps of study areas, showing (from left to right) study area locations relative to British Columbia and Alberta, and topographic relief of both regions. 1 = Vancouver Island, 2 = Georgia Strait, 3 = Coast Mountains, and 4 = Interior Plateau.

## B. Model

In the initial GAN formulation, the Critic network estimates the probability of a sample being from the training set; the Generator attempts to make it more challenging to distinguish training samples, while the Critic tries to improve its ability at discriminating. This approach often led to instability during training, as it requires that both networks learn at approximately the same rate and can lead to vanishing gradients. [9] Addressed this challenge with the introduction of the Wasserstein GAN, where the Critic estimates the Wasserstein distance between the generated and training samples. Intuitively, the Wasserstein distance represents the amount of mass required to transform one distribution to another, and is a distance metric between the distributions. In our case, the Wasserstein distance estimates the distance between the high dimensional distributions of the training data and the generated output. During training, the Generator attempts to minimize the Wasserstein distance between the generated fields and the training data, thus increasing the distributional similarity between them. Following recent literature in downscaling and computer vision (refs xxx), we adopt this approach.

One appealing aspect of GANs compared to more standard Convolutional Neural Networks, especially for climate downscaling, is that they do not solely rely on a pixel-wise error metric as the loss function. Downscaling is an underdetermined problem, meaning that there is a distribution on HR fields consistent with a single set of LR conditioning fields. Thus, while we require convergence in large-scale structure between the downscaled results and the LR conditioning fields, we should accept slight differences in fine-scale structure between the generated output and the training data. Using a pixelwise metric such as mean absolute error can overly constrain models by forcing them to match the training data too closely, a phenomenon known as the double penalty problem. In such cases, the model will often converge on the conditional median, producing blurry output. In contrast, the adversarial loss calculated by the GAN's Critic network (in our case, the Wasserstein distance) is not a pixel wise metric, and aims for convergence in distribution, which is a desirable property. However, [10] showed that only using the adversarial loss in the Generator training procedure often leads to unstable training and poor convergence of large-scale structures. They suggested adding a pixel-wise loss back in as the content-loss, to reward convergence in realization at large-scales. We follow this approach, and our Generator loss function is composed of both the adversarial loss, and a pixelwise content loss. A more detailed discussed of this issue is presented in [3].

This study uses the stochastic GAN architecture developed by [4], which was based on the deterministic GAN described in [3]. The architecture makes extensive use of convolutional layers, which are designed to extract representative features from images [11]. In the Generator network, we use Residual in Residual Dense Blocks (RRDB), which contain stacked convolutional layers followed by leaky rectified linear units to add non-linearity. For upsampling, we use three pixel-shuffle blocks [12]. Following [4], we inject Gaussian noise fields into the convolutional filters inside each RRDB. We also include a HR input stream to allow inclusion of HR covariates, such as topography. This stream uses the same structure of RRDB as that of the LR inputs, but skips the up sampling step. Once the up sampling has occurred on the LR stream, all inputs have the same dimension, and

are concatenated. We also include all covariates as inputs to the Critic network, using a LR input stream for the LR covariates. Intuitively, including the conditioning information should allow the Critic to better estimate the conditional distributions.

Unless otherwise specified, all models presented in this study use the best model from [4], with stochastic sampling and CRPS as a content loss. Stochastic sampling, adapted from [5] creates multiple realizations of each training sample, and computes the content loss across these realizations. CRPS is a probabilistic measure, which aims for convergence in pixel wise distributions. As our aim is to sample from the HR conditional distributions, it is natural to use a probabilistic metric. Since stochastic sampling and the CRPS content loss are only applicable in a stochastic setting, when using a deterministic Generator, we employed standard training with MAE as a content loss, as in [3].

### C. Training

We trained individual models for each of temperature, specific humidity, and precipitation using the framework described above. We used two years of hourly data as a trainingset (2003 and 2006), and one year as an out-of-sample test set (2005). Initial tests showed that model results did not improve substantially using more than two years of training data, so we chose this size for computational efficiency. Models were trained until metrics on the test dataset stabilised ( $\leq 250$  epochs). We then saved the Generator from the final epoch for analysis. Multivariate models for predicting all variables were trained in a similar way, with the HR training data created by stacking the individual variables as separate channels. We trained all models on an NVIDIA RTX 4090 GPU; models took on average 48 hours to train [4] found that the stochastic GAN was better able to capture wind component extremes (i.e., the tails of the distribution) than the deterministic GAN. A common challenge with precipitation downscaling is underestimation of high-precipitation events [13]. We thus created a deterministic GAN similar to that considered in [4] by removing the noise injection from the convolutional layers in the Generator and using mean absolute error as the content loss metric.

### D. HR Topography

To test the importance of including HR topography as an input to the network, we trained models using a) HR topography, b) LR topography interpolated to the HR grid, and c) LR topography. The LR interpolated topography experiment was done as a control for network architecture we kept the Generator architecture identical, but fed the network LR information. To create the LR topography model, we adjusted the Generator network by including an up sampling block in the topography stream. We chose this strategy to keep the network architecture as consistent as possible between experiments. We kept all other features consistent.

