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

Thomas Vandal
Northeastern University, Civil and Environmental Engineering
Boston, MA
vandal.t@husky.neu.edu

Evan Kodra
risQ Inc.
Cambridge, MA
evan.kodra@risq.io

Jennifer Dy
Northeastern University, Electrical and Computer Engineering
Boston, MA
j.dy@neu.edu

Sangram Ganguly
Bay Area Environmental Research Institute / NASA Ames Research Center
Moffett Field, CA
sangram.ganguly@nasa.gov

Ramakrishna Nemani
NASA Advanced Supercomputing Division/ NASA Ames Research Center
Moffett Field, CA
rama.nemani@nasa.gov

Auroop R Ganguly
Northeastern University, Civil and Environmental Engineering
Boston, MA
a.ganguly@neu.edu

ABSTRACT

Deep Learning (DL) methods have been transforming computer vision with innovative adaptations to other domains including climate change. For DL to pervade Science and Engineering (S&E) applications where risk management is a core component, well-characterized uncertainty estimates must accompany predictions. However, S&E observations and model-simulations often follow heavily skewed distributions and are not well modeled with DL approaches, since they usually optimize a Gaussian, or Euclidean, likelihood loss. Recent developments in Bayesian Deep Learning (BDL), which attempts to capture uncertainties from noisy observations, aleatoric, and from unknown model parameters, epistemic, provide us a foundation. Here we present a discrete-continuous BDL model with Gaussian and lognormal likelihoods for uncertainty quantification (UQ). We demonstrate the approach by developing UQ estimates on "DeepSD", a super-resolution based DL model for Statistical Downscaling (SD) in climate applied to precipitation, which follows an extremely skewed distribution. We find that the discrete-continuous models outperform a basic Gaussian distribution in terms of predictive accuracy and uncertainty calibration. Furthermore, we find that the lognormal distribution, which can handle skewed distributions, produces quality uncertainty estimates at the extremes. Such results may be important across S&E, as well as other domains such as finance and economics, where extremes are often of significant interest. Furthermore, to our knowledge, this is the first UQ model in SD where both aleatoric and epistemic uncertainties are characterized.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Reconstruction*; • **Applied computing** → **Earth and atmospheric sciences**;

KEYWORDS

Bayesian Deep Learning, Uncertainty Quantification, Climate Downscaling, Super-resolution, Precipitation Estimation

ACM Reference Format:
Thomas Vandal, Evan Kodra, Jennifer Dy, Sangram Ganguly, Ramakrishna Nemani, and Auroop R Ganguly. 2018. Quantifying Uncertainty in Discrete-Continuous and Skewed Data with Bayesian Deep Learning. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219996

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219996

1 INTRODUCTION

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 [41] and detect tropical cyclones and atmospheric rivers [35]. Remote sensing models such as DeepSAT [3], a satellite image classification framework, also leverage computer vision technologies. Physicists are using deep learning for detecting particles in high energy physics [1] and in transportation deep learning has aided in traffic flow prediction [30] and modeling network congestion [31]. Scientists have even used convolutional neural networks to approximate the Navier-Stokes equations of unsteady fluid forces [34]. 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 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 it's potential in S&E applications, our models must be resilient to non-normal and discrete-continuous distributions.

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

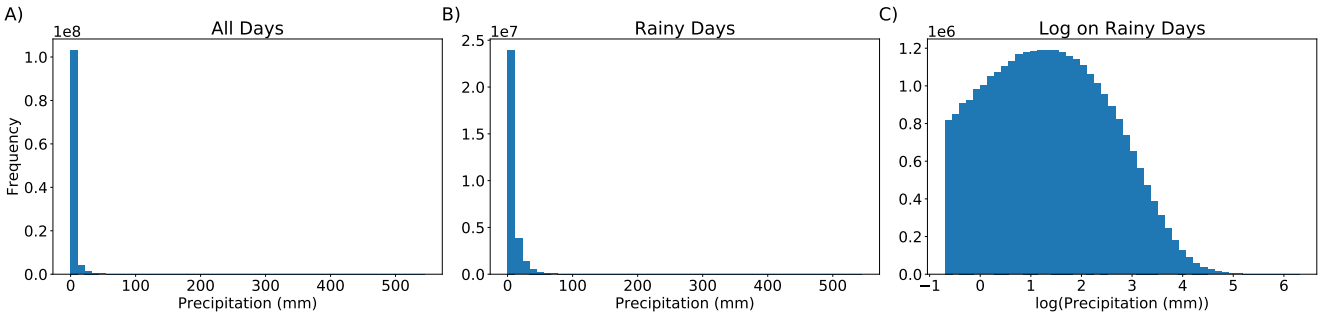

Figure 1: Histogram of daily precipitation on the Contiguous United States from 2006 to 2015. A) All precipitation data points. B) Precipitation distribution on rainy days only. C) Log distribution of precipitation on rainy days.

