Deep Learning Downscaling of Climate Variables to Convection-permitting Scales

by

Kiri Shea Daust
B.Sc., University of Victoria, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the School of Earth and Ocean Sciences

Deep Learning Downscaling of Climate Variables to Convection-permitting Scales

by

Kiri Shea Daust
B.Sc., University of Victoria, 2022

Supervisory Committee

Dr. A. Monahan, Supervisor
(School of Earth and Ocean Science)

Dr. C. Mahony, Committee Member
(Ministry of Forests)

Dr. A. Cannon, Committee Member
(School of Earth and Ocean Science)

Supervisory Committee

Dr. A. Monahan, Supervisor
(School of Earth and Ocean Science)

Dr. C. Mahony, Committee Member
(Ministry of Forests)

Dr. A. Cannon, Committee Member
(School of Earth and Ocean Science)

ABSTRACT

Adapting to the changing climate requires accurate local climate information, a computationally challenging problem. Recent studies have used Generative Adversarial Networks (GANs), a type of deep learning, to learn complex distributions and downscale climate variables efficiently. Capturing variability while downscaling is crucial for estimating uncertainty and characterising extreme events—critical information for climate adaptation. Since downscaling is an undetermined problem, many fine-scale states are physically consistent with the coarse-resolution state. To address this ill-posed problem, downscaling techniques should be stochastic, able to sample realisations from a high-resolution distribution conditioned on low-resolution input. Previous stochastic downscaling attempts have found substantial underdispersion, with models failing to represent the full distribution. I propose approaches to improve the stochastic calibration of GANs in three ways: a) injecting noise inside the network, b) adjusting the training process to explicitly account for the stochasticity, and c) using a probabilistic loss metric. I tested models first on a synthetic dataset with known distributional properties, and then on a realistic downscaling scenario, predicting high-resolution wind components from low-resolution climate covariates. Injecting noise, on its own, substantially improved the quality of conditional and full distributions in tests with synthetic data, but performed less well for wind

field downscaling, where models remained underdispersed. For wind downscaling, I found that adjusting the training method and including the probabilistic loss improved calibration. The best model, with all three changes, showed much improved skill at capturing the full variability of the high-resolution distribution and thus at characterising extremes.

Investigating the stochastic GAN framework with other variables, I show that it successfully downscales temperature, specific humidity, and precipitation. I also find that the stochastic framework substantially improves the downscaling of extreme precipitation. Next, I find that while multivariate downscaling can improve dependence structures between downscaled variables, it leads to blurry downscaling of individual variables. I demonstrate that including high-resolution topography as an input improves spatial structure for most variables. Finally, I test the generalisability of the GAN framework to a new location with a different climate, and show that while the GAN performs well for temperature and humidity, it fails for precipitation due to mismatches between the low- and high-resolution data. These results represent important techniques and insights towards operational GAN-based downscaling.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

This thesis would not exist without help from a number of extraordinary people. First, a huge thank you to my supervisor, Dr. Adam Monahan, for being so supportive, having excellent ideas, and being a great teacher. Thank you to Nic Annau, who started the GAN downscaling project, for excellent discussions, lending me his code, and setting everything up for the GANs! I look forward to more collaboration. Also, I am very grateful to Dr. Colin Mahony, not only for his supervision as part of the committee, but for making it possible to continue this research while working with the Ministry of Forests. And finally, a huge thank you to my family; my parents, Dr. Karen Price and David Daust for support, discussion, and getting me interested in science to begin with, my partner, Chloe Leroy, for her constant support and motivation, and of course Lucy, for her cuddles and fluffiness. This research was funded by the BC Ministry of Forests' ClimatEx project, the BC Graduate Scholarship Award, and the UVic Graduate Awards.

DEDICATION

To Mummy, for always encouraging curiosity and teaching me statistics!

Chapter 1

Introduction

I grew up in a family of ecologists and modellers who frequently use gridded climate data as a tool for ecological modelling and decision making. Climate data is used as an input to nearly all ecological models, but it is often taken for granted. While there are a range of types, scales, and qualities of climate data, I have found accessing accurate and local-scale climate data to be a consistent challenge. In our changing climate, reliable climate information is crucial.

There are two main sources of climate information for local adaptation: observational data (usually from weather stations), and climate model output [10]. Earth System Models (ESMs), resolve large-scale physical processes across the globe in grids of the surface and atmosphere. These large-scale models can be combined with station observations and remote sensing data to correct model biases, and create reanalysis datasets, providing good information for historic climates [15]. ESMs can also be used to project future climates, crucial for planning and decision-making, However, ESMs operate at too large a spatial scale to be directly applicable to most local-scale projects, as they do not resolve convective processes. Numerical climate models can be run at higher spatial and temporal resolutions to resolve finer-scale processes. Convection permitting models are an important class of higher resolution numerical weather models, which use a spatial resolution of 3-6 km and can resolve some large convective processes [37]. However, due to their computationally complexity, convection-permitting models are not generally applicable to large regions or time scales. Thus, downscaling of ESM output to a higher resolution is often necessary.

Climate downscaling is challenging: low-resolution (LR) inputs contain insufficient information to fully specify a downscaled output; i.e., there are many potential high-resolution (HR) climate states consistent with a given set of LR information,

leading to an underdetermined problem [34]. Nevertheless, various methods of climate downscaling have existed for decades. While some methods perform well for certain applications, existing techniques are either a) highly computationally expensive, b) perform poorly at projecting extremes, or c) show limited success for variables with spatially inhomogeneous statistics, such as precipitation [7]. Recently, research teams have developed new downscaling approaches using artificial intelligence; these models potentially perform much better for downscaling extremes and non-stationary variables, while being computationally feasible for large areas [e.g., 3, 14, 31]. When I learned about an opportunity to be involved in this research, I was immediately interested.

Over the past two years, my Masters' research has focused on downscaling climate variables in British Columbia, Canada from LR reanalysis data (about 30-km resolution) to HR convection permitting scales (4-km resolution) using deep learning, a form of artificial intelligence which uses many-layered artificial neural networks to learn complex patterns. Specifically, I use a Generative Adversarial Networks (GAN), an algorithm which has been well used in computer vision fields, and was recently adopted by the climate downscaling community. This research follows that of Annau et al. [3], who developed a deterministic GAN framework for successfully downscaling wind components to convection permitting scales.

My research investigates using a probabilistic framework for downscaling, where the GAN attempts to learn the distribution of HR downscaled variables conditional on (i.e., consistent with) a set of LR climate fields, and then sample multiple realisations from the HR distribution for each set of LR fields. Since downscaling is an underdetermined problem, it makes sense to consider it in a probabilistic framework. Not only does the probabilistic framework allow quantification of downscaling uncertainty, but sampling the tails of the distributions should obtain better estimates of extremes.

In **Chapter 2**, I develop a stochastic GAN framework which allows the model to sample multiple realisations from the HR distribution. I first investigate the ability of the model to sample representative realisations using synthetic data, and then apply the model to a realistic wind downscaling scenario. My results show that the stochastic GAN framework can both quantify variability and improve estimates of extremes for wind components.

Chapter 3 follows directly from Chapter 2, investigating questions needed to apply the stochastic GAN framework to operational downscaling. Specifically, I extend

the stochastic GAN to temperature, specific humidity, and precipitation, and test its performance in different geographic locations. My results suggest that the stochastic GAN framework could likely be applicable to operational downscaling projects, although challenges with mismatched LR and HR training data will require more consideration.

Chapter 2 is under review for publication in Artificial Intelligence for Earth Systems, and we intend to submit the contents of Chapter 3 to the same journal. This research is part of the ClimatEx program, a collaborative project funded by the BC Ministry of Forests, with the goal of improving the quality of and access to downscaled climate variables over BC and North America. I have been fortunate to work with colleagues at UBC, UVic, and the provincial government as part of this research.

Chapter 2

Capturing Climatic Variability: Using Deep Learning for Stochastic Downscaling

The contents of this chapter are under review for publication in Artificial Intelligence for Earth Systems as *Capturing Climatic Variability: Using Deep Learning for Stochastic Downscaling* (Daust and Monahan, 2024).

2.1 Introduction

As society tries to adapt to Earth's changing climate, access to accurate, local-scale climate information is essential. Earth System Models (ESMs) provide state-of-the-art projections on a global scale, but provide insufficient spatial resolution for regional analyses. Thus, downscaling - creating high-resolution climate information from low-resolution, large-scale data - is an important practical tool. Spatially and temporally high resolution fields of climate data over large scales are important for many applications, including simulation of extreme events at local scales [e.g. fires, floods, storms; 10], local planning, and making climate informed ecological decisions such as tree species selection [26]. Hence, accurate and computationally efficient downscaling methods are crucial for local climate adaptation.

Strategies for downscaling coarse-resolution climate model simulations to regional scales can be broadly classified into dynamic downscaling and statistical downscaling. Dynamic downscaling employs a limited-area numerical climate model to resolve fine-

scale features, driven by large-scale weather patterns from the low resolution ESM [37]. Dynamic downscaling can capture complex spatial patterns at smaller scales, and can provide high-accuracy downscaling. However, it is highly computationally intensive, and can only be used for relatively short time periods. Statistical downscaling develops statistical relationships between low-resolution (LR) and high-resolution (HR) climate variables. Statistical downscaling can be applied either at individual points, often using a combination of bias correction and transformation of statistical moments [27], or can be used to downscale entire fields. Common techniques for the latter include parametric approaches (e.g., Gaussian Process/kriging), where covariances are specified to allow analytic solutions. While these approaches have shown some success, climate-field downscaling methods make strong assumptions about the distribution and homogeneity of statistics, which are often not satisfied [7]. Also, many current methodologies struggle to accurately downscale spatially complex variables (i.e., variables with non-linear dependence on elevation or spatially heterogeneous dependence structures) and capture extremes [14]. Recently, a new strategy for statistical downscaling over climate fields has been developed that uses deep (i.e., many layered) learning algorithms to learn a mapping from LR to HR paired climate fields [3]. It has been found that this strategy can produce downscaled fields with much higher accuracy than traditional statistical downscaling, and does not require the prohibitive computation required for dynamic downscaling.

Downscaling is intrinsically an underdetermined problem with a distribution of possible HR realisations physically consistent with any given LR input [2]. This is especially true since weather is sensitive to initial conditions: minute differences can result in drastically different outcomes due to development of internal variability [25]. Stochastic Weather Generators, which attempt to sample from the distribution of weather states, have been used to try and account for this variability [44]. Being able to capture the full variability of a downscaling problem is crucial for quantifying uncertainty, and for characterising extreme events. Ideally, downscaling techniques would allow sampling from the HR distribution, conditioned on the LR input.

Deep learning methods are a promising approach for such distributional downscaling problems. Generative Adversarial Networks (GANs) have been successful in various generative AI fields, especially computer vision applications [13]. GANs generally use two separate deep neural networks during training: a Generator network which is given input and attempts to create plausible counterfeits of the training data; and a discriminator or Critic network which is provided training data mixed

with Generator output and attempts to distinguish between the counterfeits and the real data. During training, the two networks play a minimax game: the Generator tries to improve its output to "fool" the discriminator and the discriminator tries to improve its ability at distinguishing between real and generated samples. In the last few years, GANs have been introduced to deep-learning-based downscaling and have shown success in drawing realizations from high-dimensional non-Gaussian distributions with complicated dependence structures . Conditional GANs, developed by Mirza and Osindero [30] allow the GAN to draw realisations from distributions, conditioned on covariates.

Much of the development of GANs for climate downscaling builds on work from the computer vision field of super resolution. Most studies in computer vision use conditional GANs, where the networks are provided LR information and learn to sample from the HR conditional distribution [19]. With the introduction of GANs to climate downscaling, difficulties with instability during training [42] were improved by the introduction of the Wasserstein GAN [WGAN; 4]. In the WGAN, instead of using a Discriminator network to estimate the probability of individual realisations being real, a Critic network estimates the Wasserstein distance between the true HR distribution and the generated distribution. Intuitively, the Wasserstein distance is a critic score, specifying the quality of the downscaled fields compared to the training data. Not only does this substantially improve stability during training, but for downscaling, it conceptually makes sense to focus on convergence in distribution of generated and truth fields.

The initial formulation of (unconditional) GANs used a stochastic approach where the only input to the Generator was Gaussian noise, generating different realisations for each different noise input. With the development of conditional GANs [30, 19], and the subsequent Super Resolution GAN (SRGAN) and Enhanced Super Resolution GAN (ESRGAN) frameworks from the field of super resolution, the noise input was replaced by the conditioning fields, leading to a semi-deterministic network. In this setting, each trained Generator would still draw a realisation from the conditional distribution, but it would always draw the same realisation for each set of conditioning fields (theoretically, one could draw a different realisation by training a new model). A few studies successfully used variations of this architecture for downscaling climate fields; for example, Stengel et al. [38] adapted the SRGAN model to create very high-resolution downscaled wind fields by first downscaling to a moderate resolution, and then further downscaling to the final HR. Recently, studies have investigated meth-

ods for allowing explicit stochasticity in conditional GANs, usually using variations of adding noise covariates, stacked with the LR conditioning fields. Price and Rasp [31] concatenated a noise layer part way through their Generator network, while Harris et al. [14] concatenated multiple noise inputs with the conditioning information at the beginning of the network. Both studies found that the stochastic results were underdispersive: trained models were unable to capture the full range of variability, often only sampling from the centre of the conditional distribution. Recent advancements in Super Resolution [e.g. nESRGAN+, 32] have improved stochastic calibration, but have not yet adapted it to climate downscaling. Furthermore, many downscaling studies using stochastic GANs have focused their analyses on image quality; to our knowledge, no research in climate downscaling has fully investigated the ability of stochastic GANs to learn and sample from the conditional HR distribution.

We aim to fill this gap by improving stochastic GAN frameworks for climate downscaling. Most of this work builds on Annau et al. [3]. While their model showed success for downscaling wind fields, it was not stochastic as each model only produced one output. We use a similar model architecture as Annau et al. [3] adapted for full stochasticity. An obvious challenge with testing distributional quality in a downscaling setting is the lack of a truth conditional distribution, as in most applications we only have access to a single truth realisation for each timestep. To address this challenge, we first consider an idealised approach based on synthetic data with known distributional properties. Based on these experiments, we test a "noise injection" method, where hundreds of noise fields, at different spatial resolutions, are injected into the latent layers of the network. This approach provides excellent stochastic calibration on the synthetic data. We then test our modification on a real-world downscaling problem, predicting HR wind components from LR conditioning data. Challenges with underdisperison on the wind data lead to development of an updated loss function using a probabilistic error function and modification of the training method to fully utilise the stochasticity. Our final model is successful at capturing variability, and improves estimates of moderate extremes.

2.2 Methods

All models in this paper use the same basic super-resolution structure. We train the models on paired sets of LR conditioning fields (covariates) and HR truth fields. The GAN then learns a mapping to the HR fields from the input covariates. For

consistency, we keep the same resolution and size of fields across all models: HR fields are 128×128 pixels, and LR fields are 16×16 pixels, resulting in a downscaling factor of eight.

2.2.1 Data

Synthetic Data

Evaluating the distributional quality of stochastic realisations on realistic downscaling problems is challenging, as there is rarely more than one realisation of the "ground truth" to compare with for a given sample of the conditioning field. While certain metrics can provide information on model calibration, it is often not possible to directly compare conditional distributions. We thus created a simple synthetic dataset with known distribution properties. To make the HR fields, we added a mean zero Gaussian field with a specified covariance structure to a specified non-stationary mean (exponential in one axis and sigmoidal in the other). This sum was then squared to generate a field with pointwise chi-square marginal distributions. That is, we drew realisations from r_{ij} where

$$Y \sim N(\vec{0}, \Sigma) \tag{2.1}$$

$$s_{ij} = \frac{5e^{x_i}}{1 + e^{-8y_j}} + Y_{ij} \tag{2.2}$$

$$r_{ij} = s_{ij}^2 \tag{2.3}$$

where r is the output HR field, x and y are the axis values, and Σ is a 128 x 128 covariance matrix with correlations decreasing linearly to zero over four pixels along both directions. To ensure large scale structure varied between samples, we randomly scaled and reflected the x and y axis, as follows: for each field, we drew realisation of random variables A_1, A_2, B_1, B_2 from a uniform distribution over -1,0,1, such that $A_1 \neq A_2$ and $B_1 \neq B_2$ and then rescaled x and y as

$$x_n = A_1 + \frac{n(A_2 - A_1)}{128}, n \in \{0, 1, ..., 128\} \tag{2.4}$$

$$y_n = B_1 + \frac{n(B_2 - B_1)}{128}, n \in \{0, 1, ..., 128\}. \tag{2.5}$$

To create the LR input fields, we spatially averaged 8 x 8 regions of the HR fields.

GANs commonly struggle to capture multi-modal distributions and tend to con-

verge on the conditional mean, a phenomenon known as mode collapse [e.g., 35]. To test the ability of our GANs to learn multi-modal distributions, we generated a second set of realisations of bimodal fields using a Gaussian Mixture Model where δ is a Bernoulli random variable:

$$X \sim N(\vec{5}, \Sigma) \tag{2.6}$$

$$Y \sim N(\vec{1}, \Sigma) \tag{2.7}$$

$$\delta \sim B(n = 1, p = 0.35) \tag{2.8}$$

$$s_{ij} = \frac{5e^{x_i}}{1 + e^{-8y_j}} + X_{ij}^{\delta} \cdot Y_{ij}^{1-\delta} \tag{2.9}$$

$$r_{ij} = s_{ij}^2. \tag{2.10}$$

For all synthetic data experiments, training used 5000 pairs of fields; 2000 additional pairs were reserved for testing. To compare marginal distributions, we created a set of 500 fields with the same large-scale spatial pattern (i.e. same x and y scale and rotation), so that the only difference was the added Gaussian/mixture field. These could then be interpreted as ensembles of truth realisations given the same conditioning field, and were used to test generated pixel-wise marginal distributions. Note that since all LR input fields were create by coarsening the HR fields, these datasets result in pure super resolution setups, with no need to bias correct the LR inputs.

To investigate the effects of spatial complexity on the generated fields, we created three further synthetic datasets with three different levels of spatial heterogeneity. We used a field of complex topography from the south-coast of British Columbia, and added the same Gaussian field described above. To vary spatial complexity, we scaled the topography with 3 different weights. That is,

$$s_{ij} = wZ_{ij} + Y_{ij}, w \in \{0.1, 1, 10\} \tag{2.11}$$

$$r_{ij} = s_{ij}^2 \tag{2.12}$$

where Z is the topography field, scaled to zero mean and unit standard variance. We used weights of 0.1 (field dominated by original dataset) for low heterogeneity, 1 for moderate heterogeneity, and 10 for high heterogeneity (dominated by added topography).

Convection-Permitting Regional Model Case Study

Since this research aims to improve deep-learning based downscaling of climate data, it is important to test results on more realistic settings. Here, we modelled HR zonal (u) and meridional (v) wind components using LR wind components, pressure, temperature, and HR topography. Our architecture follows that of Annau et al. [3] with the exceptions that: (i) HR topography is included as a covariate in the Generator, and (ii) all covariates are also passed to the Critic. Wind is an important climate variable for various applications, but it is often challenging to model due to having complex mesoscale patterns. We consider a square region covering the coastal mountains, in southwestern Canada (49° to 53° N, 122° to 126° W), as its high degree of topographic complexity represents a realistically challenging downscaling scenario. HR wind fields were obtained from WRF runs produced for the WCA (Western Canada) simulation driven by ERA-Interim [21], which contains hourly data for a 14 year period. LR covariates were from ERA5 [15], a state of the art global reanalysis product. While ERA-Interim was used to drive the WRF model at the boundary conditions, both ERA5 and ERA-Interim represent the same realisation of the climate system, so it is reasonable to use ERA5 in the paired LR data.

This HR and LR pairing represents a practical application of downscaling, where the LR and HR fields are from different models [as in 3]. Many previous studies in deep-learning based downscaling have used idealised pairings, where the LR fields are created by coarsening the HR fields, resulting in perfectly matched pairing (note that this was the approach we used for the synthetic data experiments). As a consequence of natural internal variability, some of the meteorological features on scales common to both resolutions will differ between the LR and HR fields, so our model has to account for such differences, in addition to downscaling.

To preprocess the data, we first transformed the WRF fields to the ERA5 projection, and then remapped the HR fields to the specified downscaling factor. We then standardised the data to mean zero and unit variance across time and space, and within each covariate, as a standard normalization technique in machine learning studies [3]. As the standardisation occurs across all timesteps, the diurnal and seasonal patterns are preserved. Finally, we selected three apparently unexceptional years (no El Niño-Southern Oscillation (ENSO) or evident seasonal extreme wind events) for training (2003, 2008, and 2013), and two years as an unseen test set (2005 and 2012). Hourly data over three years resulted in 26304 samples; sample number

was limited by computational constraints. Data were randomly shuffled during the training process. All models were trained on a single NVIDIA RTX 4090 GPU with 24 Gb VRAM.

2.2.2 Model

This work utilizes conditional GANs [30], which have shown success at learning the mapping between low resolution variables and the desired output variables. Specifically, we use the Wasserstein Conditional GAN formulation [4], where the Critic network learns to estimate the Wasserstein distance between the high-dimensional distributions of the generated and true fields ($\mathbb{P}_\partial$ and $\mathbb{P}_r$ respectively).

Architecture

Most GAN network architectures employ dense convolutional blocks, which have been shown to be excellent at extracting representative features from images. Our network architecture is based on the Enhanced Super-Resolution GAN [ESRGAN; 43] setup, using Residual in Residual Dense Blocks (RRDB) as the main convolutional blocks in the Generator. RRDBs, introduced by Zhang et al. [45] employ densely connected convolutional layers to extract important features, while also maintaining direct connections from previous layers to all convolutional layers to create a contiguous memory. Specifically, we adapt the architecture employed in Annau et al. [3] to allow noise input and multiple covariate streams, as follows.

Unlike other applications of super resolution, climate downscaling often has access to pertinent HR information during training. A common example of such HR information is topography, which influences local climate strongly. While many previous studies have included topography as a covariate, it has often been input at low resolution with climate covariates, discarding potentially useful information. We therefore created a Generator architecture which allowed us to fully utilise the HR covariates. Using a strategy similar to Depthwise Separable networks [16] we created two input streams within the Generator, one stream for each resolution. Each stream has equivalent convolutional blocks applied in parallel, and after the LR stream has passed through the upsampling blocks to increase the resolution, the two streams are concatenated and passed through a final convolutional block (figure 2.1).