### E. Analysis and Quality Metrics

Quality assessment in image generation problems often poses a challenge, because there are multiple, often competing, metrics that could be used. Commonly used metrics assess pixel wise error of realizations, and while these are useful, they can overly penalise underdetermine downscaling results due to the double penalty problem. Thus, it is generally better to compare statistics between the generated fields and truth fields from the test set, instead of comparing individual realizations. Pixel wise comparisons of statistics and distributions (e.g., medians and quantiles) are the simplest examples of such comparisons. However, these metrics on their own do not tell a complete picture. It is often important to know how well spatial structures of different scales (i.e., textures) match between the generated and truth fields. For this task, we used a Radially Averaged Spectral Power metric (RASP), which calculates the 2D spectral power at each wavenumber, averages power over all angles from the centre of wavenumber zero, and standardises the power at each wavenumber to the corresponding power in the truth field. RASP values greater than one then represent too much spatial variance at the given scale, while values less than one represent too little.

The metrics described above investigate the quality of the full distribution  $P(HR)$ . Especially with a stochastic GAN, it is also important to investigate the conditional distribution,  $P(HR|LR)$ . These two distributions are related through distributions of pixel-values by month, the PDF of generated values closely matched the distribution of WRF values, for both January and July, although showed a slight difference in January. Maps of pixel-wise quantile differences show good

$$p(HR) = \int p(HR|LR)p(LR) dLR \approx \frac{1}{n} \sum_{k \in LR} p(HR|LR_k) \tag{1}$$

calibration for the median and 0.99 quantiles, but more bias in 0.01 quantiles. For the 0.01 quantile, the GAN underestimated values in the mountains and ocean, and overestimated value.

To test the stochastic calibration of the conditional distribution of generated fields, we used CDFs of rank histograms calculated over one year of samples. As it is not usually possible to access multiple realisation of the same truth field, an ensemble of stochastic realisations have to be compared to a single truth field. For a properly calibrated model, the truth field should be indistinguishable from any of the generated realisations, and thus the distribution of ranks of the truth value in the ensemble values should be uniform. For ease of model comparison, we plotted CDFs of the rank histograms, to avoid the sensitivity of histogram bin width. To investigate the stochastic calibration of individual conditional distributions, we also present rank histogram maps, where each pixel represent the rank of the truth field out of the ensemble of generated values for that pixel. In the plateau region northeast of the coast mountains, with biases up to about 3 K.

### III. RESULTS

In this section, we first investigate extension of GAN downscaling from wind components to three other important variables: temperature, humidity, and precipitation. As correct dependence between variables is important, we then consider the impact of multivariate and univariate downscaling. We then present a sensitivity analyses, showing the importance of including HR topography. Finally, we test spatial generalisability of this GAN framework, by training and testing on a new region in Northeastern BC.

#### A. Extension to temperature, humidity, and precipitation

In this section, we assess the accuracy of downscaling for temperature, humidity, and precipitation. For each variable, we chose two hourly timesteps, that represent the 0.1 and 0.9 quantile, averaged over the field (i.e., a cold and a warm hour). We also show pixel-wise statistics across all timesteps, and compare PDFs of pixel values for January and July.

Temperature downscaling generally performed well, and succeeded in capturing HR details (figure 2). Consideration of HR realisations for the representative cold hour showed slight underestimation of most realisations in the Straight of Georgia and missed some fine-scale details in the Northeast corner. Realisations of the warm hour showed excellent agreement with the ground-truth, picking up stronger elevation gradients than the cold sample, especially in the Coast Mountains and the continental plateau in the Northeast. In these two samples, the model was successful at capturing the more stable temperature of the ocean compared to the surrounding land, as well as sharp transitions between the ocean and continent. Conditional standard deviation across stochastic ensemble members ranged from about 0.5 to 1.5 K, except for very low values in mountains and the Georgia Straight in the cold sample. Considering

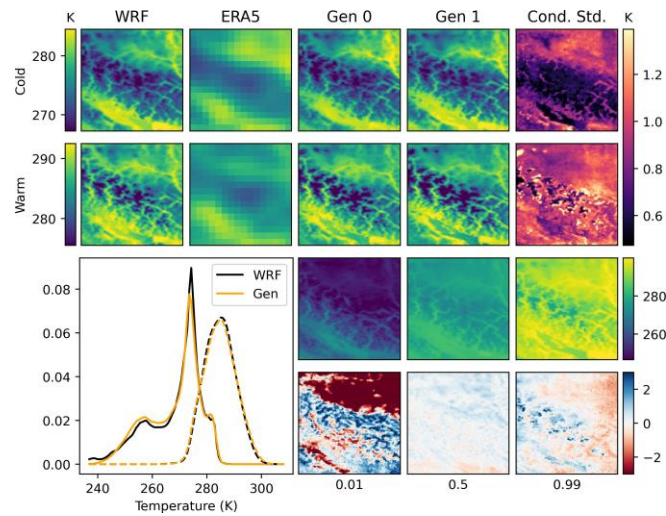


Fig. 2. Evaluation of univariate GAN downscaling of temperature for the Southwest study area. Top two rows show respectively an example cold and warm sample, with the WRF field, LR conditioning field (showing the continental outline), two stochastic realisations, and the conditional pixel wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.9 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences (truthgenerated). The left-bottom panel shows overall PDFs of pixel values for January samples (solid) and July samples (dashed).

