Abstract—Global Climate Models (GCMs) contain uncertainty from internal variability of the system, parameterizations, and constraints from unknown physical processes. Furthermore, statistical downscaling of GCMs exasperates the uncertainty further. As downscaled datasets are often leveraged for climate change adaptation and planning, quantifying the uncertainty of our statistical model is crucial. In this work, we show how uncertainty quantification from Bayesian deep learning approaches can be leveraged in climate applications. We review and compare three approaches, Monte-carlo dropout, Concrete dropout, and Alpha-divergence based dropout, for statistically downsampling precipitation in Orlando Florida. In our experiments, we find that Concrete and Monte-carlo dropouts perform well but Alpha-divergence based dropout is less effective. While further experimentation is needed, Concrete dropout is a promising approach which extends Monte-carlo dropout by optimizing dropout probabilities.

I. MOTIVATION

Downscaling is a process of enhancing the spatial (or temporal) resolution of global climate models (GCMs) for the purpose of climate change adaptation and planning. The statistical approach to downscaling aims to learn a functional mapping between low- and high-resolution historical datasets, which can then be applied directly to coarse resolution GCM outputs. A wide range of methods have been explored for statistical downscaling, ranging from bias correction techniques [1] and nearest neighbor approaches [2], [3] to sparse linear regressions [4] and neural networks [5], [6]. Intercomparison studies have shown that each method tends to perform well for certain regions, seasons, and climate variables but have difficulty generalizing [4]. The lack of consensus between statistical downscaling approaches highlights the difficulty of the problem and the need for uncertainty quantification in downscaled projections.

Given the difficulty of credibly downscaling GCMs, it is crucial to provide stakeholders with a range possible events. For example, a stakeholder may be interested in various metrics of extreme precipitation, including maximum daily return values. These metrics can be extracted from downscaled precipitation from GCM outputs under a set of scenarios. However, we know that these downscaled values have some amount of uncertainty which should be quantified appropriately. The concept of uncertainty quantification applied to statistical downscaling can be leveraged to compute uncertainties on the daily scale which can then be aggregated when computing a chosen metric. In this work, we refer to uncertainty quantification as the uncertainty over our prediction (rather than uncertainty over the parameters).

Bayesian models are well suited for uncertainty quantification and can be leveraged in statistical downscaling. Given the recent hype in deep learning as well as a recent publication using convolutional neural networks for statistical downscaling [6], we begin by studying the use of Bayesian deep learning (BDL) for uncertainty quantification in the climate domain.

Bayesian neural networks, originally studied in the early 1990’s [7], applies a prior over the weights and biases of the network and aims to learn the posterior distribution given the data. As is common in many Bayesian modeling problems, the formulation is intractable and unable to scale to multiple hidden layers. BDL has attempted to solve this problem by estimating the posterior distribution using approximate variational inference and leveraging stochastic regularization techniques such as dropout [8]. Similar to more basic deep learning architectures, the tuning of hyper-parameters can greatly influence a model’s results. In the remainder of this paper, we discuss the only three approaches which have been proposed for BDL and experiment in
downscaling daily precipitation in Orlando Florida.

II. METHOD

The technique of using dropout as a stochastic regularization technique has been crucial for the major advances in deep learning. Dropout is implemented by randomly setting weights in the neural network to zero during training, usually sampled as a Bernoulli random variable. This process then reduces overfitting by more evenly spreading out information at each layer. The further study of BDL has provided a basic theoretical understand of why dropout effectively reduces overfitting of deep networks and how it can provide robust uncertainty estimates [9]. In this section, we review the fundamental concepts of BDL as well as two variants.

