Quantile-Regression-Ensemble: A Deep Learning Algorithm for Downscaling Extreme Precipitation

Thomas Bailie[1], **Yun Sing Koh**[1], **Neelesh Rampal**[2], **Peter B. Gibson**[2]

[1] School of Computer Science, The University of Auckland, New Zealand
[2] National Institute of Water and Atmospheric Research, New Zealand
tbai869@aucklanduni.ac.nz, y.koh@auckland.ac.nz, Neelesh.Rampal@niwa.co.nz, Peter.Gibson@niwa.co.nz

Abstract

Global Climate Models (GCMs) simulate low resolution climate projections on a global scale. The native resolution of GCMs is generally too low for societal-level decision-making. To enhance the spatial resolution, downscaling is often applied to GCM output. Statistical downscaling techniques, in particular, are well-established as a cost-effective approach. They require significantly less computational time than physics-based dynamical downscaling. In recent years, deep learning has gained prominence in statistical downscaling, demonstrating significantly lower error rates compared to traditional statistical methods. However, a drawback of regression-based deep learning techniques is their tendency to overfit to the mean sample intensity. Extreme values as a result are often underestimated. Problematically, extreme events have the largest societal impact. We propose Quantile-Regression-Ensemble (QRE), an innovative deep learning algorithm inspired by boosting methods. Its primary objective is to avoid trade-offs between fitting to sample means and extreme values by training independent models on a partitioned dataset. Our QRE is robust to redundant models and not susceptible to explosive ensemble weights, ensuring a reliable training process. QRE achieves lower Mean Squared Error (MSE) compared to various baseline models. In particular, our algorithm has a lower error for high-intensity precipitation events over New Zealand, highlighting the ability to represent extreme events accurately.

Introduction

Given the large-scale socioeconomic impacts that extreme weather events have on society, detailed information on their spatial patterns, and how they may change under climate change, is highly sought out by decision makers and stakeholders. For instance, extreme precipitation events can have severe economic costs (Frame et al. 2020; Merz et al. 2010) and cause large population displacement due to flooding, with developing countries being particularly vulnerable. End of century climate projections can be obtained via a GCM. Despite their sophistication, however, the low spatial resolution of their output imposes large uncertainties at local scales. Nevertheless, climate projections produced by GCMs are invaluable for macro-level decision making. A potential solution is to enhance the resolution of climate projections over a smaller geographical area is by coupling GCM outputs with a Regional Climate Model (RCM). However, the high computational cost of RCMs limits their practicality when downscaling GCM projections. In contrast to RCMs, statistical downscaling methods directly construct a map between low-resolution climate variables (*e.g.*, humidity or pressure) from a GCM to a high-resolution climate variable such as precipitation, running orders of magnitude faster than RCM orientated methods. Deep learning has recently emerged as a prominent method in statistical downscaling. Existing literature is rich in approaches that successfully use statistical downscaling (Baño-Medina et al. 2022; Adewoyin et al. 2021; Vandal et al. 2018; Oyama et al. 2023). We adopt a similar methodology, mapping low-resolution meteorological variables to high-resolution precipitation over New Zealand. However, deep-learning-based implementations often underestimate high-intensity extreme events when trained over the observational record (Baño-Medina et al. 2022; Rampal et al. 2022). To address this concern, we introduce the Quantile-Regression-Ensemble (QRE) algorithm as depicted in Figure 1, which trains a series of members on distinct subsets of precipitation observations corresponding to specific intensity levels. As a result, members can achieve significantly lower error MSE within

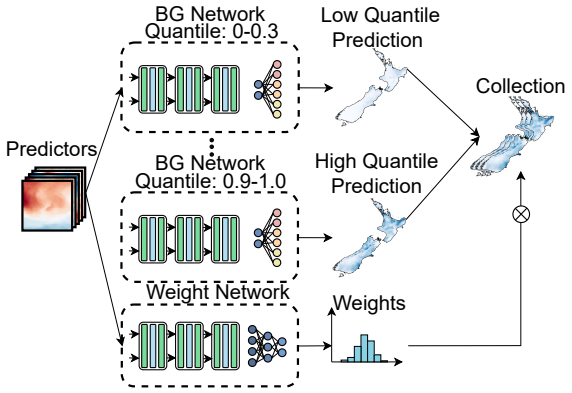

Figure 1: An overview of Quantile-Regression-Ensemble. The ensemble is formed by training each member on distinct quantiles of precipitation intensity. Samples are softly assigned to a member via the Weight Network.

their original intensity than the same model trained over the entire dataset. Our approach employs dynamic weight assignment, matching incoming data to the appropriate member using the Weight Network, thus forming an ensemble. QRE provides a robust solution to the *regression to the mean* phenomenon, a problem experienced by many deep-learning based downscaling approaches. We evaluate our algorithm's performance across different precipitation intensity ranges and over several key regions in New Zealand. Our contributions are as follows:

- We developed Quantile-Regression-Ensemble, an ensemble-based algorithm that trains members over a partition of the dataset, allowing each member to capture a specific intensity level of precipitation. In rigorous evaluations across various domains, QRE demonstrated a significant increase in accuracy for extreme precipitation levels.
- QRE accurately predicts extreme precipitation events, giving decision-makers high-resolution information on climate scenarios and enhancing preparedness against extreme events. Because of its compatibility with many regression models, QRE emerges as a viable tool in climate projections and disaster preparedness.
- QRE's enhanced accuracy over extreme values indicates its ability to aggregate local information from members via a classifier into high-quality global information.