Specific humidity proved to be a more challenging variable to accurately downscale (figure 3). In the dry sample especially, the model substantially underestimated the humidity over the Georgia Strait, and while it did predict higher humidity at lower elevations (e.g., over valleys), these were not as clear as in the WRF fields. In the moist sample, while humidity values across the field matched well with WRF, the spatial patterns in the generated fields were more blurred and sharp gradients were not as well represented. Most realizations of the moist sample showed a dry bias on the Eastern side of the field. Overall, the conditional standard deviation was much lower for the dry sample than the moist sample. The moist sample showed low deviation in the ocean, but relatively high variability in the mountains near the coast. Distributions of pixel values for January and July showed substantially more bias than with temperature, especially for July, where the PDF of generated values shows under dispersion.

Pixelwise 0.01 quantiles show that the models overestimated humidity through most of the field, except in the Strait of Georgia, where humidity was always underestimated. Median values were similar, but showed a slight underestimation across the field. For 0.99 quantiles (very moist) the models largely underestimated humidity, except on the tops of mountains.

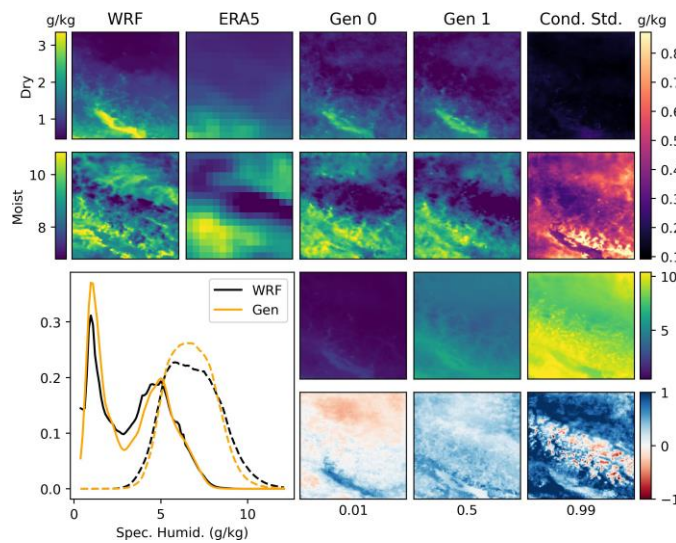


Fig. 3. Evaluation of univariate GAN downscaling of specific humidity for the Southwest study area. Top two rows show respectively an example dry and moist sample, with the WRF field, LR conditioning field (showing the continental outline), two stochastic realisations, and the conditional pixelwise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.9 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences (truthgenerated). The left-bottom panel shows overall PDFs of pixel values for January samples (solid) and July samples (dashed).

The GAN also performed well at downscaling precipitation (figure 4). The light-rain-hour sample shows a large degree of variation between realisations, as is desired. The heavy rain sample also shows high conditional standard deviation, although relative to the mean, not as much as the light rain sample. With precipitation especially, it is noticeable that single realisations of the generated fields often show different patterns than the WRF field. However, as our goal is to sample from the distribution of possible downscalings, this is expected.

Pixel value distributions were very similar in January and July, and were combined in the PDF for better visual interpretation. Generated and WRF distributions matched well, although the GAN slightly underestimated extreme precipitation events. This underestimation of heavy precipitation is also apparent in the pixel wise quantile difference maps, which show good calibration for 0.01 quantiles and medians, but predominantly underestimation of 0.99 quantiles.

We found that covariate choice was especially important for precipitation. Initial models, which only included LR precipitation, temperature, evaporation, and pressure produced fuzzy and biased downscaling. Addition of CAPE and wind components substantially improved results.

Stochastic GANs did a much better job of capturing extreme precipitation than equivalent deterministic models (figure 5). While both models slightly underestimated the probability .