In the standard super-resolution formulation, the Critic network is only given samples of the predictands (either generated or from the training data). However,

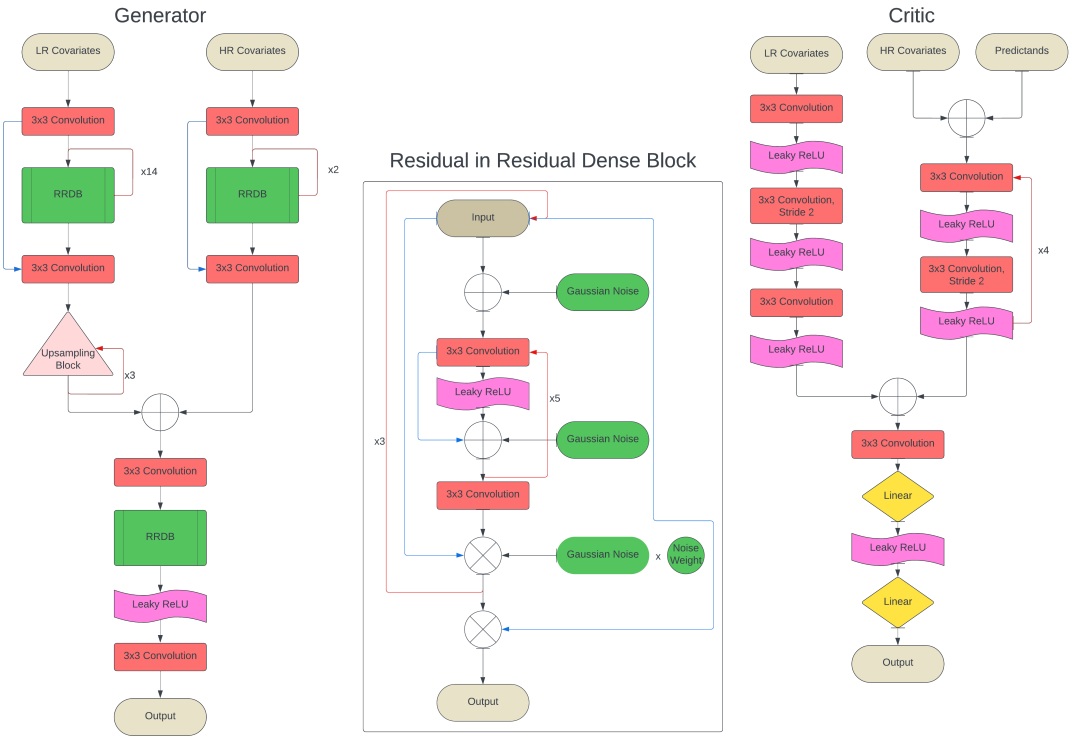

Figure 2.1: Architecture of GAN networks showing Residual in Residual Dense Block (RRDB) with noise injection. Green denotes locations where noise is added into the network. Rectified Linear Units (ReLU) are used to introduce non-linearity.

Harris et al. [14] found that passing all available covariates to the Critic network can improve its predictions, and in the Wasserstein GAN formulation, it is important that the Critic network is able to make relevant estimates of the Wasserstein distance. Given that we want the Generator to sample from a conditional distribution, we thus want the Critic to make use of conditioning information when quantifying this distribution. Similarly to the Generator, we adapted the Critic network to include a separate input stream for the LR covariates, which is concatenated after the HR stream has been downsampling via strided convolution (figure 2.1).

Noise Injection

To improve the model's ability to sample across the entire range of the conditional distribution, we adjusted the Generator architecture to inject noise directly into the latent representations produced by the convolutional layers. Specifically, we based

our approach on nESRGAN+ [32], and concatenate uncorrelated, mean zero, unit variance Gaussian noise fields with the latent layers inside each Dense Block (figure 2.1). With our architecture, this leads to six noise injection instances in each Residual Dense Block, and 18 noise injections in each Residual in Residual Dense Block (RRDB). Our full noise Generator contains 14 RRDB in the LR input stream, 2 in the HR input stream, and one after concatenation, resulting in 252 LR noise layers and 54 HR noise layers.

To test the effect of the number of noise injection layers, we replaced some of the stochastic RRDB with deterministic RRDB (i.e., RRDB with no noise injection). Altogether, we tested three different levels of noise injection: low (2 stochastic RRDB-LR, 0 stochastic RRDB-HR), moderate (7 stochastic RRDB-LR, 0 stochastic RRDB-HR), and full noise (14 stochastic RRDB-LR, 2 stochastic RRDB-HR). As a baseline model, we also considered the more standard noise-covariate approach [similar to that used in 20], for which a Gaussian noise field is concatenated with the LR input covariates before passing through the Generator network.

Losses

In Wasserstein GAN training, the Critic tries to maximise the difference in the distributional distance between true fields and generated fields, while the Generator attempts to minimize this distance (i.e., it attempts to make it difficult for the Critic to distinguish the generated fields from the training data). Previous studies have found that solely relying on the Critic loss (adversarial loss) for the Generator training can lead to instability, such that the training process does not converge [43]. Here, we use the common approach of adding an extra content loss term to the Generator - a pixel-wise error metric between the training data and the generated fields, intended to aid convergence at large scales.

We experimented with two different content loss functions: mean absolute error (MAE), and cumulative rank probability score (CRPS). MAE is commonly used as a content loss in deep learning, but is a deterministic measure. CRPS is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \Big(F(y) - H(y - x) \Big)^2 dy \tag{2.13}$$

where F represents the CDF of the predicted distribution, H is the Heaviside function, and x is the ground truth value. The CRPS metric returns lower values to

distributions whose mass is centred around the ground truth value. In a deterministic setting, the PDF of y is a delta functtion, and the CRPS reduces to MAE, so is appropriate to use as a content loss. To apply the CRPS, we calculated an empirical CRPS metric for each pixel, using the ground truth value and values from the stochastic sampling realisations, and then took the mean across pixels.

Training

Since our goal is to sample realisations from the entire conditional distribution of HR fields, we do not want the loss function to force the Generator to create copies of the training data - we consider the ground truth as one realisation of the conditional distribution. Ideally, we would expect generated realisation to have similar statistics and large-scale features as the training data, but not be identical. Over-reliance on content loss (particularly deterministic measures) can degrade model performance, as it overly penalises small deviations in feature location/presence. In situations where features are spatially shifted, pixel wise error metrics such as the content loss will penalize the model twice: once for the feature not occurring where it does in the ground truth, and once for the feature occurring where it is not in the ground truth. This double-penalty problem is a well-known issue with pixel-based losses in generative networks [34], and results in overly blurry output, where the model converges on the conditional median, thus suppressing small-scale features and extremes [3]. While the adversarial loss in GANs (in our case the Wasserstein distance) is not a pixel-wise metric and does not constrain the network in the same way, the use of content losses can suppress variability.

We considered two training techniques in our models to address the double penalty problem while rewarding convergence at large scales: frequency separation, and stochastic sampling. Frequency separation, introduced by Annau et al. [3], coarsens HR fields to only provide low frequencies to the content loss, whereas the full HR fields are used for the adversarial loss. Theoretically, this approach allows the model to more freely develop high frequency patterns by removing some of the content loss constraints. Stochastic sampling is an approach modified from Harris et al. [14], where multiple stochastic realisations are passed to the content loss, which uses an ensemble metric to assess calibration (Alg 1). In each Generator training step, we generate six stochastic realisations of each field in the batch, and pass these to the content loss. In the MAE loss case, this approach averages over generated fine scale

features and only applies the content loss on the patterns which are expected to be consistent across realisations. Both frequency separation and stochastic sampling allow generated variability at small scales but encourage convergence at large scales.

With the stochastic sampling method, our Generator loss function is

$$L_G = \underbrace{-\mathbb{E}_{G(x)\sim\mathbb{P}_g} C[G(x)]}_{\text{Adversarial Loss}} + \underbrace{\alpha\mathbb{E}_{y\sim\mathbb{P}_r} l_c[y, G(x)_1, G(x)_2, ..., G(x)_6]}_{\text{Content Loss}} \tag{2.14}$$

where x represents LR conditioning fields, y is the HR ground truth, α is the content loss weighting, l_c is the content loss function, and G and C represent the Generator and Critic network respectively. Note that an ensemble of i stochastic realisations are passed to the content loss.

Algorithm 1 Pseudocode for Stochastic Sampling algorithm, using CRPS metric. Note that we chose six stochastic realisations as the maximum number that fit in GPU memory.

$nRealisation \leftarrow 6$
for $i \in 1 : nBatch$ **do**
  $LRBatch = LRBatches_i$
  $HRBatch = HRBatches_i$
  $adversarialLoss \leftarrow -\text{mean}(\text{Critic}(LRBatch))$
  for $j \in 1 : nSample$ **do**
    $subBatch \leftarrow \text{repeat}(LRBatch_j, nRealisation)$
    $subRealisations \leftarrow \text{Generator}(subBatch, invariant)$
    $CRPSBatch_j \leftarrow \text{mean}(\text{pixelwiseCRPS}(subRealisations, HRBatch_j))$
  end for
  $contentLoss \leftarrow \text{mean}(CRPSBatch)$
  $loss \leftarrow adversarialLoss + contentLoss$
end for

Throughout this paper, we will use the following naming conventions to specify models:

$$\text{Training}^{\text{Content Loss}}_{\text{Noise Level}},$$

where Training is represented respectively by F or S for frequency separation and stochastic sampling, and noise level is represented by NC for noise covariate, and low, $medium$, and $full$ for noise injection. For example, S^{CRPS}_{full} specifies a model using stochastic sampling, the CRPS content loss, and full noise injection. Note that we have not investigated all combinations of Generator parameters, as some only make sense in combination or in specific settings. Model training parameters are presented

in table A.1.

2.2.3 Validation

Quality assessment in image generation problems often poses a challenge, because there are multiple, often competing, metrics that could be used. Potential metric priorities include convergence in realization (pixelwise or at large scales), or convergence in statistical features such as spatial covariance or pixelwise marginal distributions. In general, we will consider a combination of these factors, depending on the problem. As noted earlier, while deterministic pixel-wise error metrics are important, they should not be relied on too much due to the double-penalty problem. Following from Harris et al. [14], Annau et al. [3] and Ravuri et al. [33], we used a Radially Averaged Spectral Power metric (RASP) for comparing spatial variance of different scales (or alternatively, the covariance structure) between the generated fields and the ground truth. We calculated RASP by first performing a 2-dimensional Fourier transform on each field, averaging the amplitudes within annular rings centred at wavenumber zero, and then averaging power densities across at least 1000 fields. Ideally, spectral power at each spatial scale should be the same in the generated and truth fields; to aid in visual comparison, we standardised amplitudes at each wavenumber by the amplitude of the ground truth field in the corresponding bin. A value below one represents less spectral power at the given wavenumber in the generated field than in the ground truth, and a value above one suggests more spectral power than in the ground truth field. This quantity allows assessment of biases across spatial scales.

To assess the quality of stochastic realisations for a given conditional distribution, we compared pixel-wise marginal distributions (where possible), and used rank histograms. For the synthetic data, we estimated the true marginal distributions using Kernel Density Estimates of 500 realisations sampled from the truth distributions, and compared these pixelwise to the equivalent marginal distributions of 500 stochastic realisations from the trained model. To test pixelwise convergence across the whole domain, we calculated the empirical Kolmogorov-Smirnov statistic (KS) at each pixel and investigated the distribution of these KS statistics for a given model. The KS statistic is defined as

$$D = \sup_x |G(x) - T(x)|$$

where G and T represent the CDFs of the generated and true distributions, respectively.

We calculated rank histograms by generating 96 stochastic realisations of HR fields for each of 50 randomly selected LR conditioning fields, and determined the pixelwise rank of the truth field in the ensemble of generated fields. That is, for each pixel, we calculated

$$k_n = \text{rank}(x, \langle g_1, ..., g_{96} \rangle) \tag{2.15}$$

where x is the truth value for the pixel, and g are the ensemble forecast members. We then used CDFs of histograms to investigate the distribution of the ranks. If the model is well calibrated, the truth field should be indistinguishable from any ensemble member, and so the rank histogram should estimate a uniform distribution, corresponding to a linear CDF. Conversely, if the CDF has more weight at the tails, corresponding to a U-shaped histogram, then the model is underdispersive (most truth points fall outside the range of generated realisations).

The rank histogram described above is computed using the rank of each pixel. However, since calibration of extremes is often most important, we used a modified rank histogram to assess performance with regards to spatial extremes. Here, we randomly selected 400 samples from the test set, generated 100 stochastic realisations of each, and then calculated the 0.999 and 0.001 quantiles across the 16384 pixels in each field. We then produced a rank histogram of true quantiles compared to the 96 generated quantiles, for each sample:

$$m_n = \text{rank}(q(x), \langle q(g_1), ..., q(g_{96}) \rangle) \tag{2.16}$$

where q represents the quantile over the field.

For analyses investigating distributions of extremes, we tested multiple percentiles (0.1 and 99.9; 0.01 and 99.99; and 0.001 and 99.999). These choices allowed us to investigate near tails of the distributions, without having to invoke extreme value theorems. Unless specified, results were similar across all extreme percentiles. Generally, we show 0.01 and 99.99 quantiles as a representation of nearer extremes.

2.3 Results

In this section, we investigate results pertaining to two main classes of distributions: the distribution of HR fields conditioned on LR covariates, and full HR fields across

samples. Consideration of the conditional distribution, $p(\text{HR}|\text{LR})$ allows investigation of the stochastic calibration: given a set of LR covariates, what is the distribution of the HR fields? The full distribution,

$$p(\text{HR}) = \int p(\text{HR}|\text{LR})p(\text{LR})\ d\text{LR} \approx \frac{1}{n}\sum_{k\in\text{LR}} p(\text{HR}|\text{LR}_k) \tag{2.17}$$

represents the full distribution across LR conditioning sets, where n is the number of conditioning sets (in our case, timesteps) being considered. Throughout the results, we focus on four models: F_{nc}^{MAE} (the baseline; partial frequency separation with noise covariates and MAE content loss), F_{full}^{MAE} (partial frequency separation with full noise injection and MAE content loss), S_{full}^{MAE} (stochastic sampling with full noise injection and MAE content loss), and S_{full}^{CRPS} (stochastic sampling with full noise injection and CRPS content loss).

2.3.1 Synthetic Data

Noise Injection

Models with full noise injection performed substantially better overall than the baseline models at matching the true marginal distribution. A representative example of pixelwise marginal distributions for the truth and generated fields is shown in figure 2.2a. The baseline F_{NC}^{MAE} model produced highly underdispersive distributions, while noise injection models were able to match the true distributions well. This result held across all pixels: KS statistics comparing the true marginal distributions with those from noise injection models were significantly smaller than with the baseline model (figure 2.2b). Both S class models had slightly larger median KS statistics than F_{full}^{MAE} , but still showed good distributional matching.

Decreasing the number of noise injection layers in the Generator decreased performance of conditional distributions (figure 2.2b). Low noise injection (F_{low}^{MAE}) showed slight improvement over the baseline model, but still produced underdispersive results. Medium noise injection (F_{med}^{MAE}) had improved statistics, but was still generally underdispersive, and full noise injection (F_{full}^{MAE}) created results with the best agreement of pixelwise marginal distributions.

Rank histograms provide another tool for investigating calibration of conditional distributions, averaged across multiple conditioning fields. Rank histograms showed good calibration for all models with full noise injection, and severe underdispersion for

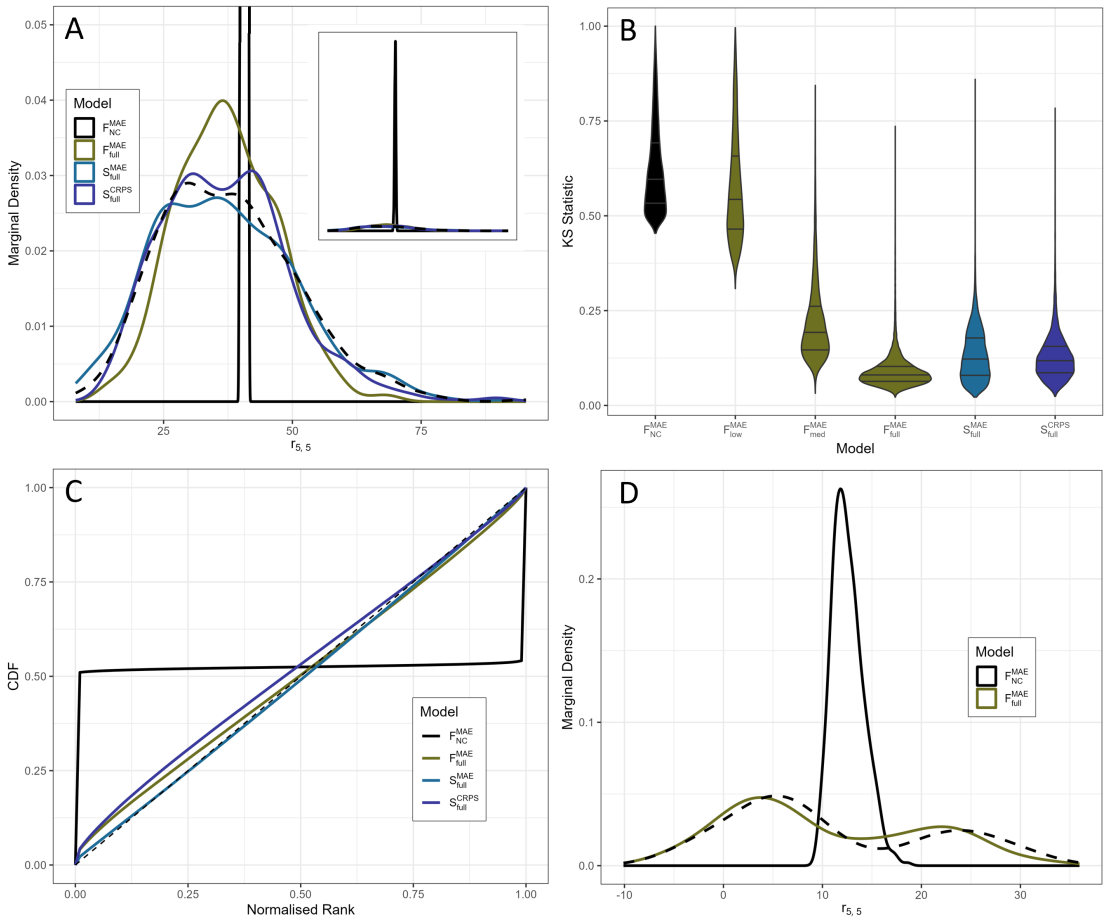

Figure 2.2: a) Kernel density estimates (KDEs) of marginal distributions of $p(\text{HR}|\text{LR})$ for the unimodal synthetic dataset for one example pixel (i = 5, j = 5) for the true distribution and generated distributions. KDEs are based on 500 realisations for a single conditioning field for each distribution. Dashed line shows true marginal distribution. Inset figure shows full y-axis range. b) Violin plot showing KS statistic values comparing generated marginal conditional distributions to ground truth distributions for all pixels. Statistics are calculated for each pixel individually, using 500 realisations of a single conditioning field. Lines show 0.25, 0.5, and 0.75 quantiles, respectively. c) CDF of rank histogram on unimodal synthetic data, with four models, showing calibration of conditional distributions. Dashed line shows reference uniform distribution. Rank histograms were calculated across 100 randomly selected conditioning fields, with 96 HR realisations generated for each. d) KDEs of marginal conditional distributions for one example pixel of a bimodal dataset, comparing true (dashed line) and generated distributions. Distributions were estimated using the same approach as in a).

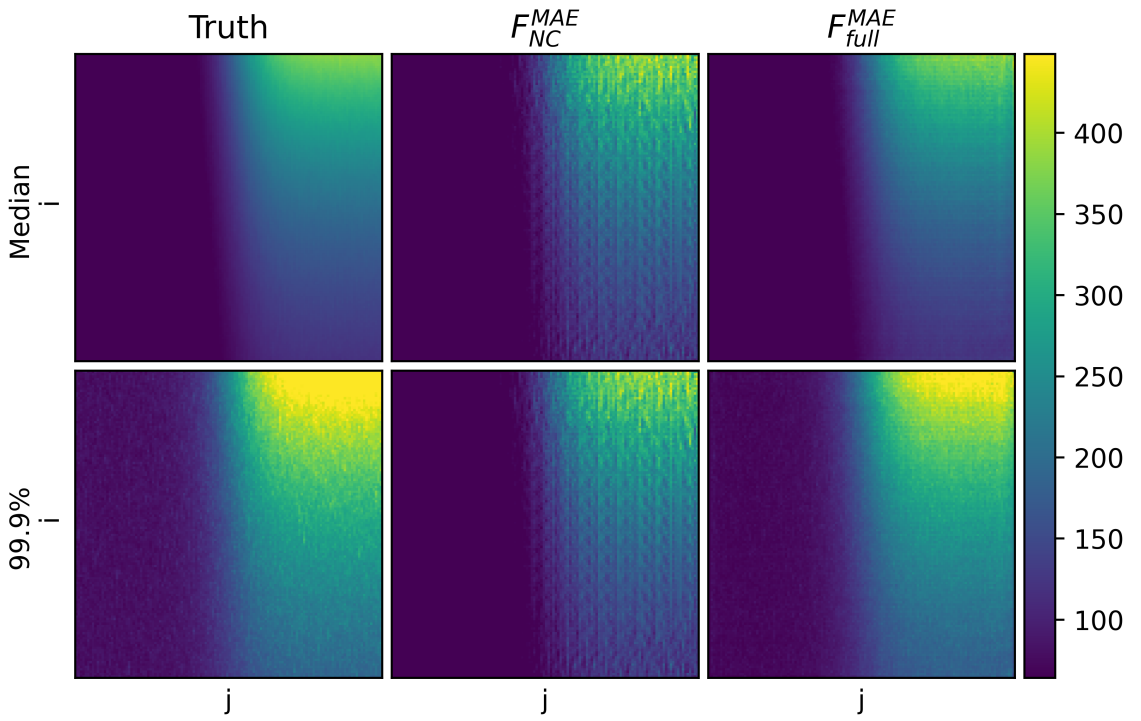

Figure 2.3: Spatial fields of median and 99.9 percentiles of the full distributions across samples for ground truth, and generated data from two models, using the unimodal synthetic dataset (equations 2.1 to 2.3).

the baseline model (figure 2.2c). The S_{full}^{MAE} model showed almost perfect calibration, but all noise injection models performed well.

