

Statistical Downscaling of Global Climate Models with Image Super-resolution and
Uncertainty Quantification

A Dissertation Presented

By

Thomas James Vandal

to

The Department of Civil and Environmental Engineering

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the field of

Interdisciplinary Engineering

Northeastern University
Boston, Massachusetts

August 2018

ABSTRACT

High-resolution probabilistic projections of precipitation and temperature under climate change are crucial for stakeholders to make well-informed decisions in mitigating and adapting to more intense, longer duration, and more frequent extreme weather events. General circulation models (GCMs) provide us with the data to study climate change at the continental spatial scales, but are too coarse for local adaption. Furthermore, ensembles of multiple models, initial conditions, and emission trajectories must be harnessed for well quantified probabilistic estimates. Statistical downscaling, an approach that learns a functional mapping between low- and high-resolution GCMs, can be used to generate high-resolution ensemble projections in a computationally efficient manner. However, this process exacerbates, at a local scale, uncertainties inherently found in GCMs. Hence, it is crucial for our statistical downscaling methods to incorporate and quantify uncertainties, including both epistemic, or parameter misunderstanding, and aleatoric, or observational, uncertainties. In this work, we present a Bayesian deep learning and image super-resolution approach for statistical downscaling using discrete-continuous and non-normal likelihoods. Promising results for downscaling daily precipitation in the contiguous United States measured on predictive accuracy and uncertainty quantification are presented. Future work on stacking Bayesian deep learning networks and harnessing ensembles of high-resolution GCMs is discussed.

ACKNOWLEDGMENTS

It is my pleasure to thank the many people who have guided and helped me through the PhD process. Firstly, I'd like to thank my wife Amanda Mercer for being my rock as well as my parents Ann and James Vandal and sister Jessica Vandal for their continued support.

I'd also like to thank my advisors Prof. Auroop R. Ganguly and Dr. Evan Kodra for their mentoring, guidance, and putting up with my nonsense. I thank my committee members Prof. Jennifer Dy, Prof. Edward Beighley, Prof. Tarik Gouhier, and Dr. Sangram Ganguly for their advising.

I thank my co-authors and collaborators for all of our discussions that directed this work (alphabetically): Dr. Arindam Banerjee, Udit Bhatia, Prof. Jennifer Dy, Dr. Forrest Hoffman, Prof. Auroop Ganguly, Dr. Sangram Ganguly, Dr. Lucas Joppa, Dr. Evan Kodra, Prof. Vipin Kumar, Dr. Daniel McDuff, Andrew Michaelis, Dr. Ramakrishna Nemani, Prahbat, Dr. Steve Sain, and Samuel Zimmerman.

I'd like to thank my colleagues in SDS lab for their support (alphabetically): Udit Bhatia, Kevin Clark, Debasish Das, Kate Duffy, Babak Fard, Dr. Devashish Kumar, Matthew Mage, Bharat Sharma, Konduri Shashank, Dr. Daiwei Wang, and Mary Warner.

Lastly, I thank all the sponsors of this work including the Bay Area Environmental Research Institute, NASA Earth Exchange (NEX), and the National Science Foundation CyberSEES (1442728), CRISP (1735505), and BIGDATA (1447587) grants.

TABLE OF CONTENTS

1	Introduction	1
	1.1 Statistical Downscaling	2
	1.2 Machine Learning in Statistical Downscaling	4
	1.2.1 Multi-task Learning for Statistical Downscaling	4
	1.2.2 Deep Learning for Statistical Downscaling	5
2	Intercomparison of Machine Learning Methods for Statistical Downscaling: The Case of Daily and Extreme Precipitation	7
	2.1 Statistical Downscaling Methods	7
	2.1.1 PP Approaches	8
	2.1.2 MOS: Bias Corrected Spatial Disaggregation	13
	2.1.3 Hybrid: Bias Corrected Spatial Disaggregation with MSSL (BCSD-MSSL)	14
	2.2 Data	14
	2.2.1 United States Unified Gauge-Based Analysis of Precipitation	15
	2.2.2 NASA Modern-Era Retrospective Analysis for Research and Applications 2 (MERRA-2)	15
	2.3 Experiments and Evaluation	16
	2.4 Results	18
	2.4.1 Daily Anomalies	18
	2.4.2 Large Climate Trends	20
	2.4.3 Extreme Events	21
	2.5 Discussion	23
3	DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution	25
	3.1 Introduction	25
	3.1.1 Key Contributions	26
	3.1.2 Organization of the Paper	27
	3.2 Earth Science Data	27
	3.3 Statistical Downscaling	29
	3.4 Related Work	29
	3.5 Methodology	31
	3.5.1 Super-resolution CNN	32
	3.5.2 Stacked SRCNN	33
	3.5.3 DeepSD	34
	3.6 Application of DeepSD	35
	3.6.1 Training DeepSD	35
	3.6.2 Comparison	36

3.6.3	Scalability on NASA's NEX	40
3.7	Discussion	41
4	Quantifying Uncertainty in Discrete-Continuous and Skewed Data with Bayesian Deep Learning	43
4.0.1	Key Contributions	45
4.1	Precipitation Estimation	45
4.1.1	Statistical Downscaling	45
4.1.2	Climate Data	47
4.2	Background	48
4.2.1	DeepSD	48
4.2.2	Bayesian Deep Learning	49
4.3	Bayesian Deep Learning for Skewed Distributions	51
4.3.1	Gaussian Likelihood	51
4.3.2	Discrete-Continuous Gaussian Likelihood	52
4.3.3	Discrete-Continuous Lognormal Likelihood	54
4.4	Precipitation Downscaling	55
4.4.1	Predictive Ability	56
4.4.2	Uncertainty Quantification	58
4.5	Discussion	60
5	Conclusions and Future Work	62

LIST OF FIGURES

2.1 Left) The deep autoencoder architecture used to pretrain neural network. Right) The pretrained encoder is fine-tuned for a supervised task.	12
2.2 Each map presents the spatial bias, or directional error, of the model. White represents no bias produced by the model while red and blue respectively show positive and negative biases.	19
2.3 Daily Root Mean Square Error (RMSE) is computed for each downscaling location and method. Each boxplot presents the distribution of all RMSEs for the respective method. The box shows the quartiles while the whiskers show the remaining distribution, with outliers displayed by points.	20
2.4 (Best viewed in color) The simple daily intensity index (Annual Precipitation/Number of Precipitation Days) averaged per year.	23
3.1 Prism Observed Precipitation: A) Low resolution at 1.0° ($\sim 100\text{km}$). B) High resolution at $1/8^\circ$ ($\sim 12.5\text{km}$).	28
3.2 Augmented SRCNN Architecture. From the left to right: Precipitation and Elevation sub-image pair, filters learned in layer 1, layer 1 activations, layer 2 filters, layer 2 activations, layer 3 filters, and HR precipitation label.	32
3.3 Layer by layer resolution enhancement from DeepSD using stacked SRCNNs. Top Row: Elevation, Bottom Row: Precipitation. Columns: 1.0° , $1/2^\circ$, $1/4^\circ$ and $1/8^\circ$ spatial resolutions.	34
3.4 Daily Root Mean Square Error (RMSE) computed at each location for years 2006 to 2014 (test set) in CONUS for Left) DeepSD and Right) BCSD. Red corresponds to high RMSE while blue corresponds to low RMSE.	37

3.5 Comparison of DeepSD, SRCNN, and BCSD for increasingly extreme precipitation. At each location in CONUS, all precipitation events above a percentile threshold (x-axis) are selected. Percentile thresholds between 90 and 99.9 are used. A) RMSE, B) Correlation, C) Bias, and D) Skill are computed at each location and averages over CONUS. The confidence bounds of each metric are taken from the 25th and 75th quantiles.	39
4.1 Histogram of daily precipitation on the Contiguous United States from 2006 to 2015. A) All precipitation data points. B) Precipitation distribution on rainy days only. C) Log distribution of precipitation on rainy days.	44
4.2 Prism Observed Precipitation: Left) Low resolution at 64km. Right) High resolution at 16km.	46
4.3 Residual SRCNN Architecture used for DeepSD with a skip connection between precipitation and the output layer.	49
4.4 Dropout probabilities learned using Concrete Dropout for both hidden layers. . . .	56
4.5 Daily Root Mean Square Error (RMSE) computed at each location for years 2006 to 2015 (test set) in CONUS. A) Gaussian, B) Conditional-Gaussian, and C) Conditional-Lognormal. Red corresponds to high RMSE while blue corresponds to low RMSE.	56
4.6 Precision recall curve of classifying rainy days in conditional models.	58
4.7 Calibration is computed as the frequency of predictions within a given probability range. This probability is varied on the x-axis with the corresponding frequency on the y-axis. Columns represent each model Gaussian, DC-Gaussian and Lognormal. Calibration plots on the first row compute per pixel with the shaded area representing the 80% confidence interval of calibration. The second row depicts calibration root mean square error (RMSE) per location.	58
4.8 Uncertainty widths based on quantiles from their predictive distributions. The points are observations versus the expected value. The bands correspond to 50%, 80%, and 90% predictive intervals.	59

LIST OF TABLES

2.1 Daily Projection Results (hold-out dataset with years 2005-2014). Daily statistical metrics averaged over space for annual, winter, and summer projections. Bias measures the directional error from each model. Correlation (larger is better) and RMSE (lower is better) describe the models ability to capture daily fluctuations in precipitation. The skill score statistic measure the model's ability to estimate the observed probability distribution.	18
2.2 Large scale projection results based on total monthly and yearly rainfall amounts. Bin width for monthly skill is set to 10mm and 100mm on the annual scale.	21
2.3 Statistics for ClimDEX Indices	22
3.1 Comparison of predictive ability between all six methods for 1000 randomly selected locations in CONUS. Runtime is computed as the amount of time to downscale 1 year of CONUS.	36
3.2 Comparison of Predictive Ability between DeepSD and BCSD for each season, Winter, Summer, Spring, and Fall. Values are computed at each location in CONUS and averaged.	38
4.1 Predictive accuracy statistics computed pixel-wise and aggregated. Daily intensity index (SDII) and yearly precipitation events greater than 20mm (R20) measure each model's ability to capture precipitation extremes. R20-Err and SDII-Err measures the difference between observed indicies and predicted indicies (closer to 0 is better).	57

Chapter 1

Introduction

The sustainability of infrastructure, ecosystems, and public health depends on a predictable and stable climate. Key infrastructure allowing society to function, including power plants and transportation systems, are built to sustain specific levels of climate extremes and perform optimally in its expected climate. Studies have shown that the changing climate has had, and will continue to have, significant impacts on critical infrastructure [35, 87]. Furthermore, climate change is having dramatic negative effects to ecosystems, from aquatic species to forests ecosystems, caused by increases in greenhouse gases and temperatures [118, 88, 44]. Increases in frequency and duration of heat waves, droughts, and flooding is damaging public health [42, 31].

Global Circulation Models (GCMs) are used to understand the effects of the changing climate by simulating known physical processes up to two hundred years into the future. The computational resources required to simulate the global climate on a large scale is enormous, limiting models to coarse spatial and temporal scale projections. Most often, the critical systems society depends on exist at the regional and local scale, where projections are most limited. Downscaling techniques are applied to provide climate projections at finer spatial scales, exploiting GCMs to build higher resolution outputs.

Statistical and dynamical are the two classes of techniques used for downscaling. The first approach, dynamical downscaling are based on Regional Climate Models (RCMs) which embeds sub-grid parameters and processes within boundary conditions of coarse resolution GCMs. This is particularly useful for better simulating convective and extreme precipitation events [62], which are often underestimated in statistical downscaling [12]. However, RCMs are typically sensitive to

their boundary conditions which can dramatically change the results [27]. For this reason, in order to obtain credible projections, multi-initial condition ensembles should be considered. However, a limitation of RCMs is it's high computational requirements that limits the use of multi-model and multi-initial condition ensembles.

The second approach, statistical downscaling (SD) aims to learn a statistical relationship between coarse scale climate variables and high resolution observations, which we will discuss in further depth in the following sections. SD is commonly applied to both RCMs, most often for bias correction, and GCMs. A key advantage of SD is it's computational efficiency and ability scale to climate model ensembles. However, GCMs have difficulty modeling convective precipitation, hence the application of SD to GCMs is expected to not well capture extremes [68]. In contrast, Ahmed et al. compared six statistically downscaled GCMs and four RCMs to a regional impact assessment of temperature and precipitation over the northeast United States and found no considerable differences in results between SD and using RCMs [3]. This suggests that SD directly from GCMs is sufficient over our region of interest.

1.1 Statistical Downscaling

Many statistical approaches can be used to relate large-scale climate processes to local-scale projections which can be grouped into three main categories: Perfect Prognosis (PP), Model Output Statistics (MOS), and Weather Generators (WG) [95, 81]. Regression and analog methods are categorized as PP approaches, which rely on large-scale observational data, often reanalysis datasets, as well as local-scale observations. On the other hand, MOS approaches rely only on the local-scale observations as well as the climate model outputs, either GCMs or RCMs. Quantile mapping and bias correction techniques are commonly applied MOS approaches in hydrological impact studies [125]. PP and MOS approaches can also be used together to improve predictability and statistical coherence. Lastly, WG are used to stochastically generate time series and spatial fields that resemble those observed. For instance, WG are often applied to better understand future extreme weather events [100].

As discussed in Maraun et al., PP approaches comprise of four key steps including prediction selection and transformation followed by a statistical model and model selection [81]. Selecting the

well-informed large-scale climate predictors that well represent the local-scale variable of interest is crucial to credible downscaling. The climate predictors of interest can be selected using knowledge of physical processes as well as locating those highly correlated with the predictand. This means that the selected predictors should be well simulated by GCMs to ensure statistically credible future projections. Hence, a statistical assumption that our predictors are related to the predictand is made.

Predictors selected for statistical downscaling are often high dimensional because the features extracted for each climate variable occur on a grid. With many variables and vertical pressure levels, as we'll discuss below, the total number of features can be in the thousands. Many approaches have been developed for high-dimensional feature selection and regression. Transformations from a high- to low-dimensional feature space are often applied such as principle component analysis [47, 107, 36]. Benestad et al. found that using principle component analysis improved performance while reducing sensitivity to the predictor domain when downscaling station data [9]. Rather than compressing all the features, some of which may not have relevant dependencies with the predictor, one can select a subset of these features using dependency measures. A few methods for features selection include correlation analysis [122], sparse regression [110, 43], and Bayesian models [24]. A combination of feature selection and transformation can also be applied. In this work, we compare multiple approaches to reduce dimensionality through both sparse regression and principle component analysis.

Given the set of features, we can define a statistical relationship to estimate the local-scale climate variable by learning from an observational dataset, either station based or gridded. Regression methods are most commonly used ranging in complexity, interpretability, and scalability. Linear and Generalized Linear models are applied most often and provide the basis of automated statistical downscaling [46]. Non-linear approaches, such as support vector machines [36] and artificial neural networks [105, 21], are often applied to capture more variability in the data but often overfit. To reduce overfitting, in both linear and non-linear approaches, regularization techniques can simplify the model by utilizing regularization terms, such as l_1 or l_2 penalties. Method of analogs is also used for SD, which is a nearest neighbor approach that find the k most similar observed examples given a new low-resolution projection [49, 90]. However, in this work we focus on regression approaches as they are more closely related to corresponding machine learning techniques.

Alternatively, MOS techniques use statistical attributes of observed climate variables but do not

necessarily rely on lower resolution observations. In addition, the climate variable being downscaling should be well simulated by GCMs to ensure credibility. In particular, precipitation is not yet well simulated and should be used cautiously in practice [97]. Bias Correction Spatial Dissagregation (BCSD), developed specifically for downscaling precipitation and temperature from GCMs, is a well studied MOS method that uses quantile mapping on the GCM followed by spatial scaling [125]. Though widely applied, bias correction approaches have been criticized for generalizing into the future and magnifying model errors that can cause impausible projections [80].

The selection of SD method is largely dependent on the use case and data availability. The needs of the user depending on the application, which can vary between climate variables, seasons, extreme events, temporal variability and scales, spatial coherence, physical consistency, and resilience to non-stationarity [81]. Characteristics and intercomparison studies are relied on for method selection. For instance, Burger et al. presented an intercomparison on five state-of-the-art methods for downscaling temperature and precipitation at a daily temporal resolution to quantify extreme events [12]. Another study by Gutmann et al. presented an intercomparison of methods on daily and monthly aggregated precipitation [41]. More recently, Maraun et al. developed the VALUE framework to objectively compare SD approaches using marginal, temporal, spatial, and multi-variate statistical indices [82]. One study by Gutiérrez et al., leveraging the VALUE framework, compare PP, MOS, and WG downscaling methods using all four categories of statistical indices and found downscaling methods generally reduce model bias but no method is superior. Similarly, Hetrig et al. used the VALUE framework to extreme events in europe and found large variations in skill between PP approaches. This study using marginal statistical attributed found in the VALUE framework to compare machine learning approaches and extend on these previous studies.

1.2 Machine Learning in Statistical Downscaling

1.2.1 Multi-task Learning for Statistical Downscaling

Traditionally, SD has focused on downscaling a locations independently without accounting for clear spatial dependencies in the system. Fortunately, numerous machine learning advances may aid SD in exploiting such dependencies. Many of these advancements focus on an approach known as multi-task learning, aiming to learn multiple tasks simultaneously rather than in isolation. A

wide variety of studies have shown that exploiting related tasks through multi-task learning (MTL) greatly outperforms single-task models, from computer vision [131] to biology [66]. Consider the work presented by [28] in which increasing the number of tasks leads to more significant feature selection and lower test error through the inclusion of task relatedness and regularization terms in the objective function. MTL has also displayed the ability to uncover and exploit structure between task relationships [133, 18, 4].

Recently Goncalves et al. presented a novel method, Multi-task Sparse Structure Learning (MSSL) [39] and applied it to GCM ensembles in South America. MSSL aims to exploit sparsity in both the set of covariates as well as the structure between tasks, such as set of similar predictands, through alternating optimization of weight and precision (inverse covariance) matrices. Goncalves et al. results showed significant improvements in test error over Linear Regression and Multi-model Regression with Spatial Smoothing when applied to climate model ensembles over South America. Along with a lower error, MSSL captured spatial structure including long range teleconnections between some coastal cities. The ability to harness this spatial structure and task relatedness within a GCM ensembles drives our attention toward MTL in other climate applications.

Consider, in SD, each location in a region as a task with an identical set of possible covariates. These tasks are related through strong unknown spatial dependencies which can be harnessed for SD projections. In the common high dimensional cases of SD, sparse features learned will provide greater significance as presented by [28]. Furthermore, the structure between locations will be learned and may aid projections. MSSL, presented by [39], accounts for sparse feature selection and structure between tasks.

1.2.2 Deep Learning for Statistical Downscaling

In the past years, interest in data science and machine learning has drastically increased corresponding to higher performing, more advanced, methods. The vast majority of the state of the art work has been developed by researchers in the domains of computer vision, speech recognition, and natural language processing. The computer vision applications are vast, ranging from medical imaging to self driving cars. Methods developed in this area focus on spatial dependencies within images, temporal dependencies in consecutive images (eg. videos), dimensionality reduction, and others. Developments in speech recognition and natural language processing often build off each other, due

to their temporal structure. The high dimensionality in all of these datasets continue to provide challenges, leading to solutions which may aid in climate problems.

In particular, Convolutional Neural Networks (CNNs) are a specific deep learning architecture developed to extract spatial information from images by using kernel filters [71]. CNNs have been successfully applied to a variety of problems including image recognition [70], image super-resolution [26], and image segmentation [94]. Earth science datasets, such as climate model outputs and satellite based observations, are often interpreted as a visual representation just like images. Hence, CNNs are a natural fit to the spatio-temporal data found in earth science datasets.