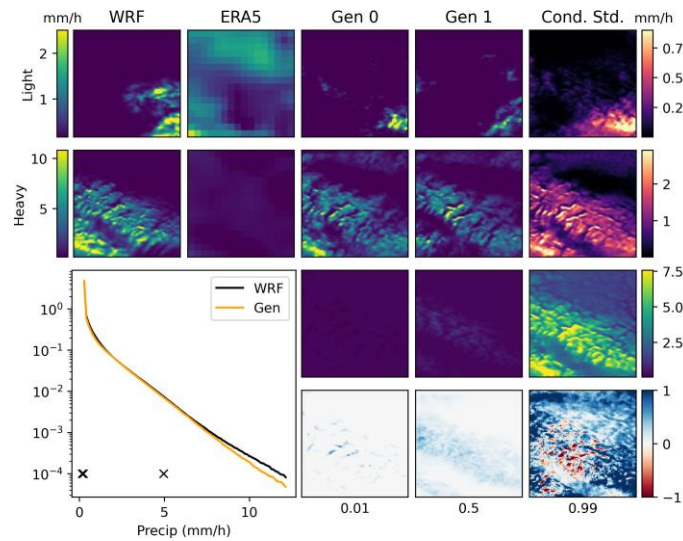


Fig. 4. Evaluation of univariate GAN downscaling of precipitation for the Southwest study area. Top two rows show respectively an example light rain and heavy rain sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.9 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences (truth - generated). The left- bottom panel shows overall PDFs of pixel values for months January to July combined. Crosses indicate the location of 0.01, 0.5, and 0.99 quantiles. All timesteps that had zero precipitation in the WRF field were removed prior to analysis.

high precipitation compared to WRF, the stochastic model matched the distribution of the WRF data well, while the deterministic model did not predict any precipitation values < 18 mm/h. Both models show a low-precipitation bias for light precipitation (< 2.5 mm/h). All three variables showed good.

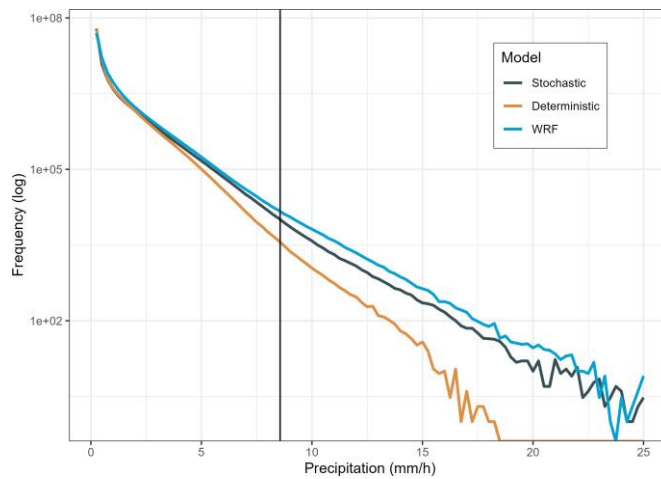


Fig. 5. Distributions of pixel values for precipitation fields, comparing WRF, deterministic generated, and stochastic generated. Distributions were estimated from all pixels of one year of hourly samples. Note the y-axis is shown on a log scale.

stochastic calibration and similar spectral power to WRF fields (figure 6). Median RASP estimates for generated fields showed more than 80% similarity to corresponding WRF estimates for temperature and humidity. Precipitation also showed good calibration through most of the field, but had a high-power bias at high wavenumbers, corresponding to an overabundance of very fine-scale textures (figure 6). Precipitation also displayed much more variability between samples than temperature or humidity. This is expected, as precipitation fields vary more across samples than do the other variables.

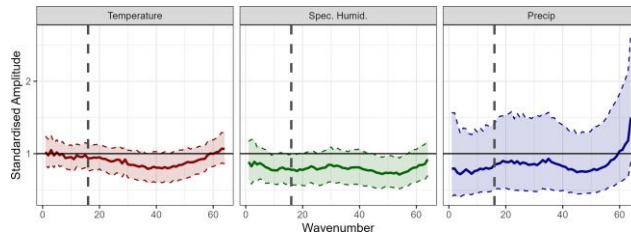


Fig. 6. RASP metric for the three variables. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields. Solid lines show median spectral power, shaded region show inter-quartile range.

Considering calibration of conditional distributions for individual samples, all variables showed slight under dispersion, with ranks concentrated at either end of the scale (figure 7a) instead of being uniformly distributed across possible ranks. Temperature showed good calibration for the median and 0.99 quantile sample, but underestimation over most of the range for the 0.01 quantile sample. Specific humidity showed the most consistent under dispersion, especially in the 0.01 and median samples. Precipitation generally was better calibrated, but showed some underestimation in high precipitation areas. These results also match the rank histogram across timesteps; temperature and humidity both showed slight under dispersion of conditional distributions, and precipitation showed under- estimation of high values (figure 7b).

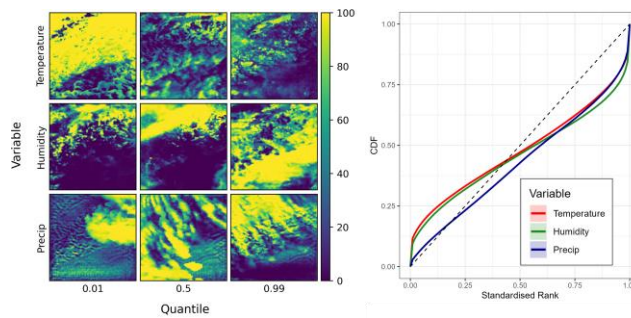


Fig. 7. a) Rank histogram maps for individual samples showing, for each pixel, the rank of the WRF pixel compared to an ensemble of 100 realisations.

b) CDF of rank histograms showing stochastic calibration of conditional distributions for univariate models of temperature, specific humidity, and precipitation. Rank histograms were calculated across 100 randomly selected conditioning fields, with 100 HR realisations of each. Dashed line shows reference uniform CDF.