Learning multimodal distributions is challenging for GANs; they tend to show mode collapse, in which distributions are collapsed to the conditional mean [35]. We found that using the bimodal dataset (equations 2.6 to 2.10), the F_{full}^{MAE} model could learn both modes of the marginal distributions. The baseline models usually showed mode collapse and could not meaningfully recreate any of the marginal distributions (figure 2.2d).

Investigating results of the full distribution, p(HR), statistics of generated fields had fewer artifacts and were closer to the ground truth statistics using the F_{full}^{MAE} model than the baseline model (figure 2.3). Even after training metrics had converged, the baseline model showed noticeable traces of the convolutional filters as checkerboard artifacts, which were not apparent in the F_{full}^{MAE} model. The baseline model also substantially underestimated the 99.9 percentiles, especially at the highest values. These were better captured (although not perfectly) by the F_{full}^{MAE} model.

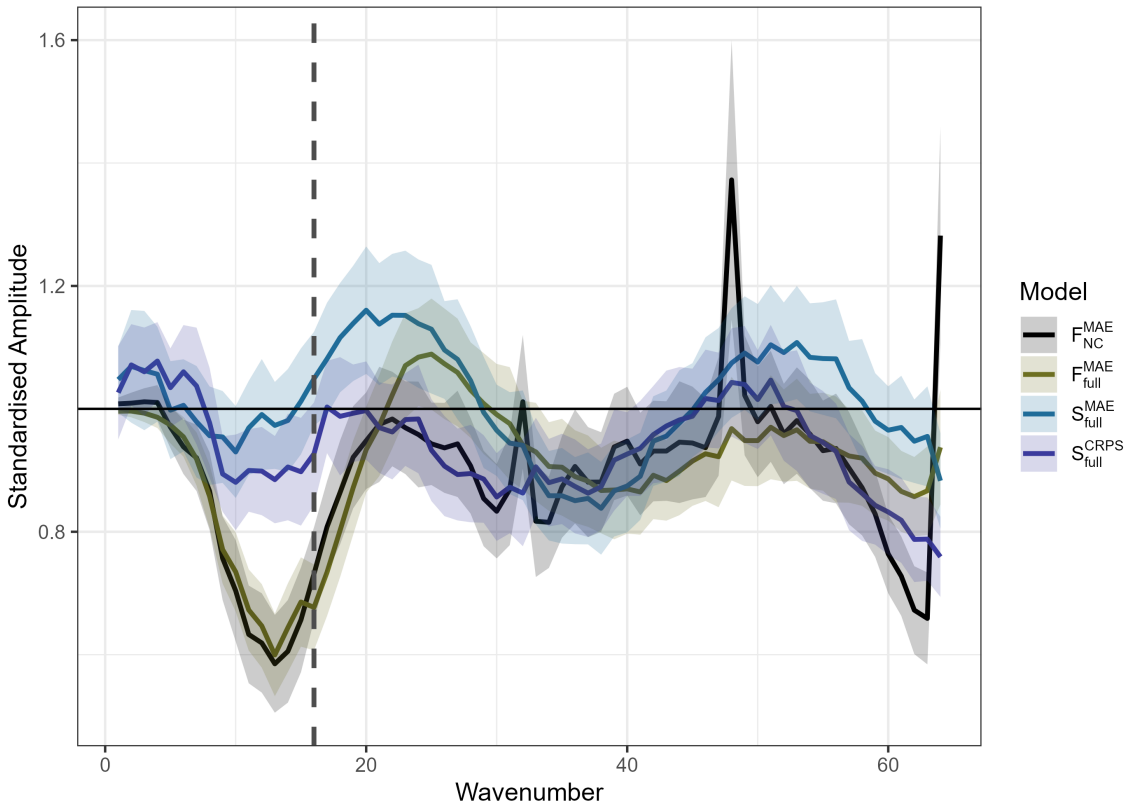

Figure 2.4: Radially averaged spectral power (RASP) for four models. Values are standardised to amplitudes of ground truth wavenumbers, so perfectly matched spectral power occurs at one. Solid lines and shaded regions respectively show mean and +/- one standard deviations across 1200 randomly selected samples. Dashed line indicates wavenumber corresponding to LR pixel size.

As well as better capturing pixelwise marginal variability, all noise injection models performed better at representing covariance patterns. The RASP metrics (figure 2.4) demonstrate that the baseline model showed too little power for a range of low wavenumbers, and then spurious spikes at other wavenumbers. The F^{MAE}_{full} did not show the spikes, but still had a lower power bias at low wavenumbers. Both S class models substantially improved the representation of power at low wavenumbers, and in general were the closest to the true spectral power accross all wavenumbers. There was no obvious difference in spectral power between the two S class models using this metric.

Overall, synthetic data experiments showed that models using full noise injection were substantially better at capturing conditional distributions than the base-

line model. In general, all models with noise injection performed comparably well. There was also noticeable improvement in the quality of the full distributions using noise injection, and the S class models showed improved ability to represent spatial dependence.

2.3.2 Wind Downscaling Case Study

As above, we first present results for conditional distributions, before moving to the full distributions across time. Note that while all models predicted both zonal (eastward) and meridional (northward) wind components, results were generally similar, and unless otherwise stated, we only show results for meridional components.

Even with noise injection, the F class models applied to a realistic downscaling problem produced underdispersive results (figure 2.5). While the F_{full}^{MAE} model showed substantial improvement over the baseline model, it did not fully capture the conditional variability. Both S_{full}^{MAE} and S_{full}^{CRPS} models had better calibration than the F class models; the S_{full}^{CRPS} model, while still showing some underdispersion, performed best among the models considered (figure 2.5).

With the S_{full}^{CRPS} model, individual realisations were visually realistic, and showed noticeable differences in spatial patterns given a single set of conditioning fields (figure A.5). Examples of standard deviation fields showed that the conditional distribution varies substantially with the state of the conditioning fields. Realisations of moderate wind component extremes did not show noticeable bias, and were similar to the ground truth WRF fields. None of the other models performed as well (see A); all models, and especially the F_{NC}^{MAE}, had substantially lower conditional standard deviations. The S_{full}^{MAE} model generated realisations that were overly fuzzy, and the F_{NC}^{MAE} model generated realisations that did not match the WRF fields as well.

Looking at the full distribution, spectral power was also better calibrated with the S class models (figure 2.7). The baseline model produced low spectral power at both low and high wavenumbers; performance of the F_{full}^{MAE} model was better but followed similar patterns and showed a substantial high bias at intermediate wavenumbers. The S_{full}^{CRPS} model generally had the most similar distribution of spectral power to the truth fields, although it showed a modest high power bias at low and high wavenumbers, especially for meridional wind components. All models using the MAE loss function showed low-power bias at high wavenumbers, consistent with the suppression of fine-scale features by this deterministic metric.

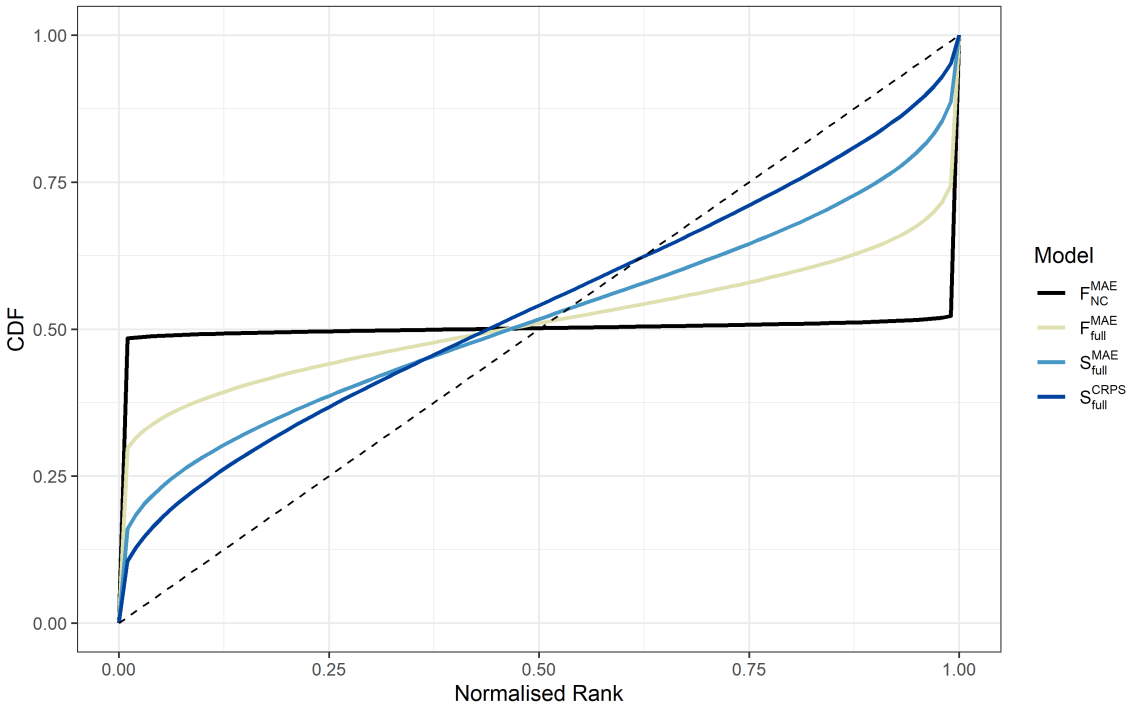

Figure 2.5: CDFs of rank histograms for meridional wind components, using four different models. Rank histograms were calculated across 100 randomly selected conditioning fields, with 96 HR realisations Generator for each. Dashed line shows reference uniform CDF.

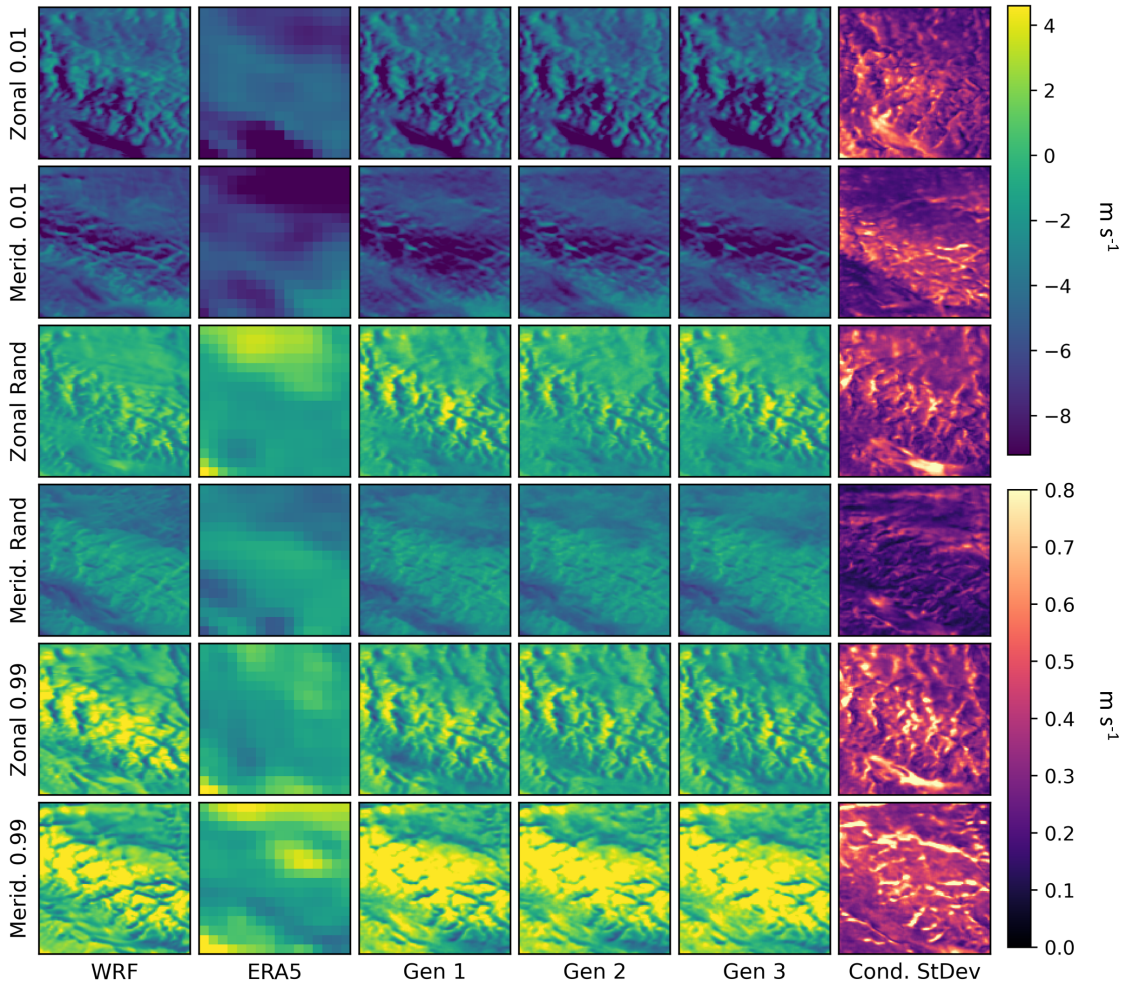

Figure 2.6: Example meridional and zonal wind fields for coastal BC using the S_{full}^{CRPS} model. First two rows show hours representative of 0.01 quantiles, middle two rows show randomly selected hours from the full test set, and bottom rows show hours representative of 0.99 quantiles. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisations.

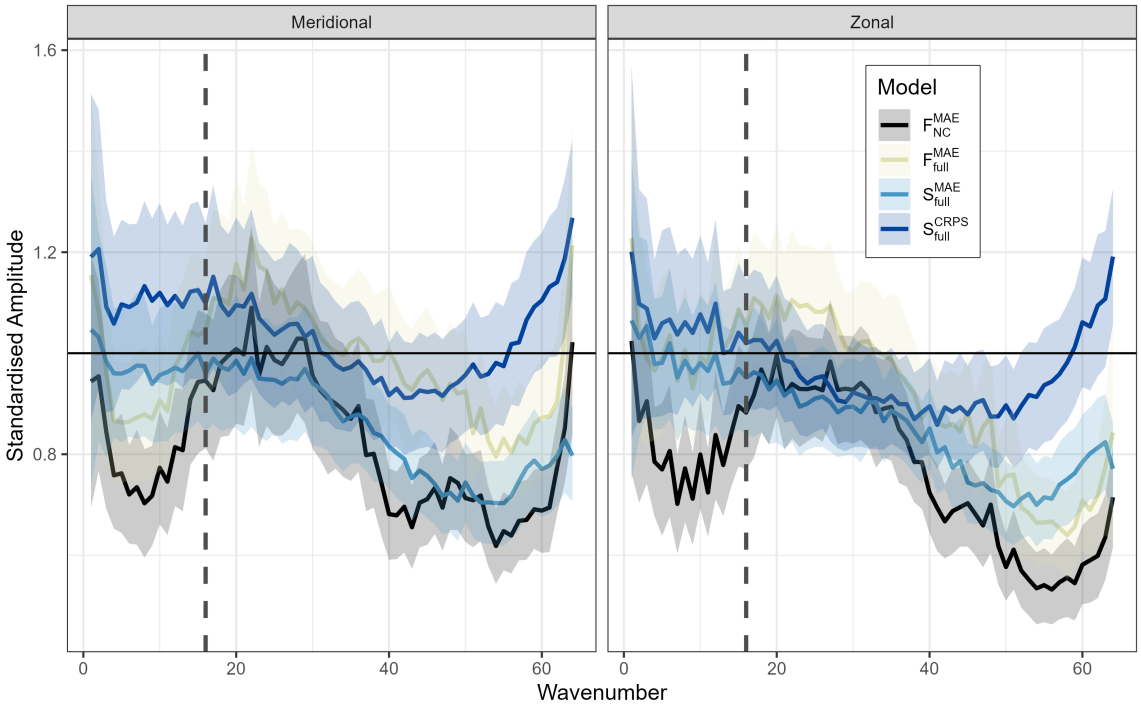

Figure 2.7: RASP metric (mean +/- 1 SD) standardised to ground truth values for zonal and meridional wind fields. Spectral powers are calculated across 1200 randomly selected fields. Dashed line shows wavenumber corresponding to LR grid size.

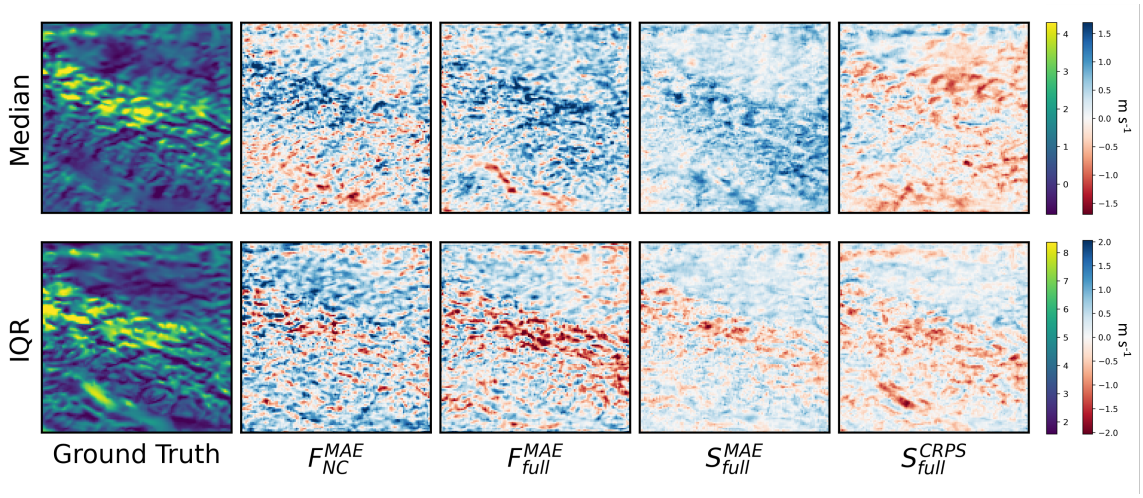

Figure 2.8: Pixelwise median and inter-quartile range (IQR) of full distribution of the test dataset for meridional wind fields. The first column shows truth statistics, followed by difference fields for each of the four models (truth - model).

All models produced visually realistic downscaled HR fields, and pixelwise marginal statistics of full distributions generally matched the true statistics well (figure 2.8). Differences in median and inter-quartile range were spatially smoother with the S class models, suggesting these models were able to capture fine spatial patterns better. The baseline model substantially underestimated inter-quartile range (IQR) in many locations. This bias was improved with the F_{full}^{MAE} model, although it overestimated IQR in certain areas, and the S class models further improved the results.

Investigating the tail behaviour of the full distributions, we found that models that were better at learning conditional distributions performed better at predicting accurate extremes (figure 2.9). Comparing moderately large marginal extremes (99.99 and 0.01 percentiles) across pixels, the S_{full}^{CRPS} model had the least biased estimate of extremes while biases from the F_{nc}^{MAE} were largest, underestimating 99.99 percentiles and overestimating 0.01 percentiles (figure 2.9a). This pattern is also apparent in difference maps of the extremes (cf. 0.01 percentile in figure 2.9b). The map for F_{nc}^{MAE} is almost negative everywhere, whereas the map for S_{full}^{CRPS} shows reduced systematic bias and is fairly well centred around zero for the 0.01 percentiles. We also investigated less extreme 0.01 and 99.9 percentiles, and more extreme 0.001 and 99.999 percentiles; all extreme percentiles showed the same pattern.

Calibration of spatial extremes was also improved in the S class models (figure

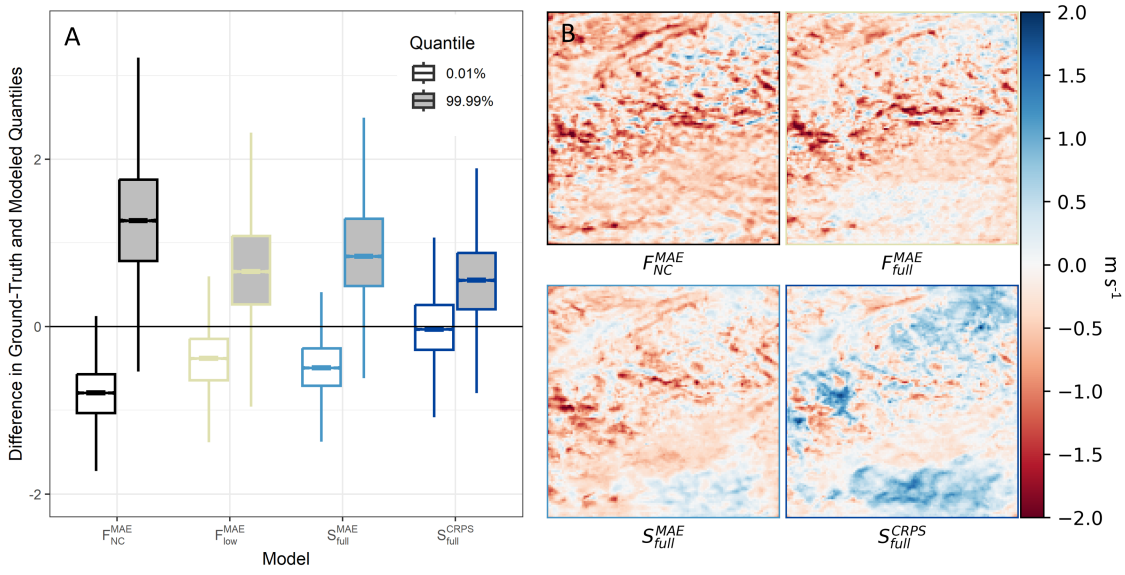

Figure 2.9: Calibration of moderate extremes for meridional wind fields over full distributions. a) Boxplots of distributions of difference in 99.99 and 0.01 percentiles of ground truth and generated realisations for four models, based on 500 realisations for each of 350 randomly selected conditioning fields. Values below zero represent model overestimation; values above zero represent underestimation. b) Difference maps of 0.01 percentiles of ground truth and generated realisations for four models.