The remainder of this dissertation consists of four major parts. First, we compare a set of machine learning approaches to traditional statistical downscaling approaches. Second, a new approach to statistical downscaling using image super-resolution which we call DeepSD is presented. Third, DeepSD is extended to incorporate uncertainty quantification using Bayesian Deep Learning. Lastly, we conclude and discuss future research directions.

Chapter 2

Intercomparison of Machine Learning Methods for Statistical Downscaling: The Case of Daily and Extreme Precipitation

In this study we aim to compare traditional statistical downscaling approaches, BCSD, Multiple Linear Regression, Lasso, and Support Vector Machines, against new approaches in machine learning, Multi-task Sparse Structure Learning and Autoencoder Artificial Neural Network (AEs). During experimentation we apply common training architectures as part of the automated statistical downscaling framework. Results are then analyzed with a variety of metrics including, Root Mean Square Error (RMSE), bias, skill of estimating underlying distributions, correlation, and extreme indices.

2.1 Statistical Downscaling Methods

In this section, we describe the application of five PP methods, one MOS, and a combination MOS+PP. In particular, the methods focus downscaling of daily precipitation. Regression constants are ignored to simplify notation.

2.1.1 PP Approaches

As discussed above, machine learning methods are often well suited for PP downscaling for both regression tasks and feature selection. In particular, to downscale daily precipitation, it is advantageous to use a model that first estimate the existence of precipitation and apply a regression if needed, as applied in the Automated Statistical Downscaling (ASD) framework[[46]]. Such a model can be defined as:

$$\hat{y}(X) = \begin{cases} 0, & \text{if } g(X) < 0.5 \\ f(X), & \text{otherwise} \end{cases} \quad (2.1)$$

where $g_k(X) = p(y_k > 0 | X)$ such that predictors $X \in \mathcal{R}^{N \times d}$ and precipitation labels $y_k \in \mathcal{R}^N$ at location k of K with N samples and d covariates. The following five subsections presents methods based on this framework with varying classification and regression models.

Ordinary Least Squares with PCA (PCAOLS)

Linear models are widely used in statistical downscaling due to their simplicity and interpretability [[122, 46]]. However, linear models will still overfit high-dimensional regression tasks where $d \gg N$. Principle Component Analysis, which applies an orthogonal transformation of the feature space to "components", is used to reduce the dimensionality of the predictors [[123]]. More specifically, we select the minimum number of components that capture 98% ($c_{0.98}$) of explained variability in the feature space. We denote these principle components as $\bar{X}_{0.98}$.

A logistic regression is used for classifying the occurrence of precipitation is defined and optimized as:

$$g(X) = \frac{1}{1 + e^{-X\alpha}} \quad (2.2)$$

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left(\sum_{k=1}^N \log(1 + e^{-y_k g(X)}) \right) \quad (2.3)$$

where α represents the parameter coefficients at grid point k . Similarly, a linear regression for precipitous days is written and optimized as:

$$f(X) = X\beta \quad (2.4)$$

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|y - f(X)\|_2^2) \quad (2.5)$$

with parameters β . Plugging these into Equation 2.1 with input $\bar{X}_{0.98}$ gives us our first PP model.

Elastic-Net (ELNET)

Rather than transforming the entire feature space, as done with PCA, we can benefit by selecting a subset of features that explain variability in y . Sparsity constraints can be placed on the feature space to reduce or eliminate the influence of covariates which do not affect the output. Lasso is a widely used method that uses an L_1 norm sparsity constraint to enforce the coefficients of non-influential covariates to zero [[110]]. Similarly, Ridge regression uses an L_2 norm to produce a similar effect [[54]]. Elastic-Net, which we will apply in this study, uses a linear combination of L_1 and L_2 norms[[135]]. The parameters of the ELNET are learned using the following optimization objective:

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|y_k - f(X; \beta)\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2) \quad (2.6)$$

The L_1 norm forces uninformative covariate coefficients to zero while the L_2 norm enforces smoothness while allowing correlated covariates to persist. Cross-validation is applied with a grid-search to find the optimal parameter values for λ_1 and λ_2 . High-dimensional Elastic-Net is much less computational than stepwise regression techniques and most often leads to more generalizable models. A similar approach is applied to the classification step by using a logistic regression with an L_1 normalization term:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^N \log(1 + e^{-y_i g(X_i)}) + \lambda_1 \|\beta\|_1 \right). \quad (2.7)$$

Previous studies have considered the use of Lasso for SD [[43]] but to our knowledge, none have considered Elastic-Net.

Multi-task Sparse Structure Learning (MSSL)

Recent work in Multi-task Learning aims to exploit structure in the set of predictands while keeping a sparse feature set, much like Lasso and Elastic-Net. Multi-task Sparse Structure Learning (MSSL)

in particular learns the structure between predictands while enforcing sparse feature selection [[39]]. Goncalves et al. presented MSSL's exceptional ability to predict temperature through ensembles of GCMs while learning interesting teleconnections between locations [[39]]. Moreover, the generalized framework of MSSL allows for implementation of classification and regression models.

In the case of downscaling, each of the K locations in space are tasks in our machine learning model which are learned simultaneously. This is contrary to the other PP approaches that train models per location independently. We use the notation defined above and denote our predictand $Y \in \mathbb{R}^{n \times K}$ and regression coefficients $\beta \in \mathbb{R}^{d \times K}$. Furthermore, we define an inverse precision matrix, $\Omega \in \mathbb{R}^{K \times K}$, to quantify similarity between tasks. As proposed in [[39]], we write minimize the following objective:

As proposed in [[39]], optimization over the precision matrix, Ω , is defined as

$$\min_{\beta, \Omega > 0} \left\{ \frac{1}{2} \sum_{k=1}^K \| X\beta_k - Y_k \|_2^2 - \frac{K}{2} \log|\Omega| + Tr(\beta\Omega\beta^T) + \lambda \| \Omega \|_1 + \gamma \| \beta \|_1 \right\}. \quad (2.8)$$

The L_1 regularization parameters λ and γ enforce sparsity over Ω and β . The trace operation affects the rows of β and enforces similarity. Alternating minimization is applied to (2.8)

1. Initialize $\Omega^0 = I_k, \beta^0 = 0_{d \times k}$

2. for t=1,2,3,... do

$$\beta^{t+1} | \Omega^t = \min_{\beta} \left\{ \frac{1}{2} \sum_{k=1}^K \| X_k \beta_k - Y_k \|_2^2 + Tr(\beta\Omega\beta^T) + \gamma \| \beta \|_1 \right\} \quad (2.9)$$

$$\Omega^{t+1} | \beta^{t+1} = \min_{\Omega} \left\{ Tr(\beta\Omega\beta^T) - \frac{K}{2} \log|\Omega| + \lambda \| \Omega \|_1 \right\} \quad (2.10)$$

2.9 and 2.10 are independently approximated through Alternating Direction Method of Multipliers (ADMM). Furthermore, by assuming the predictors of each task is identical (as it is for SD), 2.9 is updated using Distributed-ADMM across the feature space [[10]].

A nearly identical model using a logistic regression can be optimized as:

$$\min_{\alpha, \Omega \succ 0} \left\{ \frac{1}{2} \sum_{k=1}^K \left(\sum_{i=1}^N \log(1 + e^{-y_i g(X_i)}) \right) - \frac{K}{2} \log|\Omega| + \text{Tr}(\alpha \Omega \alpha^T) + \lambda \|\Omega\|_1 + \gamma \|\alpha\|_1 \right\} \quad (2.11)$$

where α denotes the model parameters. As we can see, MSSL enforces similarity between rows of β by learning the structure Ω . For example, two locations which are nearby in space may tend to exhibit similar properties. MSSL will exploit these properties and impose similarity in their corresponding linear weights. By enforcing similarity in linear weights, we are encouraging smoothness of SD projections between highly correlated locations. L_1 regularization over β and Ω jointly encourages sparseness and does not force structure. The parameters encouraging sparseness, γ and λ , are chosen from a validation set using the grid-search technique. These steps are applied for both regression and classification.

Support Vector Machine Regression with PCA (PCASVR)

Ghosh et al. introduced a coupled approach of PCA and Support Vector Machine Regression (SVR) for statistical downscaling [[37, 36]]. The use of SVR for downscaling aims to capture non-linear effects in the data. Identical to the MLR model, we select components that capture 98% ($c_{0.98}$) of explained variability as inputs to the SVR. SVR is used to define the transfer function between the principle components and observed precipitation. Given a set of covariates (the chosen principle components) $\bar{X}_{0.98}$, the support vector regression is defined as [[103]]:

$$f(x) = \beta_0 + \sum_{i=1}^N \beta_i x_i^T x \quad (2.12)$$

with parameters $\beta_i > 0$. The support vectors are selected during training by optimizing the number of points from the training data to define the relationship between the predictand (y) and predictors (X) using a hinge loss:

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left(\frac{1}{2} \|W\|_2^2 + C \sum_{i=1}^N \max(0, |y_i - f(x_i)| - \epsilon) \right) \quad (2.13)$$

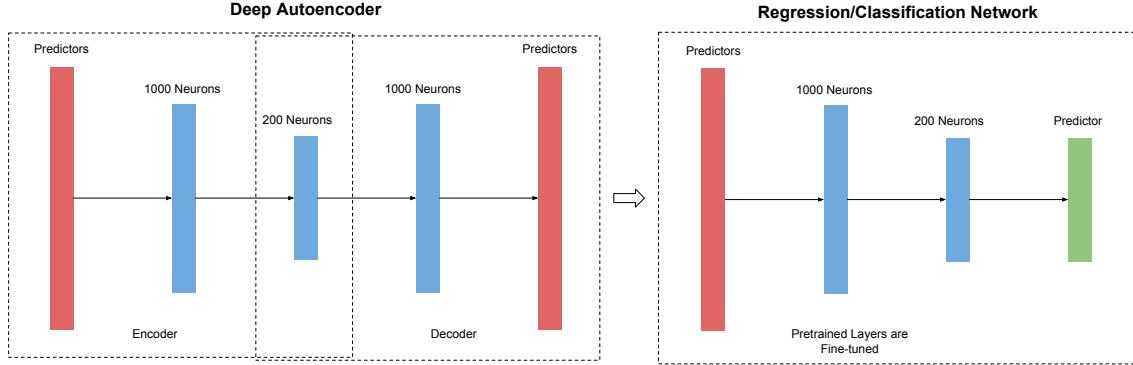


Figure 2.1: Left) The deep autoencoder architecture used to pretrain neural network. Right) The pretrained encoder is fine-tuned for a supervised task.

where parameters C and ϵ are set during training, which we set to 1.0 and 0.1 respectively, corresponding to regularization and loss sensitivity. A linear kernel function is applied to limit overfitting to the training set. Furthermore, support vector classifier was used for classification of rainy versus non-rainy days.

Autoencoder Networks (AE)

Artificial Neural Networks (ANN) have been widely applied to SD with mixed results [[105, 98, 12]], to name a few. In the past, ANNs had difficulty converging to a local minimum. Recent progress in deep learning has renewed interest in ANNs and are beginning to have impressive results in many applications, including image classification and speech recognition [[70, 50, 8]]. Recent success of deep learning can be attributed to its ability to learn high-level abstract representations which generalize across examples. Autoencoder Networks (AEs) present a straightforward architecture to learn abstract representations by training a neural network to predict the input with a bottleneck layer where the number of hidden units is lower than the input dimension [[51]]. Once trained, the middle layer (ie. bottleneck) will contain a lower rank representation of the input data. The encoder layers of the autoencoder network can then be used to initialize a supervised neural network. Hinton and Salakhutdinov show that autoencoders separate the feature space more effectively than PCA on three datasets including handwritten digits, document categories, and human faces [[51]].

This approach can be used for SD by training an autoencoder to learn a low-rank approximation

of the input, much like the PCA implementations above. The learned autoencoder is then used to initialize a neural network extended with an output layer. Two supervised neural networks are then trained, one for classification using a sigmoid cross entropy loss and a second for regression using a euclidean loss. In our application, there are 8835 features which guides the selected architecture with hidden layers of size 1000 and 200 units each, respectively (See Figure 2.1). Each layer is connected by rectified linear units (ReLU). The bottleneck of 200 units is selected to correspond to a similar number of components selected by PCA. In both supervised networks, the 200 units are densely connected to each high resolution location, applying a sigmoid activation for classification. The learned representation from the autoencoder is fine-tuned to learn a more useful representation for predicting precipitation.

For training, we use the Adam Optimizer [[67]] with a learning rate of $1e - 4$ for 1000 epochs and a batch size of 100 examples. During training dropout is applied before each hidden layer which aids in learning quicker and more general representations. At test time dropout is not applied.

2.1.2 MOS: Bias Corrected Spatial Disaggregation

BCSD [[125]] is widely used in the downscaling community due to its simplicity [[2, 12, 124, 84]]. Most commonly, GCM data is bias corrected followed by spatial disaggregation on monthly data and then temporally disaggregated to daily projections. Temporal disaggregation is performed by selecting a month at random and adjusting the daily values to reproduce it's statistical distribution, ignoring daily GCM projections. Thrasher et al. presented a process applying BCSD directly to daily projections [[109]], removing the step of temporal disaggregation. We apply the following steps to downscale the reanalysis precipitation dataset:

- 1) Bias correction of daily projections using observed precipitation. Observed precipitation is remapped to match the reanalysis grid. For each day of the year values are pooled, ± 15 days, from the reanalysis and observed datasets to build a quantile mapping. With the quantile mapping computed, the reanalysis data points are mapped, bias corrected, to the same distribution as the observed data. When applying this method to daily precipitation detrending the data is not necessary because of the lack of trend and is therefore not applied.

- 2) Spatial disaggregation of the bias-corrected reanalysis data. Coarse resolution reanalysis is then bilinearly interpolated to the same grid as the observation dataset. To preserve spatial details

of the fine-grained observations, the average precipitation of each day of the year is computed from the observation and set as scaling factors. These scaling factors are then multiplied to the daily interpolated GCM projections to provide downscaled GCM projections.

2.1.3 Hybrid: Bias Corrected Spatial Disaggregation with MSSL (BCSD-MSSL)

To further understand the use of BCSD in Statistical Downscaling, we propose a technique to estimate the errors introduced in BCSD. As presented above, BCSD utilizes a relatively simple quantile mapping approach to statistical downscaling following by interpolation and spatial scaling. Following the BCSD estimates of the observed climate, we compute the presented errors, which may be consistent and have a predictive signal. Modeling such errors using the transfer function approaches above, such as MSSL, may uncover this signal and improve BCSD projections. To apply this technique, the following steps are taken:

1. Apply BCSD to the coarse scale climate variable and compute the errors.
2. Excluding a hold out dataset, use MSSL where they predictand is the computed errors and the predictands are from a different set of climate variables, such as Temperature, Wind, Sea Level Pressure, etc.
3. Subtract the expected errors modeled by step 2 from BCSD projections in step 1.

The transfer function learned in step 2 is then applicable to future observations.

2.2 Data

The Northeast United States, a region with increasing climate extremes [[108, 57]], is selected to compare the seven SD approaches presented above. As studied by Thibeault and Seth, the northeast has experienced increased heavy precipitation days which is expected to continue under climate change [[108]]. Similarly, Tryhorn et al. found similar increases in extreme precipitation but found that agreement varied between downscaling methods [[112]]. Hence, this region provides a good testbed for comparing SD methods for daily precipitation.

2.2.1 United States Unified Gauge-Based Analysis of Precipitation

High resolution gridded precipitation datasets often provide high uncertainties due to a lack of gauge based observations, poor quality control, and interpolation procedures. Fortunately, precipitation gauge data in the continental United States is dense with high temporal resolution (hourly and daily). The NOAA Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Precipitation exploits the dense network of rain gauges to provide a quality controlled high resolution (0.25° by 0.25°) gridded daily precipitation dataset from 1948 to the current date. State of the art quality control [[19]] and interpolation [[127]] techniques are applied giving us high confidence in the data. We select all locations within the Northeastern United States watershed.

2.2.2 NASA Modern-Era Retrospective Analysis for Research and Applications 2 (MERRA-2)

Reanalysis datasets are often used as proxies to GCMs for statistical downscaling when comparing methods due to their low resolution gridded nature with a range of pressure levels and climate variables. Uncertainties and biases occur in each dataset, but state-of-the-art reanalysis datasets attempt to mitigate these issues. NASA's MERRA-2 reanalysis dataset [[93]] was chosen after consideration of NCEP Reanalysis I/II [[56]] and ERA-Interim [[25]] datasets. [[69]] showed the reduced bias of MERRA and ERA-Interim over NCEP Reanalysis II, which is most often used in SD studies. MERRA-2 provides a significant temporal resolution from 1980 to present with relatively high spatial resolution (0.50° by 0.625°).

Only variables available from the CCSM4 GCM model are selected as covariates for our SD models. Temperature, vertical wind, horizontal wind, and specific humidity are chosen from pressure levels 500hpa, 700hpa, and 850hpa. At the surface level, temperature, sea level pressure, and specific humidity are chosen as covariates. To most closely resemble CCSM4, each variable is spatially upscaled to 1.00° to 1.25° at a daily resolution. A large box centralized around the Northeastern Region ranging from 35°N to 50°N latitude and 110°W to 50°W longitude is used for each variable. When applying the BCSD model, we use a spatially upscaled Land Precipitation MERRA-2 Reanalysis dataset at a daily temporal resolution. Bilinear interpolation is applied over the coast to allow for quantile mapping of coastal locations as needed.

2.3 Experiments and Evaluation

In-depth evaluation of downscaling techniques is crucial in testing and understanding their credibility. The implicit assumptions in SD must be clearly understood and tested when applicable. Firstly, SD models assume that the chosen predictors credibly represent the variability in the predictands. This assumption is partially validated through the choice of predictors presented above, which physically represents variability of precipitation. The remainder of the assumption must be tested through experimentation and statistical tests between downscaled projections and observations. The second assumption then requires the statistical attributes of predictands and predictors to be valid outside of the data used for statistical modeling. A hold out set will be used to test the feasibility of this assumption at daily, monthly, and annually temporal resolutions. Third, the climate change signal must be incorporated in the predictors through GCMs. Predictands chosen for this experiment are available through CMIP5 CCSM4 simulations. It is understood that precipitation is not well simulated by GCMs and therefore not used in ASD models [[97]].

To test these assumptions, we provide in-depth experiments, analysis, and statistical metrics for each method presented above. The years 1980-2004 are used for training and years 2005-2014 are used for testing, taken from the overlapping time period of MERRA-2 and CPC Precipitation. For each PP method, we chose all covariates from each variable, pressure level, and grid point presented above, totaling 12,781 covariates. Each method applies either dimensionality reduction or regularization techniques to reduce complexity of this high dimensional dataset, as presented in Section 2.1. Separate models are trained for each season (DJF, MAM, JJA, SON) and used to project the corresponding observations.

Analysis and evaluation of downscaled projections aim to cover three themes:

1. Ability to capture daily anomalies.
2. Ability to respond to large scale climate trends on monthly and yearly temporal scales.
3. Ability to capture extreme precipitation events.

