# Contour Regression: A Distribution-Regularized Regression Framework for Climate Modeling

**Zubin Abraham[1]*, Pang-Ning Tan[1], Perdinan[2], Julie A. Winkler[1], Shiyuan Zhong[1] and Malgorzata Liszewska[3]**

[1]*Department of Computer Science, Michigan State University, East Lansing, MI 48823, USA*

[2]*Geophysics and Meteorology, Bogor Agricultural University, Indonesia*

[3]*Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland*

Received 10 September 2013; revised 10 January 2014; accepted 11 February 2014
DOI:10.1002/sam.11222
Published online in Wiley Online Library (wileyonlinelibrary.com).

**Abstract:** Regression methods are commonly used to learn the mapping from a set of predictor variables to a continuous-valued target variable such that their prediction errors are minimized. However, minimizing the errors alone may not be sufficient for some applications, such as climate modeling, which require the overall predicted distribution to resemble the actual observed distribution. On the other hand, histogram equalization methods, such as quantile mapping, are often used in climate modeling to alter the distribution of input data to fit the distribution of observed data, but they provide no guarantee of accurate predictions. This paper presents a flexible regression framework known as *contour regression* that simultaneously minimizes the prediction error and removes biases in the predicted distribution. The framework is applicable to linear, nonlinear, and conditional quantile models and can utilize data from heterogenous sources. We demonstrate the effectiveness of the framework in fitting the daily minimum and maximum temperatures as well as precipitation for 14 climate stations in Michigan. The framework showed marked improvement over standard regression methods in terms of minimizing their distribution bias.  Statistical Analysis and Data Mining, 2014

**Keywords:** regularized regression; contour regression

## 1. INTRODUCTION

Many postprocessing tasks in climate modeling can be cast into a regression problem—from bias correcting the climate scenarios generated by numerically simulated climate models to downscaling the coarse-scale scenarios into a finer resolution for use in climate change impact assessment studies [1–3]. In addition to minimizing the errors of the regression outputs, these tasks require the predicted distribution to preserve specific characteristics of the actual distribution. As most regression methods focus on minimizing point-wise prediction errors alone, their overall predicted distribution does not always resemble the actual distribution.

As an illustration, consider a two-dimensional data set, where the response variable $y$ is generated from 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 10 000 training examples, multiple linear regression (MLR) accurately learns the model parameters as $\hat{\Omega} = [0.99, 1.96, 5.05]$. However, it fails to replicate the shape of the original distribution, as can be seen from the cumulative distribution function (CDF) plots given in Fig. 1. We consider the difference in area between the two CDFs as *bias of the predicted distribution*.

As another example, Fig. 2 compares the histograms of daily maximum temperature observed at a climate station in Michigan and the predicted outputs of MLR. The width of the histogram obtained using MLR is narrower than the actual distribution of observation data, which implies MLR has underestimated the frequency of days with extreme high and low maximum temperatures. Alternative methods such as quantile mapping (QM) [4,5] can be applied to correct the distribution bias of any input data to match the desired output distribution. However, their prediction accuracy is typically worse than regression methods. For the example

*Correspondence to:* Zubin Abraham (abraha84@msu.edu)

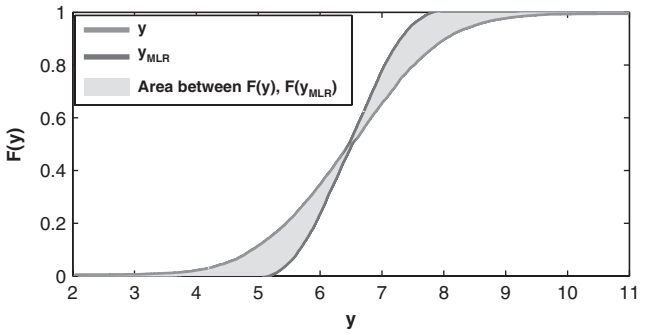


Fig. 1 Area between the CDF of $y$ and $y_{MLR}$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

