Distribution Regularized Regression Framework for Climate Modeling

Zubin Abraham* Pang-Ning Tan† Perdinan‡ Julie A. Winkler§ Shiyuan Zhong¶
Malgorzata Liszewska‖

Abstract

Regression-based approaches are widely used in climate modeling to capture the relationship between a climate variable of interest and a set of predictor variables. These approaches are often designed to minimize the overall prediction errors. However, some climate modeling applications emphasize more on fitting the distribution properties of the observed data. For example, histogram equalization techniques such as quantile mapping have been successfully used to debias outputs from computer-simulated climate models to obtain more realistic projections of future climate scenarios. In this paper, we show the limitations of current regression-based approaches in terms of preserving the distribution of observed climate data and present a multi-objective regression framework that simultaneously fits the distribution properties and minimizes the prediction error. The framework is highly flexible and can be applied to linear, nonlinear, and conditional quantile models. The paper demonstrates the effectiveness of the framework in modeling the daily minimum and maximum temperature as well as precipitation for climate stations in the Great Lakes region. The framework showed marked improvement over traditional regression-based approaches in all 14 climate stations evaluated.

Keywords: Regression; Regularization.

1 Introduction

There are numerous climate modeling applications that can be cast into a regression problem, from projecting future climate scenarios to downscaling the coarse-scale outputs from global/regional climate models for climate change impact assessment and adaptation studies [2, 12, 17]. In addition to minimizing the residuals of the predicted outputs, some of these applications emphasize preserving specific characteristics of the predicted distribution. However, as most regression-based approaches are designed to optimize the former, they tend to perform poorly on the latter criterion.

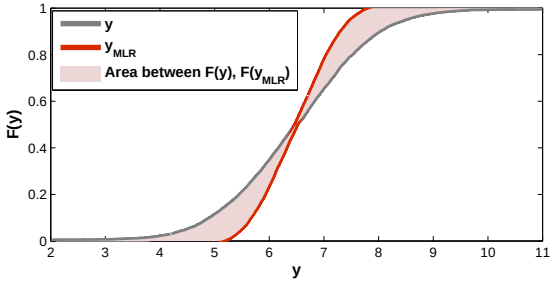

Figure 1: Area between the CDF of y and y_{MLR}.

As an illustration, consider a two-dimensional regression problem, where the response variable y is related to the predictor variables $\mathbf{x}$ according to the following equation: $y = \omega^T\mathbf{x} + \omega_0 + \epsilon(0, \sigma^2)$, where $\Omega = [\omega_2\omega_1\omega_0] = [1, 2, 5]$. Using the least square (maximum likelihood) estimation approach, multiple linear regression (MLR) was able to fairly accurately estimate Ω as [0.99, 1.96, 5.05]. Yet, it fared poorly in terms of replicating the shape of the original distribution of y as seen from its cumulative distribution function (CDF) plots given in Figure 1. Even though the regression model was trained using ten thousand data points, it is clear from Figure 1 that MLR fails to replicate the shape of the cumulative distribution for y, particularly the tails of the distribution.

As another example, Figure 2 compares the histograms of daily maximum temperature observed at a climate station in Michigan and the predicted outputs of MLR. In this case, the standard deviation of MLR's predicted outputs differs quite substantially from that of observation data. In spite of minimizing the sum of squared prediction error, regression-based approaches such as MLR fared poorly in preserving the overall shape of the distribution compared to non-regression based approaches such as quantile mapping (QM), which had an RMSE value 25% worse than that of MLR but gives a better fit to the distribution of maximum temperature. As a consequence, distribution-

*Michigan State University, abraha84@msu.edu.
†Michigan State University, ptan@cse.msu.edu.
‡Michigan State University, perdinan@msu.edu.
§Michigan State University, winkler@msu.edu.
¶Michigan State University, zhongs@msu.edu.
‖University of Warsaw, m.liszewska@icm.edu.pl.

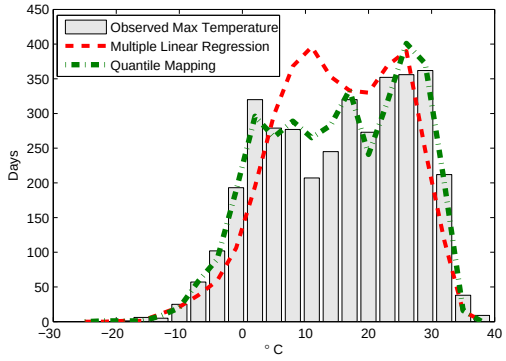

Figure 2: Histogram of predicted daily maximum temperature at a weather station in Michigan, 1990-1999.

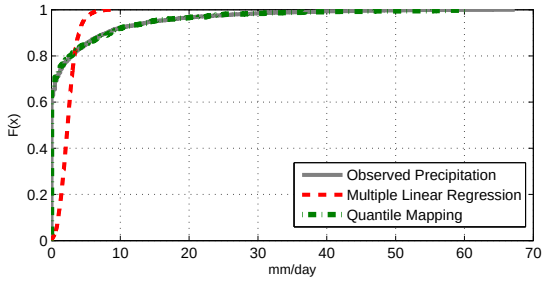

Figure 3: CDF of predicted daily precipitation at a weather station in Michigan, 1990-1999.

driven approaches [16, 15] have been used to correct the distribution characteristics of the data to better match the observed climate variable. However, their prediction accuracy is typically worse than regression-based approaches.