Similar evaluation techniques were applied in recent intercomparison studies of SD [[12, 41]]. Evaluation of daily anomalies are tested through comparison of bias (Projected - Observed), Root Mean Square Error (RMSE), correlations, and a skill score [[89]]. The skill score presented by

[[89]] measures how similar two probability density functions are from a range of 0 to 1 where 1 corresponds to identical distributions and is defined as:

$$\text{Skill Score} = \sum_{i=1}^B \text{minimum}(Z_i^{(obs)}, Z_i^{(predicted)}) \quad (2.14)$$

where n is the number of bins and Z_i is the proportion of samples to fall in bin i . As this score depends on bin width, the scores cannot be compared between temporal scales. Furthermore, it should be noted that at the daily time scale this skill score may misrepresent skill because of the skewed distribution of daily precipitation. Statistics are presented for winter (DJF), summer (JJA), and annually to understand season credibility. Statistics for spring and fall are computed but not presented in order to minimize overlapping climate states and simply results. Each of the measures are computed independently in space then averaged to a single metric. Large scale climate trends are tested by aggregating daily precipitation to monthly and annual temporal scales. The aggregated projections are then compared using the same metrics as computed for daily anomalies [[89]]. Due to the limited number of data points in the monthly and yearly projections, we estimate each measure using the entire set of projections and observations.

Climate indices are used for evaluation of SD models' ability to estimate extreme events. Four metrics from ClimDEX (<http://www.clim-dex.org>), chosen to encompass a range of extremes, will be utilized for evaluation, as presented in [[12]]. These are:

1. CWD - Annual maximum length of consecutive wet days $\geq 1\text{mm}$
2. R20 - Annual number of very heavy wet days $\geq 20\text{mm}$
3. RX5day - Annual consecutive maximum 5 day precip
4. SDII - Simple Daily intensity index = Annual total / precip days $\geq 1\text{mm}$

Metrics will be computed on observations and downscaled estimates followed by annual (or monthly) comparisons. For example, correlating the maximum number of consecutive wet days per year between observations and downscaled estimates measures each SD models' ability to capture yearly anomalies. The skill score will also be utilized to understand abilities of reproducing statistical distributions.

Table 2.1: Daily Projection Results (hold-out dataset with years 2005-2014). Daily statistical metrics averaged over space for annual, winter, and summer projections. Bias measures the directional error from each model. Correlation (larger is better) and RMSE (lower is better) describe the models ability to capture daily fluctuations in precipitation. The skill score statistic measure the model's ability to estimate the observed probability distribution.

<i>SD Method</i>	Bias (mm/day)			Correlation			RMSE (mm/day)			Skill Score		
	Annual	DJF	JJA	Annual	DJF	JJA	Annual	DJF	JJA	Annual	DJF	JJA
<i>PCAOLS</i>	-0.89	-0.71	-1.16	0.55	0.60	0.49	7.04	5.52	7.49	0.82	0.81	0.76
<i>ELNET</i>	-0.88	-0.66	-1.16	0.64	0.69	0.55	6.53	4.99	7.23	0.84	0.85	0.78
<i>MSSL</i>	-1.58	-1.20	-2.05	0.62	0.64	0.54	6.81	5.49	7.43	0.92	0.90	0.88
<i>AE</i>	-0.86	-0.59	-1.38	0.60	0.65	0.49	6.84	5.37	7.84	0.93	0.92	0.90
<i>PCASVR</i>	0.37	0.04	0.20	0.33	0.39	0.31	11.02	7.89	10.52	0.91	0.87	0.87
<i>BCSD</i>	-0.44	-0.36	-0.36	0.52	0.49	0.46	7.50	6.49	8.12	0.93	0.92	0.89
<i>BCSD-MSSL</i>	-0.16	-0.10	-0.02	0.58	0.60	0.50	6.93	5.61	7.71	0.79	0.80	0.74

2.4 Results

Results presented below are evaluated using a hold-out set, years 2005-2014. Each model's ability to capture daily anomalies, large scale climate trends, and extreme events are presented. Our goal is to understand a SD model's overall ability to provide credible projections rather than one versus one comparisons, therefore statistical significance was not computed when comparing statistics.

2.4.1 Daily Anomalies

Evaluation of daily anomalies depends on a model's ability to estimate daily precipitation given the state of the system. This is equivalent to analyzing the error between projections and observations. Four statistical measures are used to evaluate these errors: Bias, Pearson Correlation, Skill score (Equation 2.14), and Root Mean Square Error (RMSE), as presented in Figure 2.2, Figure 2.3, and Table 2.1). All daily precipitation measures are computed independently in space and averaged to provide a single value. This approach is taken to summarize the measures as simply as possible. Figure 2.2 shows the spatial representation of annual bias in Table 2.1.

Overall, methods tend to underestimate precipitation annually and seasonally with only PCASVR overestimating. MSSL in particular underestimates daily precipitation by -1.58 mm on average while the PCAOLS, ELNET, and AE PP approaches each have nearly 50% less bias. As expected, the bias correction MOS approaches have much lower bias, especially when coupled with MSSL. In Figure 2.2 we see that there is systematic bias over space for each method other than PCASVR, which has no discernible pattern. This suggests that the support vectors for PCA are not consistent

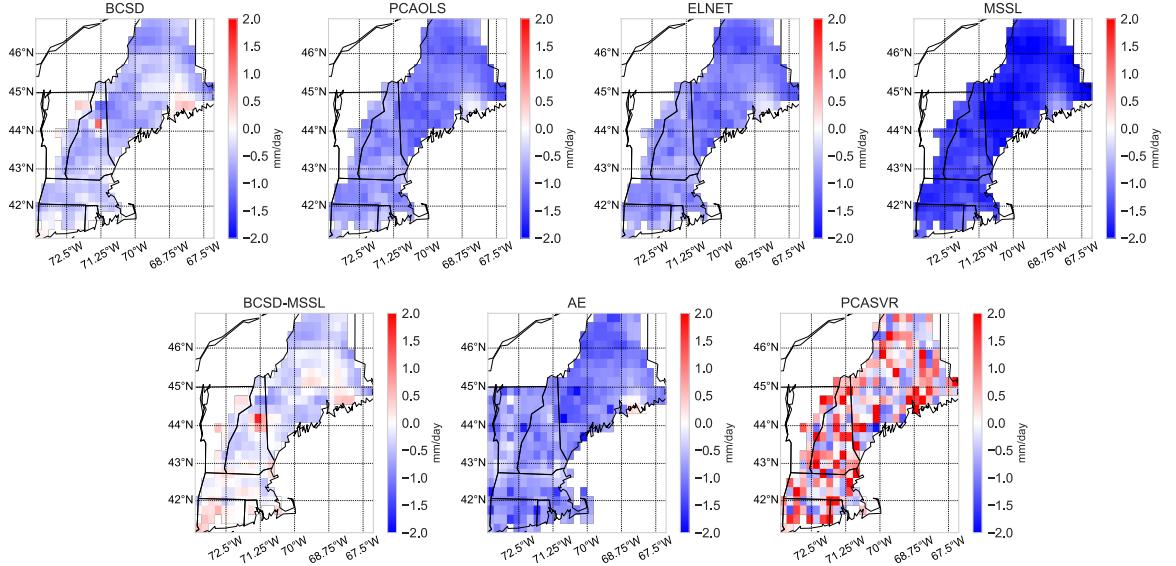


Figure 2.2: Each map presents the spatial bias, or directional error, of the model. White represents no bias produced by the model while red and blue respectively show positive and negative biases.

over space.

RMSE, presented in Figure 2.3 and Table 2.1, measures the overall ability of prediction by squaring the absolute errors. The boxplot in Figure 2.3, where the box present the quartiles and whiskers the remaining distributions with outliers as points, shows the distribution of RMSE annually over space. The sparse models of ELNET and MSSL have similar error distributions and outperform others. This may suggest that the L_1 sparsity constraints in ELNET and MSSL aid in feature selection. While AE's daily RMSE is similar to MSSL on average, the distribution of errors over space, as shown in the boxplots, is much wider. The estimation of error produced by BCSD-MSSL aids in lowering the RMSE of plain BCSD. PCAOLS reasonably minimizes RMSE while PCASVR severely under-performs compared to all other models. Regression models applied minimize error during optimization while BCSD does not. Seasonally, winter is easier to project with summer being more challenging.

Correlation measures in Table 2.1 presents a high linear relationship between projections and observations for the models ELNET (0.64 annually) and MSSL (0.62 annually). We find that BCSD has a lower correlation even in the presence of error correction in BCSD-MSSL. PCASVR provides low correlations, averaging 0.33 annually, but PCAOLS performs substantially better at 0.55.

The skill score is used to measure a model's ability to reproduce the underlying distribution of

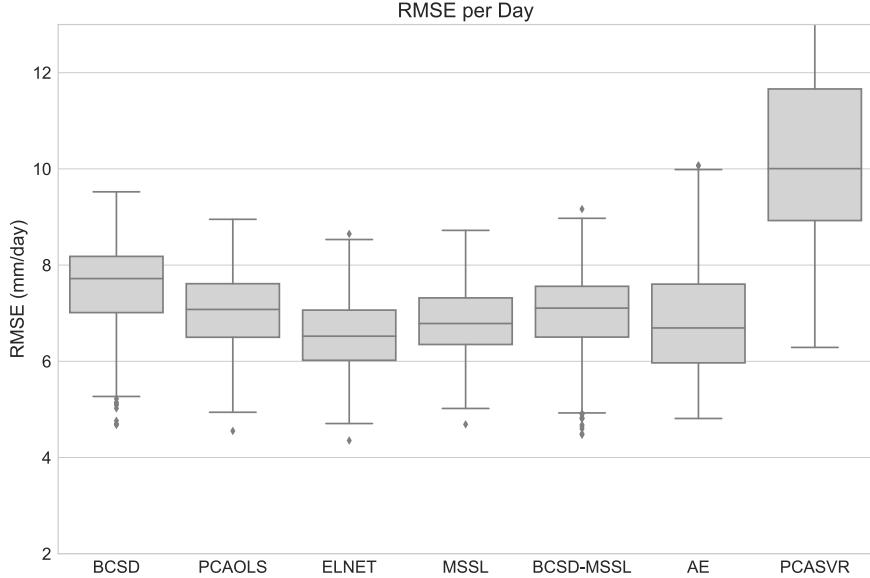


Figure 2.3: Daily Root Mean Square Error (RMSE) is computed for each downscaling location and method. Each boxplot presents the distribution of all RMSEs for the respective method. The box shows the quartiles while the whiskers show the remaining distribution, with outliers displayed by points..

observed precipitation where a higher value is better between 0 and 1. As expected, the quantile mapping used in the BCSD produces well calibrated skill scores, though BCSD-MSSL reduces this skill. AE produces similar distributions as BCSD while PCASVR and MSSL follow. PCAOLS and ELNET, both simple linear models, have difficulty reproducing the observed distributions. This suggests that non-linearity and complexity in PP approaches may better fit the underlying distributions, though as discussed above, we must be cautious when making conclusions using skill at a daily temporal scale.

2.4.2 Large Climate Trends

A SD model's ability to downscale extremes from reanalysis depends on both the response to observed anomalies and ability to reproduce the underlying distribution. Resulting correlation measures present the response to observed anomalies, shown in Figure 2.4 and Table 2.3. We find that BCSD has higher correlations for three metrics, namely consecutive wet days, very heavy wet days, and daily intensity index along with a similar results from 5-day maximum precipitation. Furthermore, modeling BCSD's expected errors with BCSD-MSSL decreases the ability to estimate the chosen extreme indices. The non-linear methods are split, once again, with AE performing well and

Table 2.2: Large scale projection results based on total monthly and yearly rainfall amounts. Bin width for monthly skill is set to 10mm and 100mm on the annual scale.

<i>Time-frame</i>	Bias		Correlation		RMSE		Skill	
	Month	Year	Month	Year	Month	Year		
<i>PCAOLS</i>	-27.14	-325.71	0.62	0.35	50.01	362.73	0.40	0.15
<i>ELNET</i>	-26.80	-321.55	0.71	0.56	46.96	353.67	0.42	0.15
<i>MSSL</i>	-48.16	-577.89	0.67	0.48	62.63	597.80	0.30	0.02
<i>AE</i>	-26.12	-313.44	0.65	0.37	50.10	362.16	0.38	0.23
<i>PCASVR</i>	11.20	134.39	0.38	0.22	92.17	414.40	0.32	0.39
<i>BCSD</i>	-13.31	-159.67	0.87	0.83	31.97	204.78	0.47	0.47
<i>BCSD-MSSL</i>	-4.76	-57.13	0.85	0.79	31.24	155.04	0.46	0.56

PCASVR failing to capture chosen extremes. The linear methods, PCAOLS, ELNET, and MSSL, provide similar correlative performance.

Analysis of a SD model’s ability to capture large scale climate trends can be done by aggregating daily precipitation to monthly and annual temporal scales. In our analysis of monthly precipitation, we compute Bias, RMSE, Skill, and Correlation independently for each month and average over all grid points. These monthly statistics are then averaged and reported in Table 2.2. Metrics are also computed on the annual scale as reported in Table 2.2.

As expected, the tested MOS methods outperformed PP methods for large scale trends as the underlying MOS covariate is precipitation but is not used in PP approaches. We also find that BCSD-MSSL generally improves beyond BCSD on large scale trends, likely due to inclusion of more climate variables. Within the PP methods, PCAOLS, ELNET, and AE have similar results in monthly and annual projections with monthly skills approaching MOS. The large negative biases shown on the daily scale are compounded at monthly and annual scales, negatively affecting results. In particular, MSSL Bias is 50 mm/day which shifts it’s distribution far enough to decrease skill to 0.02. This effect can be seen for the PCAOLS, ELNET, and AE. On the other hand, while PCASVR has difficulty reducing bias and error, its predicted values cover the distribution more completely.

2.4.3 Extreme Events

A SD model’s ability to downscale extremes from reanalysis depends on both the response to observed anomalies and ability to reproduce the underlying distribution. Resulting correlation measures present the response to observed anomalies, shown in Figure 2.4 and Table 2.3. We find that

Table 2.3: Statistics for ClimDEX Indices

<i>Metric</i>	Correlation				Skill Score			
	CWD	R20	RX5day	SDII	CWD	R20	RX5day	SDII
<i>model</i>								
<i>PCAOLS</i>	0.25	0.65	0.44	0.67	0.69	0.60	0.65	0.44
<i>ELNET</i>	0.36	0.71	0.57	0.64	0.79	0.62	0.63	0.35
<i>MSSL</i>	0.33	0.84	0.56	0.52	0.90	0.63	0.57	0.16
<i>PCASVR</i>	0.24	0.81	0.19	0.25	0.78	0.89	0.80	0.65
<i>AE</i>	0.24	0.68	0.58	0.48	0.91	0.67	0.79	0.53
<i>BCSD</i>	0.43	0.83	0.73	0.70	0.71	0.80	0.84	0.44
<i>BCSD-MSSL</i>	0.25	0.83	0.70	0.69	0.41	0.75	0.84	0.08

BCSD has higher correlations for three metrics, namely consecutive wet days, very heavy wet days, and daily intensity index along with a similar results from 5-day maximum precipitation. Furthermore, modeling BCSD’s expected errors with BCSD-MSSL decreases the ability to estimate the chosen extreme indices. The non-linear methods are split, once again, with AE performing well and PCASVR failing to capture chosen extremes. The linear methods, PCAOLS, ELNET, and MSSL, provide similar correlative performance.

A skill score is used to quantify each method’s ability to estimate an indices statistical distribution, presented in Table 2.3. Contrary to correlative results, PCASVR outperforms the other methods on two metrics, very heavy wet days and daily intensity index, with better than average scores on the other two metrics. BCSD also performs reasonably well in terms of skill scores while BCSD-MSSL suffers from the added complexity. MSSL estimates the number of consecutive wet days well but is less skilled on other metrics. The very complex AE model has a reasonable ability to capture the underlying distribution.

Figure 2.4 displays a combination of correlative power and magnitude estimate of the daily intensity index. The SDII metric is computed from total annual precipitation and number of wet days. A low SDII metric corresponds to either a relatively large number of estimated wet days or low annual precipitation. We find that the on average methods underestimate this intensity. In particular, both BCSD and BCSD-MSSL severely underestimate SDII. While PCASVR overestimates SDII, its skill at representing the observed SDII distribution is largest (see Table 2.3).

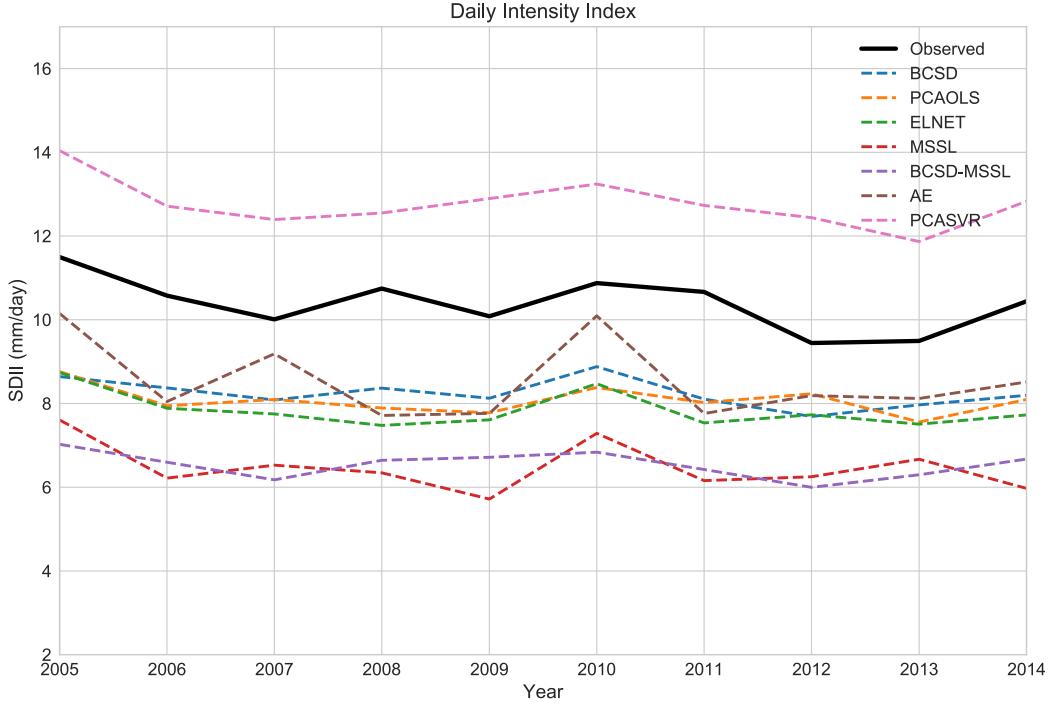


Figure 2.4: (Best viewed in color) The simple daily intensity index (Annual Precipitation/Number of Precipitation Days) averaged per year.

2.5 Discussion

The ability of statistical downscaling methods to produce credible results is necessary for a multitude of applications. Despite numerous studies experimenting with a wide range of models for statistical downscaling, none have clearly outperformed others. In our study, we experiment with the off-the-shelf applicability of machine learning advances to statistical downscaling in comparison to traditional approaches.

Multi-task Sparse Structure learning, an approach that exploits similarity between tasks, was expected to increase accuracy beyond automated statistical downscaling approaches. We find that MSSL does not provide improvements beyond ELNET and other PP approaches. Furthermore, the parameter set, estimated through cross-validation, attributed no structure aiding prediction.

The recent popularity in deep learning along with its ability to learn complex high-level representations through autoencoding motivated us to experiment with basic architectures for statistical downscaling. AEs benefit greatly by implicitly learning abstract non-linear features based on the target variable. This approach had moderate success relative to the simpler methods. More ex-

perimentation with AEs and other deep learning architectures may provide valuable downscaling results.