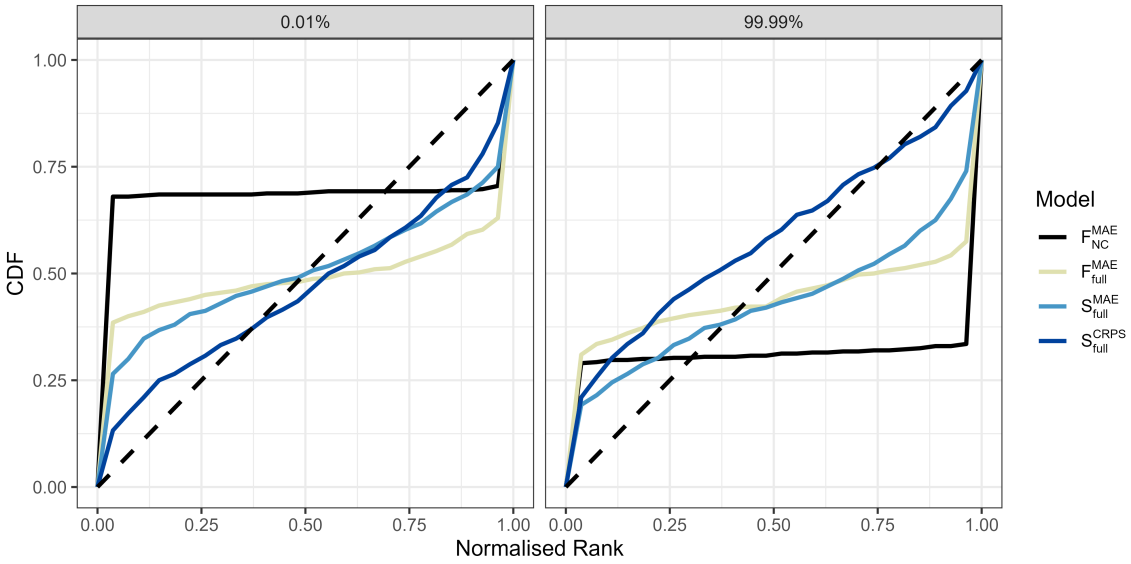

Figure 2.10: CDFs of rank histograms based on 0.01 and 99.99 percentiles of meridional wind fields over 400 conditioning fields, with 96 realisations of each field. Dashed lines represent CDF of a uniform distribution.

2.10). Here, rank histograms represent stochastic calibration of models in regards to spatial extreme values (i.e., large or small quantiles of values across the domain). Rank histograms of 0.01 and 99.99 percentiles across wind fields showed that the S_{full}^{CRPS} model was the least underdispersive, and had less bias than the other models (figure 2.10). By contrast, the rank histogram of the baseline F_{NC}^{MAE} model showed the model systematically overpredicted 0.01 percentiles, and underpredicted 99.99 percentiles.

To investigate why the F_{full}^{MAE} model showed excellent calibration on the synthetic dataset, but was underdispersive when used to generate high-resolution wind fields, we tested the effect of increased spatial heterogeneity on synthetic data models (cf. Section 2a). Models trained on datasets with high heterogeneity showed a greater degree of underdispersion than those with low or moderate heterogeneity (figure 2.11). Investigating the KS statistics of the pixelwise marginal distributions showed that distributions from datasets with high heterogeneity had very large ranges in quality, whereas those from low heterogeneity data were more consistent and better matches on average (figure 2.11b). The rank histograms of the results from these datasets showed similar patterns - the low heterogeneity model was relatively well calibrated, and the high heterogeneity model was underdispersive. Interestingly, while the moderate level model did not show as much underdispersion, it showed the most bias,

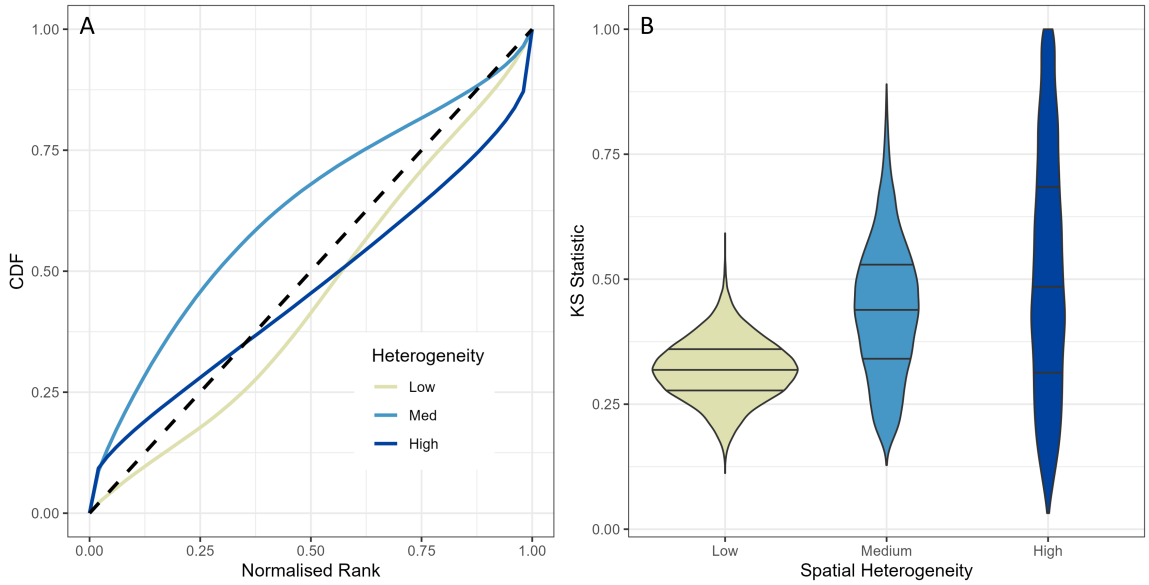

Figure 2.11: Comparison of stochastic calibration of the F^{MAE}_{full} model based on synthetic data with low, moderate and high spatial heterogeneity. a) CDFs of rank histograms based on all pixels of 50 random conditiong fields with 96 realisations of each. b) Distribution of pixel-wise KS statistics between generated and true marginal distributions, using 500 stochastic realisations for a single conditioning field.

more often than not overestimating values. These results suggest that the F class models struggle to produce good stochastic calibrations when there is a high degree of spatial heterogeneity in the fields - as is often the case in a realistic setting.

2.4 Discussion and Conclusions

This paper discusses three main classes of GAN-based downscaling models distinguished by: noise type (noise covariate vs noise injection), training method (frequency separation vs stochastic sampling) and content loss type (deterministic MAE vs probabilistic CRPS). We aim to improve stochastic calibration, creating models that can successfully sample from the full range of the conditional distribution (i.e., for a given large-scale atmospheric state, we want the model to generate the full range of local weather possibilities). We first present a novel network architecture, where many layers of noise at different resolutions are injected into the Generator. Compared to the baseline F_{NC}^{MAE} model, our architecture performs better at capturing the conditional variability of the data (thus reducing underdispersion), and achieves good calibration on synthetic data. We then introduce a stochastic training method which greatly improves stochastic calibration and spatial structure, especially when combined with the probabilistic CRPS metric in the S_{full}^{CRPS} model. The S_{full}^{CRPS} model shows improved skill at estimating marginal conditional distributions, as well as marginal and spatial statistics of the full distribution. Using this model, we successfully downscale wind fields and show that ensembles of generated realisations are well calibrated. Although individual realisations of wind fields are often visually quite similar, we do not expect wind fields to have as large a conditional distribution as some other fields, especially precipitation. All models we investigate here use a GAN architecture with a separate stream for HR topography, as initial tests showed substantially improved results. The overall advantages of this approach is tested in detail in Chapter 3.

Most conditional GANs in super resolution are deterministic, and recent attempts at reintroducing stochasticity have added noise fields as additional covariates [e.g., 20]. Our approach of injecting noise directly into the convolutional layers fundamentally differs in that it adds noise to latent representations deep inside the network, instead of to the input. When noise is introduced as a covariate at the beginning of the network, we hypothesise that the network will learn low weights for the noise layers in order to optimise the loss function. By adding noise to the latent representations, we slightly alter features inside the network, leading to better representation of conditional and full distributions. It would be interesting to investigate the mathematical basis of noise injection and its relationship to calibration of the marginal distributions. We adopted noise injection as a pragmatic approach in the absence of theoretical guidance. Our approach is similar to the nESRGAN+ architecture [32]

which injects noise inside the Residual in Residual Dense Blocks, but we inject noise one level deeper, inside the Dense Blocks, to alter the output of the basic convolutional layers. In contrast to the nESRGAN+, we also use noise injection at both the low and high resolution, allowing for more scales of stochasticity. Interestingly, we never experienced problems with overdisperison as we increased the number of noise injection layers; marginal distributions were closest with the maximum possible number of noise injection layers for a given architecture.

Stochastic sampling is an approach adapted from Harris et al. [14], in which the content loss function is calculated on a set of stochastic realisations. While the F^{MAE}_{full} model showed excellent calibration on the synthetic data and much improved performance than the F^{MAE}_{NC} model, it still failed to fully capture the variability in the wind downscaling application. The S class models, especially in combination with the CRPS loss function, resulted in substantially improved performance. This improvement could be the result of a few different factors. First, by using multiple realisations of each field in the loss function, the network has more information to use for backpropagation. Second, CRPS is a stochastic metric which aims to quantify distributional matching; it seems reasonable that it thus improves the stochastic calibration. In some cases, the S^{CRPS}_{full} model produced too much variability at very fine spatial scales. Since CRPS is a pixelwise metric, we hypothesise that the improved pixelwise calibration may be a cause of this pixel-scale variability. In contrast to our study, Harris et al. [14] did not find an improvement with using CRPS in the loss function. This difference in results could be due to differences in Generator architecture. As Harris et al. [14] did not use noise injection, perhaps the generated stochasticity was less suitable to optimisation with the CRPS metric.

All models considered in this study use a separation of spatial scales in the content loss in an effort to address the double penalty problem inherent in the ill-conditioned nature of climate downscaling. In our synthetic data experiments, we found that partial frequency separation (PFS), as described in Annau et al. [3], resulted in well calibrated output. However, this method did not perform as well in the realistic setting of wind component downscaling, motivating consideration of the stochastic sampling approach. Fundamentally, PFS and stochastic sampling have similar goals: allowing the adversarial loss freedom to create small-scale features, while rewarding consistency between generated and conditioning fields at large scales. While PFS achieves this by only sending low frequency information to the content loss, the stochastic sampling approach applies the content loss function across an ensemble of

stochastic realisations, thus "smoothing out" the smaller scale features of the generated fields. The stochastic sampling approach is likely more accurate than PFS - instead of arbitrarily choosing a frequency for separation, the sample conditional means define the transition from conditioning scales to sampling scales. Indeed, we found that for downscaling wind components, stochastic sampling always outperformed models using PFS. A practical challenge with stochastic sampling is that it uses more computational resources during training than PFS models, as each of the stochastic realisations have to be used during backpropagation. Most notably, stochastic sampling nearly doubled the amount of memory required during training, and it increased training time by about 50% (using a stochastic batch size of six). In practice, the choice of training approach will likely depend on the desired outcome and computational resources available. While the stochastic sampling models performed substantially better at capturing conditional and full distributions, they only performed slightly better at capturing the spatial patterns of single conditioning fields. Thus, if the goal is to produce downscaling without needing to capture conditional variability, it may be prudent to use PFS and reduce training requirements. It is of course possible that the improvement gained from using stochastic sampling on capturing the conditional distributions will depend strongly on the fields being considered. It should be noted that we did not conduct formal hyperparameter tuning in this study, and used similar values to those in Annau et al. [3]. While we believe the architecture and setup are similar enough that these settings are reasonable, it may be prudent to tune important hyperparameters with our stochastic framework.

Wind fields show a high level of spatial heterogeneity, which we expect is responsible for the difficulties experienced by the F_{full}^{MAE} model in capturing the conditional variability. Our experiment with spatial heterogeneity showed that even with synthetic data, increasing heterogeneity lead to increased underdispersion and bias in conditional means. High heterogeneity will generally lead pixel-wise metrics to be more sensitive - slight shifts in features from the truth fields will tend to result in poor pixel-wise metrics compared to similar shifts in fields with low heterogeneity. Future work could consider wind fields and other pertinent physical fields across areas with different degrees of spatial heterogeneity. It will also be important to consider the impact of stochasticity on temporal dependence. In the current framework, correlations between timesteps are only introduced through the LR conditioning fields. Since substantial stochasticity occurs at sub-grid scales, there is no mechanism for enforcing consistency of small features between timesteps. Future research could adapt the

network to include a recurrent architecture, such as the convolutional gated recurrent unit used by Leinonen et al. [20].

It will also be important to properly account for temporal dependence of the generated realizations. In the current framework, correlations between timesteps are only introduced through the LR conditioning fields. Since substantial stochasticity occurs at sub-grid scales, there is no mechanism for enforcing consistency of small features between timesteps. Future research could adapt the network to include a recurrent architecture, such as the convolutional gated recurrent unit used by Leinonen et al. [20]. It would also be interesting to incorporate temporal dependence into the injected noise, although this approach may not lead to improved temporal dependence in the final downscaled fields.

From a theoretical standpoint, stochastic downscaling is an appealing approach as it provides a way to quantify the range of solutions to the underdetermined problem of climate downscaling. In addition, we have found that improved distributional estimates lead to better representation of extreme events, both spatially and temporally. A model which is able to accurately sample from a distribution will sometimes draw samples from the tails of the distribution, whereas models with substantial underdispersion will tend to only sample from the conditional mean. Harris et al. [14] found that their models were underdispersive when applied to extremes; it would be interesting to see if the improvements made here could improve their analysis of precipitation downscaling.

An important avenue for future research involves comparing stochastic downscaling across different classes of deep-learning models. In particular, diffusion models have recently been introduced to many applications, and recent research by Mardani et al. [29] has shown impressive results using diffusion models for stochastic atmospheric downscaling. While diffusion models have been favoured for being easier to train than GANs, the Wasserstein GAN is a conceptually different approach to training, and in our experience shows excellent stability. While diffusion models may have benefits over Wasserstein GANs (an in particular, may be able to address the slight underdispersion we observed even in the final model), they are also substantially more computationally intensive to train. For reference, the model developed in Mardani et al. [29] took 21,504 GPU hour to train, while our GAN models only required about 36 hours. Depending on the situation, Wasserstein GANs may be more suitable due to their relative efficiency.

Modelling extreme events is of utmost importance to climate adaptation, and these

events are often more challenging to model than averages [39]. Infrastructure needs to handle precipitation and wind extremes; most heat-related human health issues occur during extreme heat waves [17]. Generally, statistical downscaling has not been successful at capturing extreme events, and while dynamical downscaling can perform better, it is too computationally intensive for some practical applications. Our study has shown that by improving the ability of GANs to make distributional estimates, we are able to obtain better estimates of extremes, both spatially and temporally, often with a marginal increase in computational cost. Hence, deep-learning based downscaling shows promise as a statistical downscaling strategy with the ability to more accurately capture extremes. Further research will be required to determine whether these results generalise to a non-stationary climate (e.g., across time periods). If so, deep-learning downscaling could become an essential part of climate adaptation for estimating future extremes.

Chapter 3

Generative Adversarial Networks for Deep Learning Downscaling of Temperature, Humidity, and Precipitation

3.1 Introduction

Accurate, local-scale climate information is essential support for societal efforts to adapt to anthropogenic climate change. Earth System Models (ESMs) provide state-of-the-art projections on a global scale, but at insufficient spatial resolution for local planning [24]. Thus, downscaling - creating high-resolution climate information from low-resolution, large-scale model output - is an important practical tool for planning and decision making. Spatially and temporally high resolution fields of climate data over large scales are important for many applications, including simulation of extreme events at local scales [e.g. fires, floods, storms; 10], local infrastructure planning, and extreme event attribution [41]. Unfortunately, accurate downscaling is often limited by computational constraints. Ideally, we would like downscaling techniques to be computationally efficient, accurate over regions with different climates, and capable of capturing extremes. The latter requirement also suggests that we should be able to generate ensembles of downscaled fields, to sample the full range of meteorological states.

Strategies for downscaling coarse-resolution climate model simulations to regional

scales can be broadly classified into dynamic downscaling and statistical downscaling. Dynamical downscaling employs a limited-area numerical climate model to resolve fine-scale features, driven by large-scale weather patterns from the low-resolution ESM [37]. Statistical downscaling develops statistical relationships between low-resolution (LR) and high-resolution (HR) climate variables, using a variety of strategies ranging from simpler bias correction methods [28], to more complex methods using constructed analogs [1] or multivariate quantile mapping [6]. Both dynamical and statistical downscaling approaches have strengths and weaknesses. Statistical downscaling is generally much more computationally efficient than dynamical downscaling, but can have limited capacity to downscale variables with complex dependence structures and/or non-stationary evolution through time. Also, traditional statistical downscaling usually requires observations to calibrate models properly.

Dynamical downscaling is necessary for capturing effects that violate the assumption of temporal or spatial stationarity (e.g., elevation dependent warming) as it uses physical models of the atmosphere. It can be implemented at multiple spatial scales, and different classes of downscaling often lend themselves to different model specifications. Regional climate models (RCMs), a form of dynamic downscaling, have existed for decades, and are usually used to downscale ESM or reanalysis products to a 20-40 km resolution. More recent convection-permitting models usually operate on a 1 - 4 km resolution, and have numerous advantages in representing fine-scale weather patterns which could not otherwise be effectively captured, including lake effects, valley effects on winds, valley inversions and elevational gradients in mountainous terrain [23]. However, due to the computational expense of convection-permitting models, they are of limited use for exploring climate change uncertainty, and are often not practical for operational products.

Over the past few years, deep learning has been introduced as a new statistical downscaling strategy which can potentially capitalise on the benefits of both traditional statistical and dynamical downscaling. Specifically, deep neural networks can be trained on HR output from dynamical downscaling, and the network learns to emulate the dynamical downscaling, given a suite of LR climate information. So far, deep-learning methods have demonstrated potential for creating downscaled fields of similar quality to dynamical methods, but with the efficiency of standard statistical methods [31]. The deep-learning downscaling we present in this paper attempts to emulate the results of convection-permitting models, downscaling to high spatial resolution at hourly timesteps.

An inherent challenge with downscaling of any form is its underdetermined nature. A set of LR climate fields does not have sufficient information to fully specify a single corresponding HR field; rather, there exists a conditional distribution of possible HR fields that are physically consistent with the LR input. In order to quantify the uncertainty inherent in the downscaling process, it is necessary to sample multiple realisations from this conditional distribution. In a dynamic downscaling context, sampling from the HR conditional distribution is theoretically simple, since each new run with slightly different initial conditions will create different realisations. However, dynamical downscaling to convection-permitting scales is generally too computationally expensive to allow more than one run. In traditional statistical downscaling methods, the probabilistic nature has to be parametrically modeled. A simple example of this would involve sampling from the error distribution of a linear regression. While there are a growing number of probabilistic downscaling techniques that allow sampling from the conditional HR distribution [e.g., 11], most deep-learning downscaling to date has not been explicitly probabilistic. However, recent research in Chapter 2, and by Leinonen et al. [20], Harris et al. [14] has worked on creating fully probabilistic models which can sample multiple realisations from the HR conditional distribution, conditional on the LR climate input fields. This study further explores the applications of stochastic deep-learning downscaling.

Most of the recent research in generative deep learning based downscaling has employed conditional Wasserstein Generative Adversarial Networks (GANs). GANs, first introduced by Goodfellow et al. [13], adapted into a conditional version by Mirza and Osindero [30] and adapted to the Wasserstein GAN (wGAN) by Arjovsky et al. [4], 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 assesses the distributional similarity of the generated output compared to the training data. In the wGAN, the Critic estimates the Wasserstein distance between the distribution of generated samples, and the distribution of training data. Theoretically, the GAN will learn to sample from the conditional distribution of the HR variables, conditioned on the LR fields. Annau et al. [3] developed a GAN framework which produced accurate downscaling of wind components. Chapter 2 adapted this framework to be stochastic, allowing the GAN to sample multiple realisations from the learned conditional distribution. Our research 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 stochastic GAN framework developed in Chapter 2.

While substantial research has investigated GAN downscaling, there are multiple questions that should be addressed prior to use in an operational setting. First, most studies to date have focused on downscaling single variables. Many studies [e.g., 14, 20, 31] have focused on precipitation, and Annau et al. [3] and Chapter 2 focused solely on downscaling wind components. However, operational downscaling 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. The stochastic GAN developed in Chapter 2 produced accurate downscaling of wind components, but its ability to extend to other variables has not previously been assessed. Multiple studies [14, 20] have shown that GANs can struggle to capture extreme precipitation events, a task which is crucial for adaptation planning [12]. Our research in Chapter 2 found that the 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 measures of 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 [14, 20, 31], and while Annau et al. [3] and Chapter 2 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, so are not a challenging multivariate downscaling 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. Harris et al. [14] 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. Chapter 2 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 relatively small domains for computational reasons, and the portability of GAN frameworks over large spatial regions has yet to be assessed. Applying GAN downscaling over large contiguous areas poses many challenges. Computational constraints aside, it is necessary to assess how well a GAN framework (i.e., the model architecture and training method) developed in one region performs in other locations, as most research to-date has developed and tested GAN frameworks in a single region. Showing that a GAN framework can succeed in multiple locations is a first step to developing models capable of downscaling large regions.

This current study aims to address 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 in southwest British Columbia (including Vancouver Island, the Coast Mountains, and the Interior Plateau) considered in Chapter 2. We then investigate the advantages and disadvantages of univariate versus multivariate prediction, and assess 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 performance of the stochastic GAN framework developed in Chapter 2.

3.2 Methods

All GANs in this paper use the same basic structure. We train the models on paired sets of LR conditioning fields (covariates), time-invariant HR surface features, and HR truth fields. The GAN then learns a mapping to the HR fields from the input covariates. For consistency, we keep the same resolution and size of fields across all models: HR domains are 128 x 128 pixels (corresponding to approximately 270 x 450

km or 4° x 4°), and LR domains are 16x16 pixels, corresponding to a downscaling factor of eight.

3.2.1 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 Weather Research and Forecasting (WRF) model output over a Western Canada domain [21] 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 a broader suite of covariates than ERA-interim, and is a state of the art reanalysis product. Although there will be some differences between ERA5 and ERA-Iterim, they represent the same realisation 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 biases on common scales between the WRF output and the ERA5 conditioning fields. In addition, 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 ERA5 to WRF as well as downscale.

All GANs use a subset of nine LR covariates: temperature (2m), specific humidity (2m), relative humidity, precipitation, wind components (10m), convective available potential energy (CAPE), surface evaporation, and surface pressure. Table 3.1 lists the suite of covariates used for each experiment. We also include HR topography as an invariant field in all experiments. For the majority of our analyses, we consider a rectangular region in Southwestern BC, Canada (49° to 53° N, 122° to 126° W; henceforth called the Southwest region), as its topographic complexity represents a realistically challenging downscaling scenario (figure 3.1). Chapter 2 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 standardise all variables to mean zero and unit standard deviation prior to training using a spatially and temporally invariant time-space mean and standard deviation for each variable.