Raw projections of climate variables are often obtained from General Circulation Models (GCM) and more recently from Regional Climate Models (RCM) that incorporate complex topography, land cover, and other regional forcings into the physical models. These raw climate projections need to be further post-processed to meet the requirements of impact assessment studies. In addition to the previously mentioned requirements from the climate variables, empirical downscaling of the output from the climate models to a finer resolution is often needed to bridge the mismatch in spatial or temporal scale between the model output and the scale desired, since the resolutions of the output from the climate models may remain too coarse for many applications where local scale information is needed. Similarly, bias correction is often needed to reduce the inherent uncertainties in the RCM outputs that may be afflicted by the systematic errors introduced by the driving GCM runs, imperfections of the RCM representation, and sampling biases due to the finite length time series used to parameterize and validate the models [10].

Since the fidelity of both the distribution characteristics and the accuracy of projections are important, we propose a framework for multivariate regression that regularizes the distribution of the response variable to simultaneous improve the accuracy of the projection as well as the shape of the distribution by jointly solving both objectives. Due to its generic nature, the framework may be applied to various types of marginal distributions as well as different objective function criteria including least square error, kernel regression and quantile regression (QR). In this paper, we also demonstrate the effectiveness of the proposed framework by downscaling and bias correcting daily temperature and precipitation to match their corresponding observations.

In summary, the main contributions of this study are:

- We identified the limitations of existing least squared error regression techniques.
- We presented a regression based framework (Contour Regression) for multivariate empirical downscaling and bias correction that address the limitation of existing approaches by simultaneously improving accuracy of projection for individual data points as well as the overall shape of the distribution.
- We demonstrated the feasibility of adapting the framework to fit various objective functions such as multivariate ordinary least squares, QR and nonlinear kernel ridge regression.
- We evaluated the framework on real world climate data and found that it consistently outperformed or was at least on-par with the baseline approaches and showed its robustness to response variables having different types of shapes of distribution.

The remainder of the paper is organized as follows. Section 2 reviews bias correction techniques evaluated in this study. Section 3 introduces the notations and concepts used in the paper. Section 4 elaborates on the proposed framework, while Section 5 provides the experimental results and discussions comparing the relative skills of the framework on real world climate data obtained from different National Center for Environmental Prediction (NCEP) driven RCMs. The relevant information on data pre-processing is also detailed in this section. This is followed by our conclusions.

2 Related Work

With the increasing availability of climate models courtesy of projects such as NARCCAP (North American Regional Climate Change Assessment Program), extensive research has been done to better utilize the long term future climate projections made by these models [6]. Most research uses these projections to focus on the impact assessment of climate change on a wide range of domains ranging from natural ecosystems [14] [11] to those related to human systems [13]. Efficient utilization of these climate models requires downscaling and bias correcting surface climate variables [7, 8, 9].

Bias correction and downscaling approaches are motivated by the need to address biases in the climate model and address the relative coarse spatial resolution of the output from the climate model. Some of the common distribution driven approaches include QM [16], Equidistant CDF Matching (EDCDFm) and the transfer functions proposed by Piani et al.(2010b). These approaches are best suited when there is no day-to-day mapping available between the climate model output and observation as is the case of downscaling from GCMs or data from RCMs driven by GCMs. Unfortunately, as mentioned earlier, these approaches underperform in terms of accuracy of prediction of individual data points. This is because these approaches do not leverage the original mapping information between the response and predictor variables during training. This drawback is all the more important, since data obtained from RCMs driven by reanalysis data have day-to-day mapping and may be used for building a regression model for downscaling and bias correction.

Regression based approaches such as MLR and analog methods [16] are examples of accuracy driven approaches. These approaches provide the best accuracy for the projection of individual data points but fare poorly in terms of capturing the shape of the distribution (Figures 2 and 3). Since many climate change impact assessment studies are interested in long-term projections and use projections from climate models that are a realization of a potential scenario, distribution of the projection is often utilized as input.

3 Preliminaries

Let $D = \{(x_i, y_i)\}_{i=1}^n$ be a labeled dataset of size n, where each $x_i \in \mathcal{R}^d$ is a vector of predictor variables and $y_i \in \mathcal{R}$ the corresponding response variable. The objective of regression is to learn a target function $f(x, \beta)$ that best estimates the response variable y. β is the parameter vector used by the target function. n represents the number of training points.