BCSD, a popular approach to statistical downscaling, outperformed the more complex models in estimating underlying statistical distributions and climate extremes. In many cases, correcting BCSD's error with MSSL increased daily correlative performance but decreased skill of estimating the distribution. From this result, we can conclude that a signal aiding in prediction was lost during quantile mapping, interpolation, or spatial scaling. Future work may study and improve each step independently to increase overall performance.

Of the seven statistical downscaling approaches studied, the traditional BCSD and PP methods outperformed Support Vector Regression, while downscaling daily precipitations. We find that BCSD is skilled at estimating the statistical distribution of daily precipitation, generating better estimates of extreme events. The expectation of AE and MSSL, two recent machine learning advances which we found most applicable to statistical downscaling, to outperform basic modeled proved false. Improvements and customization of machine learning methods is needed to provide more credible projections. In the next chapter, we present a new approach to statistical downscaling using deep learning based image super-resolution.

Chapter 3

DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution

3.1 Introduction

The lack of explicit spatial models in SD of ESMs motivated us to study the applicability of computer vision approaches, most often applied to images, to this problem. More specifically, advances in single image super-resolution (SR) correspond well to SD, which learns a mapping between low- and high-resolution images. Moreover, as SR methods attempt to generalize across images, we aim to provide downscaled climate projections to areas without high-resolution observations through what may be thought of as transfer learning. Though we will discuss this topic further in section 4 (Related Work), we found that super-resolution convolutional neural networks were able to capture spatial information in climate data to improve beyond existing methods.

Lastly, we present a framework using our super-resolution approach to downscale ensemble ESMs over the Continental United States (CONUS) at a daily temporal scale for four emission scenarios by using NASA’s Earth Exchange (NEX) platform. NEX provides a platform for scientific collaboration, knowledge sharing and research for the Earth science community. As part of NEX, along with many other earth science data products, NASA scientists have already made monthly

downscaled ESMs for CONUS up to the year 2100 at 30 arc seconds (NEX-DCP30) that are openly available to the public. However, the downscaling methodology, bias correction spatial disaggregation, has limitations and the monthly scale reduces the applicability to studying extreme events. The improvement of such data products is vital for scientists to study local impacts of climate change to resources society depends on.

3.1.1 Key Contributions

The key contributions are as follows:

- We present DeepSD, an augmented stacked super-resolution convolutional neural network for statistical downscaling of climate and earth system model simulations based on observational and topographical data.
- DeepSD outperforms a state-of-the-art statistical downscaling method used by the climate and earth science communities as well as a suite of off-the-shelf data mining and machine learning methods, in terms of both predictive performance and scalability.
- The ability of DeepSD to outperform and generalize beyond grid-by-grid predictions suggests the ability to leverage cross-grid information content in terms of similarity of learning patterns in space, while the ability to model extremes points to the possibility of improved ability beyond matching of quantiles. Taken together, this leads to the new hypothesis that methods may be able to use spatial neighborhood information to predict in regions where data may be sparse or low in quality.
- For the first time, DeepSD presents an ability to generate, in a scalable manner, downscaled products from model ensembles, specifically, simulations from different climate modeling groups across the world run with different emissions trajectories and initial conditions.
- DeepSD provides NASA Earth Exchange (NEX) a method of choice to process massive climate and earth system model ensembles to generate downscaled products at high resolutions which can then be disseminated to researchers and stakeholders.

3.1.2 Organization of the Paper

The remainder of the paper is organized as follows. Section 2 (Earth Science Data) presents necessary data used for SD along with their associated challenges. Section 3 (Statistical Downscaling) discusses the problem of SD. Section 4 (Related Work) discusses techniques previously applied to SD along with an overview of super-resolution methods. Furthermore, we discuss the relationships between images and climate data. Section 5 (Methodology) presents DeepSD, the augmented stacked super-resolution convolutional neural network formulation. In section 6 (Experiments) we compare our method to another SD technique and three off-the-shelf machine learning approaches and outline the process by which we will scale our method to many climate model simulations. In section 7 (conclusion) we briefly discuss results, limitations, and future work.

3.2 Earth Science Data

Earth science data stems from a variety of areas, including climate simulations, remote sensing through satellite observations, and station observations. The spatio-temporal nature of such data causes heavy computational and storage challenges. For instance, a single climate variable at the daily temporal and 4km spatial scales over only the United States requires 1.2GB of storage. Multiplying this effect over a large number of variables, including precipitation, temperature, and wind, globally creates high storage and processing requirements. Furthermore, analysis of these complex datasets require both technical and domain expertise.

ESM outputs, as discussed previously, are one form of earth science data which is crucial to the understanding of our changing climate. The most recent ESMs are a product of the fifth phase of the Coupled Model Intercomparison Project which simulate the climate through a dynamical system coupling effects from the atmosphere, land, and ocean [106]. However, it is well understood that holes in these models exist, including low-resolution and lack of model agreement, particularly for precipitation [97].

We can harness information in observational datasets in order to learn statistical models mapping ESM outputs to a higher resolutions. Observational datasets are available through a variety of sources, including satellite observations, station observations, and a mixture of both, namely reanalysis datasets. Often, SD models are built to downscale ESMs directly to a observational station

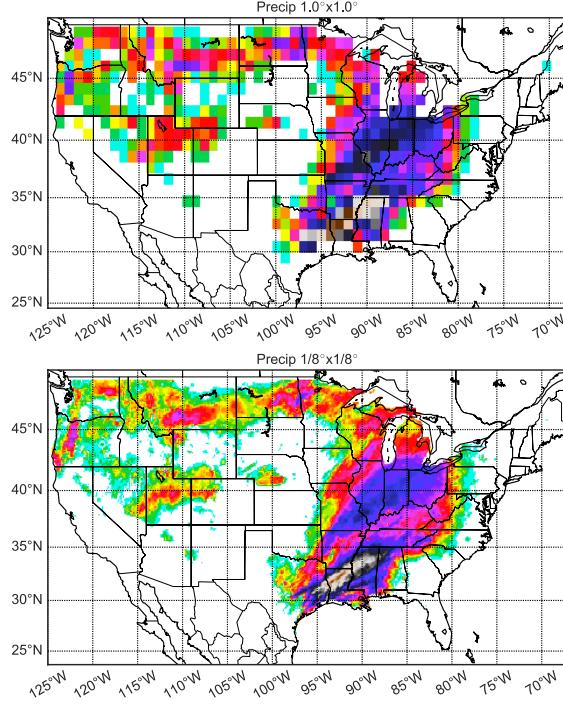


Figure 3.1: Prism Observed Precipitation: A) Low resolution at 1.0° ($\sim 100\text{km}$). B) High resolution at $1/8^{\circ}$ ($\sim 12.5\text{km}$).

while others aim to downscale to a grid based dataset. Gridded observational datasets are often built by aggregating station observations to a defined grid. For example, in our application, we obtain precipitation through the PRISM dataset at a 4km daily spatial resolution which aggregates station observations to a grid with physical and topographical information [22]. We then upscale the precipitation data to $1/8^{\circ}$ (12.5 km) as our high-resolution observations. Following, we upscale further to 1° corresponding to a low-resolution precipitation, as applied in [90]. The goal is then to learn a mapping between our low-resolution and high-resolution datasets.

Furthermore, topography has large effects on weather and climate patterns with lower temperatures, more precipitation, and higher winds [22, 23]. Taking advantage of the valuable topographical information at different scales, where $1/2^{\circ}$ may capture large scale weather patterns while $1/8^{\circ}$ spatial resolution can capture high-resolution precipitation biases.

Each of the earth science data products discussed inherently possess rich spatial dependencies, much like images. However, traditionally statistical downscaling methods, particularly regression based models, vectorize spatial data, removing this spatial structure. While colored images contain channels consisting of, for example, red, green, and blue, climate data may be represented

analogously such that the channels correspond to climate variables and topographical data. Similar approaches have been applied to satellite datasets for image classification [7] and resolution enhancement [130]. Though climate data is more complex than images due to its dynamics and chaotic nature, we propose that this representation allows scientists to approach the data in an unconventional manner and apply augmented models developed for image processing.

3.3 Statistical Downscaling

SD is the problem of mapping a low-resolution climate variable to a high-resolution projection. This mapping, which must transform a single grid point to multiple points is an ill-posed problem, one with many possible solutions (see Figure 4.2). However, we can mitigate the ill-posed problem by including static high-resolution topography data in conjunction with other low-resolution climate variables. We learn the SD model using observed climate datasets and then infer downscaled ESM projections. Spatial and temporal non-stationarity of the changing climate system challenges traditional statistical techniques. Downscaling precipitation further challenges these methods with sparse occurrences and skewed distributions. The combination of an ill-posed problem, uncertainty in the climate system, and data sparsity propagates uncertainty in downscaled climate projections further.

3.4 Related Work

As mentioned previously, SD has a rich and expansive history in the climate community. SD consists of three fundamental categories: regression models and weather classification schemes which improve spatial resolution while weather generators increase temporal resolution (ie. monthly to daily) [121]. As our interest is in increasing spatial resolution we will review regression methods and weather classification.

Regression methods applied to SD are wide in scope, both linear and non-linear, and vary based on the specific climate variable and temporal scale. For instance, downscaling daily precipitation, which we will focus on, relies on a sparse observational dataset where few days contain rainfall while the amount of rainfall in those days follow a skewed distribution. Automated Statistical

Downscaling (ASD) presents a traditional framework for this problem where a classification model is first used to classify days with precipitation followed by a regression to estimate the amount [46]. Similar approaches, among others, include quantile regression neural networks [15], bayesian model averaging[132], and expanded downscaling[11]. Each of these regression models learns a statistical relationship between observed low- and high-resolution pairs and is then applied to ESMs. Another widely used approached is Bias Corrected Spatial Disaggregation (BCSD), which begins by bias correcting a ESM to match the distribution of the high-resolution observed dataset followed by interpolation and spatial scaling to correct for local biases[124, 109]. Though BCSD is a simple approach, it has been shown to perform well compared to more complex methods [12, 83]. Furthermore, we have shown that BCSD performs similarly, or better, when compared to off-the-shelf ASD approaches [115].

Weather classification methods take a different approach to statistical downscaling through nearest neighbor estimates, grouping weather events into similar types. Given a set of observed low- and high-resolution pairs, one can compute a distance measure between an ESM and the low-resolution observations to select the nearest high-resolution estimate. Constructed analogues furthers the method by performing a regression on a group of the nearest neighbor estimates[48]. More advances, but similar approaches, have recently been presented, including Hierarchical Bayesian inference models[78].

While the approaches discussed above are often sufficient in downscaling means, they tend to fail at downscaling extreme events. For instance, ASD approaches perform reasonably well at downscaling average precipitation [46] but performs poorly at the extremes [12]. As discussed by Bürger et al. [12] and Mannshardt-Shamseldin et al. [77], as well as others, specific approaches to downscaling extremes are often required. These specialized approaches, such as those using Generalized Extreme Value theory, have been developed for this purpose [77, 45]. Ideally, a single approach to downscaling leveraging all available information would capture both averages and extremes, giving the user a more credible dataset.

To our knowledge little work has been attempted to explicitly capture spatial properties for improving downscaled projections. As computer vision approaches are built to exploit the spatial structure of images, we are motivated to understand the applicability of such methods to climate datasets. As introduced previously, we represent climate variables as channels, analogous to im-

ages, and model them similarly. However, we note that this presents an analogy and not a direct correspondence.

Using the analogy between climate datasets and images, we can relate statistical downscaling to image super-resolution, where one aims to learn a mapping from low- to high-resolution image pairs. Specifically, single image super-resolution (SR), as the name suggests, increases the resolution of a single image, rather than multiple images, from a scene.

The most successful approaches to SR have been shown to be patch based (or example-based) techniques, achieving state-of-the-art performance [111, 26, 120]. Originally proposed by Glasner et al. [38], patch based methods exploit self-similarity between images to produce exemplar patches. This approach has evolved into different variations of nearest neighbor techniques between low- and high-resolution patches through what is known as dictionary learning [30, 17, 129]. Dictionary learning approaches to SR are analogous to those presented by weather classification SD schemes. Furthermore, approaches including kernel regression [128], random forests[99], and anchored neighborhood regression [111], have been proposed for SR to improve accuracy and speed, all related to methods presented in SD literature [36, 14, 104]. Sparse-coding techniques, a form of dictionary learning, have recently shown state-of-the-art results in both speed and accuracy [111].

Convolutional neural networks were recently presented as a generalization of sparse coding, improving upon past state-of-the-art performances[26, 120]. The sparse coding generalization, non-linearity, network flexibility, and scalability to large datasets presents an opportunity to apply Super Resolution Convolutional Neural Networks to SD[26].

3.5 Methodology

This section begins by describing and formulating Super-Resolution Convolutional Neural Networks (SRCNN), as presented by [26]. We then introduce a stacked SRCNN architecture such that the output of one SRCNN is the input to the following SRCNN. DeepSD, the adaptation of a stacked SRCNNs to SD, is then introduced.

Super Resolution Convolutional Neural Network + Auxiliary Variable

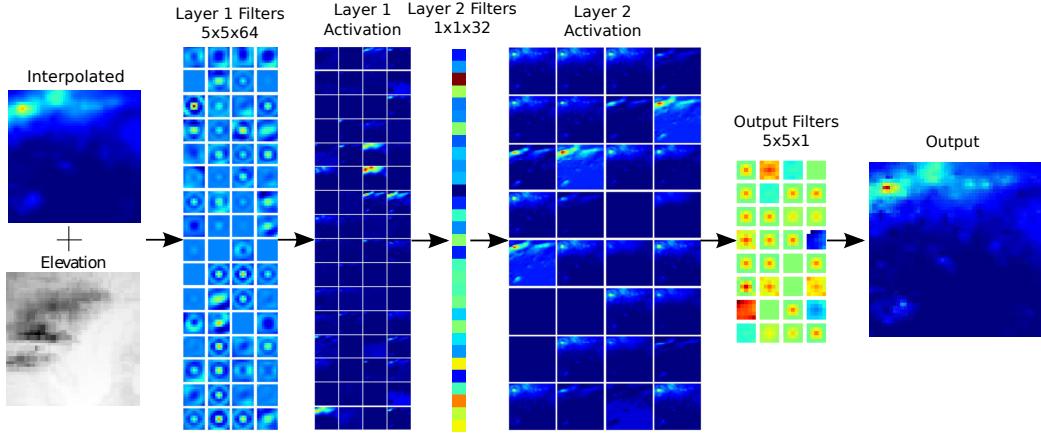


Figure 3.2: Augmented SRCNN Architecture. From the left to right: Precipitation and Elevation sub-image pair, filters learned in layer 1, layer 1 activations, layer 2 filters, layer 2 activations, layer 3 filters, and HR precipitation label.

3.5.1 Super-resolution CNN

SR methods, given a low-resolution (LR) image, aim to accurately estimate a high-resolution image (HR). As presented by Dong et al. [26], a CNN architecture can be designed to learn a functional mapping between LR and HR using three operations, patch extraction, non-linear mappings, and reconstruction. The LR input is denoted as \mathbf{X} while the HR label is denoted as \mathbf{Y} .

A three layer CNN is then constructed as follows to produce a high resolution estimate and presented in Figure 3.2. Layer 1 is formulated as

$$F_1(\mathbf{X}) = \max(0, W_1 * \mathbf{X} + B_1),$$

where ‘*’ is the convolution operation and the *max* operation applies a Rectified Linear Unit [86] while W_1 and B_1 are the filters and biases, respectively. W_1 consists of n_1 filters of size $c \times f_1 \times f_1$. The filter size, $f_1 \times f_1$, operates as an overlapping patch extraction layer where each patch is represented as a high-dimensional vector.

Correspondingly, layer 2 is a non-linear operation such that

$$F_2(\mathbf{X}) = \max(0, W_2 * F_1(\mathbf{X}) + B_2)$$

where W_2 consists of n_2 filters of size $n_1 \times f_2 \times f_2$ and B_2 is a bias vector. This non-linear operation

maps high-dimensional patch-wise vectors to another high-dimensional vector.

A third convolution layer is used to reconstruct a HR estimate such that

$$F(\mathbf{X}) = W_3 * F_2(\mathbf{X}) + B_3.$$

Here, W_3 contains 1 filter of size $n_2 \times f_3 \times f_3$. The reconstructed image $F(\mathbf{X})$ is expected to be similar to the HR image, \mathbf{Y}

This end-to-end mapping then requires us to learn the parameters $\Theta = \{W_1, W_2, W_3, B_1, B_2, B_3\}$. A Euclidean loss function with inputs $\{\mathbf{X}_i\}$ and labels $\{\mathbf{Y}_i\}$ is used where the optimization objective is defined as:

$$\operatorname{argmin}_{\Theta} \sum_{i=1}^n \|F(\mathbf{X}_i; \Theta) - \mathbf{Y}_i\|_2^2 \quad (3.1)$$

such that n is the number of training samples (batch size).

We note that convolutions in layers 1, 2, and 3 decrease the image size depending on the chosen filter sizes, f_1, f_2 , and f_3 . At test time, padding using the replication method is applied before the convolution operation to ensure the size of the prediction and ground truth correspond. During training, labels are cropped such that \mathbf{Y} and $F(\mathbf{X}_i; \Theta)$, without padding, are of equal size.

3.5.2 Stacked SRCNN

Traditional SR methods are built for resolution enhancements of factors from 2 to 4 while statistical downscaling conservatively requires resolution increases of factors from 8 to 12. Rather than enhancing resolution directly to 8-12x, as SR applications typically do, we take an alternative approach. To achieve such a large resolution improvement, we present stacked SRCNNs such that each SRCNN increases the resolution by a factor of s . This approach allows the model to learn spatial patterns at multiple scales, requiring less complexity in the spatial representations. The approach of stacking networks has been widely used in deep learning architectures, including stacked denoising autoencoders [117] and stacked RBMs for deep belief networks [52]. However, contrary to the above networks where stacking is applied in an unsupervised manner, each SRCNN is trained independently using their respective input/output resolutions and stacked at test time.

A similar approach using cascading super-resolution networks showed positive results for up-

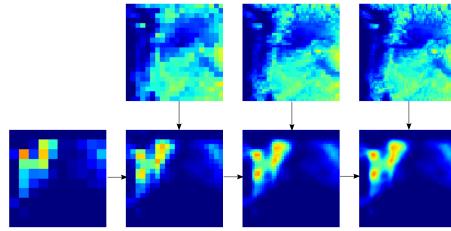


Figure 3.3: Layer by layer resolution enhancement from DeepSD using stacked SRCNNs. Top Row: Elevation, Bottom Row: Precipitation. Columns: 1.0° , $1/2^\circ$, $1/4^\circ$ and $1/8^\circ$ spatial resolutions. scaling factors below 4 [120], however through experimentation we found that cascading SRCNNs performed worse than stacked SRCNNs. The ability of arbitrarily upscaling ground truth images to lower resolution allows for input/output pairs to be produced at multiple scales to train stacked SRCNNs. However, while training a cascading model, the output of each SRCNN is the input to the following SRCNN, which may be leading to unnecessary error propagation through the network.

3.5.3 DeepSD

We now present DeepSD, an augmented and specific architecture of stacked SRCNNs, as a novel SD technique. When applying SR to images we generally only have a LR image to estimate a HR image. However, during SD, we may have underlying high-resolution data coinciding with this LR image to estimate the HR images. For instance, when downscaling precipitation we have two types on inputs including LR precipitation and static topographical features such as HR elevation and land/water masks to estimate HR precipitation. As topographical features are known beforehand at very high resolutions and generally do not change over the period of interest they can be leveraged at each scaling factor. As done when training stacked SRCNNs, each SRCNN is trained independently with it's associated input/output pairs. As presented in figure 3.3, inference is executed by starting with the lowest resolution image with it's associated HR elevation to predict the first resolution enhancement. The next resolution enhancement is estimated from the previous layer's estimate and it's associated HR elevation. This process is repeated for each trained SRCNN. Figure 3.3 illustrates this process with a precipitation event and it's various resolution improvements. We see that this stacked process allows the model to capture both regional and local patterns.