## B. Multivariate Prediction

Multivariate GANs showed improved dependence structures between pairs of downscaled variables (table I). Particularly for temperature and humidity, variables from the univariate model had much lower mutual information scores than from the WRF variables, indicating less dependence between variables. Multivariate mutual information scores were mostly slightly lower but close to WRF scores. Temperature and precipitation was the only pair of variables where the multivariate model showed too much dependence - i.e., mutual information scores were higher than for the corresponding WRF variables.

**TABLE I MUTUAL INFORMATION SCORE BETWEEN PAIRS OF VARIABLES FOR MULTIVARIATE PREDICTION, UNIVARIATE PREDICTION, AND WRF. SCORES WERE CALCULATED FOR EACH OF 600 RANDOMLY SELECTED TIMESTEPS AND AVERAGED.**

Variable 1	Variable 2	Multivariate	Univariate	WRF
Temp	Precip	0.222	0.177	0.187
Humid	Precip	0.198	0.177	0.210
Humid	Temp	0.943	0.602	1.020

While measure of dependence generally improved, marginal statistics for individual variables were worse with the full multivariate model. For temperature, humidity, and precipitation, marginal statistics generated from the full multivariate model were blurrier than those from univariate models (figure 8). Humidity showed the most severe challenges, missing a lot of fine-scale details. Precipitation, while capturing the general patterns of the marginal statistics, did not capture the high precipitation values with the multivariate model.

To assess whether these challenges were due to the inclusion of precipitation, which has substantially different spatial dependence structures, we tested a multivariate model without precipitation. This model showed improved quality, but resulting downscalings were still blurrier than those from the univariate model. This was especially obvious for humidity, which also contained traces of the convolutional filter in both multivariate models.

Power spectra of all variables showed more bias for multivariate models compared to univariate models, with the NoPrecip model in between (figure 9). Precipitation showed a substantial lowpower bias across most wavenumbers in the full multivariate model, often capturing only 25% of expected power. Humidity and temperature showed large high-power biases at high wavenumbers in both multivariate models, although to a lesser degree in the No Precip model, corresponding to the blurriness observed in figure 8. Both multivariate models also showed a spike in power at wavenumber 32, corresponding to the size of the convolutional filters. Both zonal and meridional wind components showed low-power biases throughout most wavenumbers in the full multivariate model, and high-power biases at high wavenumbers.

## C. High-Resolution Topography

To determine the importance of including HR topography as a covariate, we compared models with LR topography, inclusion of HR information.

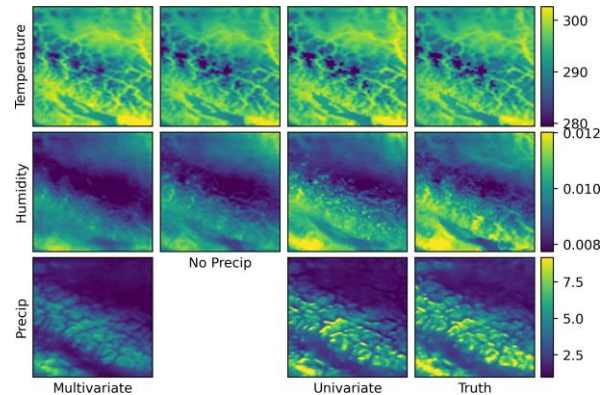


Fig. 8. 0.99 quantiles for generated temperature, specific humidity, and precipitation fields, using full multivariate prediction, multivariate prediction without precipitation, and univariate prediction. Quantiles were calculated using 3000 randomly selected timesteps.

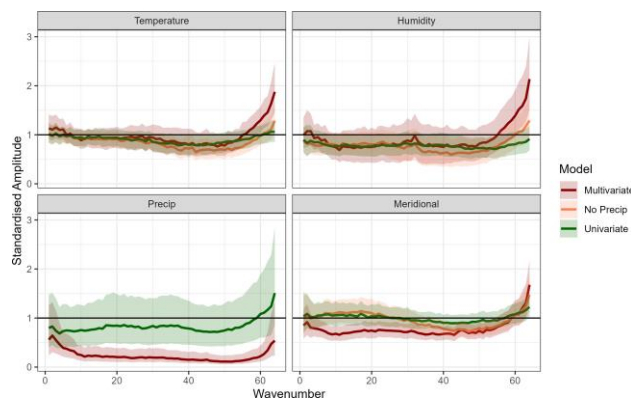


Fig. 9. Median and IQR RASP for precipitation, specific humidity, temperature, and meridional wind fields, using multivariate, no-precipitation, and univariate models. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields.

HR topography, and LR topography interpolated to HR. Considering variability at spatial scales using a RASP for temperature, humidity and precipitation showed that the HR and the interpolated topography both had better calibration of spectral power than the corresponding LR topography model (figure 10). The LR topography model generally performed well a lower wavenumbers but showed a low-power bias at high wavenumbers, for all variables. This low-power biases was more severe for temperature and humidity; the LR model for precipitation showed a fairly consistent low power bias across wavenumbers, and did not increase power at high wavenumbers as the HR models did. Interestingly, there was very little difference in spectral power between the HR model and the LR Interpolated model, suggesting that the architectural design of the network is more important than.