3.1 Multiple linear regression (MLR) MLR is the most common regression approach used for empirical downscaling of climate data. MLR uses ordinary least squares to solve a linear model of the form

$$y = x^T\beta + \boldsymbol{\epsilon}$$

where, $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is an i.i.d Gaussian error term with variance σ^2. $\boldsymbol{\beta} \in \mathcal{R}^d$ is the parameter vector. MLR minimizes the sum squared residuals $(y - X\beta)^T(y - X\beta)$ which leads to a closed-form expression for the solution

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^Ty$$

3.2 Quantile Mapping (QM) Quantile mapping is the most commonly used approach for correcting the shape of the distribution of a climate variable to match observations. It adjusts all the moments of the distribution while maintaining the rank correlation. The following equation is an example of the QM approach.

$$QM : y_i = F_Y^{-1}(F_X(x_i))$$

$F_X(x)$ is a function that corresponds to the CDF for the predictor variable 'X' and is defined by $F_X(x) = P(x \leq X)$. The above equation of QM may be rewritten as follows to help identify the correction made by QM.

$$QM : y_i = x_i + F_Y^{-1}(F_X(x_i)) - F_X^{-1}(F_X(x_i))$$

One of the main assumptions made by QM is that the data points upon which the bias correction function is to be applied come from the same distribution that describes the training sets and that the relationship between predictor and response is constant. Also, a sufficiently large enough training size is required by QM to capture the true shape of the distribution of the model and observations. A distinct advantage of QM is that no day-to-day mapped data are required.

4 Framework for Multivariate Contour Regression (CR)

Since regression based approaches have a distinct advantage in terms of prediction accuracy of individual data points but are limited by their lack of emphasis on the shape of the distribution of the projection as depicted by the area between their two CDFs in Figure 1, there is a need to regularize the area between the CDF of the target response variable and the regression result. The proposed distribution regularized framework is

$$\min_{\beta} \sum_{i=1}^{n} (\gamma\pi(f(x_i), y_i) + (1 - \gamma)\pi(f(x_i), y_{(i)}))$$

where, $y_{(i)}$ corresponds to the i-th order value of the target response variable$y$. $\pi(.,.)$ can be any generic loss

function, such as sum squared error, while $0 \leq \gamma \leq 1$ is a user defined parameter that may be used for either prioritizing fidelity in regression accuracy or its CDF.

An important required preprocessing step (elaborated in the following subsection) required, is that the predictor matrix X is pre-sorted such that $i < j \quad \forall f(x_i, \beta) \leq f(x_j, \beta)$. The choice of π determines the objective function that is to be minimized and could be as simple as ordinary least squares or a more complex user defined function. Section 4.1 elaborates on CR and describes multivariate linear contour regression (MLCR) which has an objective function that is based on ordinary least squares. Section 4.2 proposes kernel contour regression (KCR) that is a kernel-based interpretation of the CR framework. Section 4.3 proposes a quantile regression based interpretation that emphasizes the conditional quartile of the user's preference. In this study, the conditional quartile chosen corresponded to the extreme fifth percentile of the distribution.

4.1 Multiple Linear Contour Regression (MLCR) This section describes an approach for CR that is based on ordinary least square (OLS) to simultaneously regress on the response variable as well as regress on the ordered value of the response variable by minimizing the sum squared error, as shown below.

$$\sum_{i=1}^{n} (\gamma(f(x_i, \beta) - y_i)^2 + (1-\gamma)(f(x_i, \beta) - z_i)^2)$$

where, $f(X, \beta) = X\beta$ and $z_i = y_{(i)}$. This equates to minimizing

$$\gamma(y - X\beta)^T(y - X\beta) + (1-\gamma)(z - X\beta)^T(z - X\beta)$$

where the predictor matrix X is pre-sorted such that $i < j \quad \forall f(x_i, \beta) \leq f(x_j, \beta)$ and $\gamma \in [01]$ is a user defined parameter that may be used for either prioritizing fidelity in regression accuracy or shape of the distribution. It is obvious from the equation that as $\gamma \to 1$, MLCR converges to the solution of MLR as seen in Figure 4, which depicts the influence of the γ parameter on the shape of the CDF of the response variable. The closed form solution to MLCR is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1}(\gamma X^T y + (1-\gamma) X^T z)$$

Since it is often not possible to guarantee that X is pre-sorted correctly according to $f(x_i, \beta)$, we may need to iteratively solve the objective function after reordering the data points X and corresponding y, such that the new ordering of the data points conforms to $i < j \quad \forall f(x_i, \beta) \leq f(x_j, \beta)$ based on the β obtained from the previous iteration, until convergence. Convergence

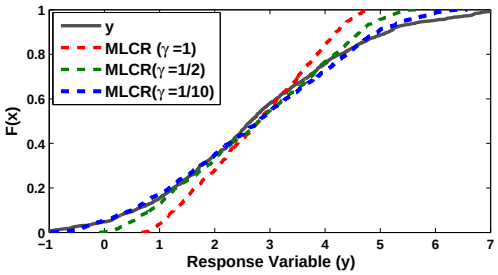

Figure 4: Influence of gamma parameter on fidelity of the response variable's cumulative distributive function.

is obtained when $\forall f(x_i, \beta) \leq f(x_j, \beta), \quad \forall i < j$. As shown in the theorem below, the following objective function converges with each iteration.

4.1.1 Proof of Convergence This section presents the proof of convergence of the iterative update algorithm. Let β_t, f_t, X_t be the regression coefficients, predicted values for the response variable and the predictor variables at the t-th iteration, while $\beta_{t+1}, f_{t+1}, X_{t+1}$ represent the regression coefficients, predicted values for the response variable and the predictor variables after the $(t+1)$-th iteration.