3.6 Application of DeepSD

Though high resolution precipitation is crucial to climate adaptation, it makes up two of the four major holes in climate science [97]. Furthermore, both statistical and dynamical downscaling approaches have been shown to add little information beyond coarse ESMs when applied to precipitation [63, 12]. This motivates our application to downscale daily precipitation over the CONUS, a region where data is credible and abundant at high resolutions. As presented above, we use daily precipitation from the PRISM dataset [22] and elevation from Global 30 Arc-Second Elevation Data Set (GTOPO30) provided by the USGS. These datasets are used to train and test DeepSD, which we compare to BCSD, a widely used statistical downscaling technique, as well as three off-the-shelf machine learning regression approaches. The years 1980 to 2005 were used for training (9496 days) while the years 2006 and 2014 (3287 days) were used for testing. Lastly, we present a scalable framework on the NASA Earth Exchange (NEX) platform to downscale 20 GCMs for multiple emission scenarios.

3.6.1 Training DeepSD

Our experiments downscale daily precipitation from 1.0° to $1/8^\circ$, an 8x resolution enhancement, using three SRCNN networks each providing a 2x resolution increase ($1.0^\circ \rightarrow 1/2^\circ \rightarrow 1/4^\circ \rightarrow 1/8^\circ$). For comparison and experimentation purposes, we also trained a single network with 8x resolution increase which we denote as SRCNN in the results.

Data preprocessing

Data for a single day at the highest resolution, $1/8^\circ$, covering CONUS is an “image” of size 208x464. Precipitation and elevation are used as input channels while precipitation is the sole output. Images are obtained at each resolution through up-sampling using a bicubic interpolation. For instance up-sampling to 1.0° decreases the image size from 208x464 to 26x58. Precipitation features for the first SRCNN, downscaling from 1.0° to $1/2^\circ$, are first up-sampled to 1.0° and then interpolated for a second time to $1/2^\circ$ in order to correspond to the output size of 52x116. This process is subsequently applied to each SRCNN depending on its corresponding resolution. During the training phase, 51x51 sub-images are extracted at a stride of 20 to provide heterogeneity in the

Model	Bias	Corr	RMSE	Skill	Runtime
	(mm/day)		(mm/day)		(secs)
Lasso	0.053	0.892	2.653	0.925	1297
ANN	0.049	0.862	3.002	0.907	2015
SVM	-1.489	0.886	3.205	0.342	27800
BCSD	-0.037	0.849	4.414	0.955	—
SRCNN	-0.699	0.894	2.949	0.833	24
DeepSD	0.022	0.914	2.529	0.947	71

Table 3.1: Comparison of predictive ability between all six methods for 1000 randomly selected locations in CONUS. Runtime is computed as the amount of time to downscale 1 year of CONUS. training set. The number of sub-images per year (1095, 9125, and 45,625) increase with resolution. Features and labels are normalized to zero mean and unit variance. Precipitation values are only available over land so we set each null value to a sufficiently low value of -5 which is then masked after downscaling accordingly.

Training Parameters

All SRCNNs are trained with the same set of parameters, selected using those found to work well by Dong et al. [26]. Layer 1 consists of 64 filters of 9x9 kernels, layer 2 consists of 32 filters of 1x1 filters, and the output layer uses a 5x5 kernel (see Figure 3.2). Higher resolution models which have a greater number of sub-images may gain from larger kernel sizes and an increased number of filters. Each network is trained using Adam optimization [67] with a learning rate of 10^{-4} for the first two layers and 10^{-5} for the last layers. Each model was trained for 10^7 iterations with a batch size of 200. Tensorflow [1] was utilized to build and train DeepSD. Training harnessed three Titan X GPUs on an NVIDIA DIGITS DevBox by independently training each SRCNN. Inference was then executed sequentially on a single Titan X GPU on the same machine.

3.6.2 Comparison

State-of-the-Art Methods

Bias Correction Spatial Disaggregation (BCSD) [124] is a simple but effective method for statistical downscaling. Many studies have compared BCSD to a variety of other downscaling methods and have found good results in estimating the underlying distribution of precipitation [12]. In our experiments we apply the daily BCSD technique [109] to precipitation over CONUS. First, the

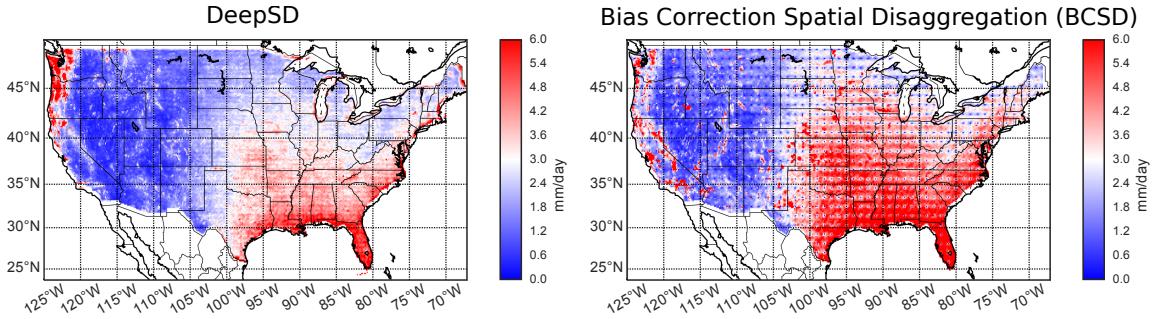


Figure 3.4: Daily Root Mean Square Error (RMSE) computed at each location for years 2006 to 2014 (test set) in CONUS for Left) DeepSD and Right) BCSD. Red corresponds to high RMSE while blue corresponds to low RMSE.

high-resolution precipitation is linearly interpolated to the low-resolution grid. Contrary to applying BCSD to ESMs, we are not required to perform quantile mapping as the distributions are identical. Next, the low-resolution precipitation is interpolated back to high-resolution such that the fine-grained information is lost. Then, scaling factors are computed by dividing high-resolution observations with the interpolated data over the training set (1980-2005). Lastly, the interpolated data is multiplied by the scaling factors to provide downscaled projections. For a more detailed description of this implementation of BCSD see [109]. The projections over the test set (2006-2014) are then used for comparison to BCSD.

A second set of methods, Automated-Statistical Downscaling (ASD) [46], is applied to compare a variety of regression techniques to DeepSD. ASD consists of two steps for downscaling precipitation: 1. Classifying rainy/non-rainy days (mm), 2. Estimating total precipitation on rainy days. Hence, this approach requires both classification and regression methods. We compare three ASD approaches using logistic and lasso regression, support vector machine (SVM) classifier and regression, and artificial neural network (ANN) classifier and regression. The Lasso penalty parameter at each location was chosen using 3-fold cross-validation. The SVMs were trained with a linear kernel and a penalty parameter of 1.0. Each ANN consists of a single layer of 100 units connected with a sigmoid function. A 9 by 9 box for the LR precipitation surrounding the downscaled location is selected as features. Each downscaled location requires individually optimized parameters making the process computationally intensive and complex. Hence, we randomly selected 1000 locations to downscale as a trade-off between complexity and statistical certainty around our results. All features and labels are normalized to zero mean and unit variance.

Season	Model	Bias	Corr	RMSE	Skill ¹
		(mm/day)		(mm/day)	
DJF	BCSD	0.02	0.89	2.36	0.95
	DeepSD	-0.03	0.95	1.53	0.94
JJA	BCSD	0.01	0.78	4.15	0.92
	DeepSD	-0.05	0.86	3.29	0.91
MAM	BCSD	0.01	0.87	3.02	0.94
	DeepSD	-0.03	0.93	2.29	0.93
SON	BCSD	0.01	0.87	3.27	0.94
	DeepSD	-0.04	0.93	2.31	0.94

Table 3.2: Comparison of Predictive Ability between DeepSD and BCSD for each season, Winter, Summer, Spring, and Fall. Values are computed at each location in CONUS and averaged.

Daily Predictability

DeepSD’s ability to provide credible projections is crucial to all stakeholders. While there are many facets to statistical downscaling, we use a few key metrics to show DeepSD’s applicability. Root mean square error (RMSE) and Pearson’s correlation are used to capture the predictive capabilities of the methods. Figure 3.4 maps this RMSE (mm/day) for each location. Bias, the average error, presents the ability to estimate the mean while a skill score metric, as presented in [89], is used to measure distribution similarity. Skill is computed as

$$skill = \sum_{i=1}^n \min(Z_o^{(i)}, Z_m^{(i)}) \quad (3.2)$$

such that Z_o and Z_m are the observed and DeepSD’s empirical probability density function while n is the number of bins. Hence, the skill score is between 0 and 1 where 1 is the best.

Our first experiment compares six approaches, DeepSD, SRCNN (DeepSD w/o stacking), BCSD, Lasso, SVM, and ANN, on their ability to capture daily predictability, presented in Table 3.1. The four metrics discussed above are computed and averaged over the 1000 randomly selected locations in CONUS where ASD methods were trained. We find that DeepSD outperforms the other approaches in terms of bias, correlation, and RMSE and closely behind BCSD in terms of skill. We also find that the stacking performed by DeepSD provides a large performance improvement beyond a single SRCNN network with an 8x resolution increase. Furthermore, we find that SVM performs poorly in testing while having the longest runtime. Similarly, the least complex ASD method, Lasso, outperforms the more complex ANN. As expected, BCSD, a method built around estimating the un-

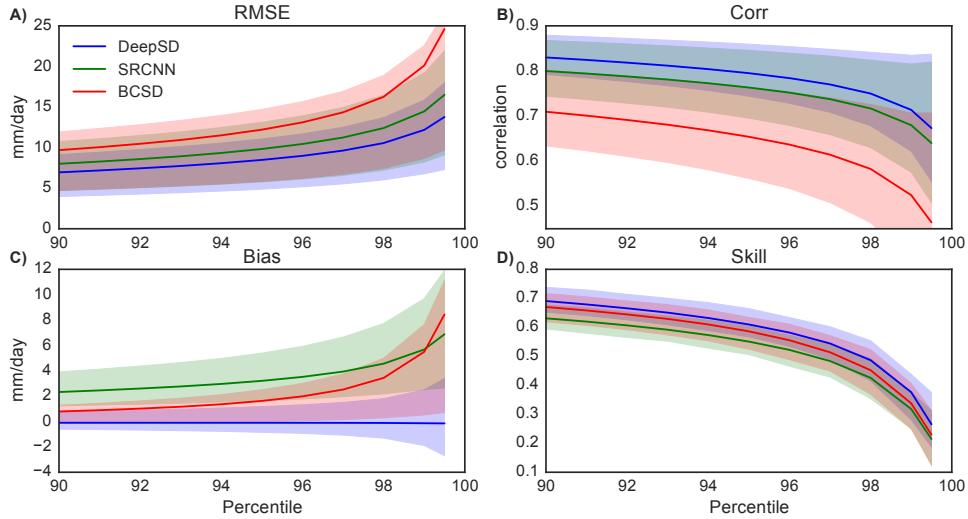


Figure 3.5: Comparison of DeepSD, SRCNN, and BCSD for increasingly extreme precipitation. At each location in CONUS, all precipitation events above a percentile threshold (x-axis) are selected. Percentile thresholds between 90 and 99.9 are used. A) RMSE, B) Correlation, C) Bias, and D) Skill are computed at each location and averages over CONUS. The confidence bounds of each metric are taken from the 25th and 75th quantiles.

derlying distribution, does well in minimizing bias and estimating the underlying distribution. For these reasons, in conjunction with our previous findings [115], the remaining experiments will limit the methods to DeepSD and BCSD.

In the next experiment compare DeepSD and BCSD, the two scalable and top performing methods from the previous experiment, with each metric over CONUS. Each metric is computed per location and season using the daily observations and downscaled estimates then averaged over CONUS, as presented in Table 3.2. We find that DeepSD has high predictive capabilities for all seasons, higher correlation and lower RMSE, when compared to BCSD. Similar results are shown in Figure 3.4 where DeepSD has a lower RMSE than BCSD for 79% of CONUS. Furthermore, we find each method's ability to estimate the underlying distribution well with low bias, < 0.5 mm/day, and a high skill score of ~ 0.98 . As BCSD is built specifically to minimize bias and match the underlying distribution, DeepSD's performance is strong. Overall, DeepSD outperforms BCSD for the chosen set metrics.

Predicting Extremes

As discussed in section 4, downscaling both averages and extreme events with a single method is challenging. Our last experiment tests this challenge by comparing DeepSD’s ability to estimate extreme precipitation events when compared to BCSD, an approach shown to perform well [12]. A varying quantile threshold approach is used to test each methods ability to capture extreme events. For instance, given a downscaled location we compute RMSE, correlation, bias, and skill for all precipitation events greater the 90th percentile. This is done for a range of percentiles between 90 and 99.9 and averaged over all locations in CONUS. Along with the mean, we select the 25th and 75th quantiles of each metric over CONUS and plot them as confidence bounds in Figure 3.5. Figure 3.5 presents BCSD’s loss of predictive capability when compared to DeepSD. We find that BCSD overestimates extremes at upper quantiles while DeepSD is relatively stable. Though RMSE, Corr, and Skill becomes worse at these extremes, DeepSD consistently outperforms BCSD, most often with thinner confidence bounds. These results show DeepSD’s ability to perform well for increasingly extreme precipitation events. DeepSD’s performance is impressive given that literature has shown that traditional techniques tend to fail when downscaling averages and extremes simultaneously. We hypothesize that capturing nearby spatial information allows DeepSD to isolate areas where extreme precipitation events are more likely than others.

3.6.3 Scalability on NASA’s NEX

Comprehensive studies of climate change requires much more than a single ESM simulation but rather multiple projections from different models, emission scenarios, and initial conditions in order to capture uncertainty. In total, CMIP5 contains more than 20 models at 4 emission scenarios (Representative Concentration Pathways (RCPs) 2.6,4.5,6.5,8.6), a variable number of initial conditions, and multiple climate variables at a daily temporal resolution. Generally, each prospective projection is available from 2006 to 2100 while retrospective projections are available from 1850 to 2005. Limiting the downscaled projections to encompass CONUS at $1/8^\circ \times 1/8^\circ$, a single simulation requires 134MB. Following the current timeframes of downscaled projections on NASA’s NEX platform, downscaling from the year 1950 to 2005 requires 7.4GB of storage while each prospective run needs 13GB. Hence, the final dataset size is 1.2TB. When downscaled further to PRISM’s

native resolution, $1/16^\circ \times 1/16^\circ$, the dataset size increases to approximately 5TB. Furthermore, the dataset scales linearly as more variables are added, including temperature minimum and maximum.

We test computational scalability by computing the amount of time taken to downscale 1 year of CONUS, presented in Table 3.1. For DeepSD, this includes the 3 feed-forward processes and their corresponding interpolations computed on a single GPU in NVIDIA DIGITS DevBox. Runtime for each of the ASD methods is estimated and scaled from the length of time to downscale the 1000 selected locations using 40 CPUs in parallel (Intel Xeon CPU E5-2680 2.8GHz). BCSD’s runtime is computed as the amount of time taken to quantile map, interpolate, and scale a years worth of data. We find that BCSD and DeepSD widely outperform the ASD approaches. While BCSD provides the quickest runtime, DeepSD is still scalable.

Though DeepSD is a highly scalable method, due to a single feed-forward neural network architecture, generating such large datasets still requires heavy computational power. However, storage and compute resources are satisfied by dedicated access to the Pleiades supercomputer housed in NASA’s Advanced Supercomputer Division (HECC) at NASA Ames. High resolution projections can be quickly computed using GPU’s, which are available on each node, in coordination with the Message Passing Interface (MPI). High-resolution projections are then stored on NEX’s filesystem which has currently 2.3PB of rapid-access storage in addition to large scale tape storage accessible on the HECC platform. In this paper we present a methodology for statistical downscaling, DeepSD, that leverages recent advances in image super-resolution and convolutional neural networks. DeepSD differs from previous SD methods by explicitly capturing spatial structure while improving scalability. A brief comparison with baseline SD techniques, BCSD and ASD, shows promising results in predictive capabilities when downscaling precipitation over the continental United States. Lastly, we describe how DeepSD can be scaled using NASA’s Earth Exchange platform to provide an ensemble of downscaled climate projections from more than 20 ESMs.

3.7 Discussion

Though DeepSD shows promise for SD, there are still some limitations in our experimentation regarding spatial and temporal generalization. An advantage of DeepSD is that a single trained model is able to downscale spatial heterogeneous regions. However, we do not test predictability in regions

where the model was not trained. Future work will examine this hypothesis to understand DeepSD’s credibility in regions with few observations. Second, we do not test temporal non-stationarity, a longstanding problem in statistical downscaling. Evaluation under non-stationarity can be tested using approaches presented by Salvi et al. [96], such that training and testing data is split between cold/warm years. As there is a single model for all locations, including cold and warm climates, we hypothesize that DeepSD is capable of capturing non-stationarity.

Furthermore, future work can improve multiple facets of DeepSD. For instance, the inclusion of more variables such as temperature, wind, and humidity at different pressure levels of the atmosphere may capture more climate patterns. Also, downscaling multiple climate variables simultaneously could be explored to find similar spatial patterns in the high-resolution datasets, such as high temperatures and increased precipitation. Most importantly, DeepSD fails to capture uncertainty around its downscaled projections, a key factor in adapting to climate change. Recent advances in Bayesian Deep Learning concepts [32] are presented in the following chapter to allow us to quantify uncertainty in deep networks.

Chapter 4

Quantifying Uncertainty in Discrete-Continuous and Skewed Data with Bayesian Deep Learning

Science and Engineering (S&E) applications are beginning to leverage the recent advancements in artificial intelligence through deep learning. In climate applications, deep learning is being used to make high-resolution climate projections [116] and detect tropical cyclones and atmospheric rivers [91]. Remote sensing models such as DeepSAT [7], a satellite image classification framework, also leverage computer vision technologies. Physicists are using deep learning for detecting particles in high energy physics [5] and in transportation deep learning has aided in traffic flow prediction [74] and modeling network congestion [75]. Scientists have even used convolutional neural networks to approximate the Navier-Stokes equations of unsteady fluid forces [85]. However, for many of these applications, the underlying data follow non-normal and discrete-continuous distributions. For example, when modeling precipitation, we see most days have no precipitation at all with heavily skewed amounts on the rainy days, as shown in Figure 4.1. Furthermore, climate is a complex nonlinear dynamical system, while precipitation processes in particular exhibit extreme space-time variability as well as thresholds and intermittence, thus precipitation data cannot be assumed to be Gaussian. Hence, for deep learning to be harnessed to its potential in S&E applications, our models must be resilient to non-normal and discrete-continuous distributions.