Table 3.1: List of covariates used for each experiment.

Experiments	LR Covariates	HR Covariates
Univariate Temperature and Humidity	zonal wind, meridional wind, temperature, spec.humidity, relative humidity, pressure, evaporation	topography
Univariate Precipitation, Multivariate, HR Topography	zonal wind, meridional wind, temperature, spec.humidity, precipitation, relative humidity, pressure, evaporation, CAPE	topography
Northeast Region Wind, Temperature, Humidity	zonal wind, meridional wind, temperature, spec.humidity, precipitation, pressure, CAPE	topography
Northeast Region Precipitation	zonal wind, meridional wind, temperature, spec.humidity, precipitation, pressure, CAPE	topography, land use index

To investigate the performance of this GAN framework in 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 information about water bodies and surface vegetation, as a second HR invariant field.

3.2.2 Model

We use the Wasserstein GAN introduced by Arjovsky et al. [4], where the Critic estimates the Wasserstein distance between the generated and training samples. Compared to the original GAN formulation, the Wasserstein GAN addresses many common challenges related to vanishing gradients and training instability [9]. Intuitively, the Wasserstein distance represents the work 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. Meanwhile, the Critic refines its estimation of the Wasserstein distance, and is able to better characterise differences between generated and truth fields.

One appealing aspect of GANs compared to more standard Convolutional Neu-

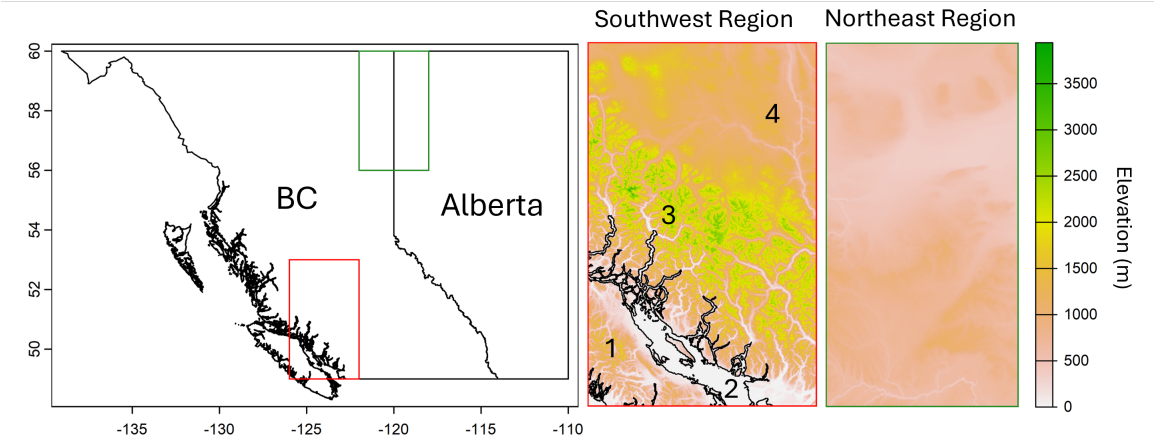

Figure 3.1: Maps of study areas, showing (from left to right) study area locations relative to British Columbia and Alberta, and topographic relief of both regions. For the Southwest Region panel: 1 = Vancouver Island, 2 = Georgia Straight, 3 = Coast Mountains, and 4 = Interior Plateau.

ral Networks, especially for climate downscaling, is that they do not solely rely on a deterministic pixel-wise error metric as the loss function. Since downscaling is an underdetermined problem, using a deterministic 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 deterministic pixel-wise metric, and aims for convergence in distribution, which is a desireable property. However, Wang et al. [43] 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 realisation at large-scales. We follow this approach, and our Generator loss function is composed of both the adversarial loss, and a pixel-wise content loss. A more detailed discussed of this issue is presented in Annau et al. [3].

This study uses the stochastic GAN architecture developed in Chapter 2, which was based on the deterministic GAN described in Annau et al. [3]. The architecture makes extensive use of convolutional layers, which are designed to extract representative features from images [22]. 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 [36]. Following Chapter 2, 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 upsampling step. Once the upsampling 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 model found to perform best in Chapter 2, with stochastic sampling and cumulative ranked probability score (CRPS) as a content loss. Stochastic sampling, adapted from Harris et al. [14] creates multiple realisations of each training sample, and computes the content loss across these realisations. CRPS is a probabilistic measure, so its use in training emphasizes convergence in distributions. As our aim is to sample from the

HR conditional distributions, it is natural to use a probabilistic metric.

3.2.3 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 training set (2003 and 2006), and one year as an out-of-sample test set (2005). Initial tests showed that model performance on the test set 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.

In Chapter 2, we found that the stochastic GAN was better able to capture wind component extremes (i.e., the tails of the distribution) than the determinstic GAN. A common challenge with precipitation downscaling is underestimation of high-precipitation events [18]. As a point of comparison we created a deterministic GAN similar to that considered in Annau et al. [3] by removing the noise injection from the convolutional layers in the Generator and using mean absolute error as the content loss metric.

3.2.4 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 upsampling block in the topography stream. We chose this strategy to keep the network architecture as consistent as possible between experiments.

3.2.5 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 pixelwise error of realisations, and while these are useful, they can overly-penalise underdetermined downscaling results due to the double penalty problem. Thus, it is generally better to compare statistics of the generated fields and truth fields from the test set, instead of comparing individual realisations. Pixelwise comparisons of distributions (e.g., medians and quantiles) are simple and easy to interpret 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 across 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 discrete 2D spectral power spectrum and averages power over all angles from the centre of wavenumber zero. For ease of interpretation, we normalise the power at each wavenumber to the corresponding power in the truth field. Normalised 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

$$p(HR) = \int p(HR|LR)p(LR)\ dLR \approx \frac{1}{n}\sum_{k\in LR} p(HR|LR_k). \tag{3.1}$$

As it is not usually possible to access multiple realisations of the same truth field, an ensemble of stochastic realisations have to be compared to a single truth field.To test the stochastic calibration of the conditional distribution of generated fields, we used CDFs of rank histograms calculated over one year of samples [40]. 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 to histogram bin width. To investigate the stochastic calibration of conditional distributions for specific sets of conditioning fields, 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.

3.3 Results

In this section, we first investigate extension of GAN downscaling from wind components to three other variables: temperature, humidity, and precipitation. As capturing dependence between variables correctly is important, we then compare multivariate and univariate downscaling. We then present a sensitivity analyses, showing the benefits of using an HR stream in the Generator. Finally, we test spatial portability of this GAN framework, by training and testing on a second region with different topography and climate.

3.3.1 Univariate downscaling of temperature, humidity, and precipitation

In this section, we assess the accuracy of univariate GAN downscaling for temperature, humidity, and precipitation. For illustrative purposes, we chose two representative hours that represent the 0.1 and 0.9 quantiles of domain averaged variables (e.g., a cold and a warm hour for the whole domain) and show a pair off HR realisations. We also show pixel-wise statistics across all timesteps, and compare PDFs of pixel values.

Temperature downscaling generally performed well and succeeded in capturing HR details (figure 3.2). HR realisations for the representative cold hour were often slightly too cold in the Straight of Georgia and missed some fine-scale details in the Northeast corner. Realisations for 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 Interior Plateau in the Northeast. In these two samples, the model was successful at capturing the more consistent temperature of the ocean compared to the surrounding land, as well as sharp transitions between the ocean and continent. Conditional standard deviation across ensemble members for this pair of representative hours ranged from about 0.5 to 1.5 K, with lower values in mountains and the Georgia Straight in the cold sample. Note that the conditional standard deviation fields differ in non-trivial ways between hours, which is physically reasonable but often difficult to achieve with classical parametric downscaling techniques. Considering distributions of pixel-values by month, the PDF of generated values closely matched the distribution of WRF values, for both January and July, although with somewhat larger difference in January. Maps of pixel-wise quantile

differences showed good 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 values in the Interior Plateau, with biases of up to about 3 K.

Specific humidity downscaling was generally less accurate, although succeeded in capturing spatial structure and general features (figure 3.3). In the dry hour especially, the model substantially underestimated the humidity over the Georgia Straight in most realisations, and while it did predict higher humidity at lower elevations (e.g., within valleys), these were not as clear as in the WRF fields. In the moist case, while humidity values across the field matched well with WRF, there were localised biases - most realisations showed a dry bias on the Eastern side of the field.

Conditional standard deviation showed a moderate degree of variation between realisations, usually ranging from 8-12% of the deviation in the fields. Unsurprisingly, conditional standard deviation was much lower for the dry case than the moist case. The moist case showed low variability over the ocean, but relatively high variability in the mountains near the coast. Distributions of pixel values for January and July were substantially more biased than with temperature, especially for July, where the PDF of generated values were underdispersive.

Pixelwise 0.01 quantiles showed that the models had a postive bias at the low tail of the humidity distribution through most of the field, except in the Straight of Georgia, where quantiles were systematically underestimated. Truth and generated median fields were similar, but showed a slight underestimation across the field. For 0.99 quantiles the models generally underestimated humidity, except on the tops of mountains. The most severe biases were up to about 20% of the variability within a single field.

The GAN performed well at downscaling precipitation (figure 3.4). Compared to temperature and specific humidity, there was substantial diversity between realisations for a given set of conditioning fields. This is consistent with expectations for sampling from the conditional distribution of HR precipitation. The light-rain hour especially showed a large degree of variation between realisations, as indicated by the high conditional standard deviation. The heavy rain hour also showed a large conditional standard deviation, although not so large in relative terms as the dry hour.

As pixel value distributions were very similar in January and July, we combined them in a single PDF. Generated and WRF distributions matched well, although the

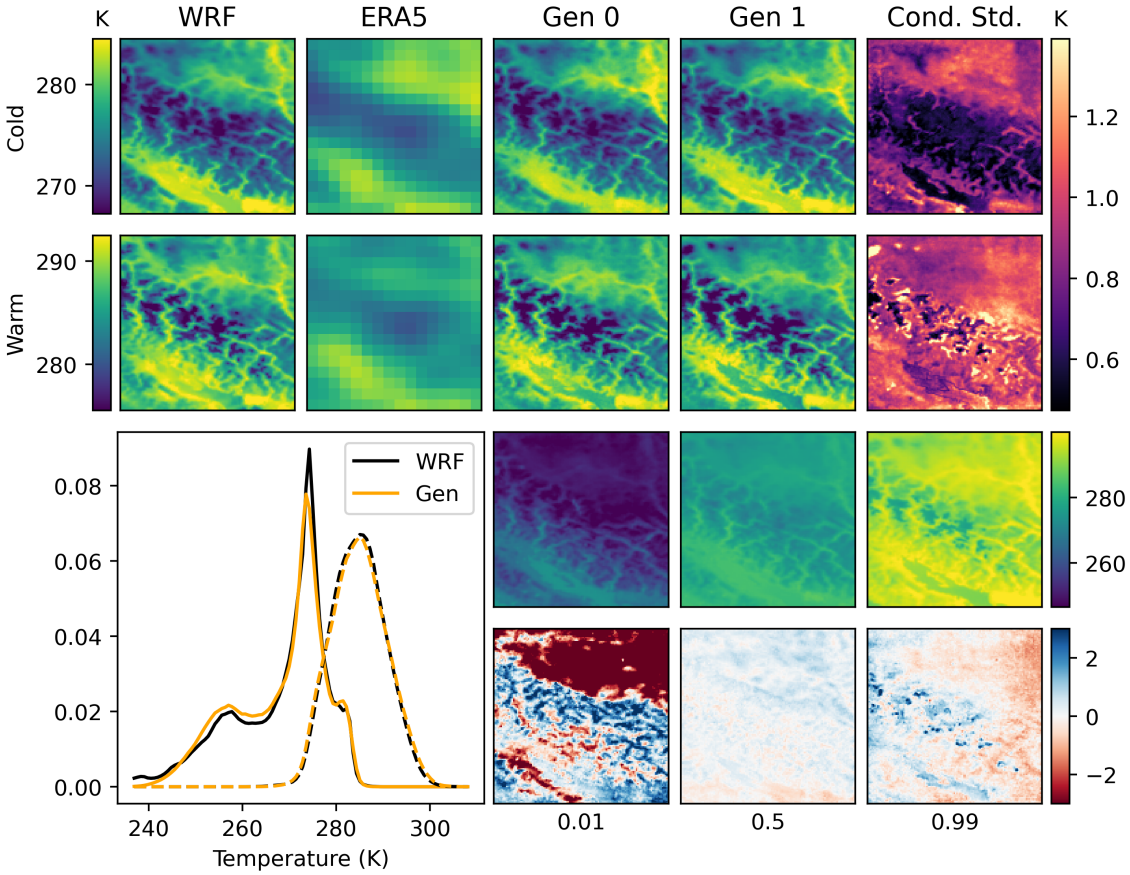

Figure 3.2: Evaluation of univariate GAN downscaling of temperature for the Southwest study area. Top two rows show respectively an example cold and warm sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.99 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences relative to the ground truth (truth - generated). The left-bottom panel shows overall PDFs (across space and time) of pixel values for January samples (solid) and July samples (dashed).

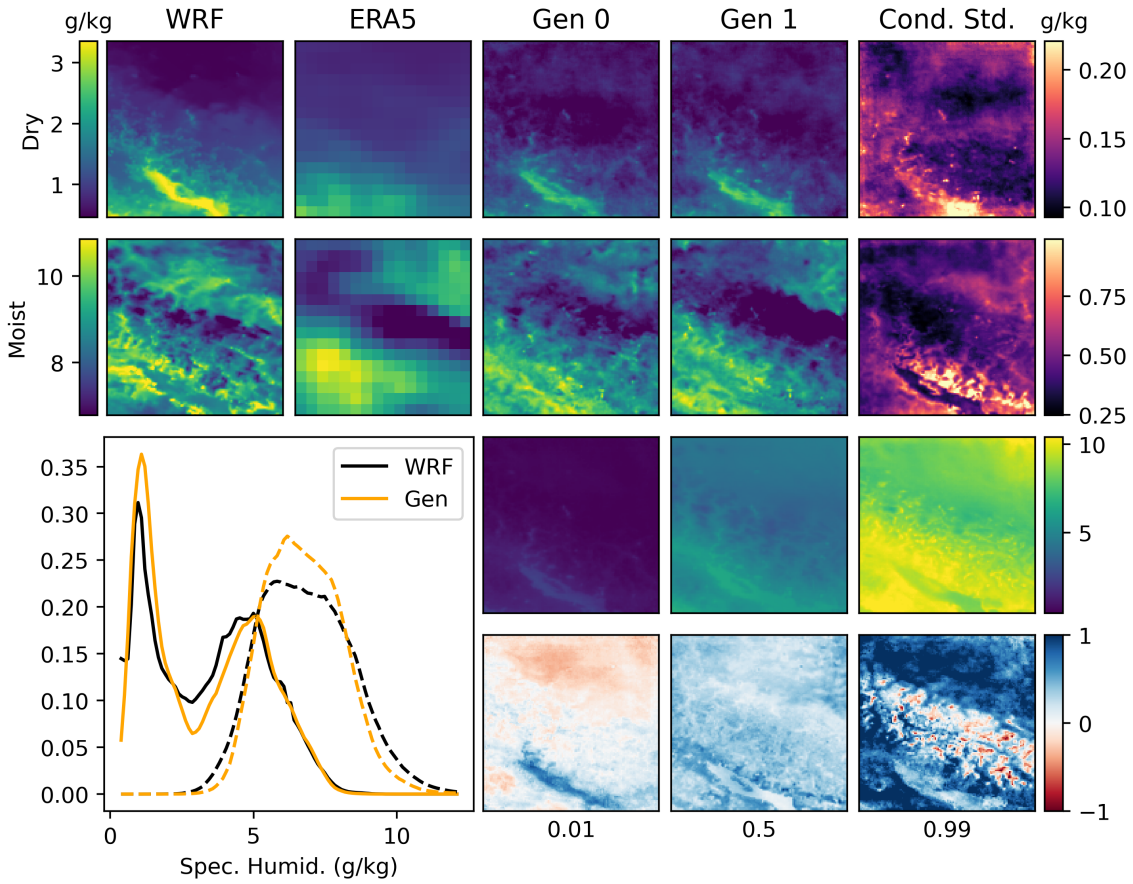

Figure 3.3: Evaluation of univariate GAN downscaling of specific humidity for the Southwest study area. Top two rows show respectively an example dry and moist sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.99 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences relative to the ground truth (truth - generated). The left-bottom panel shows overall PDFs (across space and time) of pixel values for January samples (solid) and July samples (dashed).

GAN underestimated extreme precipitation events, and slightly underestimated drizzle. This underestimation of heavy precipitation was also apparent in the pixelwise quantile difference maps, which showed very slight underestimation for 0.01 and 0.5 quantiles, but larger underestimation of 0.99 quantiles.

A common challenge with precipitation downscaling is the creation of drizzle on samples with no precipitation in the conditioning fields. The GAN did not experience this challenge; in general, realisations of dry timesteps did not show any precipitation. Thus, for futher analysis, we excluded any timesteps that had zero precipitation in the WRF ground truth fields.

We found that covariate choice was especially important for precipitation. Initial precipitation models, which only included LR precipitation, temperature, evaporation, and surface pressure produced fuzzy and biased downscaled fields. Addition of CAPE and wind components substantially improved results. There was no obvious difference in downscaling accuracy for temperature with the addition of these covariates, and only slight improvements for humidity.

Stochastic GANs did a much better job of capturing extreme precipitation than equivalent deterministic models (figure 3.5). While both models slightly underestimated the probability of 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 rate values > 18 mm/h. Both models show a low-precipitation bias for light precipitation (< 2.5 mm/h).

All three variables showed similar spectral power to WRF fields (figure 3.6). Median RASP estimates for generated fields showed good overall agreement with corresponding WRF estimates, particularly for temperature, with a slight systematic low power bias in general for humidity. Precipitation also showed good spectral calibration across most scales, but had a high-power bias at high wavenumbers, corresponding to an overabundance of near grid-scale features. Precipitation RASP ratios also displayed much more variability between samples than temperature or humidity.

Rank histograms for temperature and humidity aggregated across timesteps showed minor underdispersion of conditional distributions, and precipitation showed underestimation of high values (figure 3.7a). These results were consistent with rank histogram maps we used to investigate the calibration of conditional distributions for individual cases (representative of 0.01, 0.5, and 0.99 quantiles across the the whole test set). While these cases represent individual snapshots, they are broadly representative of patterns in those quantiles (figure 3.7b). All variables showed slight

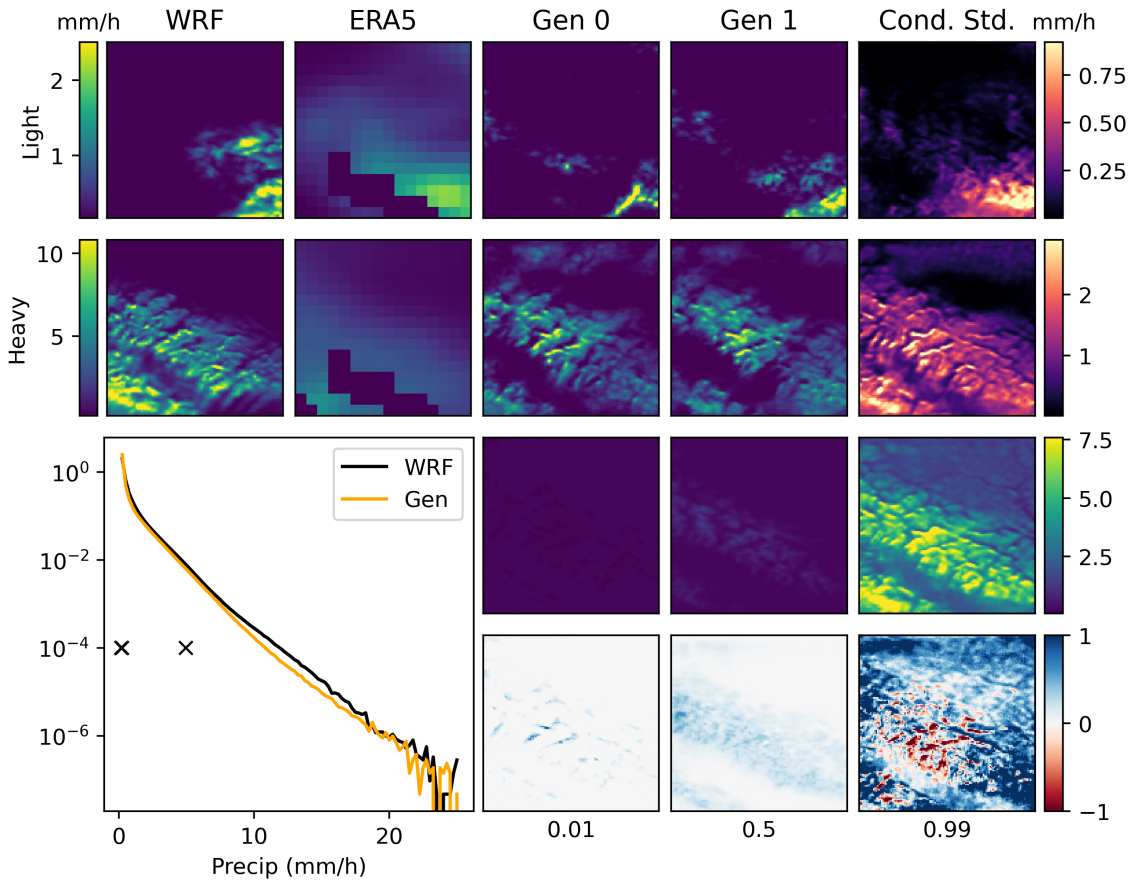

Figure 3.4: Evaluation of univariate GAN downscaling of precipitation for the Southwest study area. Top two rows show respectively an example light rain and heavy rain sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.99 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.