PROPOSITION 1. *Assuming that the indices of the predictor variables are fixed,* $L(\beta_t, f_t, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_t)$.

Proof. For a fixed X_t, $L(\beta_{t+1}, f_{t+1}, X_t) \leq L(\beta_t, f_t, X_t)$ since the β_{t+1} is obtained from a closed form solution of ordinary least squares and by definition is the solution that minimizes the objective function. In the worst case, $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_t, f_t, X_t)$.

PROPOSITION 2. *Assuming that the regression coefficients* β *are fixed,* $L(\beta_{t+1}, f_{t+1}, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_{t+1})$

Proof. Let $L(\beta_{t+1}, f_{t+1}, X_t) = L^y_{t+1} + L^z_t$ where, L^y_{t+1} refers to the first half of the loss function that regresses on y and L^z_t refers to the second half of the loss function that regresses on z. Since, the change in ordering of X from t-th to the $t+1$-th iteration doesn't impact the L^y component of the loss function, and $L(\beta_{t+1}, f_{t+1}, X_{t+1}) = L^y_{t+1} + L^z_{t+1}$, we shall concentrate on L^z. $L^z_t = (1-\gamma)\sum_{i=1}^{n}(f(x_i, \beta) - z_i)^2$ which can be rewritten as

$$L^z_t = \sum_{i=1}^{n}(f_i^2 + z_i^2 + 2 f_i z_i)$$

$(1-\gamma)$ being a constant, is ignored for simplicity. Given that β and values for f are fixed, $L^z_{t+1} = \sum_{i=1}^{n}(f^2_{(i)} + z_i^2 + 2f_{(i)}z_i)$.

$$\Rightarrow L^z_t - L^z_{t+1} = \sum_{i=1}^{n}(f_{(i)}z_i - f_i z_i)$$

And since, $\sum_{i=1}^{n} a_{(i)}b_{(i)} \geq \sum_{i=1}^{n} a_i b_i \quad \forall a \in R^n, b \in R^n$ we have $\sum_{i=1}^{n}(f_{(i)}z_i) \geq \sum_{i=1}^{n}(f_i z_i)$, since by definition, $z_i = z_{(i)}$.

$$\Rightarrow L^z_t - L^z_{t+1} \geq 0$$
$$\Rightarrow L(\beta_{t+1}, f_{t+1}, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_{t+1})$$

THEOREM 4.1. *The objective function $L(\beta)$ is monotonically non-increasing given the update formula for β, f and X.*

Proof. The update formula iteratively modifies the objective function as follows: $L(\beta_t, f_t, X_t) \Rightarrow L(\beta_{t+1}, f_{t+1}, X_t) \Rightarrow L(\beta_{t+1}, f_{t+1}, X_{t+1})$. Using the above propositions, we have $L(\beta_t, f_t, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_t)$ and $L(\beta_{t+1}, f_{t+1}, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_{t+1})$.

$$\Rightarrow L(\beta_{t+1}, f_{t+1}, X_{t+1}) \leq L(\beta_t, f_t, X_t)$$

LEMMA 4.1. *The objective function will eventually converge, as the value of the loss function is always non-negative and since we know $L(\beta)$ is monotonically decreasing.*

4.2 Kernel Contour Regression (KCR) A variant of MLR, called ridge regularization is used to mitigate over-fitting in regression. Ridge regression also provides a formulation to overcome the hurdle of a singular covariance matrix X^TX that MLR might be faced with during optimization. Unlike the loss function of MLR, the loss function for ridge regression is

$$(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta,$$

and its corresponding closed-form expression for the solution is

$$\hat{\boldsymbol{\beta}} = (X^TX + \lambda I)^{-1}X^Ty$$

where, the ridge coefficient $\lambda > 0$ results in a non-singular matrix $X^TX + \lambda I$ always being invertible. The dual ridge regression is given by the equation

$$\hat{\boldsymbol{\alpha}} = y^T(G + \lambda I)^{-1}X$$

where, $G = XX^T$. By mapping ϕ the predictor variable X to a higher dimension feature space F, i.e.,

$$\phi : X \in R^d \rightarrow F \subseteq R^N$$

where $N >> d$, one can transform the regularized least square regression to feature space F using the Kernel K. Similarly, the predictor variables of CR can be mapped to a higher dimension feature space F by using the ridge counterpart of MLCR.

$$\boldsymbol{\beta} = (\phi(X)^T\phi(X) + \lambda I)^{-1}(\gamma\phi(X)^Ty + (1-\gamma)\phi(X)^Tz)$$
$$\Rightarrow \beta = \lambda^{-1}\phi(X)^T(\gamma y + (1-\gamma)z - \phi(X)\beta) = \phi(X)^T\alpha$$
$$\Rightarrow \alpha = (G + \lambda I)^{-1}(\gamma y + (1-\gamma)z)$$

where, $G = \phi(X)\phi(X)^T$, $G_{ij} = \langle\phi(x_i), \phi(x_j)^T\rangle = K(x_i, x_j)$.