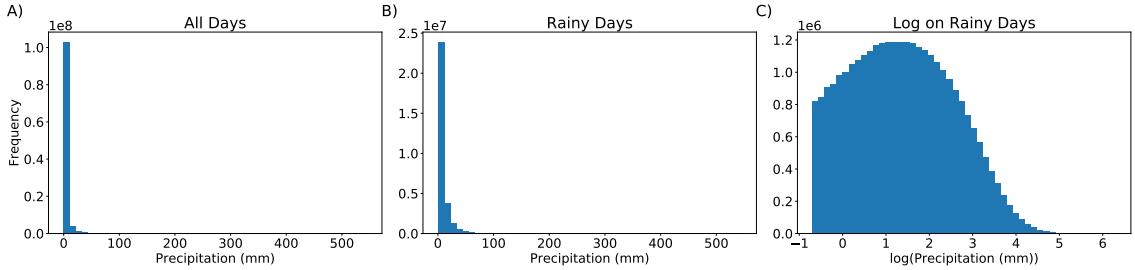


Figure 4.1: Histogram of daily precipitation on the Contiguous United States from 2006 to 2015. A) All precipitation data points. B) Precipitation distribution on rainy days only. C) Log distribution of precipitation on rainy days.

Uncertainty quantification is another requirement for wide adoption of deep learning in S&E, particularly for risk management decisions. Twenty years ago, Jaeger et al. stated, “uncertainties in climate change are so pervasive and far reaching that the tools for handling uncertainty provided by decision analysis are no longer sufficient [55].” As expected, uncertainty has been a particular interest of climate and computer scientists to inform social and infrastructure adaptation to increasing weather extremes and natural disasters [58, 73]. For example, Kay et al. studied six different sources of uncertainty of climate change impacts on a flood frequency model [59]. These uncertainties included future greenhouse gas scenarios, global climate models (GCMs) structure and parameters, downscaling GCMs, and hydrological model structure and parameters. Hence, quantifying the uncertainty from each of these processes is critical for understanding the system’s uncertainty. This provides us with the problem of quantifying uncertainty in discrete-continuous and non-normal distributions.

Recent work in Bayesian Deep Learning (BDL) provides a foundation for modeling uncertainty in deep networks which may be applicable to many S&E applications [32, 34, 61, 134]. The simplicity of implementing BDL on an already defined deep neural network makes it an attractive approach. With a well-defined likelihood function, BDL is able to capture both aleatoric and epistemic uncertainty [61]. *Epistemic* uncertainty comes from noise in the model’s parameters which can be reduced by increasing the dataset size. On the other side, *Aleatoric* uncertainty accounts for the noise in the observed data, resulting in uncertainty which cannot be reduced. Examples of aleatoric uncertainty are measurement error and sensor malfunctions. Aleatoric uncertainty can either be homoscedastic, constant uncertainty for different inputs, or heteroscedastic, uncertainty depending on the input. Heteroscedastic is especially important in skewed distributions, where the

tails often contain orders of magnitude increased variability. Variants of these methods have already been successfully applied to applications such as scene understanding [60] and medical image segmentation [119].

While BDL has been applied to few domains, these models generally assume a Gaussian probability distribution on the prediction. However, as we discussed in S&E applications, such an assumption may fail to hold. This motivates us to extend BDL further to aperiodic non-normal distributions by defining alternative density functions based on domain understanding. In particular, we focus on a precipitation estimation problem called statistical downscaling, which we will discuss in Section 2. In section 3, we review “DeepSD”, our statistical downscaling method [116], and Bayesian Deep Learning Concepts. In section 4, we present two BDL discrete-continuous (DC) likelihood models, using Gaussian and lognormal distributions, to model categorical and continuous data. Following in Section 5, we compare predictive accuracy and uncertainty calibration in statistical downscaling. Lastly, Section 6 summarizes results and discusses future research directions.

4.0.1 Key Contributions

1. A discrete-continuous bayesian deep learning model is presented for uncertainty quantification in science and engineering.
2. We show that a discrete-continuous model with a lognormal likelihood can model fat-tailed skewed distributions, which occur often in science and engineering applications.
3. The first model to capture heteroscedastic, and epistemic, uncertainties in statistical downscaling is presented.

4.1 Precipitation Estimation

4.1.1 Statistical Downscaling

Downscaling, either statistical or dynamical, is a widely used process for producing high-resolution projections from coarse global climate models (GCMs) [47, 29, 81]. Dynamical downscaling, often

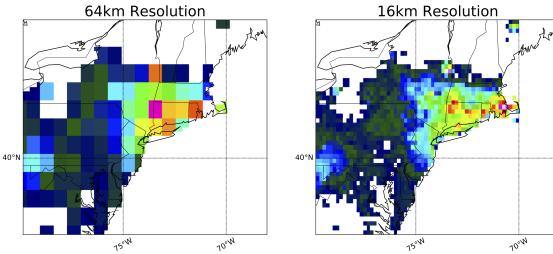


Figure 4.2: Prism Observed Precipitation: Left) Low resolution at 64km. Right) High resolution at 16km.

referred to as regional climate models, are physics based numerical models encoding localized sub-grid processes within GCM boundary conditions to generate high-resolution projections. Similar to GCMs, dynamical downscaling are computational expensive and simply cannot scale to ensemble modeling. Statistical downscaling is a relatively efficient solution which aims to use observed data to learn a functional mapping between low- and high-resolution GCMs, illustrated in Figure 4.2. Uncertainty in GCMs is exacerbated by both observational data and parameters in the functional mapping, motivating a probabilistic approach.

GCMs through the Fifth Coupled Model Intercomparison Project (CMIP5) provides scientist with valuable data to study the effects of climate change under varying greenhouse gas emission scenarios [106]. GCMs are complex non-linear dynamical systems that model physical processes governing the atmosphere up to the year 2200 (some to 2300). GCMs are gridded datasets with spatial resolutions around 100km and contain a range of variables including temperature, precipitation, wind, and pressure at multiple pressure levels above the earth’s surface. More than 20 research groups around the world contributed to CMIP5 by developing their own models and encoding their understanding of the climate system. Within CMIP5, each GCM is simulated under three or four emission scenarios and multiple initial conditions. This suite of climate model simulations are then used to get probabilistic forecasts of variables of interest, such as precipitation and temperature extremes [101]. While the suite of models gives us the tools to study large scale climate trends, localized projections are required for adaptation.

Many statistical models have been explored for downscaling, from bias correction spatial disaggregation (BCSD) [16] and automated statistical downscaling (ASD) [46] to neural networks [105] and nearest neighbor models [48]. Multiple studies have compared different sets of statistical down-

scaling approaches on various climate variables and varying temporal and spatial scales showing that no approach consistently outperforms the others [12, 41, 115]. Recently, Vandal et al. presented improved results with an alternative approach to downscaling by representing the data as "images" and adapting a deep learning based super-resolution model called DeepSD [116]. DeepSD showed superior performance in downscaling daily precipitation in the contiguous United States (CONUS) when compared to ASD and BCSD.

Even though uncertainty is crucial in statistical downscaling, it is rarely considered in downscaling studies. For instance, all the downscaled climate projections used in the latest US National Climate Assessment report (CSSR), produced on the NASA Earth Exchange, come with no uncertainty estimates. Though widely used in climate impact assessments, a recurrent complaint from the users is a lack of uncertainty characterization in these projections. What users often request are estimates of geographic and seasonal uncertainties such that the adaptation decisions can be made with robust knowledge [126]. Khan et al. presented one study that assessed monthly uncertainty from confidence based intervals of daily predictions [64]. However, this approach only quantifies epistemic uncertainty and therefore cannot estimate a full probability distribution. To the best of the authors' knowledge, no studies have modeled aleatoric (heteroscedastic) uncertainty in statistical downscaling, presenting a limitation to adaptation.

4.1.2 Climate Data

A wide variety of data sources exists for studying the earth's climate, from satellite and observations to climate models. Above we discussed some of the complexities and uncertainty associated with ensembles of GCMs as well as their corresponding storage and computational requirements. While the end goal is to statistically downscale GCMs, we must first learn a statistical function to apply a low- to high-resolution mapping. Fortunately, one can use observed datasets that are widely available and directly transfer the trained model to GCMs. Such observation datasets stem from gauges, satellite imagery, and radar systems. In downscaling, one typically will use either in-situ gauge estimates or a gridded data product. As we wish to obtain a complete high-resolution GCM, a gridded data product is required. Such gridded-data products are generally referred to as reanalysis datasets, which use a combination of data sources with physical characteristics aggregated to a well estimated data source. For simplicity, the remainder of this paper we will refer to reanalysis datasets

as observations.

In SD, it is important for our dataset to have high spatial resolution at a daily time temporal scale spanning as many years as possible. Given these constraints, we choose to use precipitation from the Prism dataset made available by Oregon State University with a 4km spatial resolution at a daily temporal scale [22]. The underlying data in Prism is estimated from a combination of gauges measuring many climate variables and topographical information. To train our model, the data is upscaled from 4km to the desired low-resolution. For example, to train a neural network to downscale from 64km to 16km, we upscale Prism to 16km and 64km and learn the mapping between the two (see Figure 4.2).

For the reader, it may be useful to think about this dataset as an image where precipitation is a channel analogous to traditional RGB channels. Similarly, more variables can be added to our dataset which therefore increases the number of channels. However, it is important to be aware that the underlying spatio-temporal dynamics in the chaotic climate system makes this dataset more complex than images. In our experiments with DeepSD, we included an elevation from the Global 30 Arc-Second Elevation Data Set (GTOPO30) provided by the USGS.

4.2 Background

4.2.1 DeepSD

The statistical downscaling approach taken by DeepSD differs from more traditional approaches, which generally do not capture spatial dependencies in both the input and output. For example Automated Statistical Downscaling (ASD) [46] learns regression models from low-resolution to each high-resolution point independently, failing to preserve spatial dependencies in the output and requiring substantial computational resources to learn thousands of regression models. In contrast, DeepSD represents the data as low- and high-resolution image pairs and adapts super-resolution convolutional neural networks (SRCNN) [26] by including high-resolution auxiliary variables, such as elevation, to correct for biases. These auxiliary variables allows one to use a single trained neural network within the training domain. This super-resolution problem is essentially a pixel-wise regression such that $\mathbf{Y} = F(\mathbf{X}; \Theta)$ where \mathbf{Y} is high-resolution with input $\mathbf{X} = [\mathbf{X}_{lr}, \mathbf{X}_{aux}]$ and F a convolutional neural network parameterized by Θ . F can then be learned by optimizing the

loss function:

$$\mathcal{L} = \frac{1}{2N} \sum_{i \in S} \|F(\mathbf{X}_i; \Theta) - \mathbf{Y}_i\|_2^2 \quad (4.1)$$

where S is a subset n examples. Based on recent state-of-the-art results in super-resolution [65, 72], we modify the SRCNN architecture to include a residual connection between the precipitation input channel and output layer, as shown in Figure 4.3.

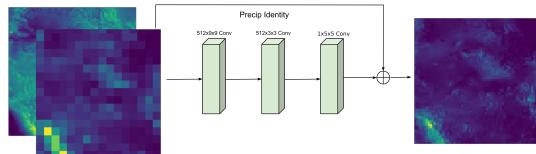


Figure 4.3: Residual SRCNN Architecture used for DeepSD with a skip connection between precipitation and the output layer.

As discussed above, the resolution enhancement of 8x or more needed in statistical downscaling is much greater than the 2-4x enhancements used for images. DeepSD uses stacked SRCNNs, each improving resolution by 2x allowing the model to capture regional and local weather patterns, depending on the level. For instance, to downscale from 100km to 12.5km, DeepSD first trains models independently (or with transfer learning) to downscale from 100km to 50km, 50km to 25km, and 25km to 12.5km. During inference, these models are simply stacked on each other where the output of one plus the next corresponding auxiliary variables are inputs to the next. In the case of downscaling precipitation, inputs may include LR precipitation and HR elevation to predict HR precipitation. In this work, we focus on uncertainty quantification for a single stacked network which can then be translated to stacking multiple Bayesian neural networks.

4.2.2 Bayesian Deep Learning

In the early 1990's Mackay [76] introduced a Bayesian neural networks (BNNs) by replacing deterministic weights with distributions. However, as is common with many Bayesian modeling problems, direct inference on BNNs is intractable for networks of more than a one or two hidden layers. Many studies have attempted to reduce the computational requirements using various approximations [40, 6, 53]. Most recently, Gal and Ghahramani presented a practical variational approach to

approximate the posterior distribution in deep neural networks using dropout and monte carlo sampling [33, 32]. Kendall and Gal then followed this work for computer vision applications to include both aleatoric and epistemic uncertainties in a single model [61].

To begin, we define weights of our neural network as $\omega = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L\}$ such that $\mathbf{W} \sim \mathcal{N}(0, I)$ and L being the number of layers in our network. Given random outputs of a BNN denoted by $f^\omega(\mathbf{x})$, the likelihood can be written as $p(\mathbf{y}|f^\omega(\mathbf{x}))$. Then, given data \mathbf{X} and \mathbf{Y} , as defined above, we infer the posterior $p(\omega|\mathbf{X}, \mathbf{Y})$ to find a distribution of parameters that best describe the data. For a regression task assuming a predictive Gaussian posterior, $p(\mathbf{y}|f^\omega(\mathbf{x})) = \mathcal{N}(\hat{\mathbf{y}}, \hat{\sigma}^2)$ with random outputs:

$$[\hat{\mathbf{y}}, \hat{\sigma}^2] = f^\omega(\mathbf{x}).$$

Applying variational inference to the weights, we can define an approximate and tractable distribution $q_\Theta(\omega) = \prod_{l=1}^L q_{\mathbf{M}_l}(\mathbf{W}_l)$ where $q_{\mathbf{M}_l}(\mathbf{W}_l) = \mathbf{M}_l \times \text{diag}[\text{Bernoulli}(1-p_l)^{K_l}]$ parameterized by $\Theta_l = \{\mathbf{M}_l, p_l\}$ containing the weight mean of shape $K_l \times K_{l+1}$, K_l being the number of hidden units in layer l , and dropout probability p_l . Following, we aim to minimize the Kullback-Leibler (KL) divergence between $q_\Theta(\omega)$ to the true posterior, $p(\omega|\mathbf{X}, \mathbf{Y})$. The optimization objective of the variational interpretation can be written as [33]:

$$\hat{\mathcal{L}}(\Theta) = -\frac{1}{M} \sum_{i \in S} \log p(\mathbf{y}_i|f^\omega(\mathbf{x}_i)) + \frac{1}{N} \text{KL}(q_\Theta(\omega)||p(\omega)) \quad (4.2)$$

$$= \hat{\mathcal{L}}_x(\Theta) + \frac{1}{N} \text{KL}(q_\Theta(\omega)||p(\omega)) \quad (4.3)$$

where S is a set of M data points. To obtain well calibrated uncertainty estimates, it is crucial to select a well estimated p_l . Rather than setting p_l to be constant, we can learn it using a concrete distribution prior which gives us a continuous approximation of the Bernoulli distribution [34]. As presented by Gal et al., the KL divergence term is then written as:

$$\text{KL}(q_{\Theta}(\omega) \parallel p(\omega)) = \sum_{l=1}^L \text{KL}(q_{\mathbf{M}_l}(\mathbf{W}_l) \parallel p(\mathbf{W}_l)) \quad (4.4)$$

$$\text{KL}(q_{\mathbf{M}_l}(\mathbf{W}) \parallel p(\mathbf{W})) \propto \frac{l^2(1-p_l)}{2} \|\mathbf{M}_l\| - K_l \mathcal{H}(p_l) \quad (4.5)$$

where

$$\mathcal{H}(p) = -p \log p - (1-p) \log (1-p) \quad (4.6)$$

is the entropy of a Bernoulli random variable with probability p . We note that given this entropy term, the learning dropout probability cannot exceed 0.5, a desired effect. For brevity, we encourage the reader to refer to [34] for the concrete dropout optimization. In the remainder of this paper, we will use this concrete dropout formulation within all presented models.

4.3 Bayesian Deep Learning for Skewed Distributions

In this section we describe three candidate Bayesian deep learning models to quantify uncertainty in super-resolution based downscaling. We begin by formalizing the use of BDL within the SRCNN architecture assuming a normal predictive distribution, identical to the pixel-wise depth regression in [61]. This approach is further extended to a discrete-continuous model that conditions the amount of precipitation given an occurrence of precipitation. This leverages the domain knowledge that the vast majority of data samples are non-rainy days which are easy to predict and contain little information for the regression. Such a technique was used by Sloughter et al. using a discrete-continuous gamma distribution [102]. Lastly, we show that a lognormal distribution can be applied directly in BDL and derive its corresponding log-likelihood loss and unbiased parameter estimates.

4.3.1 Gaussian Likelihood

Super-resolution is an ill-posed pixel-wise regression problem such that BDL can be directly applied, as Kendall and Gal showed for predicting depth in computer vision [61]. As discussed in previous sections, it is crucial to capture both aleatoric and epistemic uncertainties in downscaling. As shown in section 3.1 of [61], we must measure the aleatoric uncertainty by estimating the vari-

ance, σ^2 , in the predictive posterior while also sampling weights via dropout from the approximate posterior, $\widehat{\mathbf{W}} \sim q_{\Theta}(\mathbf{W})$. As before, we defined our Bayesian convolutional neural network \mathbf{f} :

$$[\hat{\mathbf{y}}, \hat{\sigma}^2] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}). \quad (4.7)$$

and make the assumption that $\mathbf{Y} \sim \mathcal{N}(\hat{\mathbf{y}}, \hat{\sigma}^2)$. The Gaussian log-likelihood can be written as:

$$\mathcal{L}_x(\Theta) = \frac{1}{2D} \sum_i \hat{\sigma}_i^{-2} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 + \frac{1}{2} \log \hat{\sigma}_i^2 \quad (4.8)$$

where pixel i in \mathbf{y} corresponds to input \mathbf{x} and D being the number of output pixels. The KL term is identical to that in Equation 4.4. Given this formulation, $\hat{\sigma}_i$, the variance for pixel i is implicitly learned from the data without the need for uncertainty labels. We also note that during training the substitution $s_i := \log \hat{\sigma}_i^2$ is used for stable learning using the Adam Optimization algorithm [67], a first-order gradient based optimization of stochastic objective functions.

Unbiased estimates of the first two moments can the be obtained with T Monte Carlo samples, $\{\hat{\mathbf{y}}_t, \hat{\sigma}_i^2\}$, from $\mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{x})$ with masked weights $\widehat{\mathbf{W}}_t \sim q(\mathbf{W})$:

$$\mathbb{E}[\mathbf{Y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{\mathbf{y}}_t \quad (4.9)$$

$$\text{Var}[\mathbf{Y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{\mu}_t^2 - \frac{1}{T} \sum_{t=1}^T \hat{\sigma}_t^2 + \left(\frac{1}{T} \sum_{t=1}^T \hat{\mu}_t \right)^2. \quad (4.10)$$

These first two moments provide all the necessary information to easily obtain prediction intervals with both aleatoric and epistemic uncertainties. For further details, we encourage the reader to refer to [61].