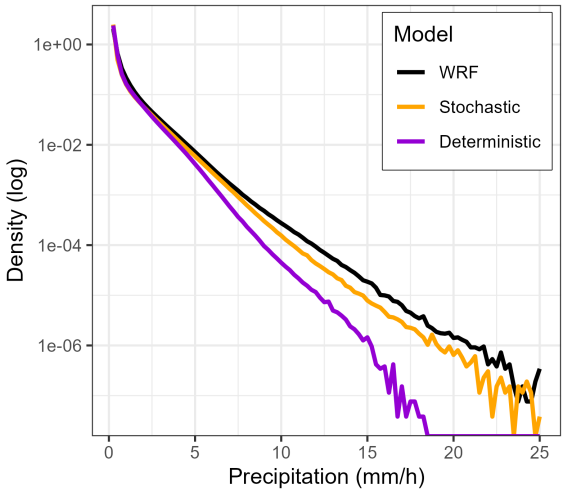

Figure 3.5: PDFs of pixel values for precipitation fields, comparing WRF, deterministic generated, and stochastic generated. Distributions were estimated from all pixels of one year of hourly samples. Note the y-axis is shown on a log scale. Vertical line shows 0.999 quantile.

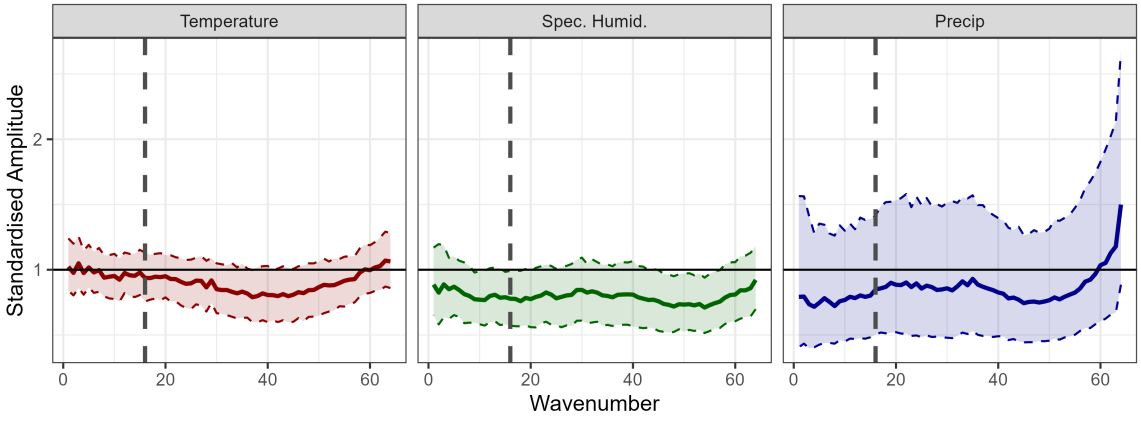

Figure 3.6: RASP metric for the three variables, with spectral powers standardised to ground truth fields. Ratios are calculated separately for each HR truth field out 1200 randomly selected fields, providing a range of estimates of spectral power. Solid lines show median spectral power ratios, shaded region show inter-quartile range. Dashed line indicates the scale of the LR gridsize.

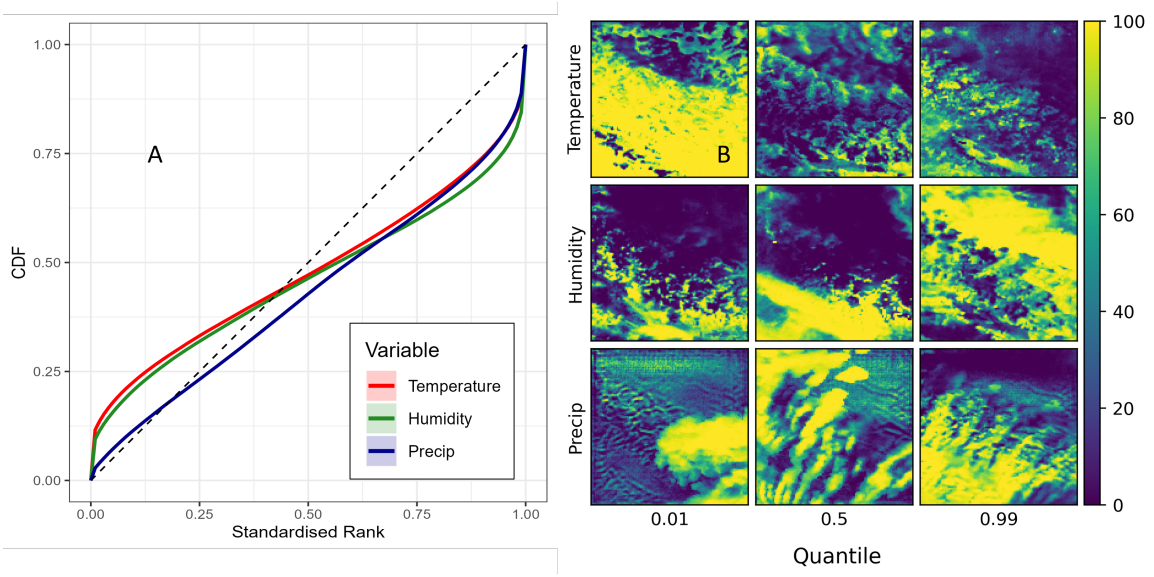

Figure 3.7: a) CDF of rank histograms showing stochastic calibration of conditional distributions for univariate models of temperature, specific humidity, and precipitation. Rank histograms were calculated across 100 randomly selected conditioning fields, with 100 HR realisations of each. Dashed line shows reference uniform CDF. b) Rank histogram maps for individual cases, representing 0.01, 0.5, and 0.99 quantiles of the dataset, showing the rank of the WRF pixel compared to an ensemble of 100 realisations.

underdispersion, with ranks concentrated at either end of the scale instead of being uniformly distributed across possible ranks. Temperature showed good calibration for the median and 0.99 quantile sample, but a cold bias over most of the range for the 0.01 quantile sample. Specific humidity showed the most consistent underdispersion, especially in the 0.01 and median samples. Precipitation generally was better calibrated, but showed some underestimation in areas that typically had high precipitation.

3.3.2 Multivariate Prediction

As expected, multivariate GANs predicting temperature, humidity, precipitation and wind components better represented dependence structures between pairs of downscaled variables (table 3.2). We used normalised mutual information (MI) scores to assess dependence; MI is a measure of mutual dependence between two variables. We

normalised it as follows:

$$\mathrm{MI}_{norm} = 1 - e^{-2\mathrm{MI}}$$

which is the equivalent of R^2 for non-Gaussian variables [8].

Particularly for temperature and humidity, pairs of variables from the univariate model had much lower mutual information scores than those from the WRF, indicating a bias in the modelling of dependence. Mutual information scores for multivariate-predicted variables were slightly lower but close to WRF scores, with one exception. Temperature and precipitation was the only pair of variables where the multivariate model had higher mutual information scores than the corresponding WRF variables.

Table 3.2: Normalised mutual information scores 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	0.364	0.296	0.312
Humid	Precip	0.338	0.298	0.343
Humid	Temp	0.850	0.699	0.870

While measures of dependence generally improved, marginal statistics and individual realisations of generated fields were worse with the full multivariate model. For temperature, humidity, and precipitation, marginal statistics generated from the full multivariate model were evidently fuzzier than those from univariate models (figure 3.8). For example, the 0.99 quantile fields of multivariate predictions showed much worse accuracy than univariate predictions (particulary for precipitation and humidity). Humidity showed the most severe bias, missing many fine-scale details. Precipitation, while capturing the general patterns of the marginal statistics, did not capture the high-precipitation values with the multivariate model. Individual realisations from the multivariate model showed substantial differences in values and textures compared to those from the univariate model (figure 3.9). Particularly for humidity, realisations from the multivariate model were very fuzzy, displaying a low power bias at large scales and a high power bias at near-grid scales.

A possible reason for the relatively poor performance of the multivariate model is the inclusion of precipitation. Precipitation has a very different spatial structure to the other variables, and the singe set of convolutional filters may not succeed in capturing all variables well. We thus tested a multivariate model without precipitation.

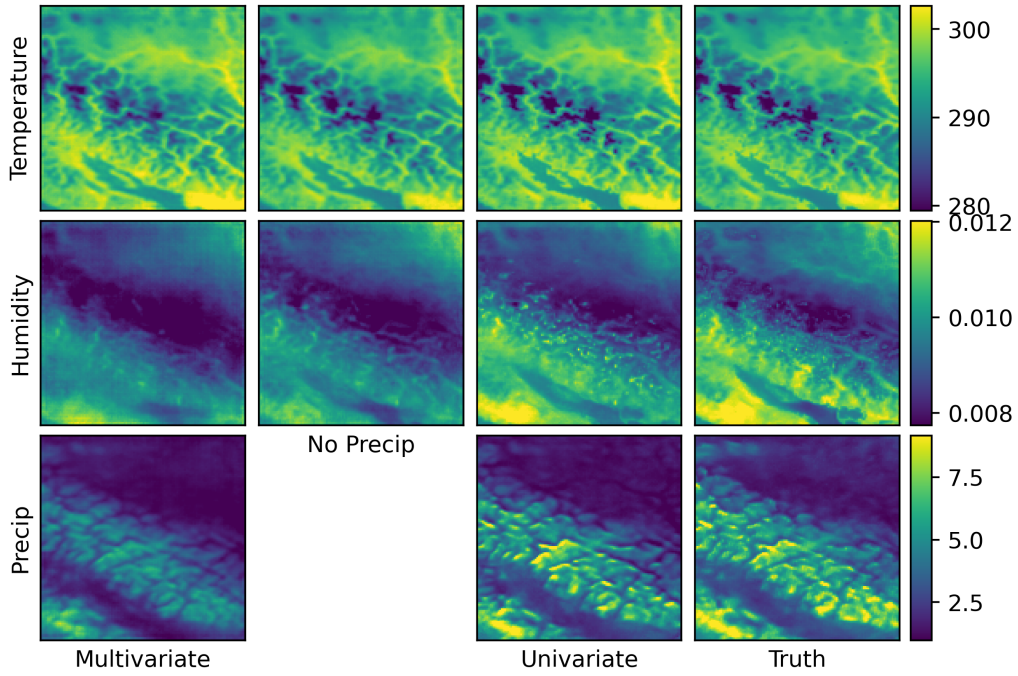

Figure 3.8: Marginal 0.99 quantiles for generated temperature, specific humidity, and precipitation fields, using full multivariate prediction, multivariate prediction without precipitation, and univariate prediction. Quantiles were calculated using 3000 randomly selected timesteps, with one realisation for each timestep.

Results from this model showed improved quality, but resulting downscaled fields and marginal statistics were still fuzzier than those from the univariate model. The fuzziness is especially apparent on the Coast Mountains, where the no-precipitation model does not capture sharp gradients well. Marginal statistics of humidity showed traces of the convolutional filter for both multivariate models. Presence of filter artifacts indicates a poorly performing model, which has not converged. Notably, there are no filter artifacts evident in either marginal statistics or realisations from the univariate models.

Power spectra of all variables showed larger bias for multivariate models compared to univariate models, with the no-precipitation model in between (figure 3.10). Precipitation showed a substantial low-power bias across most wavenumbers in the full multivariate model, at many wavenumbers capturing only 25% of the power of

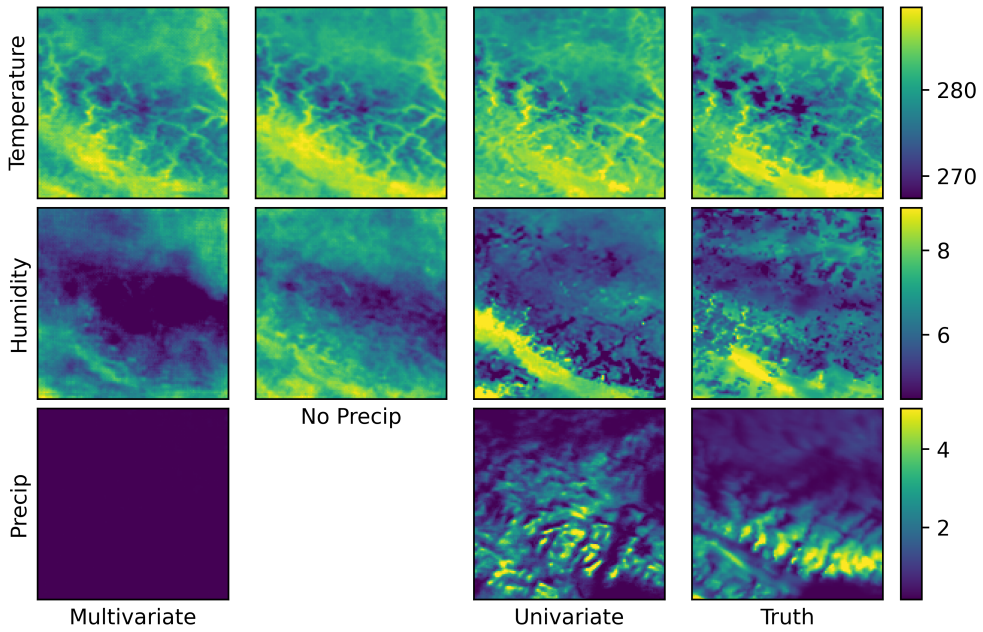

Figure 3.9: Realisations from a single representative timestep in September from the test set for temperature, specific humidity, and precipitation fields, using full multivariate prediction, multivariate prediction without precipitation, and univariate prediction.

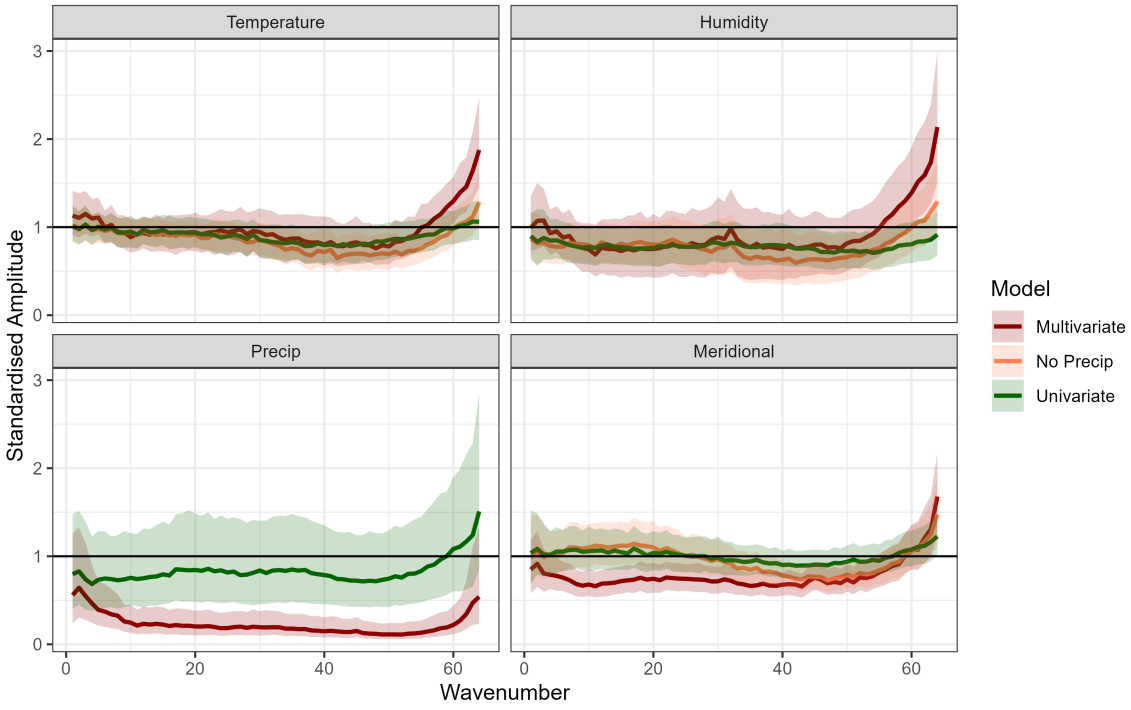

Figure 3.10: Median and IQR RASP for precipitation, specific humidity, temperature, and meridional wind fields, using multivariate, no-precipitation, and univariate models. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields.

the WRF field. Humidity and temperature showed large high-power biases at high wavenumbers in both multivariate models, although to a lesser degree in the no-precipitation model, corresponding to the fuzziness observed in figure 3.9. For humidity, 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.

3.3.3 High-Resolution Topography

To assess the importance of including HR topography as a covariate, we compared models with LR topography, HR topography, and LR topography interpolated to HR. Considering variability across spatial scales for temperature, humidity and precipitation showed that the HR and the interpolated topography both had more similar

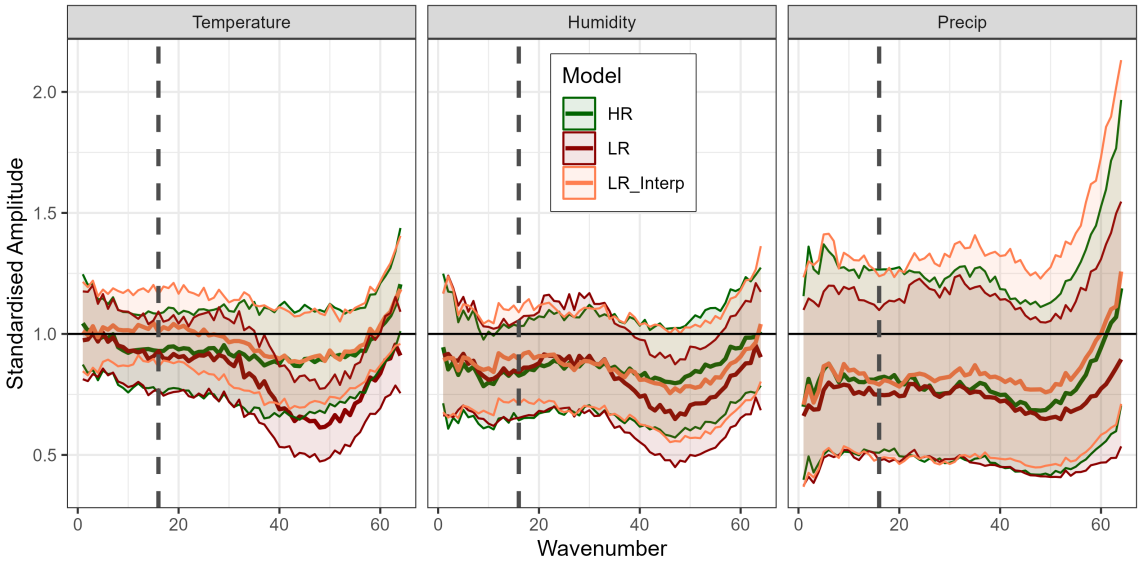

Figure 3.11: Median and IQR RASP for temperature, humidity, and precipitation using HR topography, interpolated LR topography, and LR topography. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields. Dashed line shows wavenumber corresponding to LR grid size.

spectral power than the corresponding LR topography model (figure 3.11). 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 bias 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 little systematic 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 the inclusion of HR information.

3.3.4 Portability in Space

Downscaling in the Northeast region was successful for certain variables, but was less accurate than in the Southwest region (figure 3.12). Downscaling of wind components showed similar quality to the Southwest region, whereas generated temperature and humidity fields often showed substantial differences from WRF. This was especially apparent for humidity, where generated fields were overly smooth and lacked many

of the fine-scale details of WRF. Performance was particularly poor for precipitation; generated fields often had entirely different structure than the WRF field. The fourth row in figure 3.12 shows a representative sample of precipitation, with the generated fields showing substantially different spatial structure than the WRF field.

In this region, most variables showed substantial differences in the large-scale structure of the ERA5 field compared to the WRF field. These were especially apparent for precipitation; for the sample in figure 3.12, the WRF field shows low-intensity precipitation through much of the field, while the ERA5 field shows a concentrated area of precipitation near the centre. 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 an unbiased large scale structure. This model produced substantially better downscaling, with generated precipitation patterns closely matching the WRF field (fifth row of figure 3.12).

Covariate choice was especially important in this region. In particular, we found that CAPE was an important covariate for all variables in this region, whereas in the coastal region, CAPE only improved results for precipitation.

Stochastic calibration of conditional distributions was worse for temperature, humidity and precipitation in the Northeast region than in the Southwest region (figure 3.13). Temperature and humidity both showed underdispersion, with many true samples falling outside the generated range, and precipitation was biased low, with many true samples falling above the generated range. The idealized covariate precipitation model had much better calibration than the standard precipitation model, and was one of the best calibrated models overall. Calibration of spatial structure was also much improved; RASP metrics were substantially closer to corresponding WRF fields, and had lower variability (figure 3.15). Wind components showed similar calibration to the Southwest location. RASP metrics for variables in the Northeast region showed similar median values for humidity, temperature, and wind compared to the Southwest region, but had much larger inter-quantile ranges, representing more variability in texture differences between samples (figure 3.14). Precipitation showed substantial low-power bias across at low wavenumbers, and high power bias through most of the range, where median power was typically 250% higher than corresponding WRF power. This poor match in spatial structure between generated realisations and the WRF field can be clearly seen in the fourth row of figure 3.12.