climate change are so pervasive and far reaching that the tools for handling uncertainty provided by decision analysis are no longer sufficient [20]." 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 [21, 29]. For example, Kay et al. studied six different sources of uncertainty of climate change impacts on a flood frequency model [22]. 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 [11, 13, 24, 44]. 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 [24]. *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 [23] and medical image segmentation [42].

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 [41], and Bayesian

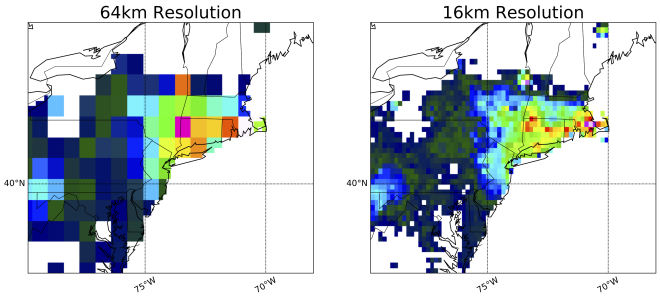

Figure 2: Prism Observed Precipitation: Left) Low resolution at 64km. Right) High resolution at 16km.

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.

1.1 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.

2 PRECIPITATION ESTIMATION

2.1 Statistical Downscaling

Downscaling, either statistical or dynamical, is a widely used process for producing high-resolution projections from coarse global climate models (GCMs) [10, 17, 33]. Dynamical downscaling, often 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 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 [39]. 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 [36]. 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) [6] and automated statistical downscaling (ASD) [16] to neural networks [38] and nearest neighbor models [18]. Multiple studies have compared different sets of statistical downscaling approaches on various climate variables and varying temporal and spatial scales showing that no approach consistently outperforms the others [5, 15, 40]. 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 [41]. 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 [43]. Khan et al. presented one study that assessed monthly uncertainty from confidence based intervals of daily predictions [25]. 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.

2.2 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 [8]. 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 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.

3 BACKGROUND

3.1 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) [16] 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) [9] 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

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:

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

where S is a subset n examples. Based on recent state-of-the-art results in super-resolution [26, 28], we modify the SRCNN architecture to include a residual connection between the precipitation input channel and output layer, as shown in Figure 3.

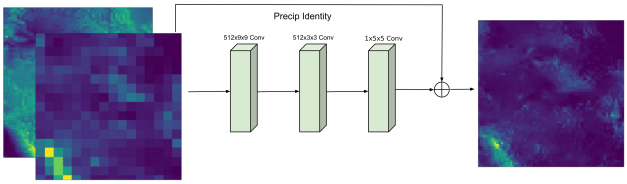

Figure 3: Residual SRCNN Architecture used for DeepSD with a skip connection between precipitation and the output layer.

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.

3.2 Bayesian Deep Learning

In the early 1990's Mackay [32] 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 [2, 14, 19]. 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 [11, 12]. Kendall and Gal then followed this work for computer vision applications to include both aleatoric and epistemic uncertainties in a single model [24].

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

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

Applying variational inference to the weights, we can define an approximate and tractable distribution $q_\Theta(\boldsymbol{\omega}) = \prod_{l=1}^{L} q_{\mathbf{M}_l}(\mathbf{W}_l)$ where $q_{\mathbf{M}_l}(\mathbf{W}_l) = \mathbf{M}_l \times \text{diag}\big[\text{Bernoulli}(1-p_l)^{K_l}\big]$ 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 $q_\Theta(\boldsymbol{\omega})$ to the true posterior, $p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$. The optimization objective of the variational interpretation can be written as [12]:

$$\hat{\mathcal{L}}(\Theta) = -\frac{1}{M} \sum_{i \in S} \log p(\mathbf{y}_i|f^{\boldsymbol{\omega}}(\mathbf{x}_i)) + \frac{1}{N} \text{KL}(q_\Theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) \tag{2}$$

$$= \hat{\mathcal{L}}_x(\Theta) + \frac{1}{N} \text{KL}(q_\Theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) \tag{3}$$