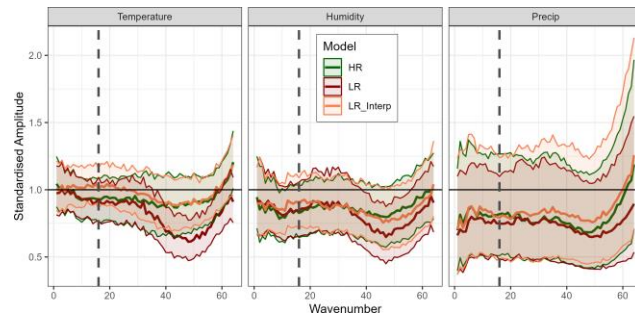


Fig. 10. Median and IQR RASP for temperature, humidity, and precipitation using HR topography, interpolated LR topography, and LR topography. Spectral powers are standardized to ground truth fields, and metrics are calculated across 1200 randomly selected fields. Dashed line shows wavenumber corresponding to LR grid size.

#### D. Generalization Across Space

Downscaling in the Northeast region was successful for certain variables, but showed more challenges than in the Coastal region (figure 11). Downscaling of wind components showed similar quality to the Coastal region, whereas generated temperature and humidity fields often showed sub spatial differences from WRF. This was especially apparent for humidity, where generated fields were overly smooth and lacked a lot of the fine scale details of WRF. Precipitation showed the most challenges; generated fields often had entirely different structure than the WRF field. The fourth row in figure 11 shows a representative sample of precipitation, with the generated fields showing patchy, high intensity precipitation across the domain. Most variables showed a substantial difference in the largescale structure of the ERA5 field compared to the WRF field. This was especially apparent for precipitation; for the sample in figure 11, the WRF field shows low-intensity precipitation through much of the field, while the ERA5 field shows a concentrated area of precipitation near the center. To determine if this mismatch between the LR and HR fields was responsible for the poor downscaling quality, we trained a model where the LR precipitation field was created by coarsening the WRF field, resulting in zero bias in the large scale structure. This model produced substantially more accurate downscaling, with generated precipitation patterns closely matching the WRF field (fifth row of figure 11). Covariate choice was especially important in this region; we found that CAPE was an important covariate for all variables in this region, whereas in the coastal region, CAPE had only improved results for precipitation. Stochastic calibration of conditional distributions was worse for temperature, humidity and standard precipitation in the Northeast region than in the Southwest region (figure 12). Temperature and humidity both showed under dispersion, with

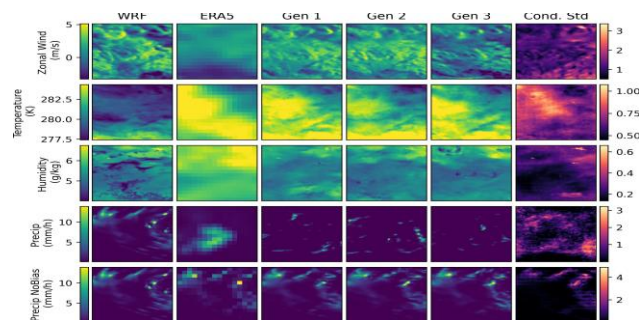


Fig. 11. Example realisations for the Northeastern region. Rows correspond to variables, and the bottom row shows a second precipitation model with no bias between the LR and HR training data. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisation smany true samples falling outside the generated range, and precipitation showed underestimation, as many

true samples fell above the generated range. The unbiased precipitation model had much better calibration than the standard precipitation model, and was one of the best calibrated models overall. Wind components showed similar calibration to the Coastal location. RASP metrics for variables in the Northeast region.

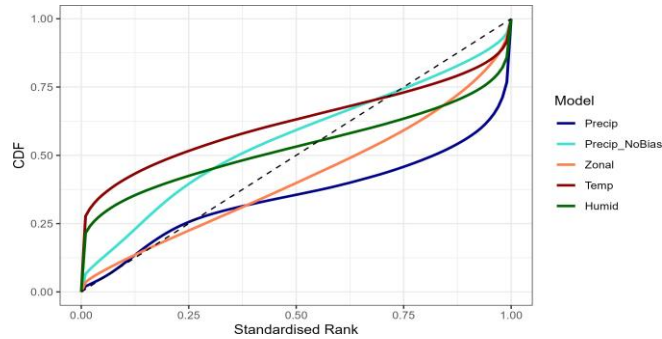


Fig. 12. CDFs of rank histograms showing stochastic calibration of models in the Northeastern region. Rank histograms were calculated across 100 randomly selected conditioning fields, with 96 HR realisations of each. Dashed line shows reference uniform CDF.

showed similar median values for humidity, temperature, and wind compared to the Coastal region, but had much larger inter-quartile ranges, representing more variability in texture bias between samples (figure 13). The standard precipitation model showed substantial low-power bias across scales, but especially at low wavenumbers where power was less than 50% of corresponding WRF power. Conversely, the unbiased precipitation model showed good calibration over most of the range, and much smaller IQR.