4.3 Quantile Contour Regression (QCR) Most regression approaches that are used for downscaling focus on predicting the conditional mean of the response variable. Predicting the conditional mean is not well suited for predicting extreme values that are better identified by the conditional quantiles that corresponds to the extreme values. Hence, unlike the common regression techniques mentioned earlier, approaches similar to quantile regression(QR) [3] are better suited to estimate the extremes of Y.

To estimate the τ^{th} conditional quantile $Q_{Y|X}(\tau)$, QR minimizes an asymmetrically weighted sum of absolute errors using the loss function:

$$\sum_{i=1}^{n}\rho_\tau(y_i - x_i^T\beta)$$

where,

$$\rho_\tau(u) = \begin{cases}\tau u & u > 0\\ (\tau-1)u & u \leq 0\end{cases}$$

and the τ^{th} quantile of a random variable Y is given by:

$$Q_Y(\tau) = F^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$$

where, $F_Y(y) = P(Y \leq y)$ is the distribution function of a real valued random variable Y and $\tau \in [0, 1]$.

Linear programming is used to solve the loss function by converting the problem to the following form.

$$\min_{u,v} \quad \tau 1_n^T u + (1-\tau)1_n^T v$$
$$\text{s.t.} \quad y - x^T\beta = u - v$$

where, $u_i \geq 0$,$v_i \geq 0$ and $\beta \in R^d$.

The objective function of QR can be adopted by CR to obtain the following loss function

$$\sum_{i=1}^{n}(\rho_{\tau_1}(y_i - x_i^T\beta) + \rho_{\tau_2}(z_i - x_i^T\beta))$$

where,

$$\rho_\tau(u) = \begin{cases} \tau u & u > 0 \\ (\tau - 1)u & u \leq 0 \end{cases}$$

which equates to

$$\min_{u,v,u',v'} \quad \tau_1 1_n^T u + (1-\tau)1_n^T v + \tau_2 1_n^T u' + (1-\tau)1_n^T v'$$

$$\text{s.t.} \quad y - x^T\beta = u - v$$

$$\text{s.t.} \quad z - x^T\beta = u' - v'$$

where, $\tau_2 = 0.5$, $u_i \geq 0$, $u_i' \geq 0$, $v_i \geq 0$, $v_i' \geq 0$ and $\beta \in R^d$.

4.3.1 Proof of Convergence Let β_t, f_t, X_t be the regression coefficients, predicted values for the response variable and the predictor variables at the t-th iteration, while $\beta_{t+1}, f_{t+1}, X_{t+1}$ be the regression coefficients, predicted values for response variable and the predictor variables after the $(t+1)$-th iteration.

PROPOSITION 3. *Assuming that the indices of the predictor variables are fixed,* $L(\beta_t, f_t, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_t)$.

Proof. For a fixed X_t, $L(\beta_{t+1}, f_{t+1}, X_t) \leq L(\beta_t, f_t, X_t)$ since β_{t+1} is the solution that minimizes the objective function. In the worst case, $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_t, f_t, X_t)$.

PROPOSITION 4. *Assuming that the regression coefficients* β *are fixed,* $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_{t+1}, f_{t+1}, X_{t+1})$

Proof. Let $L(\beta_{t+1}, f_{t+1}, X_t) = L_{t+1}^y + L_t^z$ where, L_{t+1}^y refers to the first half of the loss function that performs QR on y and L_t^z refers to the second half of the loss function that performs QR on z. Since, the change in ordering of X doesn't impact L^y we shall concentrate on L^z. Given, $L_t^z = 0.5\sum_{i=1}^n (f_i - z_i)$ and $L_{t+1}^z = 0.5\sum_{i=1}^n (f_{(i)} - z_i)$

$$\Rightarrow L_t^z = L_{t+1}^z$$

Hence, $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_{t+1}, f_{t+1}, X_{t+1})$

THEOREM 4.2. *The objective function* $L(\beta)$ *is monotonically non-increasing given the update formula for* β, f *and* X.

Proof. The update formula iteratively modifies the objective function as follows: $L(\beta_t, f_t, X_t) \Rightarrow L(\beta_{t+1}, f_{t+1}, X_t) \Rightarrow L(\beta_{t+1}, f_{t+1}, X_{t+1})$. Using the above propositions, we have $L(\beta_t, f_t, X_t) \geq L(\beta_{t+1}, f_{t+1}, X_t)$ and $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_{t+1}, f_{t+1}, X_{t+1})$.

$$\Rightarrow L(\beta_{t+1}, f_{t+1}, X_{t+1}) \leq L(\beta_t, f_t, X_t)$$

5 Experimental Results

The objective of the experiments was to evaluate the effectiveness of CR on observed climate data.

5.1 Data All the algorithms were run using climate data obtained at fourteen weather stations in Michigan, USA. Daily maximum temperature (T), minimum temperature (t), and precipitation (P) were the three climate target variables evaluated.