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 [13]. As presented by Gal et al., the KL divergence term is then written as:

$$\text{KL}(q_\Theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) = \sum_{l=1}^{L} \text{KL}(q_{\mathbf{M}_l}(\mathbf{W}_l)||p(\mathbf{W}_l)) \tag{4}$$

$$\text{KL}(q_{\mathbf{M}_l}(\mathbf{W})||p(\mathbf{W})) \propto \frac{l^2(1-p_l)}{2}||\mathbf{M}_l|| - K_l \mathcal{H}(p_l) \tag{5}$$

where

$$\mathcal{H}(p) = -p \log p - (1-p) \log (1-p) \tag{6}$$

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 [13] for the concrete dropout optimization. In the remainder of this paper, we will use this concrete dropout formulation within all presented models.

4 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 [24]. 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 el al. using a discrete-continuous gamma distribution [37]. Lastly, we show that a lognormal distribution can be applied directly in

BDL and derive its corresponding log-likelihood loss and unbiased parameter estimates.

4.1 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 [24]. As discussed in previous sections, it is crucial to capture both aleatoric and epistemic uncertainties in downscaling. As shown in section 3.1 of [24], we must measure the aleatoric uncertainty by estimating the variance, 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 $\mathbf{f}$:

$$[\hat{\mathbf{y}}, \hat{\sigma}^2] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}). \tag{7}$$

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

$$\mathcal{L}_x(\Theta) = \frac{1}{2D}\sum_i \hat{\sigma}_i^{-2}||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \frac{1}{2}\log \hat{\sigma}_i^2 \tag{8}$$

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. 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 substiution $s_i := \log \hat{\sigma}_i^2$ is used for stable learning using the Adam Optimization algorithm [27], 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, $\{\hat{\mathbf{y}}_t, \hat{\sigma}_t^2\}$, from $\mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{x})$ with masked weights $\widehat{\mathbf{W}}_t \sim q(\mathbf{W})$:

$$\mathrm{E}[\mathbf{Y}] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}_t \tag{9}$$

$$\mathrm{Var}[\mathbf{Y}] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t^2 - \frac{1}{T}\sum_{t=1}^{T} \hat{\sigma}_t^2 + \Big(\frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t\Big)^2. \tag{10}$$

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 [24].

4.2 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 $\mathbf{f}$:

$$[\hat{\mathbf{y}}, \hat{\sigma}^2, \hat{\phi}] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}) \tag{11}$$

$$\hat{p} = \mathrm{Sigmoid}(\hat{\phi}). \tag{12}$$

Splitting the distribution into discrete and continuous parts gives us:

$$p\big(\mathbf{y}|f^{\boldsymbol{\omega}}(\mathbf{x})\big) = \begin{cases} (1-\hat{p}) & \mathbf{y} = 0 \\ \hat{p}\cdot\mathcal{N}\big(\mathbf{y};\hat{\mathbf{y}},\hat{\sigma}^2\big) & \mathbf{y} > 0 \end{cases} \tag{13}$$

Plugging this in to 2 and dropping the constants gives us the loss function (for brevity, we ignore the KL term which is identical to Equation 4):

$$\begin{aligned}
\mathcal{L}_x(\Theta) &= -\frac{1}{D}\sum_i \log\Big(\mathbb{1}_{\mathbf{y}_i>0}\cdot\hat{p}_i\cdot\mathcal{N}\big(\mathbf{y}_i;\hat{\mathbf{y}}_i,\hat{\sigma}_i^2\big) + \mathbb{1}_{\mathbf{y}_i=0}\cdot(1-\hat{p}_i)\Big) \\
&= -\frac{1}{D}\sum_{i,\mathbf{y}_i>0}\Big(\log\hat{p}_i + \log\mathcal{N}\big(\mathbf{y}_i;\hat{\mathbf{y}}_i,\hat{\sigma}_i^2\big)\Big) \\
&\qquad -\frac{1}{D}\sum_{i,\mathbf{y}_i=0}\log(1-\hat{p}_i) \\
&= \frac{1}{D}\sum_i\big(\mathbb{1}_{\mathbf{y}_i>0}\cdot\hat{p}_i + (1-\mathbb{1}_{\mathbf{y}_i>0})\cdot(1-\hat{p}_i)\big) \\
&\qquad -\frac{1}{2D}\sum_{i,\mathbf{y}_i>0}\hat{\sigma}_i^{-2}||\mathbf{y}_i-\hat{\mathbf{y}}_i||^2 + \log\sigma_i^2
\end{aligned} \tag{14}$$