We define an L-layered neural network with inputs $X = \{x_1, ..., x_N\}$, label $Y = \{y_1, ..., y_N\}$, and weights $\omega = \{W_1, ..., W_L\}$ such that

$$\begin{align*}
[f^{\omega}(x)]_l &= \sqrt{\frac{1}{K_l}} W_l \sigma(...) \sqrt{\frac{1}{K_1}} \sigma(W_1 x) \\
q(x) &= N(y; \hat{y}(x), g(x)^{-1})
\end{align*}$$

(1)

where $K_l$ are the number of units in layer $l$, as defined in [10]. The probabilistic interpretation of a Bayesian neural network can be written as

$$p(y^*|x^*, X, Y) = \int p(y^*|x^*, \omega)p(\omega|X, Y)$$

$$p(y|x, \omega) = N(y; \hat{y}(x), g(x)^{-1})$$

(2)

An approximate distribution $q(\omega)$ is defined as:

$$W_i = M_i * \text{diag}(z_{i,j}^{K_i})$$

$$z_{i,j} = \text{Bernoulli}(p_i)$$

for $i = 1, ..., L$, $j = 1, ..., K_{i-1}$

given some probabilities $p_i$ and variational parameters $M_i$. The Bernoulli random variable $z_{i,j}$ performs dropout at unit $j$ for layer $i - 1$ resulting in a stochastic approximation of $W_i$. We then minimize the Kullback–Leibler divergence between the posterior and it’s variational approximation giving us the optimization objective:

$$\hat{L}_{\text{reg}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|x_i, \omega) + \frac{1}{N} \text{KL}(q(\omega)||p(\omega))$$

(3)

where $\theta$ represents the variational parameters and the term $\text{KL}(q(\omega)||p(\omega))$ is a regularization term, assuring that our approximate distribution does not deviate too far from the prior. Plugging our prior distribution $p(y|x, \omega)$ into 3, we obtain our regression objective:

$$\hat{L}_{\text{reg}}(\theta) = \frac{1}{2N} \|y_i - f(x_i)\|^2 + \log g(x_i)^{-1} + \frac{1}{N} \text{KL}(q(\omega)||p(\omega))$$

(4)

A. Monte Carlo Dropout

Monte Carlo Dropout (MC-Dropout) is the fundamental technique used in Bayesian deep learning for inference. Given a neural network trained with dropout minimizing 4, for a given new example we can use dropout to sample the posterior distribution to estimate the predictive mean and variance.

$$E[y] = \frac{1}{T} \sum_{i=1}^{T} \hat{f}(x)$$

$$\text{Var}[y] = \frac{1}{T} \sum_{i=1}^{T} g(x)^{-1} + \frac{1}{T} \sum_{i=1}^{T} \hat{f}(x)^2 - E[y]^2$$

(5)

In this approach, the probability of dropout, $p$, is held constant which can cause under- or over-estimations of uncertainty as well as less than optimal predictions. The following two approaches, Concrete Dropout [11] and Dropout with Alpha-Divergence [12], attempt to improve the model’s robustness and credibility of uncertainty estimates.

B. Concrete Dropout

The first approach we will discuss to learning a more robust uncertainty estimate is through Concrete Dropout [11]. Rather than doing a grid-search over all dropout probabilities (for each layer), which is computationally infeasible, Concrete dropout optimizes this probability. By parameterizing the variational approximation with probabilities such that $\theta = \{M_i, p_i\}_{i=1}^{L}$, we can rewrite the KL objective in 4 as:

$$\text{KL}(q(\omega)||p(\omega)) \propto \sum_{i=1}^{L} (1 - p_i)^2||M_i||^2 - K\mathcal{H}(p)$$

(6)

where

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

(7)

is the entropy of the Bernoulli random variable with probability $p$ [11]. The term $\mathcal{H}(p)$ acts as a regularizer and enforces $p \leq 0.5$. The authors show that the dropout probabilities in the lower layers tend to approach zero during training, which, in the past, has been shown to work well in various applications.
C. Dropout with Alpha-Divergences