Related Work

Statistical Downscaling. Recent advances in Deep learning-based climate downscaling have many parallels with image super-resolution (SR), which aims to reconstruct a high-resolution image from a corresponding low-resolution image (Xia et al. 2022; Dong et al. 2014; Liu et al. 2023). In climate downscaling, a deep-learning model takes low-resolution meteorological predictor variables and enhances them to generate a detailed high-resolution climate variable such as precipitation. Baño-Medina, Manzanas, and Gutiérrez (2020) trained a Convolutional Neural Network (CNN) to downscale precipitation and temperature across continental Europe. When downscaling precipitation, they trained their model on the log-likelihood of the Bernoulli-Gamma distribution. A similar approach has also been adopted to downscale precipitation over New Zealand (Rampal et al. 2022). While using the Bernoulli-Gamma loss function improved performance over the traditional MSE loss across all precipitation levels, the model nevertheless overestimates low-intensity precipitation events while underestimating high-intensity events. Once trained, a downscaling model enhances climate projections via coupling to a GCM, allowing for local information to be harnessed in decision making. Our QRE, in particular, accurately captures both high and low-intensity events, allowing access to crucial information from a decision-making perspective.

Boosting. Gradient boosting methods (Mayr et al. 2014) train multiple instances of the same model on a dataset and then weight their contributions to maximise accuracy by sequentially correcting the errors of the preceding models. Moghimi et al. (2016), introduced boosting method for CNNs for an image classification task using a variation on the GD-MCBoost algorithm, which aims to construct the collection of models such that each new member maximally decreases the training error. However, an imbalance exists for climate data between the sparsity and magnitude of samples. As a consequence, fixed ensemble weight methods are prone to overfitting.

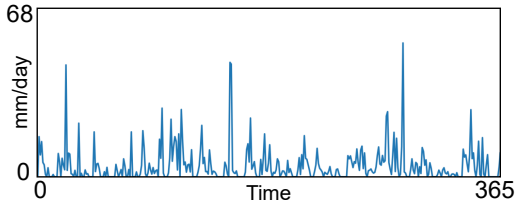

Figure 2: Precipitation time series from the VCSN data, captured over a year for a single grid cell in New Zealand.

Dataset

We use a regression-based downscaling approach, which maps low-resolution meteorological predictor variables to high-resolution precipitation observations over New Zealand. Here, our low-resolution meteorological variables are obtained from the ERA5 reanalysis dataset, the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF). Our high-resolution precipitation observations are from the Virtual Climate Station Network (VCSN).

VCSN. Our ground truth precipitation observations originate from the Virtual Climate Station Network (VCSN), which employs interpolation on surface-level observational data to produce gridded precipitation observations at a resolution of 5km (Tait et al. 2006; Tait, Sturman, and Clark 2012). Notably, bias caused by interpolation is particularly pronounced in regions sparse in the number of observational stations (Gibson et al. 2019), which is particularly evident in the Southern Alps (Tait, Sturman, and Clark 2012). In this study, we format our gridded two-dimensional precipitation observations into a single one-dimensional vector of size 11491, where each entry corresponds to a distinct grid point within New Zealand. Figure 2 shows precipitation observations from the VCSN data over a year for a grid point in New Zealand. The extremes of precipitation refer to the sudden spikes displayed by the data.

ERA5. The ERA5 reanalysis is a state-of-the-art synthesis of global observational data. The reanalysis integrates observational data from numerous sources, including station-based and satellite measurements, to generate consistent and accurate representations of the atmosphere (Hersbach et al. 2020). We focus on meteorological variables within the geographic range of $(150E^{\circ}, 170W^{\circ})$ in latitude and $(25S^{\circ}, 50S^{\circ})$ in longitude. The ERA5 reanalysis has been coarsened from its native resolution of approximately 30km to an approximate spatial resolution of about 100km - the

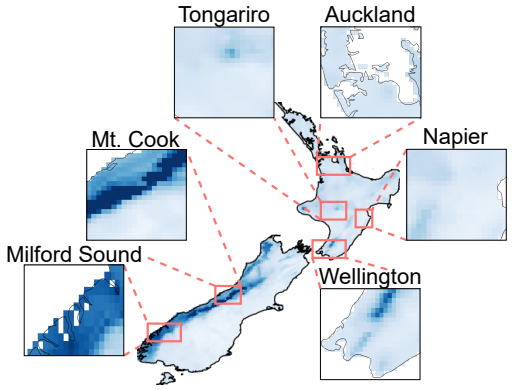

Figure 3: Spatial plot of mean precipitation over New Zealand, alongside key locations on which we evaluate our methods.

typical resolution of a GCM (Eyring et al. 2016). Once our algorithm is trained on ERA5, it can be applied to downscale any GCM, a technique described in climate downscaling as perfect prognosis. Additionally, we normalise the ERA5 data, ensuring that all samples have a mean of 0 and a standard deviation of 1. Previous research by Rampal et al. (2022) performed sensitivity testing on the geographic region and climate predictor selection. They empirically demonstrated that the set of 5 variables, at a pressure level of 850hPa, or roughly 3.5km above sea level, maximally reduces MSE over the mentioned geographical region. Specifically, these variables are air temperature (T_{850}), specific humidity (Q_{850}), zonal wind speed (U_{850}), meridional wind speed (V_{850}), and vertical wind speed (W_{850}). After coarsening, our data consequently has shape $(5, 36, 41)$. The precipitation vector has shape $(1, 11491)$.

Elevation. Alongside the ERA5 and VCSN precipitation observations, we also include the surface elevation of New Zealand at a resolution of 5km as a static predictor variable, which is also of a size of 11491. We normalise the elevation data to have a mean of 0 and a standard deviation of 1.