The predictor variables used in this study were obtained from the North American Regional Climate Change Assessment Program (NARCCAP) [1] (Table 1). Nine different data sets are used that correspond to the combination of three different RCMs and three target variables. The three RCMs used are the Canadian Regional Climate Model (CRCM), the Weather Research and Forecasting Model (WRFG) and the Regional Climate Model Version-3 (RCM3) The models were each driven by NCEP/DOE AMIP-II Reanalysis (NCEP) for a domain covering the United States and Canada. The data for the RCMs spans the period 1980-1999. The gridded RCM data have a spatial resolution of 50km. Unlike observation data that relate to a point location, RCM data are available at grid resolution with the value representing a grid-cell average.

Table 1: List of predictor variables from each RCM.

Predictor variables	Frequency
Meridional Surface Wind Speed	3 hourly
Zonal Surface Wind Speed	3 hourly
Minimum Surface Air Temperature	Daily
Maximum Surface Air Temperature	Daily
Surface Air Temperature	3 hourly
Surface Pressure	3 hourly
Precipitation	3 hourly
Surface Specific Humidity	3 hourly
500 hPa Geopotential Height	3 hourly

Since the observation data used correspond to daily values, preprocessing was also done to convert the three hour reanalysis-driven RCM data to daily values. Preprocessing was also needed for conversion of the observation data as well as data from the various RCM runs to the same units. For instance, precipitation in the observation data was in millimeters while precipitation data obtained from the various RCM runs was recorded in MKS units of $kg/m^2/s$ and needed to be converted to millimeters. In the event of missing values in the reanalysis/GCM-driven RCM simulated data the whole day corresponding to the data point was removed during the training phase of the various BCED approaches that were evaluated in this study, even if the missing value

corresponded to only a three hour time stamp for a particular day.

5.2 Experimental setup Twenty year (1980-1999) model data from the various RCM models along with the corresponding observation data were split into two parts of ten contiguous years that were used for training and testing. The results provided in this section are those observed during out-of-sample evaluation only. A 10-fold cross validation approach for comparing the performance of the various BCED models was also evaluated. But since the climate models' ability to reproduce climate variability is typically averaged over the order of ten years for the purpose of analysis, as noted by Ehret et al.(2012), and the relative performances being consistent across the two set-ups, the results of 10-fold cross validation are not included in this paper.

For the purpose of the evaluation of the relative skill in bias correction and downscaling of the proposed approach, popular BCED approaches such as MLR, Lasso, QM, PHC, LOCI, QR, kernel regression were used as baselines. For simplicity, the parameter γ was fixed across every station. Throughout this paper, we define the extreme 5 percentile of a distribution as extreme values. Consequently, 0.95 is used as τ for QR based experiments that model extreme precipitation and extreme maximum temperature, while 0.05 is used as τ for modeling extreme minimum temperature. Radial basis function (RBF) kernel was the choice of kernel used in this study. For the CR based experiments, the maximum number of iterations was set to ten.

5.3 Results The motivation behind the experiments was to evaluate the different algorithms in terms of accuracy of the prediction, the fidelity of the shape of the distribution to observation, the timing of the extreme events and the frequency with which a data point is predicted to be an extreme data point. We compared the performance of MLCR using MLR, ridge regression (Ridge), lasso regression (Lasso), QM, LOCI and fitted histogram equalization. Similarly, QCR was compared to baseline approaches such as MLR, QM, QR. Auto regressive baselines were not used as baselines as they are not well suited for long term climate projections (40-100 years into the future).

Since regression emphasizes minimizing the residuals, we started by comparing MLCR to its baseline for potential loss in root mean square error (RMSE) performance and put it in perspective of the improvement over baseline CDFs. Barring possible over-fitting, MLR should by definition of its objective function have minimum SSE among the linear regression approaches. Hence, we use MLR as a baseline to evaluate possible

Table 2: Relative performance gain of MLCR over baseline approaches.

	RMSE % loss		RMSE-CDF % gain		RMSE-CDF win-loss %	
Dataset	MLR	Lasso	MLR	Lasso	MLR	Lasso
WRFG-T	1.9	1.7	39.0	41.7	100	100
CRCM-T	2.8	2.6	25.8	28.0	100	100
RCM3-T	2.0	1.8	35.3	39.2	100	100
WRFG-t	1.0	0.6	51.4	53.7	100	100
CRCM-t	1.9	1.6	38.2	40.1	100	100
RCM3-t	1.8	1.6	53.2	56.1	100	100
WRFG-P	28.8	28.3	74.3	75.8	100	100
CRCM-P	25.8	25.0	71.1	73.2	100	100
RCM3-P	29.9	29.5	75.6	76.7	100	100

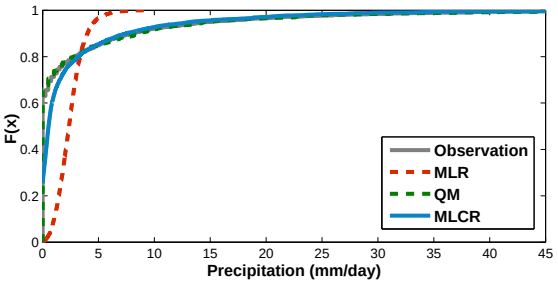

Figure 5: CDF of predicted daily precipitation at a weather station in Michigan over the years 1990-99.

deterioration in terms of RMSE by MLCR on account of MLCR's distribution regularization. MLCR showed an average deterioration in RMSE of about $< 3\%$ across the first six data sets (target variables maximum and minimum temperature) (Table 2) while improving the average error in terms of empirical cumulative distribution frequency (RMSE-CDF), around 40% (Figure 6).