As alternative apporaches to address the biases in precipitation models for this

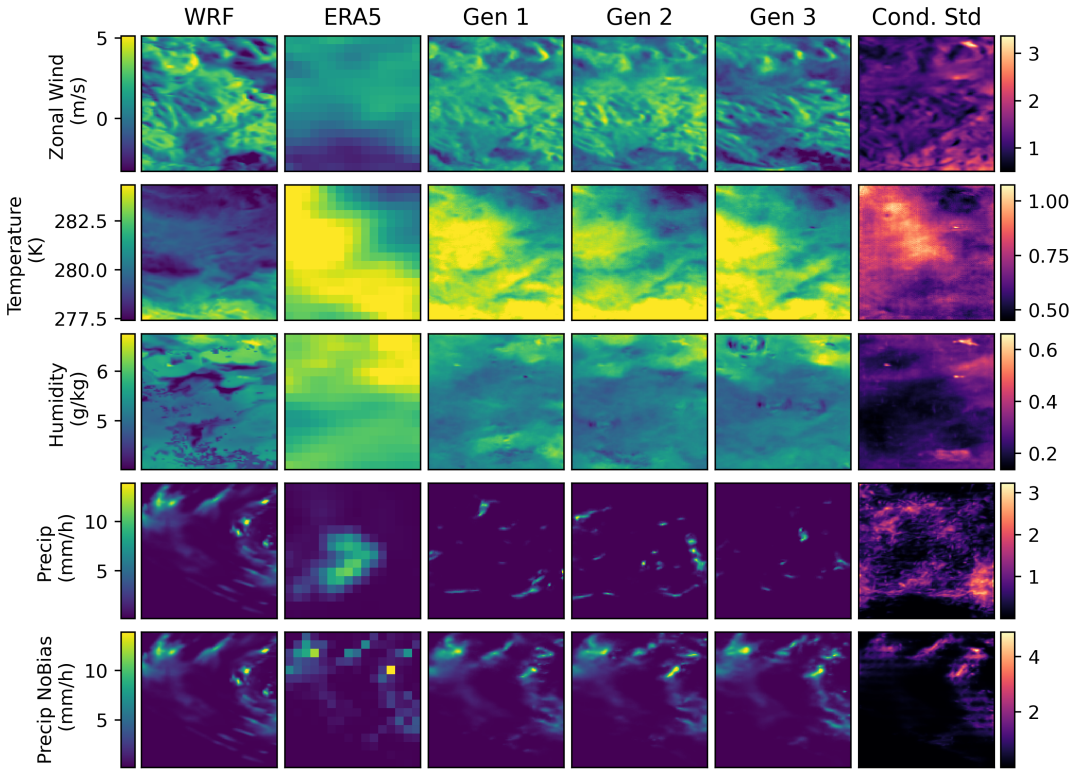

Figure 3.12: Example realisations for the Northeastern region. Rows correspond to variables, and the bottom row shows a second precipitation model with idealised LR training data. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisations

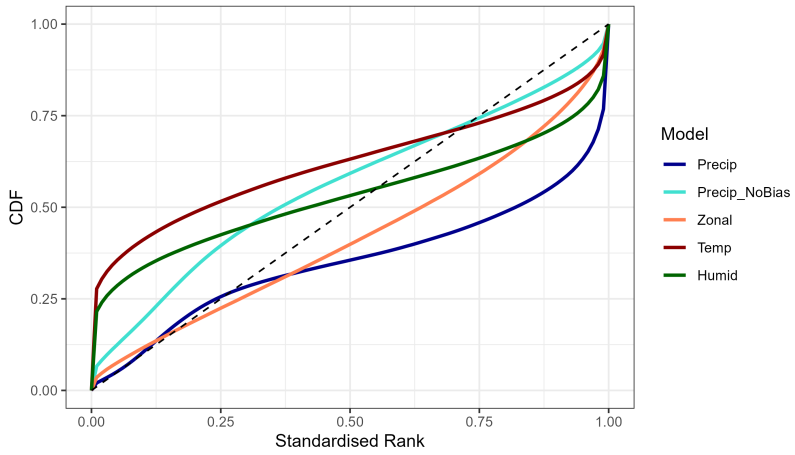

Figure 3.13: CDFs of rank histograms showing stochastic calibration of models in the Northeastern region. Rank histograms were calculated across 100 randomly selected conditioning fields, with 96 HR realisations of each. Dashed line shows reference uniform CDF.

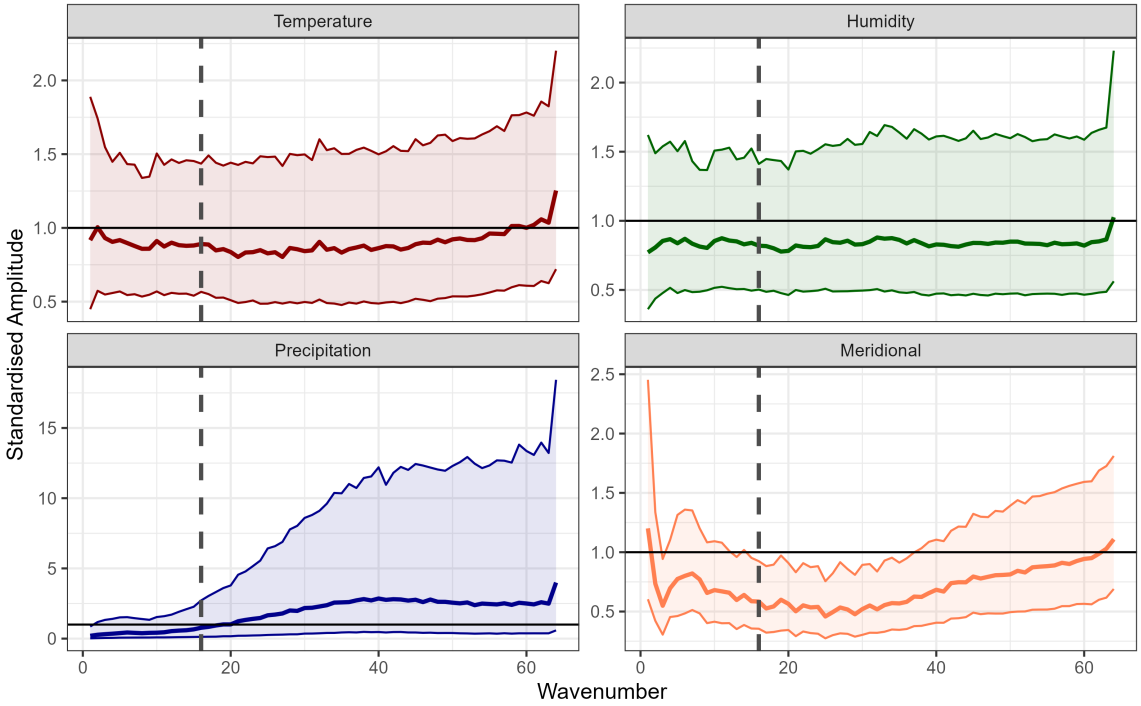

Figure 3.14: RASP metrics for humidity, precipitation, temperature, and meridional wind in the Northeast region, showing median and inter-quartile ranges. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields. Note that y-axis scales differ between plots.

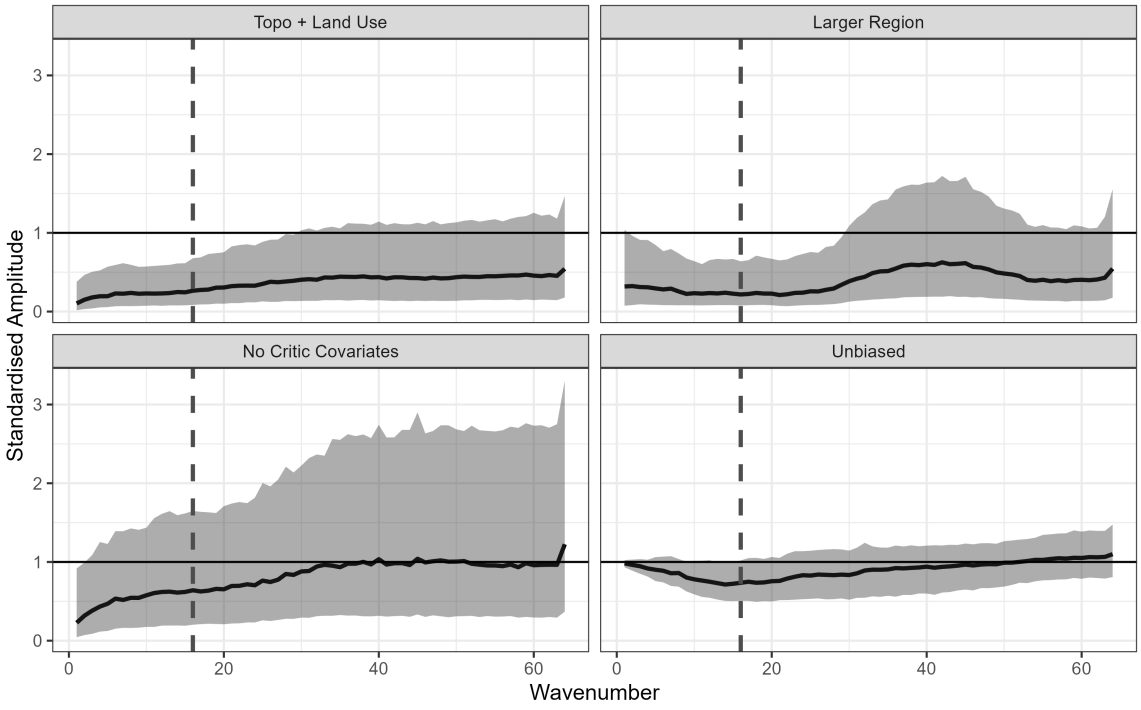

Figure 3.15: RASP metrics for four different precipitation model setups in the Northeast region, showing median and inter-quartile ranges. Spectral powers are standardised to ground truth fields, and metrics are calculated across 1200 randomly selected fields.

region, we trained a suite of models, which included a) adding an HR land use index, b) training on a larger region (224 x 224 pixels) and then analysing a subset of the region, and c) removing all LR covariates from the Critic network (figure 3.15). All models generated fields with reduced power bias than the standard precipitation model. The models with HR land use and with the larger region both had substantial low power biases across wavenumbers. The model without LR Critic covariates showed good calibration of median power at moderate and high wavenumbers, but still had a low power bias at lower wavenumbers and displayed a large amount of variability between samples.

3.4 Discussion

This paper investigates practical considerations of applying the stochastic GAN framework from Chapter 2 to operational downscaling applications. Specifically, we focus

on extension of GANs to multiple climate variables, including the accuracy of multivariate prediction, and portability of the framework to different locations. We show that the stochastic GAN framework can successfully downscale a suite of variables over the complex terrain of Southwestern British Columbia. We then find that while multivariate downscaling improves the dependence structures of downscaled variables, it tends to decrease their individual quality. Finally, we show mixed success in the Northeast region: models for temperature, humidity, and wind components produced reasonable downscaled fields, but showed more bias compared to the Southwest region. Precipitation models in particular struggled in this region, likely due to large-scale differences between the LR and HR training data.

3.4.1 Extension to temperature, humidity, and precipitation

Overall, we found that the stochastic GAN successfully downscaled temperature, specific humidity, and precipitation in the Southwest region, although humidity showed more bias in spatial structure 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 be that we did not include all relevant covariates; for example, it would be interesting to add temporal pressure tendencies as a LR covariate, as changes in pressure influence mesoscale thermal circulation.

Precipitation is an important variable, and is often more challenging to downscale as it has a substantially different distribution than other variables. We found that the stochastic GAN performed well at downscaling precipitation in the Southwest region, and was 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 our findings in Chapter 2, which showed that by sampling from the full HR distribution, the stochastic GAN was better able to estimate wind component extremes.

Overall, the stochastic GAN framework performed much worse in the Northeast region than in the Southwest region, suggesting that downscaling quality is domain dependent. We discuss this issue in more detail below.

3.4.2 Multivariate Prediction

Multivariate prediction led to improved dependence structure between dependent variables, but decreased the accuracy of the generated 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 was especially biased; the univariate model was the only model able to recreate the fine-scale features around the Straight 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 simultaneously predict variables with different distributions in challenging. Perhaps since the convolutional filters being learned for each variable 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 with an otherwise fixed model architecture means that there are fewer adjustable parameters that can be optimised specifically for a single variable. We hypothesise that this may lead to decreased flexibility for the model to adapt to specific variables. 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 break the network into separate branches before the final convolution block, one for each variable being predicted. It is also possible that some fine-scale dependence could be adding by using the same noise fields for each variable in a univariate setting. In general, we advise practitioners to use multivariate prediction for highly coupled variables (e.g., temperature and humidity, wind components), and univariate prediction for less dependent variables (e.g., precipitation), although in most cases the choice of univariate versus multivariate prediction will be situation dependent.

3.4.3 HR Topography

Including HR topography in the Generator improved spatial structures of generated 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.

3.4.4 Portability in Space

Applying the stochastic GAN framework to the Northeast location had mixed success. Generated wind components generally showed similarly high accuracy as in the Southwest region. Generated fields of temperature and humidity were reasonable but not as accurate, and precipitation models were poor, with generated fields showing very different spatial structure than the WRF fields. We hypothesise that some of the challenges in this region, especially with precipitation, were due to large differences between WRF and ERA5 structures at common scales. It was visually apparent that in many samples, the LR conditioning fields did not match the structure of the corresponding WRF fields, more so than in the Southwest domain. The idealised LR precipitation model, where we created the LR conditioning fields by coarsening the WRF fields, produced highly accurate downscaled fields, supporting our hypothesis that this mismatch is a source of the degraded performance.

Unfortunately, in a realistic setting it is generally not possible to have perfectly matched LR and HR fields as they are created by different models (and the HR fields for the prediction data do not exist). This will necessarily introduce differences due to internal variability. However, there will also likely be large-scale biases between datasets, of varying severity depending on the region. Some studies have already investigated the challenge of large-scale biases between paired training data. Price and Rasp [31] 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 hypothesise is often true with precipitation, it becomes a much more difficult problem. In regions with substantial bias, it may be possible to train models using the idealised coarsened data, and then predict using the biased LR dataset. While this technique would not perform any bias correction and would likely be underdispersive, it could perform better at downscaling than an unstable model. An initial investigation into this approach showed promise, but a detailed analysis is beyond the scope of this thesis. Adjustments in model architecture and input fields may also be useful in these biased situation. While none of the precipitation models considered in the Northeast domain performed as well as the idealised covariate model, we improved resulting spatial structure by removing the LR fields as input to the Critic network. We hypothesise that if there is too great a mismatch between HR and LR fields, providing the LR fields to the Critic hinders its ability to estimate an accurate Wasserstein distance.

Covariate choice had a large effect on downscaling accuracy in the Northeast region compared to the Southwest region. In the Northeast region, CAPE was important for all variables, including wind components, while in the Southwest region, CAPE only improved results for precipitation. This domain-dependence of variable importance is similar to results of Annau et al. [3]. Since substantial fine-scale variability can result from convection, especially in inland regions, it makes sense that CAPE is an important variable in the Northeast region. In general, different suites of LR covariates will be needed depending on the downscaling region. Therefore, to obtain accurate downscalings over large areas, it will likely be necessary to include more covariates than required for any particular smaller region, as all subdomains will require the covariates important for their specific weather patterns. We did not use a formal process for selecting final covariate sets in this study; decisions were based on physical reasoning and preliminary results using the training data. It may be worthwhile to apply a more structured variable selection technique in future studies.

3.5 Conclusions

It is becoming increasingly common for governments, industries, and other organisations to use downscaled climate data for modeling, planning, and adaptation purposes. Most of downscaled products easily available do a poor job at capturing climatic extremes, which are often the most important [5]. 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. While substantial research has occurred in this field recently, deep-learning downscaling has not be 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 and successfully represents the near extremes (0.01 and 0.99 quantiles) of their distributions. We then find that multivariate prediction improves dependence between variables but decreases accuracy, and that while the framework which is effective for Southwest British Columbia can be applied to a different region, there are some challenges related to large-scale mismatch 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.

Chapter 4

Conclusions

This thesis presents two projects, both aiming to improve GAN-based downscaling of climate variables. While GANs have shown success at downscaling climate variables, most research to date has been deterministic, and thus not easily able to capture the variability inherent in the downscaling process. To this end, I developed a stochastic GAN framework, where by injecting noise directly into the layers of the neural network, and adjusting the training metric to include a probabilistic loss function, I enabled the model to sample multiple realisations from a conditional HR distribution. In Chapter 2, I showed with synthetic data that this method of noise injection created a better calibrated model compared to the more standard method of including noise as a covariate. In other words, the model was better at sampling realisations from the full distribution, instead of only drawing from the centre of the distribution – a common challenge with GANs. Testing the new stochastic GAN on a realistic setting of downscaling wind components, I then showed that the stochastic GAN performed better at quantifying extremes than the deterministic models. Being able to accurately predict local-scale climate extremes is of crucial importance to climate adaptation.

Although substantial research has investigated GANs as a method of climate downscaling, there are various questions to be addressed prior to its use in an operational setting. My second manuscript, presented here in Chapter 3, investigates some of these. I found that for most situations, the stochastic GAN framework could be successfully applied to a suite of important variables (temperature, humidity, precipitation, and wind components) while improving estimates of moderate extremes. I then showed that while multivariate downscaling can improve dependence structures between downscaled variables, it led to poorer accuracy overall, and should likely

only be used on highly coupled variables. Finally, I tested the generalisability of the stochastic GAN framework to a different spatial region, and found that while it maintained good performance for some variables, it failed for precipitation, partly due to large-scale mismatches in the HR and LR training data.

Although predicting downscaled fields from a trained GAN is very computationally efficient, training the models requires substantial computational resources, especially GPU memory. Specifically, memory requirements increase substantially with increasing domain size, and training a convection-permitting scale model for all of Canada would not be possible with the setup I used in this thesis. However, the ability to downscale contiguous large regions is crucial for many operational products. Tiling methods, which have already shown success with GANs in medical applications, show promise for addressing this challenge, and in my opinion, are the most important avenue for future research. Particularly, it will be important to develop tiling methods that work in a stochastic setting without creating artifacts between tiles. In Chapter 2, I showed that downscaling accuracy was dependant on region; this will pose some challenges for tiling, as the resulting downscaling may only be accurate in certain locations. GAN generalisability over time will also be an important topic to study; ideally, GANs could be trained on historic/current climates, and used to downscale future LR projections (e.g., ESM output). However, the accuracy of this technique is unknown due to the non-stationarity of the climate, and should be investigated prior to operational use. Since ESMs will tend to show increased bias compared to reanalysis products (which the GANs are trained on), downscaled fields of ESMs will likely retain this bias.

Adding explicit time dependence to stochastic GANs will be important for ensuring consistency between consecutive time steps. In the current setup, temporal consistency is only maintained by the LR conditioning fields, meaning that small-scale features may not be consistent in time. Leinonen et al. [20] showed that by including a Convolutional Gated Recurrent Unit in the GAN, downscaled output showed consistency at all scales. It will be interesting to try adding such a recurrent architecture to the stochastic GAN developed here. Future research should also investigate the sensitivity of the stochastic framework to variations in hyperparameters as I did not perform a formal tuning step in this research. In particular, the learning rate of the optimiser will likely effect model performance. Finally, improving the performance of multivariate GANs will require further research. I believe a first step would involve testing new Generator architectures with separate streams for each variable (or

variables set) near the end of the network. Developing a multivariate GAN capable of producing accurate downscaled fields would help ensure consistency between variables, and reduce computational requirements for training.

Even with a highly accurate GAN downscaling, many applications will require post-processing of downscaled output prior to use. This research focuses on creating a WRF emulator; my goal is to create a computationally efficient method for producing WRF-like output. Given that WRF does not use real world observations, WRF output is often biased compared to observed climates, and requires post-processing to reduce biases. Since the GAN models presented here treat WRF as the ground truth, resulting downscaling will require at least as much bias correction as WRF output. This post-downscaling bias correction is a crucial area of research, but separate from deep-learning emulation. Hopefully, the same bias correction techniques developed for WRF can be used for GAN output. GANs can also be trained using observational datasets (e.g., radar observations of precipitation, as in Harris et al. [14]), in which case the resulting downscaled fields would not require bias correction to the same extent. Our research into stochastic GANs should also be applicable to such observational downscaling tasks.

Adaptation to anthropogenic climate change requires local-scale, accurate climate data and projections. Particularly, knowledge of extremes is crucial for infrastructure, crops, and conservation. This research shows that stochastic GANs could be a useful, computationally efficient tool for downscaling LR data to local scales, especially since they improve quantification of variability, and thus estimates of extremes. While there is still research needed, we hope that this research paves a path for deep learning downscaling to be used operationally in the near future.

Appendix A

Supplementary Material

A.1 Model Parameters

A.2 Additional Figures

Parameter	Value
Epochs	250
Batch Size	16
Optimiser	Adam
Learning Rate	0.00025
Gradient Penalty Weight	10
Adversarial Loss Weight	0.01
Content Loss Weight	20
Critic Iterations/Generator Iteration	5

Table A.1: Parameter values for GAN training.

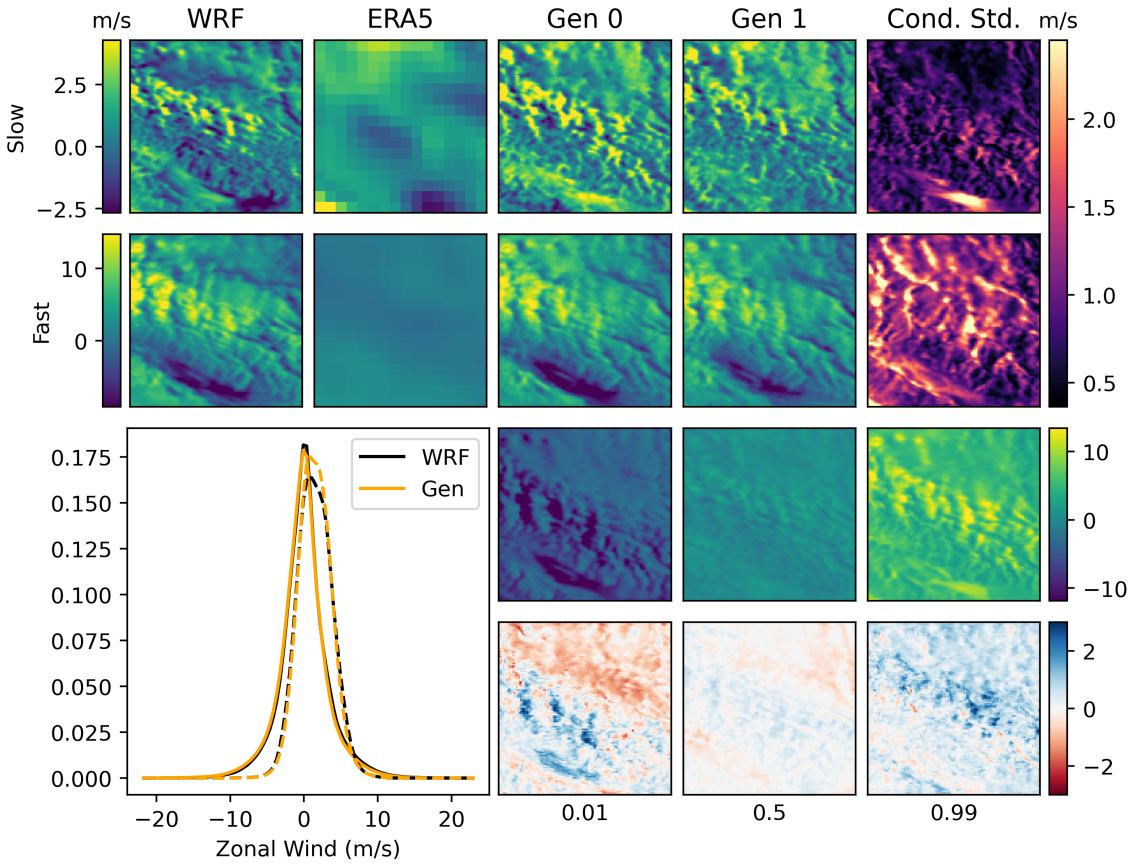

Figure A.1: Evaluation of univariate GAN downscaling of zonal wind for the Southwest study area. Top two rows show respectively an example cold and warm sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.99 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences relative to the ground truth (truth - generated). The left-bottom panel shows overall PDFs (across space and time) of pixel values for January samples (solid) and July samples (dashed).