Assessment Sites. We evaluate the performance of our algorithms across New Zealand as a whole but also focus on evaluating the performance at specific locations, shown in Figure 3. These locations span different environments and precipitation intensities, including high-elevation regions, which often experience high precipitation intensities (Mt Cook and Tongariro National Park) and a diverse range of towns (Greymouth and Napier) and major cities (Auckland and Wellington). We evaluate locations that collectively represent the entire spectrum of possible precipitation levels. We also evaluate high-elevation regions collectively, denoted as NZ^+, which denote locations with elevations greater than 1305m in the unprocessed elevation data.

Preliminaries

Bernoulli-Gamma Network Formulation. Let $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ be the dataset, where $x \in \mathcal{X}$ are the ERA5 meteorological variables and $y \in \mathcal{Y}$ the VCSN precipitation observations. To efficiently capture extreme events, each sample is set at a daily temporal resolution. Often y is sparse across its domain, with concentration being focused only on a small number of areas. A loss function like MSE focuses on the magnitude of errors and does not encourage learning of this property. The Bernoulli Gamma distribution exponentially increases with limited support, thereby encapsulating both sparsity and focal point concentration. To learn these properties, we use the log-likelihood of the Bernoulli-Gamma distribution as our loss function. Specifically, we construct our network to predict a Bernoulli-Gamma distribution for every grid point in New Zealand. At each location, we predict the inverse scale, shape and probability of precipitation, denoted as β, α and p as:

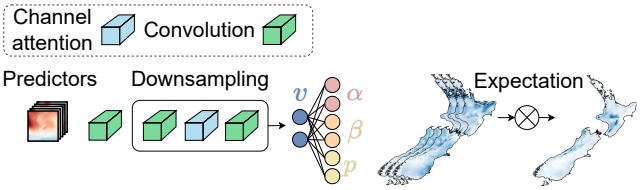

Figure 4: An illustration of the Bernoulli-Gamma network for downscaling precipitation from low-resolution meteorological fields. The network outputs the three parameters of the Bernoulli-Gamma distribution for every grid cell, the expectations of which form the final prediction.

$$h(y|\alpha,\beta,p) = \begin{cases} 1-p & y = 0 \\ \frac{p(\frac{y}{\beta})^{\alpha-1}\exp(-\frac{y}{\beta})}{\beta\Gamma(\alpha)} & y > 0. \end{cases} \quad (1)$$

The Bernoulli-Gamma loss is as follows:

$$\mathcal{L}_{BG} = -\log(h(y|\alpha(v),\beta(v),p(v))). \quad (2)$$

The network trained with the loss function in Equation 2 is known as the Bernoulli-Gamma Network (BGNet) (Baño-Medina, Manzanas, and Gutiérrez 2020). In this case, the distribution's parameters are calculated by transforming a vector v from the penultimate layer of the network using three independent dense layers, denoted as $\alpha(.)$, $\beta(.)$ and $p(.)$. The prediction for precipitation over each grid cell is given by the expected value of Equation 1. (Cannon 2008) as:

$$\mathbb{E}[h(y|\alpha,\beta,p)] = p \cdot \alpha \cdot \beta \quad (3)$$

To improve training stability, we learn using the log of the α and β parameters. An illustration of the BGNet is shown in Figure 4.

Downsampling Module. Climate processes are inherently global phenomena with strong multi-dimensional interactions between variables that cannot be constrained by isolating them to a single region of space. In previous work (Liu, Ganguly, and Dy 2020; Park et al. 2022; Miao et al. 2019; Rampal et al. 2022; Pan et al. 2019; Sun and Lan 2021), a combination of max-pooling and convolutional layers were used to downsample x. However, extracting local

features does not consider the global scale interactions between climate variables within the domain. We use a Channel Attention mechanism (Zhang et al. 2018) to incorporate this property. Specifically, we leverage a channel's average and maximum statistics as implemented in Woo et al. (2018).

Quantile Regression Ensemble

Ensemble-based deep learning methods can potentially reduce the degree of overfitting to any single intensity level of precipitation. However, no work in the existing literature has thus far proposed such a method for climate extremes. We developed Quantile-Regression-Ensemble, an algorithm to mitigate overfitting to the mean sample intensity. QRE trains multiple regression-based neural networks on segmentations of the dataset, shown in Figure 5. Consequently, each regressor is able to train on a dataset containing different levels of precipitation, allowing for a better representation of extreme precipitation events.

Data Segmentation. Precipitation varies exponentially with intensity. Some events, for instance, exhibit no precipitation at all, while others are several orders of magnitude more so than the mean intensity. To create a more balanced learning problem, we split $\mathcal{D}$ into N disjoint subsets based on $\mathbf{1} \cdot y$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$. The intention is for each set to only contain y with a similar cumulative precipitation level over New Zealand. We say that y is in quantile interval i if the cumulative precipitation across New Zealand is between in a certain intensity range corresponding to i. In particular, splits are defined by the quantile intervals $\{(q_i, q_{i+1})\}_1^N$, where $q_0 = 0.0$ and $q_N = 1.0$. For values l_i and h_i at quantiles q_i and q_{i+1} of cumulative precipitation $\{\mathbf{1} \cdot \tilde{y}\}_{\tilde{y} \in \mathcal{Y}}$, we assign a point $(x, y) \in \mathcal{D}$ to subset i if y satisfies $l_i < \mathbf{1} \cdot y < h_i$. If this is the case, we say that (x, y) corresponds to the interval $[q_i, q_{i+1})$, on which a member of the partition is denoted as:

$$\mathcal{D}^{(q_i:q_{i+1})} = \mathcal{X}^{(q_i:q_{i+1})} \times \mathcal{Y}^{(q_i:q_{i+1})}.$$