Given, RMSE-CDF $= \sqrt{\frac{\sum_{i=1}^{n}(y'_{(i)}-f'_{(i)})^2}{n}}$ and its results are in the same order as RMSE, it is clear that MLCR was able to considerably improve the shape of the distribution to better match the observations at the expense of a marginal deterioration in RMSE. This improvement was observed across all climate stations within each dataset. as shown by the 100% win-loss percentage (Table 2). Ridge and Lasso fared comparably well to MLR, while QM had the worst RMSE, as expected.

MLR fared considerably worse in terms of its CDF, when it came to modeling precipitation (Gamma distribution) (Figure 5). Since, MLR struggled to capture the shape of the precipitation distribution, we chose a smaller value for the γ parameter for MLCR

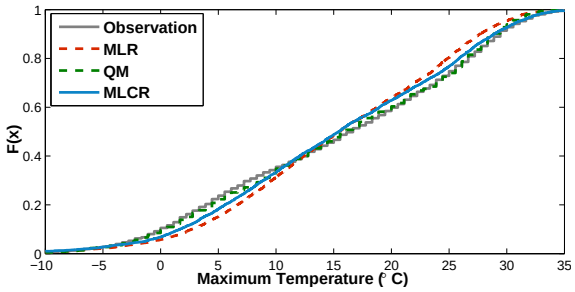

Figure 6: CDF of predicted daily maximum temperature at a weather station in Michigan, 1990-99.

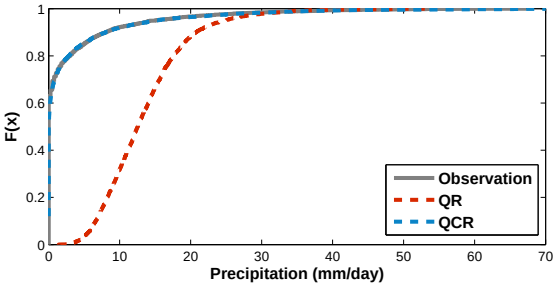

Figure 7: CDF of predicted daily precipitation at a weather station in Michigan, 1990-99.

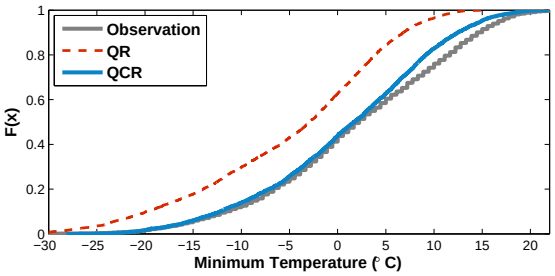

Figure 8: CDF of predicted daily minimum temperature at a weather station in Michigan, 1990-99.

than was used for the previous datasets (normal distribution) to better fit the observations' CDF. Consequently, the increase in the deterioration in terms of RMSE performance came at the expense of an impressive average RMSE-CDF improvement $> 70\%$. For evaluation of similarity of distributions, we used the Kolmogorov-Smirnov statistic (K), which for a given pair of cumulative distribution function $F_1(x)$ and $F_2(x)$ is $\max(|F_1(x) - F_2(x)|)$, the standard deviation σ, correlation(ρ) and correlation-CDF($\rho - CDF$), which measures the correlation between two CDFs. MLCR regularly outperformed the baseline regression approaches at every station (Table 3), while QM produces the most accurate standard deviation. However, MLCR was able to catch up with QR in terms of $\rho - CDF$, especially for precipitation due to the emphasis given to the distribution driven term in the experiments.

Table 3: Percentage of stations that MLCR outperformed baseline in terms of σ and $\rho - CDF$

	σ win-loss%			$\rho - CDF$ win-loss%		
Dataset	MLR	Lasso	QM	MLR	Lasso	QM
WRFG-T	100	100	0	100	100	0
CRCM-T	100	100	0	100	100	0
RCM3-T	100	100	0	100	100	0
WRFG-t	100	100	0	78.6	85.8	64.3
CRCM-t	100	100	0	92.9	100	35.8
RCM3-t	100	100	0	92.9	85.8	85.7
WRFG-P	100	100	7.1	100	100	28.6
CRCM-P	100	100	0.0	100	100	50.0
RCM3-P	100	100	7.1	100	100	64.3

5.3.1 QCR results In addition to the above-mentioned metrics for comparison, QCR was compared with baseline approaches such as MLR, QM and QR, in terms of its performance at extremes of the distributions. In terms of the RMSE for the extreme valued data point alone, QCR was able to outperform MLR, since MLR tended to underestimate the extremes. QCR also fared very well against QR (Figure 8), where the regression models emphasized the lowest τ quantile that correspond to extreme values for the target variable (minimum temperature). It is clear that QCR emphasized accuracy in the distribution of the lower quantiles of the distribution over the higher quantiles, as expected.