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:

$$\mathrm{E}[\mathbf{Y}] \approx \frac{1}{T}\sum_{t=1}^{T}\hat{p}_t\hat{\mathbf{y}}_t \tag{15}$$

$$\mathrm{Var}[\mathbf{Y}] \approx \frac{1}{T}\sum_{t=1}^{T}\hat{p}_t^2(\hat{\mathbf{y}}_t^2+\hat{\sigma}_t^2) - \Big(\frac{1}{T}\sum_{t=1}^{T}\hat{p}_t\hat{\mu}_t\Big)^2. \tag{16}$$

4.3 Discrete-Continuous Lognormal Likelihood

Precipitation events, especially extremes, are known to follow fat-tailed distributions, such as lognormal and Gamma distributions [7, 37]. 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:

$$[\hat{\mu}, \hat{\sigma}^2, \hat{\phi}] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{X}) \tag{17}$$

$$\hat{p} = \mathrm{Sigmoid}(\hat{\phi}). \tag{18}$$

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:

$$p\big(\mathbf{y}|f^{\boldsymbol{\omega}}(\mathbf{x})\big) = \begin{cases} (1-\hat{p}) & \mathbf{y} = 0 \\ \hat{p}\cdot\dfrac{1}{\mathbf{y}\hat{\sigma}\sqrt{2\pi}}\exp\Big(-\dfrac{(\log(\mathbf{y})-\hat{\mu})^2}{2\hat{\sigma}^2}\Big) & \mathbf{y} > 0 \end{cases} \tag{19}$$

This gives us the modified log-likelihood objective:

$$\begin{aligned}
\mathcal{L}_x(\Theta) &= \frac{1}{D}\sum_i\big(\mathbb{1}_{\mathbf{y}_i>0}\cdot\hat{p}_i + (1-\mathbb{1}_{\mathbf{y}_i>0})\cdot(1-\hat{p}_i)\big) \\
&\qquad -\frac{1}{2D}\sum_{i,\mathbf{y}_i>0}\hat{\sigma}_i^{-2}||\log\mathbf{y}_i-\hat{\mu}_i||^2 + \log\sigma_i^2
\end{aligned} \tag{20}$$

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

$$\mathrm{E}[\mathbf{y}] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{p}_t \exp(\hat{\mu} + \frac{1}{2}\hat{\sigma}^2) \tag{21}$$

$$\mathrm{Var}[\mathbf{Y}] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{p}_t^2 \exp(2\hat{\mu} + 2\hat{\sigma}^2) \tag{22}$$

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

$$\hat{\sigma} = \log\Big(1 + \frac{1}{2}\sqrt{\frac{4\mathrm{Var}[\mathbf{Y}]}{\mathrm{E}[\mathbf{y}]^2} + 1}\Big) \tag{23}$$

$$\hat{\mu} = \mathrm{E}[\mathbf{y}] - \frac{\hat{\sigma}^2}{2} \tag{24}$$

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.

5 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 2.2, 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 Datset (GTOPO30) provided by the USGS as an auxilary 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 >0.5mm 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 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 τ=1e-5 and prior length scale as $l = 1$ to improve uncertainly calibration performance [13]. For a pixel-wise regression the number of samples N is set as Days×Height×Width. These parameters were found to provide a good trade-off between likelihood and regularization loss terms. As shown in Figure 4, dropout rates for each model and hidden layer are close to 0.5, the largest

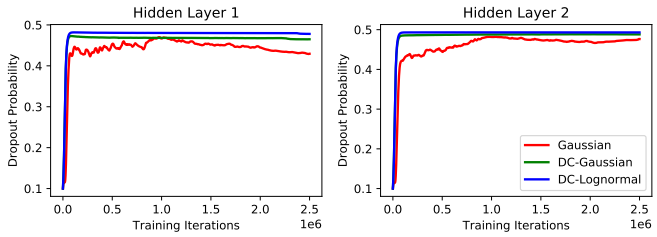

Figure 4: Dropout probabilities learned using Concrete Dropout for both hidden layers.