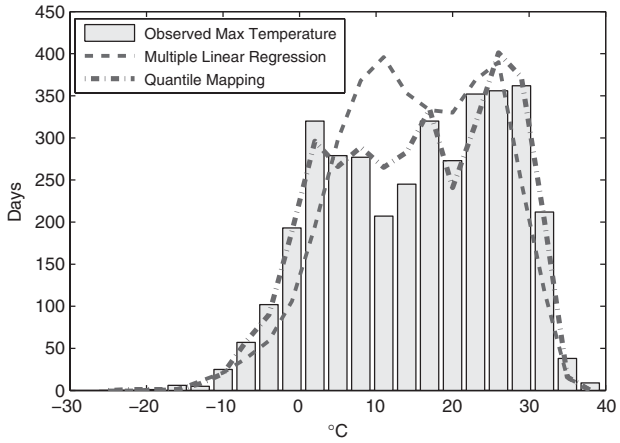


Fig. 2 Histogram of predicted daily maximum temperature at a climate station in Michigan. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

shown in Fig. 2, although QM provides a better fit to the histogram of observed data, its prediction error (RMSE) is 25% worse than that of MLR.

This paper presents a regression framework known as *contour regression* that seeks to minimize both its prediction error and bias in the predicted distribution. The framework is very flexible in that it can be easily extended to nonlinear functions and can accommodate different types of loss functions. A preliminary version of the framework was published in our conference paper [6]. In this follow-up study, we extended the framework to incorporate predictor variables from heterogeneous data sources. This is particularly useful for applications such as climate scenario development, where the predictor variables in the training and test sets are often obtained from different sources. Since their distributions can be different, the test set must be adjusted before applying the regression function estimated from the training set. The framework developed in this paper overcomes this problem by explicitly correcting the distribution mismatch between the heterogeneous data sources while minimizing the prediction error. We empirically demonstrate the effectiveness of our framework using climate data from 14 selected stations in Michigan. The framework showed marked improvement over standard regression methods in terms of minimizing biases in the predicted distribution for all the stations.

## 2. STATISTICAL POSTPROCESSING OF CLIMATE MODEL SIMULATIONS

With the increasing availability of climate models [7] courtesy of projects such as NARCCAP (North American Regional Climate Change Assessment Program) [8], extensive research has been carried out to utilize the long term future climate scenarios simulated by these models for studying the impact of climate change on human and natural ecosystems. However, there are several reasons that the scenarios must be postprocessed before they can be effectively used [9,10]. First, the scenarios are biased due to imperfections in the climate model representation and parameterization [11]. A biased scenario implies that the distribution of climate model outputs is inconsistent with observation data when simulated over historical time period. Second, the spatial resolution of the climate scenarios is often too coarse to be effectively used for climate change impact assessment studies [12,13].

Various statistical postprocessing methods have been developed to overcome these limitations. Some of these methods are *distribution-driven*, designed to adjust the distribution of the simulated model outputs to match the distribution of observation data. This includes histogram equalization techniques such as QM [4], Equidistant CDF Matching (EDCDFm), and the transfer function approaches proposed in ref. 5. These methods are applicable even when there is no day-to-day mapping between the model output and observation data. Since they focus mainly on preserving the shape of the distribution, the accuracy of their individual predictions can be poor.

By contrast, MLR and other regression methods [4] are *accuracy-driven* approaches, focusing mainly on minimizing point-wise prediction error without considering how similar is the predicted distribution to the actual distribution. These methods are trained using the climate model outputs as predictor variables and the actual observations as response variables. However, they do not account for biases between historical and future climate scenarios generated by the climate models. For example, to downscale regional climate model (RCM) simulations, the regression function is calibrated using RCM simulations driven by observation data from the National Center for Environmental Prediction (NCEP). The calibrated function is then applied to RCM simulations driven by general circulation

models (GCMs) to generate the future climate scenarios[1]. The regression function is therefore trained using predictor variables obtained from one data source but applied to predictor variables obtained from another source. Current strategy for handling this problem is to calculate a 'change factor', which represents the difference in mean or standard deviation between the two distributions. This strategy does not account for differences in higher order moments or quantiles of the distributions. The contour regression framework proposed in this paper can explicitly correct the distribution mismatch between heterogeneous data sources while minimizing the prediction error.