4.3.2 Discrete-Continuous Gaussian Likelihood

Rather than assuming a simple Gaussian distribution for all output variables, which may be heavily biased from many non-rainy days in our dataset, we can condition the model to predict whether rain occurred or not. The BNN is now formulated such that the mean, variance, and probability of precipitation are sampled respectively from \mathbf{f} :

$$[\hat{\mathbf{y}}, \hat{\sigma}^2, \hat{\phi}] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}) \quad (4.11)$$

$$\hat{p} = \text{Sigmoid}(\hat{\phi}). \quad (4.12)$$

Splitting the distribution into discrete and continuous parts gives us:

$$p(\mathbf{y}|f^\omega(\mathbf{x})) = \begin{cases} (1 - \hat{p}) & \mathbf{y} = 0 \\ \hat{p} \cdot \mathcal{N}(\mathbf{y}; \hat{\mathbf{y}}, \hat{\sigma}^2) & \mathbf{y} > 0 \end{cases} \quad (4.13)$$

Plugging this in to 4.2 and dropping the constants gives us the loss function (for brevity, we ignore the KL term which is identical to Equation 4.4):

$$\begin{aligned} \mathcal{L}_x(\Theta) &= -\frac{1}{D} \sum_i \log \left(\mathbb{1}_{\mathbf{y}_i > 0} \cdot \hat{p}_i \cdot \mathcal{N}(\mathbf{y}_i; \hat{\mathbf{y}}_i, \hat{\sigma}_i^2) + \mathbb{1}_{\mathbf{y}_i = 0} \cdot (1 - \hat{p}_i) \right) \\ &= -\frac{1}{D} \sum_{i, \mathbf{y}_i > 0} \left(\log \hat{p}_i + \log \mathcal{N}(\mathbf{y}_i; \hat{\mathbf{y}}_i, \hat{\sigma}_i^2) \right) \\ &\quad - \frac{1}{D} \sum_{i, \mathbf{y}_i = 0} \log(1 - \hat{p}_i) \\ &= \frac{1}{D} \sum_i \left(\mathbb{1}_{\mathbf{y}_i > 0} \cdot \hat{p}_i + (1 - \mathbb{1}_{\mathbf{y}_i > 0}) \cdot (1 - \hat{p}_i) \right) \\ &\quad - \frac{1}{2D} \sum_{i, \mathbf{y}_i > 0} \hat{\sigma}_i^{-2} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 + \log \sigma_i^2 \end{aligned} \quad (4.14)$$

where the first term is the cross entropy of a rainy day and the second term is the conditional Gaussian loss. Furthermore, we can write the unbiased estimates of the first two moments as:

$$\mathbb{E}[\mathbf{Y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{p}_t \hat{\mathbf{y}}_t \quad (4.15)$$

$$\text{Var}[\mathbf{Y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{p}_t^2 (\hat{\mathbf{y}}_t^2 + \hat{\sigma}_t^2) - \left(\frac{1}{T} \sum_{t=1}^T \hat{p}_t \hat{\mu}_t \right)^2. \quad (4.16)$$

4.3.3 Discrete-Continuous Lognormal Likelihood

Precipitation events, especially extremes, are known to follow fat-tailed distributions, such as log-normal and Gamma distributions [102, 20]. For this reason, as above, we aim to model precipitation using a discrete-continuous lognormal distribution. It should be noted that the lognormal distribution is undefined at 0 so a conditional is required for downscaling precipitation. To do this, we slightly modify our BNN:

$$[\hat{\mu}, \hat{\sigma}^2, \hat{\phi}] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}) \quad (4.17)$$

$$\hat{p} = \text{Sigmoid}(\hat{\phi}). \quad (4.18)$$

where $\hat{\mu}$ and $\hat{\sigma}$ are sampled parameters of the lognormal distribution. Following the same steps as above, we can define a piece-wise probability density function:

$$p(\mathbf{y}|f^\omega(\mathbf{x})) = \begin{cases} (1 - \hat{p}) & \mathbf{y} = 0 \\ \hat{p} \cdot \frac{1}{\mathbf{y}\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(\log(\mathbf{y}) - \hat{\mu})^2}{2\hat{\sigma}^2}\right) & \mathbf{y} > 0 \end{cases} \quad (4.19)$$

This gives us the modified log-likelihood objective:

$$\begin{aligned} \mathcal{L}_x(\Theta) = & \frac{1}{D} \sum_i (\mathbb{1}_{\mathbf{y}_i > 0} \cdot \hat{p}_i + (1 - \mathbb{1}_{\mathbf{y}_i > 0}) \cdot (1 - \hat{p}_i)) \\ & - \frac{1}{2D} \sum_{i,\mathbf{y}_i > 0} \hat{\sigma}_i^{-2} \|\log \mathbf{y}_i - \hat{\mu}_i\|^2 + \log \sigma_i^2 \end{aligned} \quad (4.20)$$

In practice, we optimize $\hat{s} := \exp(\hat{\sigma})$ for numerical stability. And lastly, the first two moments are derived as:

$$\mathbb{E}[\mathbf{y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{p}_t \exp(\hat{\mu} + \frac{1}{2}\hat{\sigma}^2) \quad (4.21)$$

$$\text{Var}[\mathbf{Y}] \approx \frac{1}{T} \sum_{t=1}^T \hat{p}_t^2 \exp(2\hat{\mu} + 2\hat{\sigma}^2) \quad (4.22)$$

Given these first two moments, we can derive unbiased estimates of μ and σ :

$$\hat{\sigma} = \log\left(1 + \frac{1}{2}\sqrt{\frac{4\text{Var}[\mathbf{Y}]}{\mathbf{E}[\mathbf{y}]^2} + 1}\right) \quad (4.23)$$

$$\hat{\mu} = \mathbf{E}[\mathbf{y}] - \frac{\hat{\sigma}^2}{2} \quad (4.24)$$

that can be used to compute pixel-wise probabilistic estimates. In the next section, we will apply each of the three methods to downscaling precipitation, compare their accuracies, and study their uncertainties.

4.4 Precipitation Downscaling

For our experimentation, we define our problem to downscale precipitation from 64km to 16km, a 4x resolution enhancement in a single SRCNN network. We begin with precipitation from the PRISM dataset, as presented in Section 4.1.2, at 4km which is then upscaled to 16km using bilinear interpolation. This 16km dataset are our labels and are further upscaled to 64km, generating training inputs. Furthermore, we use elevation from the Global 30 Arc-Second Elevation Dataset (GTOPO30) provided by the USGS as an auxiliary variable, also upscaled to 16km. In the end, our dataset is made up of precipitation at 64km and elevation at 16km as inputs where precipitation at 16km are the labels. In the discrete-continuous models, precipitation $> 0.5\text{mm}$ is considered a rainy day. Precipitation measured in millimeters (mm) is scaled by 1/100 for training when optimizing the Gaussian models. Elevation is normalized with the overall mean and variance. The training data is taken from years 1980 to 2005 and the test set from 2006 to 2015. Sub-images selected of size 64x64 with stride 48 are used for generating training examples.

Our super-resolution architecture is defined with two hidden layers of 512 kernels using kernel sizes 9, 3, and 5 (see Figure 4.3). The model is trained for 3×10^6 iterations using a learning rate of 10^{-4} and a batch size of 10. Three models are optimized using each of the three log-likelihood loss's defined above, Gaussian distribution as well as discrete-continuous Gaussian and lognormal distributions conditioned on a rainy day. 50 Monte Carlo passes during inference are used to measure the first two moments which then estimates the given predictive distribution's parameters.

Concrete dropout is used to optimize the dropout probability with parameters $\tau=1e-5$ and prior

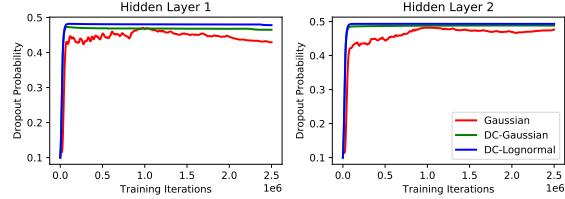


Figure 4.4: Dropout probabilities learned using Concrete Dropout for both hidden layers.

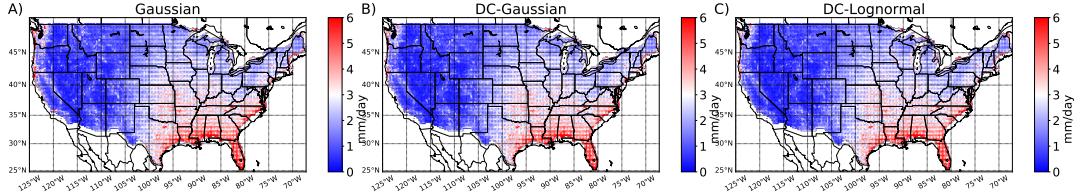


Figure 4.5: Daily Root Mean Square Error (RMSE) computed at each location for years 2006 to 2015 (test set) in CONUS. A) Gaussian, B) Conditional-Gaussian, and C) Conditional-Lognormal. Red corresponds to high RMSE while blue corresponds to low RMSE.

length scale as $l = 1$ to improve uncertainty calibration performance [34]. For a pixel-wise regression the number of samples N is set as Days \times Height \times Width. These parameters were found to provide a good trade-off between likelihood and regularization loss terms. As shown in Figure 4.4, dropout rates for each model and hidden layer are close to 0.5, the largest possible dropout rate. We find that the Gaussian distribution has difficulty converging to a dropout rate while the discrete-continuous models quickly stabilize. Furthermore, the lognormal distribution learns the largest dropout rate, suggesting a less complex model.

Validation is an important task for choosing a highly predictive and well calibrated downscaling model. In our experiments, we study each model’s ability to predict daily precipitation, calibration of uncertainty, and width of uncertainty intervals. For reproducibility, we provide the codes for training and testing on github (<https://github.com/tjvandal/discrete-continuous-bdl>).

4.4.1 Predictive Ability

We begin by comparing each model’s ability to predict the ground truth observations. Root Mean Square Error (RMSE) and bias are compared to understand the average daily effects of downscaling. To analyze extremes, we select two metrics from Climdex (<http://www.clim-dex.org>)

	Bias	RMSE	R20 Error	SDII Error
Gaussian	-0.11 ± 0.34	2.14 ± 1.31	-0.73 ± 1.94	-0.83 ± 0.93
DC-Gaussian	-0.11 ± 0.30	2.07 ± 1.28	-0.61 ± 1.67	-0.21 ± 0.78
DC-Lognormal	-0.02 ± 0.30	2.05 ± 1.27	-0.36 ± 1.63	-0.28 ± 0.81

Table 4.1: Predictive accuracy statistics computed pixel-wise and aggregated. Daily intensity index (SDII) and yearly precipitation events greater than 20mm (R20) measure each model’s ability to capture precipitation extremes. R20-Err and SDII-Err measures the difference between observed indicies and predicted indicies (closer to 0 is better).

which provides a suite of extreme precipitation indices and is often used for evaluating downscaling models [13, 115]:

1. R20 - Very heavy wet days $\geq 20\text{mm}$
2. SDII - Daily intensity index = (Annual total) / (precip days $\geq 0.5\text{ mm}$).

In our analysis, we compute each index for the test set as well as observations. Then the difference between the predicted indices and observed indices are computed, ie. $(\text{SDII}_{\text{model}} - \text{SDII}_{\text{obs}})$. These results can be seen in Table 4.1. We see a clear trend of the DC models performing better than a regular Gaussian distribution on all computed metrics. In particular, DC-Lognormal shows the lowest Bias, RMSE, and R20 error while DC-Gaussian has slightly higher errors but performs marginally better at estimating the SDII index. Furthermore, we study the predictability over space in Figure 4.5 by computing the pixel-wise RMSEs. Each model performs well in the mid-west and worse in the southeast, a region with large numbers of convective precipitation events.

We see that the DC models, DC-Lognormal in particular, have lower bias than a regular Gaussian distribution. Similarly for RMSE, DC models, lead by a DC-Gaussian, have the lowest errors. Looking more closely, we see improved performance along the coasts which are generally challenging to estimate. The convolutional operation with a 5×5 kernel in the last layer reconstructs the image using a linear combination of nearby points acting as a smoothing operation. However, when this is applied to the conditional distributions, the gradient along this edge can be increased by predicting high and low probabilities of precipitation in a close neighborhood. This insight is particularly important when applied to coastal cities.

Lastly, we look at each conditional model’s ability to classify precipitous days with precision recall curves (Figure 4.6). We see that recall does not begin to decrease until a precision of 0.8 which

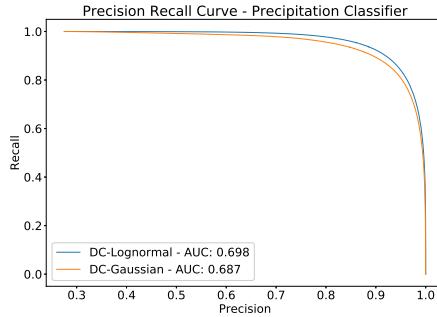


Figure 4.6: Precision recall curve of classifying rainy days in conditional models.

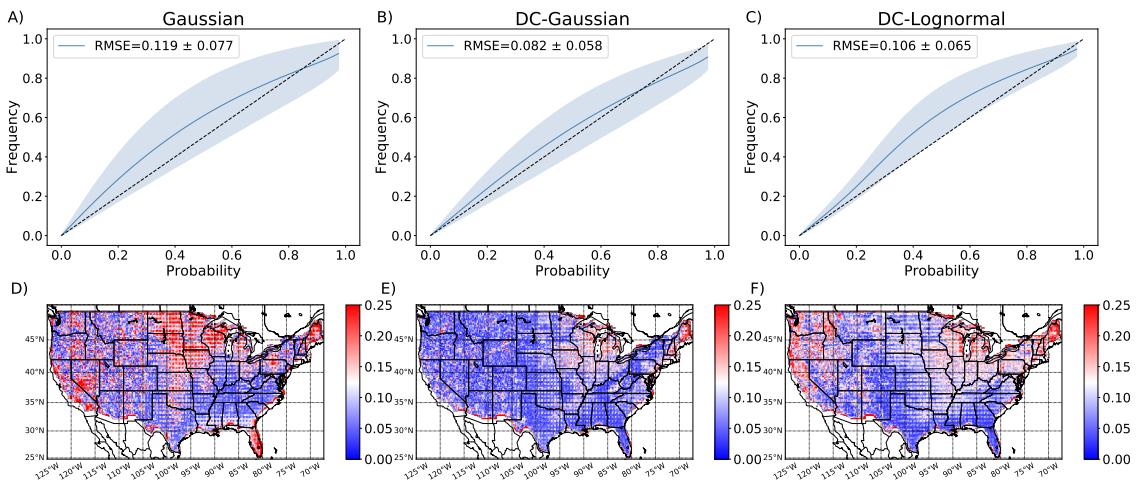


Figure 4.7: Calibration is computed as the frequency of predictions within a given probability range. This probability is varied on the x-axis with the corresponding frequency on the y-axis. Columns represent each model Gaussian, DC-Gaussian and Lognormal. Calibration plots on the first row compute per pixel with the shaded area representing the 80% confidence interval of calibration. The second row depicts calibration root mean square error (RMSE) per location.

indicates very strong classification performance. It was assumed that classification of precipitation would be easy for such a dataset.

4.4.2 Uncertainty Quantification

The remainder of our analysis focuses on each model's performance in estimating well calibrated uncertainty quantification. We limit our analysis of uncertainty to only days with precipitation ($\geq 0.5\text{mm}$) as uncertainty on non-rainy days is not of interest. The calibration metric used computes

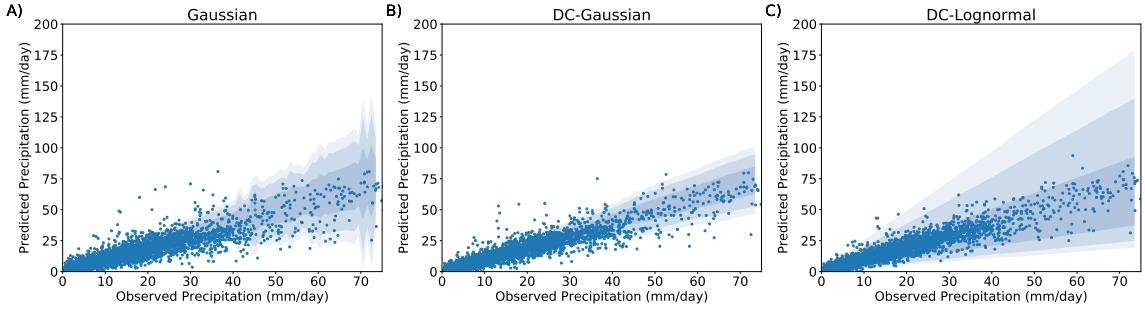


Figure 4.8: Uncertainty widths based on quantiles from their predictive distributions. The points are observations versus the expected value. The bands correspond to 50%, 80%, and 90% predictive intervals.

the frequency of observations occurring within a varying predicted probability range:

$$c(z) = \frac{1}{N} \sum_{i=1}^N \mathcal{I}_{P(y_i|f^\omega(x_i)) > (0.5 - z/2)} * \mathcal{I}_{P(y_i|f^\omega(x_i)) < (0.5 + z/2)} \quad (4.25)$$

where P is the cumulative density function of the predictive posterior and $z \in [0, 1]$ defined the predictive probability range centered at 0.5. Ideally the frequency of observations will be equal to the probability. A calibration error can then be defined as:

$$\text{RMSE}_{\text{cal}} = \sqrt{\frac{1}{K} \sum_{i=1}^K (c(i/K) - i/K)^2} \quad (4.26)$$

where K is the number bins. In our analysis, we use $K = 100$. The calibration plots for each model can be seen in Figure 4.7.

Right away we see from Figure 4.7 that the Gaussian distribution over-estimates uncertainty for most of the range with a wider range of variability between pixels. DC-Lognormal also over-estimates uncertainty but has a lower range of variability between pixels, showing more consistent performance from location to location. Overall, DC-Gaussian shows the lowest calibration error hovering right around $x = y$ but underestimates uncertainty at the tails. Though DC-Lognormal is better calibrated at the tails, one could calibrate the tails by simply forcing the variance to explode. Taking this a step further, we present calibration RMSEs per pixel in Figure 4.7 (bottom row) to visualize spatial patterns of UQ. In the Gaussian model we find weakened and more variable results at high-elevations in the west and mid-west. Each of the DC models perform well, but DC-Lognormal

also has areas of increased error in the west.

In Figure 4.8 we aim to better understand these uncertainties for increasingly intense precipitation days. At these high rainfall days our models generally under-predict precipitation, but the Gaussian models often fail to capture these extremes. While the lognormal has wider uncertainty intervals, it is able to produce a well calibrated distribution at the extremes. Furthermore, these wide intervals indicate that the model becomes less confident with decreasing domain coverage at higher intensities. This may suggest that there exists a bias-variance trade-off between the Gaussian and Log-Normal distributions.

4.5 Discussion

In this chapter we present Bayesian Deep Learning approaches incorporating discrete-continuous and skewed distributions targeted at S&E applications. The discrete-continuous models contain both a classifier to categorize an event and conditional regressor given an event’s occurrence. We derive loss functions and moments for Gaussian and lognormal DC regression models. Using precipitation as an example, we condition our model on precipitous days and predict daily precipitation on a high-resolution grid. Using the lognormal distribution, we are able to produce well-calibrated uncertainties for skewed fat-tailed distributions. To our knowledge, this is the first model for uncertainty quantification in statistical downscaling.

Through experiments, we find that this DC approach increases predictive power and uncertainty quantification performance, reducing errors with well calibrated intervals. In addition, we find that this conditional approach improves performance at the extremes, measured by daily intensity index and number of extreme precipitation days from ClimDex. Visually, we found that the DC models perform better than a regular Gaussian on the coasts, a challenge in statistical downscaling. These edge errors appear during reconstruction when the kernel partially overlaps with the coastal edge, acting as a smoothing operation. However, the DC models reduce this smoothing by increasing the expected value’s gradients.

Overall, we find that the DC distribution approaches provides strong benefits to deep super-resolution based statistical downscaling. Furthermore, while the lognormal distribution uncertainty was slightly less calibrated, it was able to produce well understood uncertainties at the extremes.

This presents a strong point, Bayesian Deep Neural Networks can well fit non-normal distributions when motivated by domain knowledge.