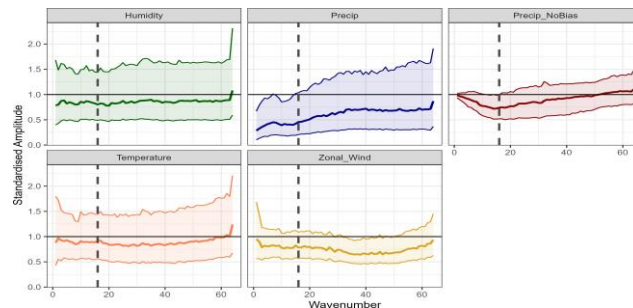


Fig. 13. RASP humidity, precipitation, temperature, and zonal wind in the Northeast region, showing median and inter-quartile ranges. Spectral powers are standardized to ground truth fields, and metrics are calculated across 1200 randomly selected fields.

#### IV. DISCUSSION

This paper investigates practical considerations of applying the stochastic GAN framework from [4] to realistic down-scaling applications. Specifically, we focus on extension of GANs to multiple climate variables, including the applicability of multivariate prediction, and generalizability to different locations. We show that the stochastic GAN framework can successfully downscale a suite of variables. We then find that while multivariate downscaling improves the dependence structures of downscaled variables, it tends to decrease the quality of individual down scalings. Finally, we show mixed success in generalising to the Northeast region: models for temperature, humidity, and wind components produced reasonable downscalings, but were less accurate than in the Southwest region. Precipitation models struggled in this region, likely due to large-scale biases between the LR and HR training data.

## A. Extension to temperature, humidity, and precipitation

Overall, we found that the stochastic GAN successfully downscaled temperature, specific humidity, and precipitation, although it was less accurate with humidity than the other variables. Challenges with downscaling humidity could be due to a variety of reasons. First, WRF humidity fields often showed very sharp gradients around valleys, which the GAN often did not capture fully. It may also be that we did not include all important covariates; for example, it would be interesting to add temporal pressure gradients as a LR covariate.

Precipitation is an important variable, and is often more challenging to downscale due to its unusual distribution. We show that the stochastic GAN performs well at downscaling precipitation, and is much better at capturing extreme precipitation events than a deterministic GAN. Since extremely heavy precipitation is likely to cause flooding and damage, being able to capture it is important. This result supports the finding of [4], who show that by sampling from the full HR distribution, the stochastic GAN was better able to estimate wind component extremes.

## B. Multivariate Prediction

Multivariate prediction lead to improved dependence structure between dependant variables, but decreased the quality of predictions of individual variables. It seems reasonable that for variables with strong dependence, multivariate prediction would lead to better consistency, as it allows fine-scale variability to be harmonised between variables. For temperature and humidity, multivariate models generated fields with mutual information scores closer to that of the WRF variables. However, variables generated from the full multivariate model were noticeably more blurry, failed to capture fine scale variability, and showed artifacts. Humidity seemed especially challenging; the univariate model was the only model able to recreate the finescale features around the Strait of Georgia, and the full multivariate model created predictions still showing artifacts of the convolutional filters. When we removed precipitation from the model and only predicted wind components, temperature, and humidity, results were improved, but still blurry. Precipitation has a very different distribution than the other variables; it seems that trying to predict variables with different distributions is challenging. Perhaps since the convolutional filters being learned for each field are so different, the final result is an overall poorer compromise. However, it is interesting that even with precipitation removed, generated fields were less accurate. Using multivariate prediction means that there are fewer tunable parameters that can be used specifically for a single variable. We hypothesize that this may lead to decreased flexibility for the model to adapt to a specific variable. An interesting avenue of future research would investigate whether adjusting the model architecture to improve flexibility could improve multivariate prediction. For example, it may be beneficial to separate the network in to separate branches near the end for each variable being prediction. In practice, it may still be advisable to use multivariate prediction for highly coupled variables (e.g., temperature and humidity, wind components), and univariate prediction for less dependant variables (e.g., precipitation).

## C. HR Topography

Including HR topography in the Generator improved spatial structures of wind components, temperature, and humidity, particularly at high wavenumbers. However, including LR interpolated topography produced downscaling of approximately similar quality, thus suggesting that network architecture may be more important than the topography resolution itself. Adding an HR input stream results in a substantially larger network, with more learnable weights at HR scales, especially since our architecture applies a RRDB to the HR input stream. Thus, even if the input has the same information, difference in architecture and the increased network size at the fine scales could allow the model to better capture fine scale details. It is interesting to note that the GANs with HR topography seemed to stabilise faster during training than the model with LR interpolated topography. This suggests that, given the correct architecture, the model can learn HR details over time, but is aided initially by having the HR information. For precipitation in the Northeast region, we also found that including a second HR covariate (land use index) improved predictions. However, since this addition slightly altered the network architecture, it is unclear whether the land use information itself was useful. Although adding a HR stream to the Generator increases network size, we believe that the substantial improvement in fine-scale structure makes this trade off worthwhile, and we suggest including HR covariates when possible.