An example of the segmentation procedure for $N = 6$ is shown on the empirical distribution in Figure 5. Notably, where $i = N - 1, N$ correspond to the smallest segmentations since density changes rapidly around the distribution's tail ends. For our purposes, it is not sufficient to only split the data but to keep a record of the predictor and quantile index pairs. Under the same condition for assigning y to partition i, we discretise y to the interval index i to form the set $\mathcal{I}$, and write $\Omega = (\mathcal{X}, \mathcal{I})$. To select the intervals, we use a Bayesian Gaussian mixture model on $\{\mathbf{1} \cdot \tilde{y}\}_{\tilde{y} \in \mathcal{Y}}$, and use the average of 100 repetitions for our quantile intervals; after each run on A, we convert the N means μ_i to its quantile index q_i in A, and set $q_N = 1.0$.

Ensemble members. We minimise the regression to the mean effect by training a member f_{ξ_i} with parameters ξ_i on each $\mathcal{D}^{(q_i:q_{i+1})}$. By removing influences of samples from other areas of the empirical distribution during training, f_{ξ_i} can learn a representation of the precipitation level unique to $\mathcal{D}^{(q_i:q_{i+1})}$. In this manner, we endeavour for each f_{ξ_i} to obtain significantly lower MSE on $\mathcal{D}^{(q_i:q_{i+1})}$ than the same model trained over $\mathcal{D}$. All models are assembled in the $\{f_{\xi_i}\}_1^N$ collection. We are concerned with cases of extreme precipitation at $i = N$, and more generally across the interval $(0.9, 1.0)$. A model that accurately predicts extreme precipitation can inform stakeholders of potential economic damage or public safety risks.

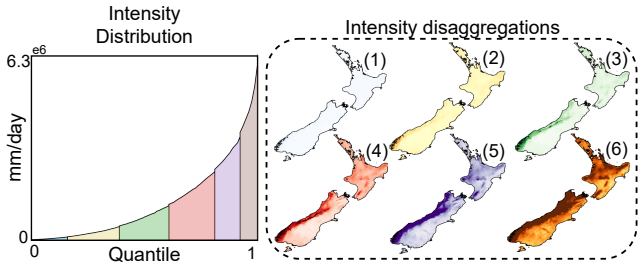

Figure 5: Example of data segmentation using 6 quantile intervals over the distribution of cumulative precipitation over New Zealand. The index of the quantile interval is denoted by (1) through to (6).

Weight Network. Here we consider the task of collecting all members into an ensemble. Other methods, such as Gradient-Boosting, carry out aggregation via assigning a fixed weight to each model such that the loss over the entire dataset is minimised (Moghimi et al. 2016). However, given the members have been trained over different datasets, there is no obvious way to compare losses or predictions between $\{f_{\xi_i}\}_1^N$. Therefore, a fixed weight strategy is likely to be inadequate; weighting models to reduce error over the entire set ignores the original intensities of each f_{ξ_i}, leading to redundant models. The ensemble's performance over the corresponding intensity intervals will likewise be substantially lower than the original member. Instead, we learn a network that dynamically assigns each ensemble member a weight. Specifically, we construct a task where (x, y) is softly assigned to a $[q_i, q_{i+1})$, essentially selecting a single f_{ξ_i} for x. Let f_θ be a network with parameters θ that, by extracting spatial characteristics from x, outputs a vector where entry i is proportional to the probability of (x, y) belonging to $[q_i, q_{i+1})$ of $\mathcal{D}$. Specifically, f_θ uses the same convolutional components as the BG-Net in Figure 4, subsequently followed by 2 dense layers. The final output is computed by a layer containing N neurons. To correctly scale predictions of f_{ξ_i} , we constrain the weights of f_θ such that $\sum f_\theta(x)_i = 1$. We therefore formulate our Weighting network ω_θ as a *Softmax* transformation of f_θ as follows:

$$\omega_\theta(x) = \text{SoftMax}(f_\theta(x)). \tag{4}$$

We train ω_θ using the Earth Mover Distance (EMD) for multi-class classification (Frogner et al. 2015) as our loss function. The EMD loss function has the ability to deal with the overlap between class distributions (Hou, Yu, and Samaras 2016; Graffeuille et al. 2022). Beyond that, EMD enforces a class ordering by incorporating the distance from the prediction to the true class. Specifically, by applying a weighting function Γ, a distance between ground truth class κ' and class κ is defined. Let Γ be the pth power of the ab-

solute distance $|\kappa - \kappa'|$ formulated as follows:

$$\Gamma(\kappa, \kappa') = |\kappa - \kappa'|^p. \tag{5}$$

From Equations 4 and 5, these loss $\mathcal{L}_\omega$ is written as:

$$\mathcal{L}_\omega = \sum_{\kappa=1}^{N} \Gamma(\kappa, \kappa')\omega_\theta(x)_\kappa. \tag{6}$$

Ensemble. We aggregate $\{f_{\xi_i}\}_1^N$ into a ensemble g by using ω_θ to assign an x to a f_{ξ_i}. The ensemble is formed via an affine summation:

$$g(x|\theta, \xi_1, \ldots, \xi_N) = \sum_{1 \le i \le N} \omega_\theta(x)_i f_{\xi_i}(x). \tag{7}$$

We note that if $\omega_\theta(x)_i = 1$, then $g(x|\theta, \xi_1, \ldots, \xi_N) = f_{\xi_i}(x)$. Suppose the same model is trained over $\mathcal{D}$. If we were to assume that f_{ξ_i} achieves lower error metrics on $\mathcal{D}^{(q_i:q_{i+1})}$, and i is the correct class, then it follows immediately that g is more accurate over that particular segmentation. Let f_{ξ_i} be BGNet as shown in Figure 4, trained using the log-likelihood of the Bernoulli-Gamma distribution, as shown in Equation 2. Here the number of bins is $N = 6$.