In the future we aim to extend this work to stacked super-resolution networks, as used in DeepSD [116], which requires sampling of between networks. Some other extensions could be the addition of more variables, extension to other skewed distributions, and larger network architectures. Finally, incorporating these theoretical advances in uncertainty characterization, the NEX team plans to use DeepSD to produce and distribute next generation of climate projections for the upcoming congressionally mandated national climate assessment.

Chapter 5

Conclusions and Future Work

In this thesis we present and explore a fundamentally new approach to the problem of generating high-resolution climate projections from low-resolution earth system models (ESMs) by incorporating recent advances in machine learning. In Chapter 1 we introduced this problem of statistical downscaling (SD) and discussed the weaknesses of current approaches along with motivating the use of highly scalable machine learning approaches [113]. Chapter 2 continues by exploring the applicability of off-the-shelf machine learning methods compared to state-of-the-art SD techniques [113]. Following, in Chapter 3, a deep learning based super-resolution approach to statistical downscaling is presented, DeepSD, and discusses improvements in predictability and scalability [116]. Lastly, Chapter 4 develops a discrete-continuous DeepSD approach to quantify uncertainty using Bayesian Deep Learning [114].

While methods and results presented in Chapters 3 and 4 introduce new tools for various problems in climate science and remote sensing, many limitations still exist. In particular, within the realm of statistical downscaling, the evaluation of DeepSD on climate models projections is required for wide adoption. For instance, our results show impressive results in both in predictability and uncertainty quantification on observed datasets, the question still remains whether such a method can improve beyond competing SD methods and coarse resolution GCMs. This question largely remains throughout the downscaling literature [97, 79, 92]. Non-stationarity in space and time is a major concern in SD as we cannot be certain that a trained statistical model on historical data can provide credible projections under climate change but few have suggested new approaches. The development of new statistical metrics and analysis tools are required to understand the added

value of SD to GCMs [96, 92]. Furthermore, a comparison of the developed SD methods to impact studies would be valuable.

The development of machine learning methods for climate science is also a crucial aspect to improving and using climate model projections. The super-resolution literature consists of a variety of approaches from very deep networks [65] to generate adverserial networks (GANs) [72]. GANs in particular could provide tremendous improvements to projecting realistic weather events such as tropical storms and other highly variable climate variables. Developments in Bayesian Deep Learning will continue to improve and present a greater number of tools for uncertainty quantification throughout climate science.

REFERENCES

- [1] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
- [2] 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.
- [3] Kazi Farzan Ahmed, Guiling Wang, John Silander, Adam M Wilson, Jenica M Allen, Radley Horton, and Richard Anyah. Statistical downscaling and bias correction of climate model outputs for climate change impact assessment in the us northeast. *Global and Planetary Change*, 100:320–332, 2013.
- [4] Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32, 2007.
- [5] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.
- [6] David Barber and Christopher M Bishop. Ensemble learning in bayesian neural networks. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 168:215–238, 1998.
- [7] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 37. ACM, 2015.

- [8] Saikat Basu, Manohar Karki, Sangram Ganguly, Robert DiBiano, Supratik Mukhopadhyay, and Ramakrishna Nemani. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. In *Proceedings of the European Symposium on Artificial Neural Networks, ESANN*, 2015.
- [9] Rasmus E Benestad, Deliang Chen, Abdelkader Mezghani, Lijun Fan, and Kajsa Parding. On using principal components to represent stations in empirical–statistical downscaling. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1):28326, 2015.
- [10] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [11] G Bürger. Expanded downscaling for generating local weather scenarios. *Climate Research*, 7(2):111–128, 1996.
- [12] G. Bürger, T. Q. Murdock, a. T. Werner, S. R. Sobie, and a. J. Cannon. Downscaling extremes—an intercomparison of multiple statistical methods for present climate. *Journal of Climate*, 25(12):4366–4388, June 2012.
- [13] G Bürger, TQ Murdock, AT Werner, SR Sobie, and AJ Cannon. Downscaling extreme-san intercomparison of multiple statistical methods for present climate. *Journal of Climate*, 25(12):4366–4388, 2012.
- [14] William R Burrows, Mario Benjamin, Stephen Beauchamp, Edward R Lord, Douglas McCollor, and Bruce Thomson. Cart decision-tree statistical analysis and prediction of summer season maximum surface ozone for the vancouver, montreal, and atlantic regions of canada. *Journal of applied meteorology*, 34(8):1848–1862, 1995.
- [15] Alex J Cannon. Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences*, 37(9):1277–1284, 2011.
- [16] Alex J Cannon and Paul H Whitfield. Downscaling recent streamflow conditions in british columbia, canada using ensemble neural network models. *Journal of Hydrology*, 259(1):136–151, 2002.

- [17] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.
- [18] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50. ACM, 2011.
- [19] Mingyue Chen, Wei Shi, Pingping Xie, Viviane Silva, Vernon E Kousky, R Wayne Higgins, and John E Janowiak. Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, 113(D4), 2008.
- [20] Hye-Kyung Cho, Kenneth P Bowman, and Gerald R North. A comparison of gamma and log-normal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11):1586–1597, 2004.
- [21] Paulin Coulibaly, Yonas B. Dibike, and François Anctil. Downscaling Precipitation and Temperature with Temporal Neural Networks. *Journal of Hydrometeorology*, 6(4):483–496, 2005.
- [22] Christopher Daly, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International journal of climatology*, 28(15):2031–2064, 2008.
- [23] Christopher Daly, Ronald P Neilson, and Donald L Phillips. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of applied meteorology*, 33(2):140–158, 1994.
- [24] Debasish Das, Auroop R. Ganguly, and Zoran Obradovic. A Bayesian sparse generalized linear model with an application to multi-scale covariate discovery for observed rainfall extremes over United States. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2014.

- [25] DP Dee, SM Uppala, AJ Simmons, Paul Berrisford, P Poli, S Kobayashi, U Andrae, MA Balmaseda, G Balsamo, P Bauer, et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, 2011.
- [26] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.
- [27] M Ekström, HJ Fowler, CG Kilsby, and PD Jones. New estimates of future changes in extreme rainfall across the uk using regional climate model integrations. 2. future estimates and use in impact studies. *Journal of Hydrology*, 300(1-4):234–251, 2005.
- [28] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.
- [29] Hayley J Fowler, Stephen Blenkinsop, and Claudia Tebaldi. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International journal of climatology*, 27(12):1547–1578, 2007.
- [30] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.
- [31] Howard Frumkin, Jeremy Hess, George Luber, Josephine Malilay, and Michael McGeehin. Climate change: the public health response. *American Journal of Public Health*, 98(3):435–445, 2008.
- [32] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, Ph. D. thesis, University of Cambridge, 2016.
- [33] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.
- [34] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *arXiv preprint arXiv:1705.07832*, 2017.

- [35] Poulomi Ganguli, Devashish Kumar, and Auroop R Ganguly. Water stress on us power production at decadal time horizons. *arXiv preprint arXiv:1511.08449*, 2015.
- [36] Subimal Ghosh. Svm-pgsl coupled approach for statistical downscaling to predict rainfall from gcm output. *Journal of Geophysical Research: Atmospheres*, 115(D22), 2010.
- [37] Subimal Ghosh and PP Mujumdar. Statistical downscaling of gcm simulations to streamflow using relevance vector machine. *Advances in water resources*, 31(1):132–146, 2008.
- [38] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356. IEEE, 2009.
- [39] Andre R Goncalves, Puja Das, Soumyadeep Chatterjee, Vidyashankar Sivakumar, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 451–460. ACM, 2014.
- [40] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.
- [41] Ethan Gutmann, Tom Pruitt, Martyn P Clark, Levi Brekke, Jeffrey R Arnold, David A Raff, and Roy M Rasmussen. An intercomparison of statistical downscaling methods used for water resource assessments in the united states. *Water Resources Research*, 50(9):7167–7186, 2014.
- [42] Andy Haines, R Sari Kovats, Diarmid Campbell-Lendrum, and Carlos Corvalán. Climate change and human health: impacts, vulnerability and public health. *Public health*, 120(7):585–596, 2006.
- [43] Dorra Hammami, Tae Sam Lee, Taha BMJ Ouarda, and Jonghyun Lee. Predictor selection for downscaling gcm data with lasso. *Journal of Geophysical Research: Atmospheres*, 117(D17), 2012.
- [44] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, SA Turubanova, Alexandra Tyukavina, David Thau, SV Stehman, SJ Goetz, TR Loveland, et al. High-

- resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853, 2013.
- [45] Muhammad Zia Hashmi, Asaad Y Shamseldin, and Bruce W Melville. Comparison of sdsm and lars-wg for simulation and downscaling of extreme precipitation events in a watershed. *Stochastic Environmental Research and Risk Assessment*, 25(4):475–484, 2011.
- [46] Masoud Hessami, Philippe Gachon, Taha BMJ Ouarda, and André St-Hilaire. Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*, 23(6):813–834, 2008.
- [47] BC Hewitson and RG Crane. Climate downscaling: techniques and application. *Climate Research*, pages 85–95, 1996.
- [48] Hugo Hidalgo, Michael Dettinger, and Daniel Cayan. Downscaling with constructed analogues: Daily precipitation and temperature fields over the united states. 2008.
- [49] Hugo G Hidalgo, Michael D Dettinger, and Daniel R Cayan. Downscaling with constructed analogues: Daily precipitation and temperature fields over the united states. *California Energy Commission PIER Final Project Report CEC-500-2007-123*, 2008.
- [50] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.
- [51] Geoffrey Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(July):504–507, 2006.
- [52] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [53] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM, 1993.

- [54] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [55] Carlo C Jaeger, Ortwin Renn, Eugene A Rosa, and Tom Webler. Decision analysis and rational action. *Human choice and climate change*, 3:141–216, 1998.
- [56] Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.
- [57] Thomas R Karl and Richard W Knight. Secular trends of precipitation amount, frequency, and intensity in the united states. *Bulletin of the American Meteorological society*, 79(2):231–241, 1998.
- [58] Richard W Katz. Techniques for estimating uncertainty in climate change scenarios and impact studies. *Climate Research*, 20(2):167–185, 2002.
- [59] AL Kay, HN Davies, VA Bell, and RG Jones. Comparison of uncertainty sources for climate change impacts: flood frequency in england. *Climatic Change*, 92(1-2):41–63, 2009.
- [60] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.
- [61] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? 2017.
- [62] Elizabeth J Kendon, Nigel M Roberts, Hayley J Fowler, Malcolm J Roberts, Steven C Chan, and Catherine A Senior. Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nature Climate Change*, 4(7):570, 2014.
- [63] Richard A Kerr. Forecasting regional climate change flunks its first test. *Science*, 339(6120):638–638, 2013.

- [64] Mohammad Sajjad Khan, Paulin Coulibaly, and Yonas Dibike. Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319(1):357–382, 2006.
- [65] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.
- [66] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. *International Conference on Machine Learning*, 2010.
- [67] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [68] Reto Knutti and Jan Sedláček. Robustness and uncertainties in the new cmip5 climate model projections. *Nature Climate Change*, 3(4):369–373, 2013.
- [69] James P Kossin. Validating atmospheric reanalysis data using tropical cyclones as thermometers. *Bulletin of the American Meteorological Society*, 96(7):1089–1096, 2015.
- [70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [71] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.
- [72] 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. *arXiv preprint*, 2017.
- [73] David B Lobell, Marshall B Burke, Claudia Tebaldi, Michael D Mastrandrea, Walter P Falcon, and Rosamond L Naylor. Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863):607–610, 2008.

- [74] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.
- [75] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, 10(3):e0119044, 2015.
- [76] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
- [77] Elizabeth C Mannhardt-Shamseldin, Richard L Smith, Stephan R Sain, Linda O Mearns, Daniel Cooley, et al. Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *The Annals of Applied Statistics*, 4(1):484–502, 2010.
- [78] Alon Manor and Sigalit Berkovic. Bayesian inference aided analog downscaling for near-surface winds in complex terrain. *Atmospheric Research*, 164:27–36, 2015.
- [79] D. Maraun, F. Wetterhall, a. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themel, V. K C Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(2009):1–34, 2010.
- [80] Douglas Maraun, Theodore G Shepherd, Martin Widmann, Giuseppe Zappa, Daniel Walton, José M Gutiérrez, Stefan Hagemann, Ingo Richter, Pedro MM Soares, Alex Hall, et al. Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7(11):764, 2017.
- [81] Douglas Maraun, F Wetterhall, AM Ireson, RE Chandler, EJ Kendon, M Widmann, S Brienen, HW Rust, T Sauter, M Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), 2010.

- [82] Douglas Maraun, Martin Widmann, José M Gutiérrez, Sven Kotlarski, Richard E Chandler, Elke Hertig, Joanna Wibig, Radan Huth, and Renate AI Wilcke. Value: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3(1):1–14, 2015.
- [83] Edwin P Maurer and Hugo G Hidalgo. Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. 2008.
- [84] Edwin P Maurer, Hugo G Hidalgo, T Das, MD Dettinger, and DR Cayan. The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. *Hydrology and Earth System Sciences*, 14(6):1125–1138, 2010.
- [85] Tharindu P Miyanawala and Rajeev K Jaiman. An efficient deep learning technique for the navier-stokes equations: Application to unsteady wake flow dynamics. *arXiv preprint arXiv:1710.09099*, 2017.
- [86] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [87] James E Neumann, Jason Price, Paul Chinowsky, Leonard Wright, Lindsay Ludwig, Richard Streeter, Russell Jones, Joel B Smith, William Perkins, Lesley Jantarasami, et al. Climate change risks to US infrastructure: impacts on roads, bridges, coastal development, and urban drainage. *Climatic Change*, 131(1):97–109, 2015.
- [88] Camille Parmesan. Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, pages 637–669, 2006.
- [89] SE Perkins, AJ Pitman, NJ Holbrook, and J McAneney. Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate*, 20(17):4356–4376, 2007.
- [90] David W Pierce, Daniel R Cayan, and Bridget L Thrasher. Statistical downscaling using localized constructed analogs (loca). *Journal of Hydrometeorology*, 15(6):2558–2585, 2014.

- [91] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Advances in Neural Information Processing Systems*, pages 3405–3416, 2017.
- [92] PN Racherla, DT Shindell, and GS Faluvegi. The added value to global model projections of climate change by dynamical downscaling: A case study over the continental us using the giss-modele2 and wrf models. *Journal of Geophysical Research: Atmospheres*, 117(D20), 2012.
- [93] Michele M Rienecker, Max J Suarez, Ronald Gelaro, Ricardo Todling, Julio Bacmeister, Emily Liu, Michael G Bosilovich, Siegfried D Schubert, Lawrence Takacs, Gi-Kong Kim, et al. Merra: Nasa’s modern-era retrospective analysis for research and applications. *Journal of Climate*, 24(14):3624–3648, 2011.
- [94] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [95] Markku Rummukainen. Methods for statistical downscaling of gcm simulations. *SMHI Rapporter Meteorologi och Klimatologi (Sweden)*. no. 80., 1997.
- [96] Kaustubh Salvi, Subimal Ghosh, and Auroop R. Ganguly. Credibility of statistical downscaling under nonstationary climate. *Climate Dynamics*, 2015.
- [97] Quirin Schiermeier. The real holes in climate science. *Nature News*, 463(7279):284–287, 2010.
- [98] Justin T Schoof and SC Pryor. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of Climatology*, 21(7):773–790, 2001.
- [99] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3799, 2015.

- [100] Mikhail A Semenov and Elaine M Barrow. Use of a stochastic weather generator in the development of climate change scenarios. *Climatic change*, 35(4):397–414, 1997.
- [101] Mikhail A Semenov and Pierre Stratonovitch. Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Climate research (Open Access for articles 4 years old and older)*, 41(1):1, 2010.
- [102] J Mc Lean Sloughter, Adrian E Raftery, Tilmann Gneiting, and Chris Fraley. Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.
- [103] Alex Smola and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.
- [104] Anne MK Stoner, Katharine Hayhoe, Xiaohui Yang, and Donald J Wuebbles. An asynchronous regional regression model for statistical downscaling of daily climate variables. *International Journal of Climatology*, 33(11):2473–2494, 2013.
- [105] James W Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.
- [106] Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485, 2012.
- [107] MJ Themeßl, Andreas Gobiet, and Armin Leuprecht. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10):1530–1544, 2011.
- [108] Jeanne M Thibeault and Anji Seth. Changing climate extremes in the northeast united states: observations and projections from cmip5. *Climatic change*, 127(2):273–287, 2014.
- [109] Bridget Thrasher, Edwin P Maurer, C McKellar, and PB Duffy. Technical note: Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrology and Earth System Sciences*, 16(9):3309–3314, 2012.

- [110] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [111] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014.
- [112] Lee Tryhorn and Art DeGaetano. A comparison of techniques for downscaling extreme precipitation over the northeastern united states. *International Journal of Climatology*, 31(13):1975–1989, 2011.
- [113] Thomas Vandal, Udit Bhatia, and Auroop R. Ganguly. Statistical downscaling in climate with state of the art scalable machine learning. In Ramakrishna Nemani Srivastava and Karsten Steinhaeuser, editors, *Large-Scale Machine Learning in the Earth Sciences*, volume 42, chapter 4. Taylor & Francis Group, 2017.
- [114] Thomas Vandal, Evan Kodra, Jennifer Dy, Sangram Ganguly, Ramakrishna Nemani, and Auroop R Ganguly. Uncertainty in discrete-continuous and skewed data with bayesian deep learning. In *Proceedings of the 53rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2018.
- [115] Thomas Vandal, Evan Kodra, and Auroop R Ganguly. Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *arXiv preprint arXiv:1702.04018*, 2017.
- [116] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1663–1672. ACM, 2017.
- [117] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network

- with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.
- [118] Gian-Reto Walther, Eric Post, Peter Convey, Annette Menzel, Camille Parmesan, Trevor JC Beebee, Jean-Marc Fromentin, Ove Hoegh-Guldberg, and Franz Bairlein. Ecological responses to recent climate change. *Nature*, 416(6879):389–395, 2002.
- [119] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Transactions on Medical Imaging*, 2018.
- [120] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–378, 2015.
- [121] RL Wilby, SP Charles, E Zorita, B Timbal, P Whetton, and LO Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods. 2004.
- [122] Robert L Wilby, Christian W Dawson, and Elaine M Barrow. Sdsma decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software*, 17(2):145–157, 2002.
- [123] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
- [124] Andrew W Wood, Lai R Leung, V Sridhar, and DP Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic change*, 62(1-3):189–216, 2004.
- [125] Andrew W Wood, Edwin P Maurer, Arun Kumar, and Dennis P Lettenmaier. Long-range experimental hydrologic forecasting for the eastern united states. *Journal of Geophysical Research: Atmospheres*, 107(D20), 2002.
- [126] Donald Wuebbles, David Fahey, Kathleen Hibbard, DJ Dokken, BC Stewart, and TK Maycock. Climate science special report: Fourth national climate assessment, volume i. 2017.

- [127] Pingping Xie, Mingyue Chen, Song Yang, Akiyo Yatagai, Tadahiro Hayasaka, Yoshihiro Fukushima, and Changming Liu. A gauge-based analysis of daily precipitation over east asia. *Journal of Hydrometeorology*, 8(3):607–626, 2007.
- [128] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1059–1066, 2013.
- [129] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.
- [130] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.
- [131] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via multi-task sparse learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2042–2049. IEEE, 2012.
- [132] Xianliang Zhang and Xiaodong Yan. A new statistical precipitation downscaling method with bayesian model averaging: a case study in china. *Climate Dynamics*, 45(9-10):2541–2555, 2015.
- [133] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *Uncertainty in Artificial Intelligence*, 2012.
- [134] Yiniao Zhu and Nicholas Zabaras. Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *arXiv preprint arXiv:1801.06879*, 2018.
- [135] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.