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).

5.1 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) which provides a suite of extreme precipitation indices and is often used for evaluating downscaling models [4, 40]:

(1) R20 - Very heavy wet days ≥ 20mm
(2) SDII - Daily intensity index = (Annual total) / (precip days ≥ 0.5 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. ($\mathrm{SDII}_{\mathrm{model}}$ - $\mathrm{SDII}_{\mathrm{obs}}$). These results can be seen in Table 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 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 5x5 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

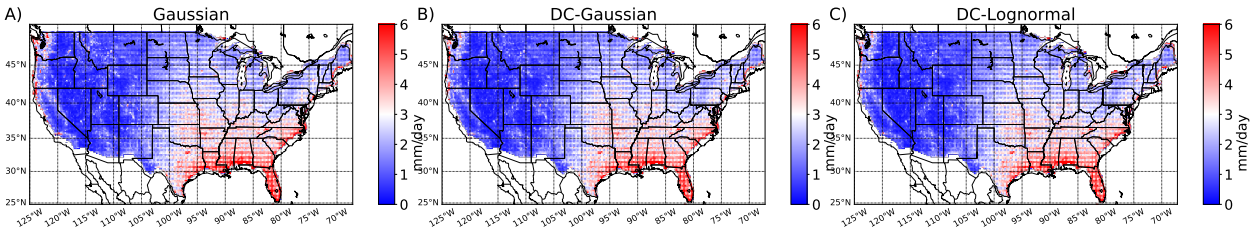

Figure 5: Daily Root Mean Square Error (RMSE) computed at each location for years 2006 to 2015 (test set) in CONUS. A) Gaussian, B) Conditional-Gaussian, and C) Conditional-Lognormal. Red corresponds to high RMSE while blue corresponds to low RMSE.

	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 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).

and low probabilities of precipitation in a close neighborhood. This insight is particularly important when applied to coastal cities.

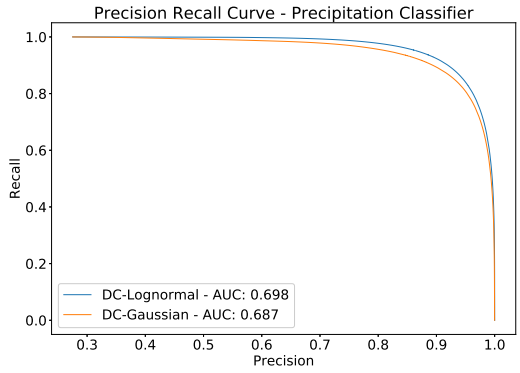

Figure 6: Precision recall curve of classifying rainy days in conditional models.

Lastly, we look at each conditional model's ability to classify precipitous days with precision recall curves (Figure 6). We see that recall does not begin to decrease until a precision of 0.8 which indicates very strong classification performance. It was assumed that classification of precipitation would be easy for such a dataset.

5.2 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 ($\geq$ 0.5mm) as uncertainty on non-rainy days is not of interest. The calibration metric used computes the frequency of observations occurring within a varying predicted probability range:

$$c(z) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{I}_{P(y_i|f^{\omega}(x_i))>(0.5-z/2)} * \mathcal{I}_{P(y_i|f^{\omega}(x_i))<(0.5+z/2)} \quad (25)$$

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:

$$\text{RMSE}_{cal} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(c(i/K) - i/K)^2} \quad (26)$$

where K is the number bins. In our analysis, we use $K = 100$. The calibration plots for each model can be seen in Figure 7.