Precision and recall of extreme events were computed to measure the timing accuracy of the prediction of extreme valued data points. F-measure, which is the harmonic mean between recall and precision values, is used as a score that summarizes the precision and recall results. It was also found that QCR had the best F-measure among the regression based approaches in terms of correctly identifying extreme values across all the stations. Figure 7 shows the performance of QCR on precipitation. In spite of larger value for the γ parameter of QCR compared with that used for MLCR, QCR performed better than MLCR in terms of correcting the overall shape of the distribution. This is because of the zero-inflated nature of precipitation, resulting in very few large valued data points, which have a larger

influence on the appearance of the CDF plot. As seen in Table 4, QCR regularly outperformed QR in terms of the other metrics such as correlation.

Table 4: Percentage of stations that QCR outperformed baseline approaches in terms of RMSE, F-measure, k statistic and correlation for data points considered extreme value.

Dataset	RMSE	F-Measure	k	ρ
WRFG-T	100	100	100	100
CRCM-T	100	100	100	92.9
RCM3-T	100	100	100	100
WRFG-t	100	100	100	64.3
CRCM-t	100	100	100	58.7
RCM3-t	100	100	100	78.6
WRFG-P	100	100	100	35.8
CRCM-P	100	100	100	28.6
RCM3-P	100	100	100	21.4

6 Conclusions

We propose a framework that regularizes the distribution characteristics of a variable to simultaneously improve the accuracy of individual data points as well as the shape of the distribution of the projections. We demonstrate the effectiveness of the framework when using a multivariate linear interpretation, a non-linear (RBF kernel), as well as a quantile driven interpretation in effectively capturing both the shape and accuracy of observed climate data of various climate stations located in Michigan, USA. In addition to consistently reducing day-to-day error of the projections, the framework is also shown to be flexible enough to capture different shapes from various distributions as shown in the case of Gaussian and Gamma distributions.

7 Acknowledgment

This work is supported by NSF Award CNH 0909378. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] *North American Regional Climate Change Assessment Program*, http://www.narccap.ucar.edu/

[2] C. Tebaldi and D. B. Lobell, *Towards probabilistic projections of climate change impacts on global crop yields*, Geophysical Research Letters, 35(8), 2008.

[3] R. Koenker and K. Hallock, *Quantile Regression*, J. Economic PerspectivesVolume 15, Number 4, (2001) pp. 143-156.

[4] A. E. Hoerl, R. W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, J. Technometrics, 12 (1970).

[5] R. Tibshirani, . *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B., 58 (1996), pp. 267–288.

[6] C. Monteleoni, G. Schmidt, and S. Saroha, *Tracking Climate Models*, NASA Conference on Intelligent Data Understanding (CIDU), 2010.

[7] S. Charles, B. Bates, I. Smith, and J. Hughes, *Statistical downscaling of daily precipitation from observed and modelled atmospheric fields*, in Hydrological Processes, (2004) pp. 1373–1394.

[8] Z. Abraham and P. N. Tan, *An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data*, in Proc of the ACM SIGKDD Int'l Conf on Data Mining, Colorado, OH, 2010.

[9] R. Wilby, S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, *Guidelines for use of climate scenarios developed from statistical downscaling methods*, Available from the DDC of IPCC TGCIA, 2004.

[10] U. Ehret, E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert, *HESS Opinions:Should we apply bias correction to global and regional climate model data?*, J. Hydrol. Earth Syst. Sci. Res., 9(2012). pp. 5355-5387.

[11] M. Fogarty, L. Incze, K. Hayhoe, D. Mountain, and J. Manning, *Potential climate change impacts on Atlantic cod (Gadus morhua) off the northeastern USA, Mitigation and Adaptation Strategies for Global Change*,13(5-6) (2008), pp 453–466.

[12] K. Hayhoe, S. Sheridan, S. Greene, and L. Kalkstein, *Climate change, heat waves, and mortality projections for Chicago*, J. Great Lakes Research, 36(2) (2010), pp. 65-73

[13] J. E. O. Norena, *Vulnerability of water resources in the face of potential climate change: generation of hydroelectric power in Colombia*, Atmosfera, 22(3) (2009),pp. 229–252.

[14] S. V. Ollinger, C. L. Goodale, K. Hayhoe, and J. P. Jenkins, *Potential effects of climate change and rising CO2 on ecosystem processes in northeastern U.S. forests*, Mitigation and Adaptation Strategies for Global Change, 13(5-6) (2008) pp. 467–485.

[15] C. Piani, G. P. Weedon, M. Best, S. M. Gomes,P. Viterbo, S. Hagemann, and J. O. Haerter, *Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models.* J. Hydrology, 395(3-4) (2010), pp. 199–215.

[16] M. J. Themeβl, A. Gobiet, and A. Leuprecht, *Empirical-statistical downscaling and error correction of daily precipitation from regional climate models.* International J. of Climatology, 31 (2010), pp. 15301544.

[17] D. Scott and G. McBoyle, *Climate change adaptation in the ski industry, Mitigation and Adaptation Strategies for Global Change*, 12(8) (2007), pp. 1411-1431.