A second approach to learning uncertainty estimates is by using Dropout with Alpha-Divergences (Dropout-AD) [12]. Using black-box $\alpha$-divergence minimization as an extension to variational inference, which tends to under-estimate uncertainty, to penalize the function in the domain where examples have not been seen by the model. As shown by generative adversarial examples, small distortions on a trained example can easily fool the neural network. As shown in [12], Black-box $\alpha$-divergence can reduce this effect. Li and Gal, using a re-parameterization trick, derive a new minimization objective:

$$
\hat{L}_\alpha(\theta) = -\frac{1}{C} \sum_{n=1}^{N} \log\text{sum-exp}[-\alpha p(y_i|x_i, \hat{\omega}_k)] + \text{KL}[q_\theta(\omega)||p_0(\omega)]
$$

where $q$ is the approximated posterior distribution parameterized by $\alpha$ and the log-sum-exp operating over $K$ samples from the approximate posterior $\hat{\omega}_k \sim q(\omega)$. In this formulation the dropout probabilities are held constant.

Experimentations in [12] suggest that $\alpha = 0.5$ provides a good balance between test accuracy and robustness to adversarial examples. However, outside of adversarial examples, the authors did not explicitly test for robust uncertainty quantification.

III. Experiments

To begin to understand the applicability of Bayesian deep learning to climate applications, we experiment on a statistical downscaling dataset. Our dataset, extracted from PRISM\(^1\), contains daily precipitation from years 1981 to 2015 in Orlando Florida. The features are a low-resolution 9x9 patch (1.25\(\degree\)) of precipitation and the label is precipitation at 1/16\(\degree\). Training and test data are split before and after year 2005, respectively.

Three Bayesian neural networks are then trained, MC-Dropout, Concrete Dropout, and Dropout with Alpha-Divergences, each consisting of 2 convolutional layers of 128 hidden units each and a fully-connected output layer. For MC-Dropout and Dropout with Alpha-Divergences, the dropout rate is set to 0.25, sufficiently large for regularization and enforcing reasonably wide predictive uncertainty. Ten Monte-Carlo samples are used during training of Alpha-divergence dropout and set $\alpha = 0.5$. 100 Monte Carlo samples are used to estimate the mean and variance of the learned distributions.

The models were trained with a batch size of 128 for 100 epochs using the Adam optimizer.

To evaluate the uncertainty estimates, we quantify what ratio of the samples fall within a certain interval around the prediction. As presented in Figure 1, we find that concrete dropout’s prediction intervals are more aligned with the ideal calibration ($y = x$) with MC-dropout performing better at wider intervals. Similar results are found for predictability, with MC-dropout’s daily RMSE being just slightly higher than Concrete dropout. Alpha-divergence based dropout has both a larger RMSE and lower calibration performance relative to the others.

When training Concrete-dropout, we find that the learned dropout probabilities are 0.1 and 0.14 for each layer, respectively. This results in slightly more overfitting to the training set and reduce generalizability on the test set. To minimize this effect, one often includes an $l_1$ or $l_2$ regularization of the weight parameters.

IV. Conclusion

The three approaches tested have shown impressive results when tested on out-of-sample data as well as adversarial examples [12] however the predictive uncertainty on our statistical downscaling dataset is relatively poor. Architecture selection is generally a challenging task requiring many experiments, especially in the case with no other literature on our application. Furthermore, this is the only Alpha-divergence based dropout applied on a regression problem that the authors are aware of.

At it’s current state, the method provides value in estimating uncertainty but should be used cautiously in practice as more experimentation is needed. The current formulations of these methods assume the output variable follows a normal distribution, but in many climate applications we know this assumption will

---

\(^1\)PRISM Climate Group, Oregon State University
not hold. The authors hypothesize that an adaptation to a well-suited distribution and network architecture will improve performance, especially at the extremes. Furthermore, the relative uncertainty estimates between multiple samples may provide value in understanding where the model is less confident, which aid in understanding where the performs poorly. Also, as is shown in the literature, these methods can provide more robust predictions when compared to vanilla dropout techniques. For these reasons, we will follow up this work with experiments in statistical downscaling with a modified likelihood distribution.

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, and National Science Foundation BIG-DATA under grant number:1447587. The authors acknowledge Upmanu Lall for his valuable comments.

REFERENCES