Right away we see from Figure 7 that the Gaussian distribution over-estimates uncertainty for most of the range with a wider range of variability between pixels. DC-Lognormal also overestimates 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 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

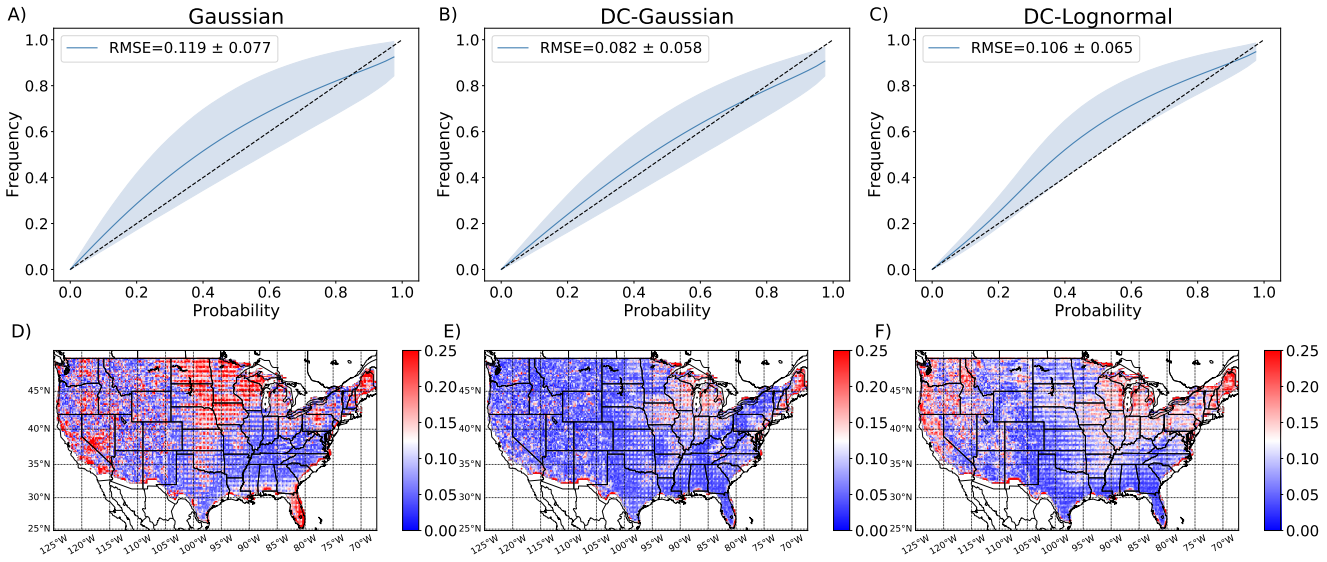

Figure 7: Calibration is computed as the frequency of predictions within a given probability range. This probability is varied on the x-axis with the corresponding frequency on the y-axis. Columns represent each model Gaussian, DC-Gaussian and Lognormal. Calibration plots on the first row compute per pixel with the shaded area representing the 80% confidence interval of calibration. The second row depicts calibration root mean square error (RMSE) per location.

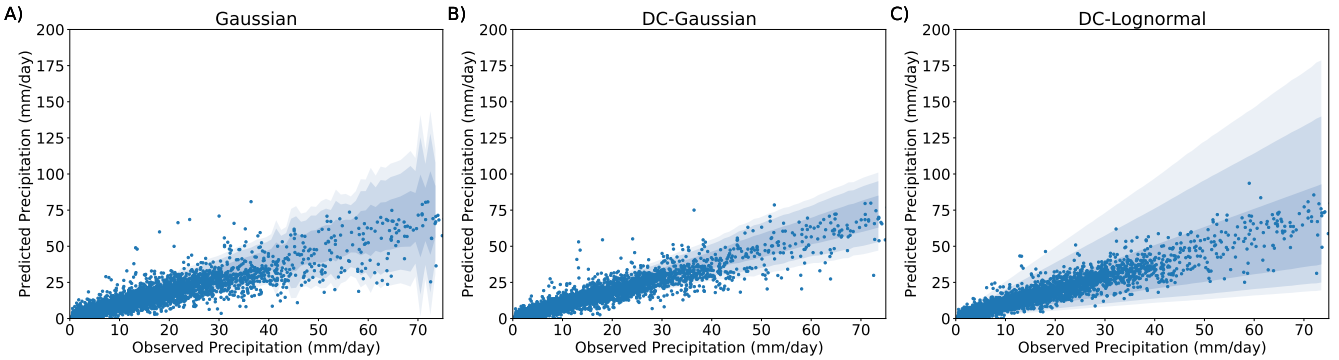

Figure 8: Uncertainty widths based on quantiles from their predictive distributions. The points are observations versus the expected value. The bands correspond to 50%, 80%, and 90% predictive intervals.

perform well, but DC-Lognormal also has areas of increased error in the west.