Pseudocode. Algorithm 1 presents the pseudocode for QRE. The function *Sorted* builds a sorted array $\mathcal{S}$ by using an arbitrary ordering between elements, which in this case is given by $\mathbf{1} \cdot y$. We use this to construct the collection $\{f_{\xi_i}\}_1^N$ between Lines 2 and 11. Specifically, we build $\mathcal{D}^{(q_i:q_{i+1})}$ between Lines 4 and 9 using points y_i and y_{i+1} closest to the quantiles q_i and q_{i+1}, which are calculated on Lines 5 and 6. Each f_{ξ_i} is initialised by calling *BGNet* on Line 3, and trained over $\mathcal{D}^{(q_i:q_{i+1})}$ by invoking the Adam optimisation scheme *Adam* on Line 10. Lines 14 to 15 are dedicated to initialising and training ω_θ. The final ensemble is assembled on Line 16. The algorithm returns the ensemble on Line 17.

Theoretical Discussion

Number of quantile intervals. The Bayesian Gaussian Mixture model used to compute the quantile intervals uses a maximum likelihood approach on the posterior distribution to converge to the N quantile intervals. Therefore, an increase in N corresponds to an increase in the overlap of the intensity distribution for adjacent quantile intervals. Classification accuracy of ω_θ subsequently decreases. However, lets we assume that each f_{ξ_i} converges to an accurate representation of the precipitation intensity of the underlying distribution of $\mathcal{D}^{(q_i:q_{i+1})}$. Then, as long as increasing N does not drastically reduce the accuracy of ω_θ over any i, the performance of QRE can be increased. However, redundant quantile intervals can occur if N grows too large; these either have low cardinality or are too similar to data from adjacent bins. Then a combination of the member overfitting and a substantial decrease in accuracy of ω_θ will lead to lower accuracy.

Complexity. Given that $\mathcal{D}$ is partitioned into a disjoint collection which covers it, each sample $(x, y) \in \mathcal{D}$ is trained over exactly once per epoch. We can therefore assume that the number of iterations for training f_{ξ_i} on $\mathcal{D}^{(q_i:q_{i+1})}$ is roughly $\frac{1}{N}$ times the cost of training the same model over $\mathcal{D}$. Suppose we assume that ω_θ has similar iteration requirements to that of f. Then, since both are trained over $\mathcal{D}$, we can assume that asymptotically the training time of QRE increases linearly with respect to the training time of f.

Algorithm 1: Quantile-Regression-Ensemble

Require: $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, $Q_{ranges} = \{(q_i, q_{i+1})\}_1^N$, $\mathcal{L}_{BG}$, $\mathcal{L}_\omega$
Ensure: Ensemble g of members and Weighting network with parameters $\theta, \xi_1, \ldots, \xi_N$

1: $\mathcal{S} \leftarrow Sorted(\mathcal{Y})$ ▷ Sort by total precipitation
2: **for** $q_i, q_{i+1} \in Q_{ranges}$ **do** ▷ Train members
3: $f_{\xi_i} \leftarrow BGNet()$ ▷ Initialise a BGNet instance
4: $\mathcal{D}^{(q_i:q_{i+1})} \leftarrow \{\}, \{\}$
5: $y_i \leftarrow S[\lfloor q_i \cdot N \rfloor]$
6: $y_{i+1} \leftarrow S[\lfloor q_{i+1} \cdot N \rfloor]$
7: $l_i \leftarrow \mathbf{1} \cdot y_i$ ▷ Upper and lower bounds
8: $h_i \leftarrow \mathbf{1} \cdot y_{i+1}$
9: $\mathcal{D}^{(q_i:q_{i+1})} \leftarrow \{(x, y) \in \mathcal{D} | l_i < \mathbf{1} \cdot y < h_i\}$
10: $f_{\xi_i} \leftarrow Adam(f_{\xi_i}, \mathcal{D}^{(q_i:q_{i+1})}, \mathcal{L}_{BG})$
11: **end for**
12: $\mathcal{I} \leftarrow \{i | l_i < \mathbf{1} \cdot y < h_i, 1 \le i \le N, y \in \mathcal{Y}\}$
13: $\Omega \leftarrow (\mathcal{X}, \mathcal{I})$
14: $f_\theta \leftarrow WeightNetwork()$ ▷ Classifier network
15: $\omega_\theta \leftarrow Adam(Softmax(f_\theta), \mathcal{I}, \mathcal{L}_\omega)$
16: $g(.|\theta, \xi_1, \ldots, \xi_N) \leftarrow \sum_{1 \le i \le N} \omega_\theta \cdot f_{\xi_i}$
17: **return** $g(.)$

Experiments

Evaluation. We evaluate techniques over data across various precipitation intensities, including the intervals QRE has been trained over. To test the generalisability of our methods, we evaluate the performance of QRE over distinct high and low precipitation intervals; $(0.9, 1.0)$ and $(0.0, 0.2)$. We also evaluated the performance across different regions of New Zealand, displayed in Table 2. When segmenting the data using both spatial and quantile coordinates, we consider the variability of precipitation levels for each particular region by performing spatial segmentation. To mitigate non-determinable effects (Quesada-Chacón, Barfus, and Bernhofer 2022), we repeated all experiments 40 times and reported the resulting mean and standard deviation. Significance is determined using the Wilcoxon signed-rank test, with the probability of rejecting the null hypothesis set to 5% (Demšar 2006). To improve reproducibility when the number of repetitions is low, each iteration of QRE uses members as the average overall runs.