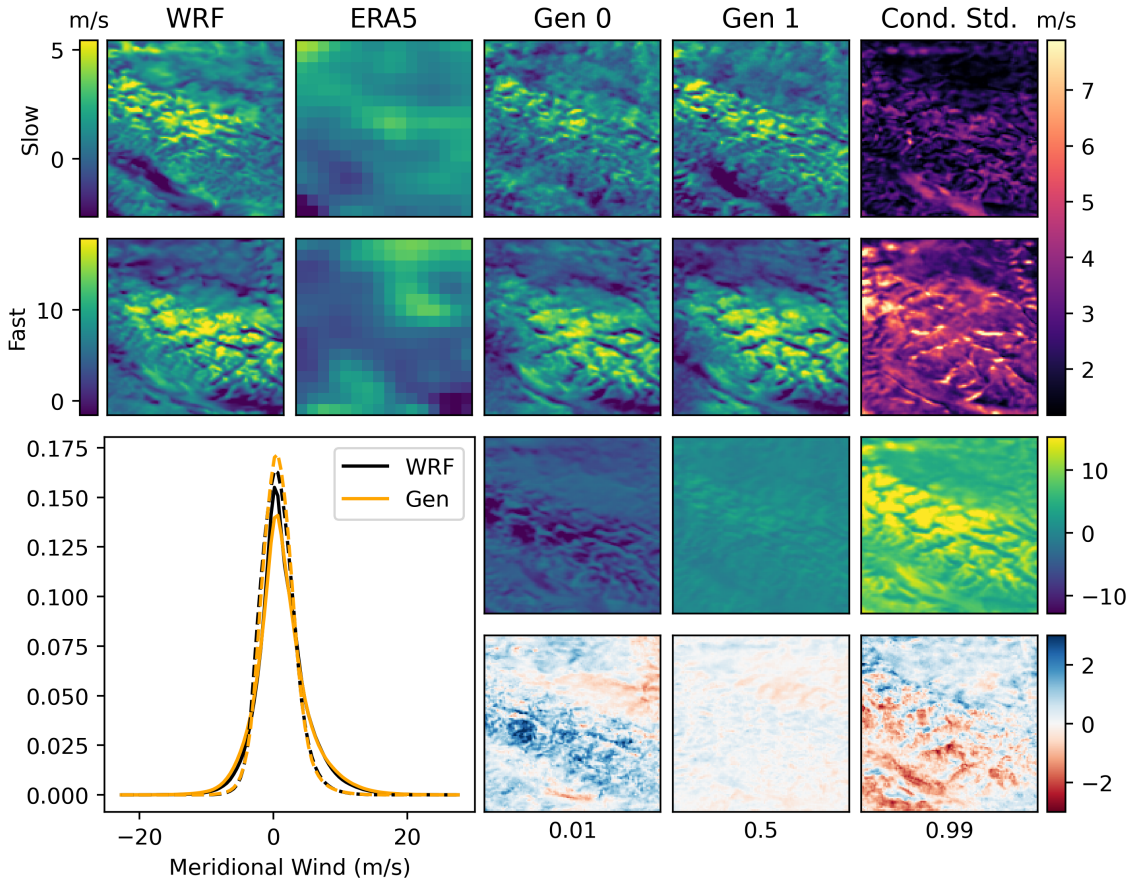

Figure A.2: Evaluation of univariate GAN downscaling of meridional wind for the Southwest study area. Top two rows show respectively an example cold and warm sample, with the WRF field, LR conditioning field, two stochastic realisations, and the conditional pixel-wise standard deviations across 100 ensemble members. The third row shows 0.01, 0.5, and 0.99 quantiles over 3000 random samples from the generated fields, and the bottom row shows the corresponding quantile differences relative to the ground truth (truth - generated). The left-bottom panel shows overall PDFs (across space and time) of pixel values for January samples (solid) and July samples (dashed).

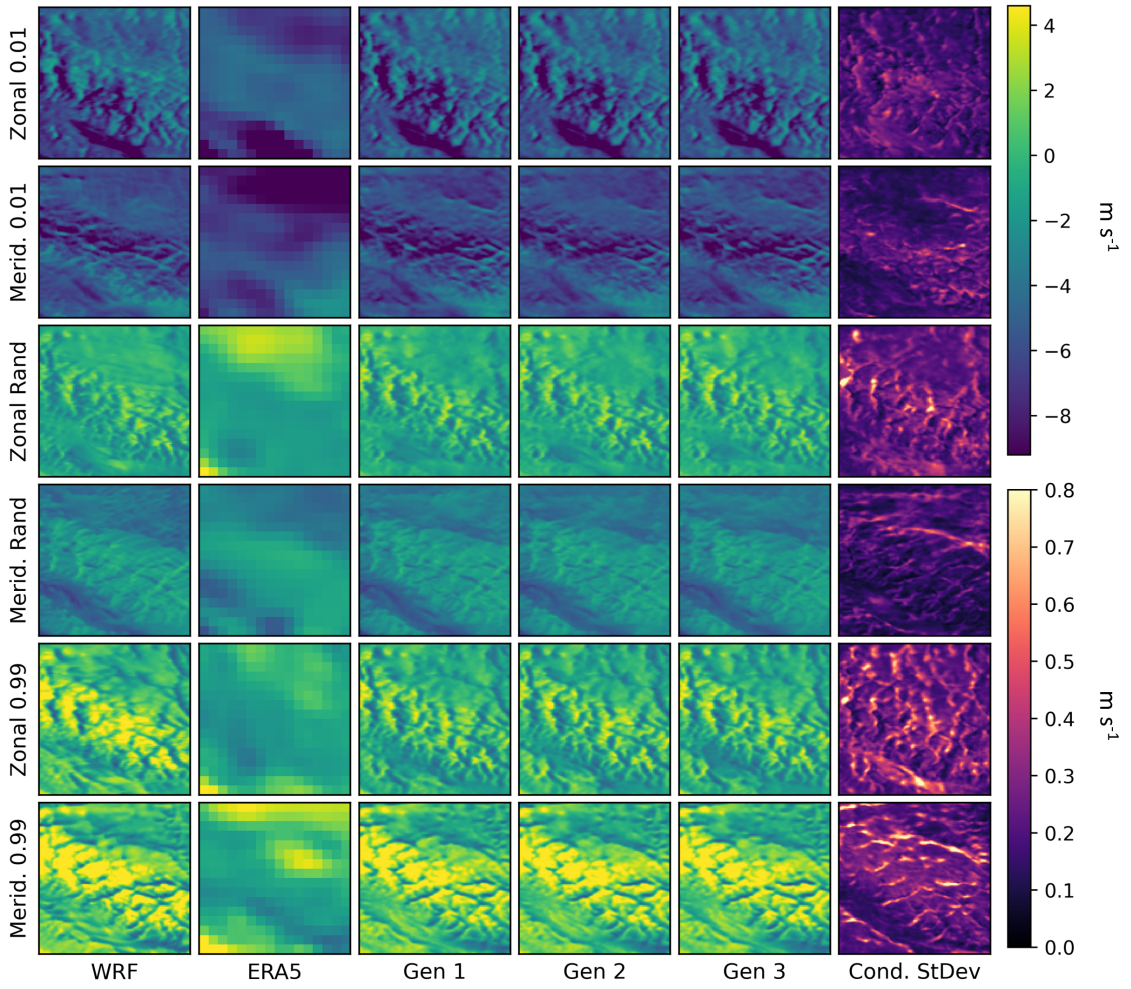

Figure A.3: Example meridional and zonal wind fields for coastal BC using the S_{full}^{MAE} model. First two rows show hours representative of 0.01 quantiles, middle two rows show randomly selected hours from the full test set, and bottom rows show hours representative of 0.99 quantiles. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisations.

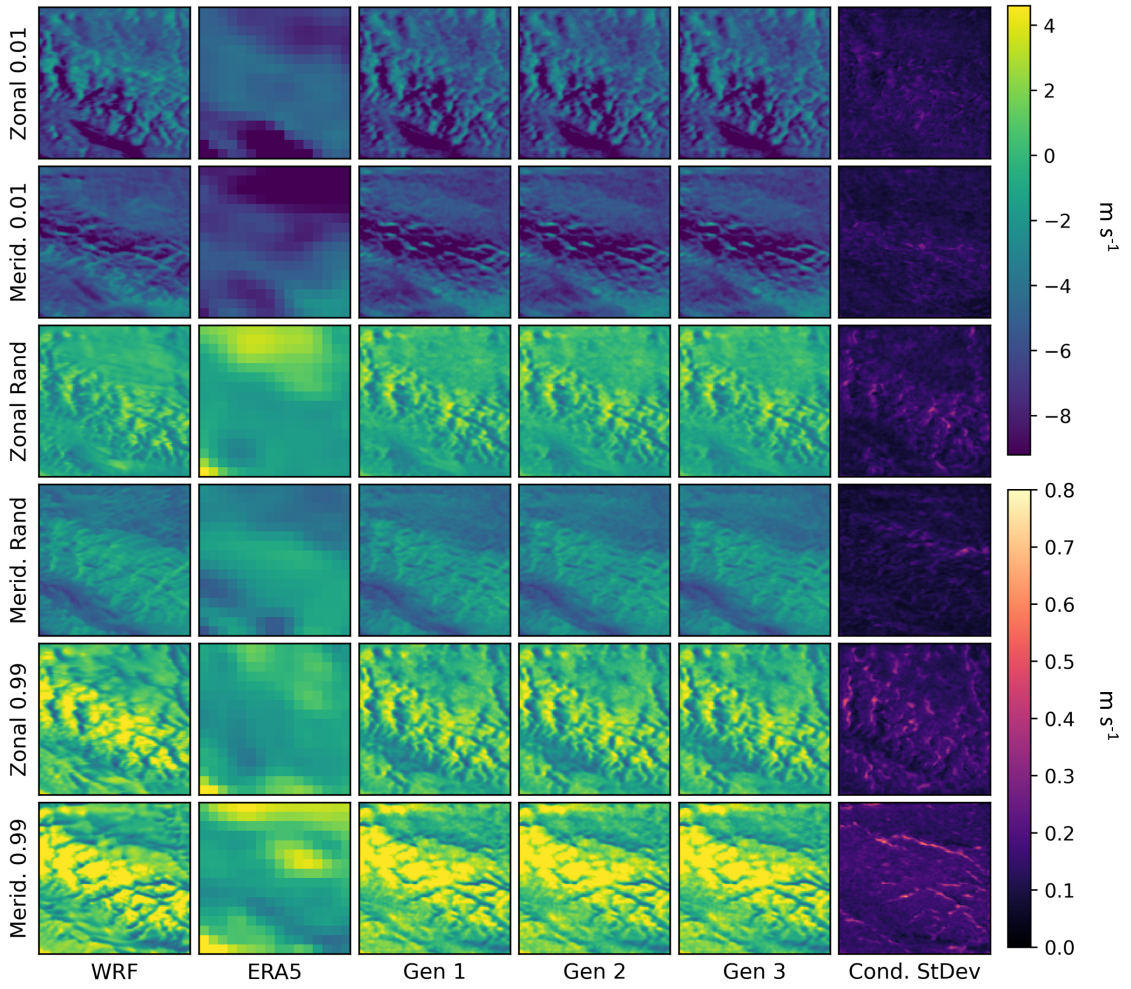

Figure A.4: Example meridional and zonal wind fields for coastal BC using the F_{full}^{MAE} model. First two rows show hours representative of 0.01 quantiles, middle two rows show randomly selected hours from the full test set, and bottom rows show hours representative of 0.99 quantiles. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisations.

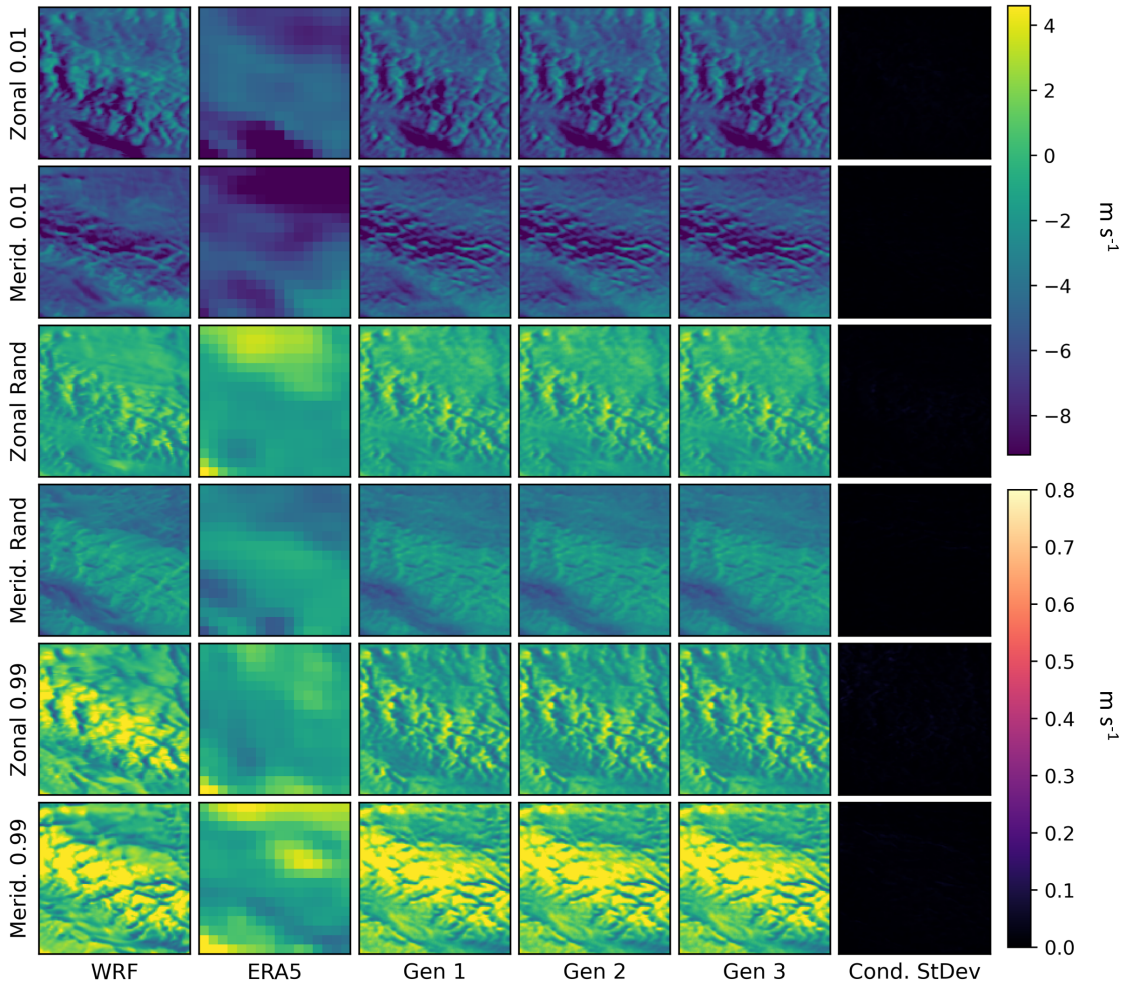

Figure A.5: Example meridional and zonal wind fields for coastal BC using the F_{NC}^{MAE} model. First two rows show hours representative of 0.01 quantiles, middle two rows show randomly selected hours from the full test set, and bottom rows show hours representative of 0.99 quantiles. Columns show, from left to right, WRF (i.e. ground truth), ERA5 (input conditioning field), three generated realisations, and the conditional standard deviations across 500 realisations.

Bibliography

[1] John T Abatzoglou and Timothy J Brown. A comparison of statistical downscaling methods suited for wildfire applications. *International journal of climatology*, 32(5):772–780, 2012.

[2] Afshin Afshari, Julian Vogel, and Ganesh Chockalingam. Statistical downscaling of SEVIRI land surface temperature to WRF near-surface air temperature using a deep learning model. *Remote Sensing*, 15(18):4447, 2023.

[3] Nicolaas J Annau, Alex J Cannon, and Adam H Monahan. Algorithmic hallucinations of near-surface winds: Statistical downscaling with generative adversarial networks to convection-permitting scales. *Artificial Intelligence for the Earth Systems*, 2(4):e230015, 2023.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[5] ZHOU Bo-Tao and Qian Jin. Changes of weather and climate extremes in the ipcc ar6. *Advances in Climate Change Research*, 17(6):713, 2021.

[6] Alex J Cannon. Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics*, 50(1):31–49, 2018.

[7] Fengrui Chen, Yu Liu, Qiang Liu, and Xi Li. Spatial downscaling of TRMM 3B43 precipitation considering spatial heterogeneity. *International Journal of Remote Sensing*, 35(9):3074–3093, 2014.

[8] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[9] Zhitong Ding, Shuqi Jiang, and Jingya Zhao. Take a close look at mode collapse and vanishing gradient in gan. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 597–602, 2022. doi: 10.1109/ICETCI55101.2022.9832406.

[10] EM Fischer, Sebastian Sippel, and Reto Knutti. Increasing probability of record-shattering climate extremes. *Nature Climate Change*, 11(8):689–695, 2021.

[11] Jose George and P Athira. A model output statistic-based probabilistic approach for statistical downscaling of temperature. *Theoretical and Applied Climatology*, pages 1–20, 2024.

[12] Nathan P Gillett, Alex J Cannon, Elizaveta Malinina, Markus Schnorbus, Faron Anslow, Qiaohong Sun, Megan Kirchmeier-Young, Francis Zwiers, Christian Seiler, Xuebin Zhang, et al. Human influence on the 2021 british columbia floods. *Weather and Climate Extremes*, 36:100441, 2022.

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[14] Lucy Harris, Andrew TT McRae, Matthew Chantry, Peter D Dueben, and Tim N Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *arXiv preprint arXiv:2204.02028*, 2022.

[15] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[16] Zetao Jiang, Yongsong Huang, and Lirui Hu. Single image super-resolution: Depthwise separable convolution super-resolution generative adversarial network. *Applied Sciences*, 10(1):375, 2020.

[17] Josiah L Kephart, Brisa N Sánchez, Jeffrey Moore, Leah H Schinasi, Maryia Bakhtsiyarava, Yang Ju, Nelson Gouveia, Waleska T Caiaffa, Iryna Dronova, Saravanan Arunachalam, et al. City-level impact of extreme temperatures and mortality in Latin America. *Nature Medicine*, 28(8):1700–1705, 2022.

[18] Bipin Kumar, Kaustubh Atey, Bhupendra Bahadur Singh, Rajib Chattopadhyay, Nachiketa Acharya, Manmeet Singh, Ravi S Nanjundiah, and Suryachandra A Rao. On the modern deep learning approaches for precipitation downscaling. *Earth Science Informatics*, 16(2):1459–1472, 2023.

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[20] Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2020.

[21] Yanping Li, Zhenhua Li, Zhe Zhang, Liang Chen, Sopan Kurkute, Lucia Scaff, and Xicai 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*, 23(11):4635–4659, 2019.

[22] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[23] Petter Lind, Danijel Belušić, Ole B Christensen, Andreas Dobler, Erik Kjellström, Oskar Landgren, David Lindstedt, Dominic Matte, Rasmus A Pedersen, Erika Toivonen, et al. Benefits and added value of convection-permitting climate modeling over fenno-scandinavia. *Climate Dynamics*, 55:1893–1912, 2020.

[24] Tania Lopez-Cantu, Marissa K Webber, and Constantine Samaras. Incorporating uncertainty from downscaled rainfall projections into climate resilience planning in us cities. *Environmental Research: Infrastructure and Sustainability*, 2(4): 045006, 2022.

[25] Philippe Lucas-Picher, Daniel Argüeso, Erwan Brisson, Yves Tramblay, Peter Berg, Aude Lemonsu, Sven Kotlarski, and Cécile Caillaud. Convection-

permitting modeling with regional climate models: Latest developments and next steps. *Wiley Interdisciplinary Reviews: Climate Change*, 12(6):e731, 2021.

[26] William H MacKenzie and Colin R Mahony. An ecological approach to climate change-informed tree species selection for reforestation. *Forest Ecology and Management*, 481:118705, 2021.

[27] Douglas Maraun. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26(6):2137–2143, 2013.

[28] Douglas Maraun. Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, 2(4):211–220, 2016.

[29] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual diffusion modeling for km-scale atmospheric downscaling. *Preprint*, 2024.

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[31] Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 10555–10571. PMLR, 2022.

[32] N. C. Rakotonirina and A. Rasoanaivo. Esrgan+ : Further improving enhanced super-resolution generative adversarial network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3637–3641, 2020.

[33] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[34] Andrea Rossa, Pertti Nurmi, and Elizabeth Ebert. Overview of methods for the verification of quantitative precipitation forecasts. In *Precipitation: Advances in measurement, estimation and prediction*, pages 419–452. Springer, 2008.

[35] Yunus Saatci and Andrew G Wilson. Bayesian gan. *Advances in neural information processing systems*, 30, 2017.

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan 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*, pages 1874–1883, 2016.

[37] William C Skamarock, Joseph B Klemp, and Jimy Dudhia. Prototypes for the WRF (weather research and forecasting) model. In *Preprints, Ninth Conf. Mesoscale Processes, J11–J15, Amer. Meteorol. Soc., Fort Lauderdale, FL*, volume 1, 2001.

[38] Karen Stengel, Andrew Glaws, Dylan Hettinger, and Ryan N King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29):16805–16815, 2020.

[39] Ross M Thompson, John Beardall, Jason Beringer, Mike Grace, and Paula Sardina. Means and extremes: building variability into community-level climate change experiments. *Ecology Letters*, 16(6):799–806, 2013.

[40] Thordis L Thorarinsdottir, Michael Scheuerer, and Christopher Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of computational and graphical statistics*, 25(1):105–122, 2016.

[41] Geert Jan Van Oldenborgh, Karin van Der Wiel, Sarah Kew, Sjoukje Philip, Friederike Otto, Robert Vautard, Andrew King, Fraser Lott, Julie Arrighi, Roop Singh, et al. Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166(1):13, 2021.

[42] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514–63537, 2020.

[43] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[44] Daniel S Wilks. Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907, 2010.

[45] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.