In Figure 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.

6 CONCLUSION

In this paper 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 [41], 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.

ACKNOWLEDGMENTS

This work was supported by NASA Earth Exchange (NEX), National Science Foundation CISE Expeditions in Computing under grant number: 1029711, National Science Foundation CyberSEES under grant number: 1442728, National Science Foundation CRISP under grant number: 1735505, and National Science Foundation BIGDATA under grant number: 1447587. The GTOPO30 dataset was distributed by the Land Processes Distributed Active Archive Center (LP DAAC), located at USGS/EROS, Sioux Falls, SD. http://lpdaac.usgs.gov. We thank Arindam Banerjee for valuable comments.

REFERENCES

[1] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.
[2] D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 168:215–238, 1998.
[3] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. 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.
[4] G. Bürger, T. Murdock, A. Werner, S. Sobie, and A. Cannon. Downscaling extremesâĂŤan intercomparison of multiple statistical methods for present climate. *Journal of Climate*, 25(12):4366–4388, 2012.
[5] 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.
[6] A. J. Cannon and P. H. Whitfield. Downscaling recent streamflow conditions in british columbia, canada using ensemble neural network models. *Journal of Hydrology*, 259(1):136–151, 2002.
[7] H.-K. Cho, K. P. Bowman, and G. R. North. A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11):1586–1597, 2004.
[8] C. Daly, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International journal of climatology*, 28(15):2031–2064, 2008.
[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.
[10] H. J. Fowler, S. Blenkinsop, and C. 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.
[11] Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, Ph. D. thesis, University of Cambridge, 2016.
[12] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.
[13] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. *arXiv preprint arXiv:1705.07832*, 2017.
[14] A. Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.
[15] E. Gutmann, T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. 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.
[16] M. Hessami, P. Gachon, T. B. Ouarda, and A. St-Hilaire. Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*, 23(6):813–834, 2008.
[17] B. Hewitson and R. Crane. Climate downscaling: techniques and application. *Climate Research*, pages 85–95, 1996.
[18] H. Hidalgo, M. Dettinger, and D. Cayan. Downscaling with constructed analogues: Daily precipitation and temperature fields over the united states. 2008.
[19] G. E. Hinton and D. 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.
[20] C. C. Jaeger, O. Renn, E. A. Rosa, and T. Webler. Decision analysis and rational action. *Human choice and climate change*, 3:141–216, 1998.
[21] R. W. Katz. Techniques for estimating uncertainty in climate change scenarios and impact studies. *Climate Research*, 20(2):167–185, 2002.
[22] A. Kay, H. Davies, V. Bell, and R. Jones. Comparison of uncertainty sources for climate change impacts: flood frequency in england. *Climatic Change*, 92(1-2):41–63, 2009.
[23] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.
[24] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.
[25] M. S. Khan, P. Coulibaly, and Y. Dibike. Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319(1):357–382, 2006.
[26] J. Kim, J. Kwon Lee, and K. 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.
[27] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.
[29] D. B. Lobell, M. B. Burke, C. Tebaldi, M. D. Mastrandrea, W. P. Falcon, and R. L. Naylor. Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863):607–610, 2008.
[30] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.
[31] X. Ma, H. Yu, Y. Wang, and Y. Wang. Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, 10(3):e0119044, 2015.
[32] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
[33] D. Maraun, F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. 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.
[34] T. P. Miyanawala and R. K. Jaiman. An efficient deep learning technique for the navier-stokes equations: Application to unsteady wake flow dynamics. *arXiv preprint arXiv:1710.09099*, 2017.

[35] E. Racah, C. Beckham, T. Maharaj, S. Kahou, M. Prabhat, and C. 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.

[36] M. A. Semenov and P. 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.

[37] J. M. L. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.

[38] J. W. Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.

[39] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[40] T. Vandal, E. Kodra, and A. R. Ganguly. Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *arXiv preprint arXiv:1702.04018*, 2017.

[41] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2017.

[42] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Transactions on Medical Imaging*, 2018.

[43] D. Wuebbles, D. Fahey, K. Hibbard, D. Dokken, B. Stewart, and T. Maycock. Climate science special report: Fourth national climate assessment, volume i. 2017.

[44] Y. Zhu and N. Zabaras. Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *arXiv preprint arXiv:1801.06879*, 2018.