The objective of our evaluation is to answer the following questions: **RQ 1** How does QRE perform under different precipitation intensity levels? **RQ 2** How does the number of members affect QRE? **RQ 3** How accurately can QRE predict precipitation?

Method	In-Domain Interval						Out-of-Domain Interval	
	0.0-0.166	0.166-0.392	0.392-0.61	0.61-0.812	0.812-0.923	0.923-1.0	0.9-1.0	0.0-0.2
Perfect-Member	0.9±0.0	6.1±0.2	22.6±0.5	61.5±2.0	123.3±5.0	312.2±8.4	-	-
QRE	1.9±0.3	9.5±0.7	29.4±0.8	67.7±1.6	**127.1±2.3**	**312.8±6.4**	**282.6±4.9**	2.4±0.3
Naive-Ensemble	1.8± 0.2	10.0± 0.6	53.9 ± 2.9	140.3± 4.4	241.0±4.7	626.9±5.9	548.2±5.0	2.2±0.2
BG-Net	2.2±0.3	10.2±0.9	30.9±1.9	69.9±3.9	133.4±6.0	326.3±9.6	297.4±8.5	2.8±0.4
BG-Net(-)	2.4±0.3	10.6±0.8	31.4±1.8	70.1±2.5	138.1±4.8	337.8±10.2	305.1±8.0	3.3±0.4
Bagging	45.5±7.5	50.3±5.5	39.0±2.2	**54.6±0.5**	140.8±2.4	643.8±10.2	545.9±8.5	45.4±7.3
Probability	27.7±3.4	29.1±2.4	**25.0±0.7**	58.0±0.8	175.2±2.5	643.8±9.7	652.8±8.1	27.5±3.3

Table 1: Comparison between BG-Net and QRE (ours) over the quantile intervals. The statistically significant MSE results according to the Wilcoxon-signed-rank test are shown in bold. Perfect-Member is the idealised case considered as the "ideal" results if perfect weightings are obtained.

Reproducibility. Figure 4 shows each convolutional layer in BGNet has a kernel size of 6. The channels in each convolutional layer are 64, 128 and 256, respectively. The coefficient r for channel attention is set to 1. The first dense layer contains 512 neurons, which during training uses dropout with a rate of 0.2 is used. The remaining branches for the distribution parameters contain 11491 neurons, for each grid point in New Zealand. The ReLu activation function is used in all layers except for the final output layer, where the α and β branches use modified tanh functions, and the p branch uses the sigmoid function. The convolutional layers of the Weight Network are the same as in BG-Net and are followed by 2 dense layers containing 100 neurons each. A dropout rate of 0.2 between both dense layers is used when training the network. The last output layer has a neuron equal to the number of members in the ensemble, where we use $N = 6$. When computing the EMD loss p is set to 1. All networks are trained using an exponentially decaying learning rate. We use a decay value of 0.7 and set the number of decay steps to 6000. When training BG-Net, the initial learning rate is 10^{-4}, whereas for the Weight Network, its value is $5 \cdot 10^{-5}$. To account for instability during training, we use early stopping with patience of 4 and 15 for BG-Net and the Weighting Network, respectively. We use the first 32 years of data for training, the next 6 for validation, and the following 8 for our test set. We train on a single NVIDIA A100 GPU with $256GB$ of RAM. The source code for our research is available at https://github.com/tomthedecoder/Quantile-Regression-Ensemble-A-Deep-Learning-Algorithm-for-Downscaling-Extreme-Precipitation.

Baseline models. We compare the effectiveness of our QRE against several baseline models: Perfect-Member, BGNet, BGNet(Interp), BGNet(Embed), Probability, and Bagging.

- *Perfect-Member.* In an ideal scenario, the Weight Network achieves perfect accuracy when assigning samples to members. We call this model Perfect-Member, which refers to the member evaluated over the same precipitation intensity it was originally trained over.
- *Variations of BGNet.* We evaluate the effectiveness of elevation data as an static predictor field. BGNet(Interp) utilises bilinear interpolation of elevation data, whereas BGNet(Embed) uses convolution and max pooling operations prior to interpolation. We include the network introduced by Rampal et al. (2022) as BG-Net(-).
- *Probability.* To highlight the importance of dynamic weight assignment from our Weight Network, we contrast against fixed weight methods. Specifically, we scale each member of QRE by a fixed scalar value. The probability technique fixes the weights of each member corresponding to the empirical probability that a sample belongs to a given quantile interval.
- *Bagging*. We use a weighting procedure that computes the mean of predictions from all members.
- *Naive-Ensemble*. We evaluate the effectiveness of members in capturing variability by fixing each member to the empirical mean over the associated particular segmentation.