## D. Generalisation in Space

Applying the stochastic GAN framework to the Northeast location showed mixed success. Wind component downscalings generally showed similar high-accuracy as in the Southwest region. Temperature and humidity downscalings were reasonable but not as accurate, and precipitation models were initially poor, with generated fields showing very different spatial structure than the WRF fields. We hypothesize that some of the challenges in this region, especially with precipitation, were due to larger biases between WRF and ERA5 structures at large scales. It was visually apparent that in many samples, the LR conditioning fields did not match the structure of the corresponding WRF fields. Our unbiased precipitation model, where we created the LR conditioning fields by coarsening the WRF fields produced highly accurate downscaling, supporting our hypothesis that this mismatch is a source of the challenges. Unfortunately, in an operational setting, it is generally not possible to have unbiased LR and HR fields, as the HR fields do not exist. Some studies have already investigated the challenge of large-scale biases between datasets. [7] developed a GAN with two stages, the first to correct biases, and the second to downscale. However, this approach is only applicable if biases are consistent across samples. If biases change between samples, which we hypothesize is often true with precipitation, it becomes a much more challenging problem. Some of the challenges we found in this region may be improved by using a larger training region; biases between LR and HR data are likely more severe at smaller scales, and by using a larger area, there may be more consistency between datasets, resulting in increased stability during training. In regions with substantial bias, it may also be possible to train models using the unbiased coarsened data, and then predict using the bias LR dataset. While this technique would then not perform any bias correction, it could perform better at downscaling than an unstable model.

We noticed that covariate choice had a large effect on downscaling accuracy in the Northeast region compared to the Southwest region. Certain covariates, which had not been necessary in the Southwest region, were important for accurate downscaling in the Northeast. Convective Available Potential Energy was important for predicting precipitation, humidity, temperature, and wind components. While we included CAPE as a covariate in our final models for both regions, CAPE only improved precipitation downscaling on the Southwest region, whereas it improved all variables in the Northeast region. Since substantial fine-scale variability can result from convection, especially in regions of flatter topography, it makes sense that CAPE is an important variable in this region. In general, different suites of LR covariates will be needed depending on the downscaling region. Therefore, to obtain accurate downscaling's over large areas, it will likely be necessary to include more covariates than required for a smaller region, as all subdomains will require the covariates important for their specific weather patterns.

## V. CONCLUSIONS

It is becoming increasingly common for governments, industries, and other organisations to use downscaled climate data for modelling, planning, and adaptation purposes. Most of downscaled products easily available do a poor job at capturing climatic extremes, which are arguably the most important. Deep learning downscaling is a promising method for improving this challenge, as it provides a computationally efficient way of downscaling LR model output to convection permitting scales, thus better capturing extremes. While substantial research has occurred in this field recently, deep-learning downscaling has not been used in a large operational setting. This paper addresses some of the challenges inherent in applying GAN based downscaling operationally. We show that the stochastic GAN framework can be extended to a suite of important variables, that including HR covariates increases accuracy, and that while the framework can be applied to a different region, there are some challenges related to large-scale bias between HR and LR fields. A final step required for operational GANs will involve overcoming the computational challenges linked to training on large spatial regions. Tiling methods, which have been applied in other deep-learning and computer vision settings, will be an important avenue of future research. Hopefully, GAN downscaling will soon be an important tool for climate adaptation.

## REFERENCES

- [1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

- [2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
- [3] N. J. Annau, A. J. Cannon, and A. H. Monahan, "Algorithmic hallucinations of near-surface winds: Statistical downscaling with generative adversarial networks to convection-permitting scales," *Artificial Intelligence for the Earth Systems*, vol. 2, no. 4, p. e230015, 2023.
- [4] K. Daust and A. Monahan, "Capturing climatic variability: Using deep learning for stochastic downscaling," arXiv preprint arXiv:2406.02587, 2024.
- [5] L. Harris, A. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, "A generative deep learning approach to stochastic downscaling of precipitation forecasts," arXiv preprint arXiv:2204.02028, 2022.
- [6] J. Leinonen, D. Nerini, and A. Berne, "Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7211–7223, 2020.
- [7] I. Price and S. Rasp, "Increasing the accuracy and resolution of precipitation forecasts using deep generative models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 555–10 571.
- [8] Y. Li, Z. Li, Z. Zhang, L. Chen, S. Kurkute, L. Scaff, and X. Pan, "High-resolution regional climate modeling and projection over western canada using a weather research forecasting model with a pseudo-global warming approach," *Hydrology and Earth System Sciences*, vol. 23, no. 11, pp. 4635–4659, 2019.
- [9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.
- [10] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.
- [11] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.
- [12] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
- [13] B. Kumar, K. Atey, B. B. Singh, R. Chattopadhyay, N. Acharya, M. Singh, R. S. Nanjundiah, and S. A. Rao, "On the modern deep learning approaches for precipitation downscaling," *Earth Science Informatics*, vol. 16, no. 2, pp. 1459–1472, 2023.