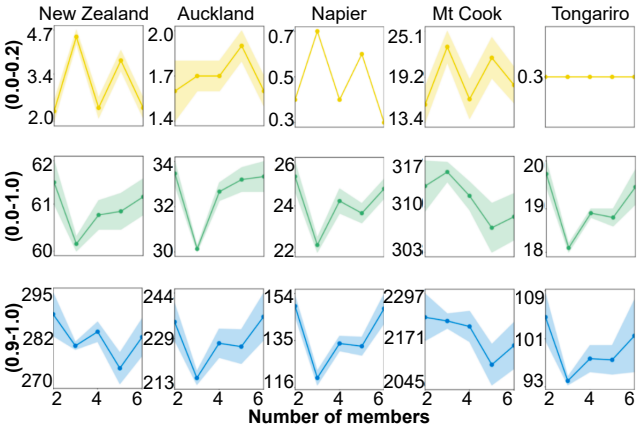

Figure 6: Sensitivity analysis of QRE to the number of ensemble members. MSE is given over various quantiles for intensity and spatial location.

Comparison of results based on varying intensities (RQ1). Table 1 compares the techniques on different quantile intervals. Firstly, we tested the data with the same intervals as the training intervals, which are denoted as the In-Domain interval. Secondly, We tested the model on different intervals to the training, denoted as Out-of-Domain

Method		High			Moderate			Low
	NZ	Mt Cook	NZ^+	Greymouth	Napier	Auckland	Wellington	Tongariro
QRE	**60.8±0.5**	**308.2±3.8**	**102.1±1.0**	**153.4±1.7**	24.5±0.4	33.2±0.5	27.7±0.3	**19.7±0.4**
BGNet	63.6±2.0	324.9±9.8	106.7±3.0	160.7±5.8	23.7±1.6	**31.6±0.9**	26.9±0.9	20.2±1.0
BGNet(-)	65.2±1.4	322.7±10.4	107.8±3.2	166.7±5.7	26.7±1.5	35.3±0.8	29.2±0.9	22.2±1.1
BG-Embed	63.3±1.5	322.0±9.0	105.8±2.9	160.2±5.7	23.7±1.2	31.8±0.9	27.0±0.9	20.2±1.0
BG-Interp	63.2±1.4	324.1±9.7	106.1±3.0	158.5±1.8	23.6±1.2	31.6±0.8	27.1±0.8	26.8±0.8

Table 2: Comparisons of MSE between QRE and baseline models across NZ and local areas, ranked via intensity. In bold are the significant areas according to the Wilcoxon-signed-rank test.

intervals. Note that we are not able to produce Perfect-Member results for Out-Domain as there exists an overlap between quantile intervals, and therefore the choice of the member is ambiguous. With the expectaion of Naive-Ensemble, we observe QRE significantly out-performs all baselines on both low and high precipitation intensity intervals, but not over those corresponding to moderate precipitation.

Sensitivity analysis based on the number of members (RQ2). We conducted a sensitivity analysis for QRE's dependence on its only hyper-parameter, the number of members or dataset partitions, N. As shown in Figure 6, there is variability in MSE across different regions and precipitation intensity levels when N is varied. We find that selecting N requires a critical balance of performance between the 3 intensity levels. In particular, our results consistently show a trade-off between performance across high and low intensity levels. Additionally, with regards to N between 2 and 4, we generally observe a positive correlation between performance over the intervals $(0.9, 1.0)$ and $(0.0, 1.0)$. However, the situation over New Zealand changes when $N \geq 4$; the MSE on $(0.0, 1.0)$ starts to increase irrespective of performance on high precipitation events. This is due to QRE constructing smaller intervals around moderate to high precipitation levels. The task of assigning weights, as a result, grows in difficulty. Furthermore, due to the ensemble aggregation method employed, training a high precipitation member may bias QRE towards that particular model, even though the model may be assigned a low-valued weight. Therefore, decreasing error on $(0.9, 1.0)$ may actually worsen performance on $(0, 0.2)$, while still not guaranteeing more accurate performance on $(0, 1.0)$. Take for instance the case of $N = 5$. This coincides with our claims made in the theoretical discussion. Taking these experiments into account, we choose $N = 6$ in our experiments as it performs reasonably on both high and low precipitation levels compared to other choices.

Comparison between QRE and baseline models over various regions (RQ3). Table 2 shows QRE results against other baselines across New Zealand and different precipitation regions, *i.e.* high (Mt Cook), moderate (Auckland, Napier) intensity levels and low (Tongariro) intensity levels. Table 2 shows QRE lowers the MSE of predictions over New Zealand by 4.8% when compared with BG-Net, and is, in this respect, significant to all other models. We can therefore infer that the aggregation of members via the Weight Network, in general, is more accurate. In the disaggregated tests, QRE performs significantly more accurately over the high and low precipitation regions of Mt Cook, NZ^+, Greymouth and Tangariro. However, QRE does not perform as accurately in the moderate-intensity regions. We observe a similar trend in Table 1. Therefore, we can deduce that QRE significantly improves performance over high and low precipitation extremes while remaining insignificant over moderate precipitation levels.

Conclusion and Future Work

High-intensity precipitation events are capable of causing large-scale socioeconomic damage. While GCMs can provide insights into these events, their resolutions are often too coarse to capture local-scale variations. Statistical downscaling offers an affordable solution to enhancing the resolution of GCMs. However, many deep learning based regression models suffer from regression to the mean precipitation intensity. To address this, here we developed Quantile-Regression-Ensemble, an algorithm that constructs an ensemble by training models over intensity-based subsets of the data, allowing for a better representation of the observed range of precipitation intensity. In experiments conducted across New Zealand and for differing precipitation intensity intervals, QRE significantly improves accuracy in prediction of high precipitation extremes, while simultaneously reducing error across all other intensity levels. QRE consequently shows potential for application in broader climate modelling, including downscaling projections of future climate scenarios.

Currently, Quantile-Regression-Ensemble successfully leverages two factors; each member achieves lower MSE than the aggregate model over its original intensity, and due to the quantile N being the most distinct class, the Weight Network achieves excellent accuracy on high precipitation extremes. However, for low precipitation levels, the Weight Network struggles to classify samples correctly, leading to a diminished performance gain across moderate precipitation regions. Future research will explore methods to mitigate this effect.

Acknowledgements

This work is supported by the New Zealand MBIE Endeavour Smart Ideas Fund (C01X2202). The authors declare no conflicts of interest.

References

Adewoyin, R. A.; Dueben, P.; Watson, P.; He, Y.; and Dutta, R. 2021. TRU-NET: a deep learning approach to high resolution prediction of rainfall. *Machine Learning*, 110: 2035–2062.

Baño-Medina, J.; Manzanas, R.; Cimadevilla, E.; Fernández, J.; González-Abad, J.; Cofiño, A. S.; and Gutiérrez, J. M. 2022. Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. *Geoscientific Model Development*, 15(17): 6747–6758.

Baño-Medina, J.; Manzanas, R.; and Gutiérrez, J. M. 2020. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4): 2109–2124.

Cannon, A. J. 2008. Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network. *Journal of Hydrometeorology*, 9(6): 1284–1300.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7: 1–30.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Eyring, V.; Bony, S.; Meehl, G. A.; Senior, C. A.; Stevens, B.; Stouffer, R. J.; and Taylor, K. E. 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958.

Frame, D. J.; Rosier, S. M.; Noy, I.; Harrington, L. J.; Carey-Smith, T.; Sparrow, S. N.; Stone, D. A.; and Dean, S. M. 2020. Climate change attribution and the economic costs of extreme weather events: a study on damages from extreme rainfall and drought. *Climatic Change*, 162: 781–797.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems*, 28.

Gibson, P. B.; Waliser, D. E.; Lee, H.; Tian, B.; and Massoud, E. 2019. Climate model evaluation in the presence of observational uncertainty: Precipitation indices over the contiguous United States. *Journal of Hydrometeorology*, 20(7): 1339–1357.

Graffeuille, O.; Koh, Y. S.; Wicker, J.; and Lehmann, M. K. 2022. Semi-supervised Conditional Density Estimation with Wasserstein Laplacian Regularisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6746–6754.

Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049.

Hou, L.; Yu, C.-P.; and Samaras, D. 2016. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

Liu, J.; Chen, C.; Tang, J.; and Wu, G. 2023. From coarse to fine: Hierarchical pixel integration for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1666–1674.

Liu, Y.; Ganguly, A. R.; and Dy, J. 2020. Climate downscaling using YNet: A deep convolutional network with skip connections and fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3145–3153.

Mayr, A.; Binder, H.; Gefeller, O.; and Schmid, M. 2014. The evolution of boosting algorithms. *Methods of information in medicine*, 53(06): 419–427.

Merz, B.; Kreibich, H.; Schwarze, R.; and Thieken, A. 2010. Review article" Assessment of economic flood damage". *Natural Hazards and Earth System Sciences*, 10(8): 1697–1724.

Miao, Q.; Pan, B.; Wang, H.; Hsu, K.; and Sorooshian, S. 2019. Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. *Water*, 11(5): 977.

Moghimi, M.; Belongie, S. J.; Saberian, M. J.; Yang, J.; Vasconcelos, N.; and Li, L.-J. 2016. Boosted convolutional neural networks. In *BMVC*, volume 5, 6.

Oyama, N.; Ishizaki, N. N.; Koide, S.; and Yoshida, H. 2023. Deep generative model super-resolves spatially correlated multiregional climate data. *Scientific Reports*, 13(1): 5992.

Pan, B.; Hsu, K.; AghaKouchak, A.; and Sorooshian, S. 2019. Improving precipitation estimation using convolutional neural network. *Water Resources Research*, 55(3): 2301–2321.

Park, S.; Singh, K.; Nellikkattil, A.; Zeller, E.; Mai, T. D.; and Cha, M. 2022. Downscaling earth system models with deep learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3733–3742.

Quesada-Chacón, D.; Barfus, K.; and Bernhofer, C. 2022. Repeatable high-resolution statistical downscaling through deep learning. *Geoscientific Model Development*, 15(19): 7353–7370.

Rampal, N.; Gibson, P. B.; Sood, A.; Stuart, S.; Fauchereau, N. C.; Brandolino, C.; Noll, B.; and Meyers, T. 2022. High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. *Weather and Climate Extremes*, 38: 100525.

Sun, L.; and Lan, Y. 2021. Statistical downscaling of daily temperature and precipitation over China using deep learning neural models: Localization and comparison with other methods. *International Journal of Climatology*, 41(2): 1128–1147.

Tait, A.; Henderson, R.; Turner, R.; and Zheng, X. 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 26(14): 2097–2115.

Tait, A.; Sturman, J.; and Clark, M. 2012. An assessment of the accuracy of interpolated daily rainfall for New Zealand. *Journal of Hydrology (New Zealand)*, 25–44.

Vandal, T.; Kodra, E.; Ganguly, S.; Michaelis, A.; Nemani, R.; and Ganguly, A. R. 2018. Generating High Resolution Climate Change Projections through Single Image Super-Resolution: An Abridged Version. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5389–5393. International Joint Conferences on Artificial Intelligence Organization.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xia, B.; Hang, Y.; Tian, Y.; Yang, W.; Liao, Q.; and Zhou, J. 2022. Efficient non-local contrastive attention for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2759–2767.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.