## 3. PRELIMINARIES

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a labeled training set of size $n$, where each $\mathbf{x}_i \in \Re^d$ is a $d$-dimensional vector of predictor variables and $y_i \in \Re$ is the corresponding response variable. The goal of regression is to learn a target function $f(\mathbf{x}, \boldsymbol{\beta})$ that best predicts the output $y$ for any given input $\mathbf{x}$. The target function is trained on $\mathcal{D}$ to learn the parameter vector $\boldsymbol{\beta}$ that minimizes a loss function $\mathcal{L}[y, f(\mathbf{x}, \boldsymbol{\beta})]$. For the remainder of this paper, we denote $\mathbf{X} = [\mathbf{x}_i^T]$ as the $(n \times d)$-dimensional input data matrix and $\mathbf{y} = [y_i]$ as the $n$-dimensional vector of response variable for training the regression function.

### 3.1. Multiple Linear Regression (MLR)

MLR is one of the most common regression approaches used for statistical downscaling of climate data. MLR uses ordinary least squares to solve a linear model of the form

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

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is a Gaussian noise term with variance $\sigma^2$. MLR assumes there is one-to-one correspondence between each $\mathbf{x}_i$ and $y_i$ in the training data. The regression parameters are estimated by minimizing the sum of squared residuals $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, which leads to the following closed-form solution

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

### 3.2. Quantile Mapping (QM)

QM can adjust the distribution of a univariate predictor variable $X$ to match the distribution of the target variable $Y$. For example, $Y$ may represent the observed maximum temperature at a climate station while $X$ is the simulated maximum temperature for its corresponding grid cell obtained from a RCM. Let $F_X$ be the CDF of $X$ and $F_Y$ be the CDF of $Y$. Assuming the CDF is continuous and strictly increasing, QM estimates the value of the target variable $y_i$ given an input $x_i$ as follows:

$$QM: \quad \hat{y}_i = F_Y^{-1}(F_X(x_i)), \tag{1}$$

where $F_Y^{-1}$ is the inverse CDF (also known as quantile function) for $Y$. QM assumes that the test data points upon which the transformation is to be applied come from the same distribution as the training data. Furthermore, a large training set is required by QM to ensure that the empirical CDFs sufficiently capture the true distribution of $X$ and $Y$. A distinct advantage of QM is that no day-to-day mapped data are required. This is particularly useful for climate scenario development, where the timing of GCM-driven RCM simulations cannot be directly mapped to the actual days of observations.

Next, we present several theoretical properties of the QM approach. Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be the set of values associated with the predictor variable $X$ and $\mathcal{Y} = \{y_i\}_{i=1}^n$ be the set of values associated with the response variable $Y$. Suppose $F_X$ and $F_Y$ are the empirical CDFs derived from $\mathcal{X}$ and $\mathcal{Y}$, respectively.

$$F_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \le x], \;\; F_Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[Y_i \le y],$$

where $\mathbf{1}[\cdot]$ is an indicator function. Let $R_i$ be the rank of a data point $x_i \in \mathcal{X}$ and $O_i$ be the rank of a data point $y_i \in \mathcal{Y}$. Thus

$$F_X(x_i) = \frac{R_i}{n}, \quad F_Y(y_i) = \frac{O_i}{n}.$$

**Proposition 1** *Let $R$ and $O$ be the multisets containing the ranks of all the data points in $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e. $R = \{R_i | F_X(x_i) = \frac{R_i}{n}, x_i \in \mathcal{X}\}$ and $O = \{O_i | F_Y(y_i) = \frac{O_i}{n}, y_i \in \mathcal{Y}\}$. If there is one-to-one correspondence (bijection) between elements of the multisets $R$ and $O$, then QM on $X$ will perfectly replicate the distribution of $Y$.*

**Proof:** Since there is a one-to-one correspondence, for each element $O_i \in O$, there is exactly one element $R_j \in R$ such that $O_i = R_j$. QM will map the data point $x_i$ to its predicted value $\hat{y}_i$ based on the following sequence of transformation:

$$x_i \xrightarrow{F_X} \frac{O_i}{n} = \frac{R_j}{n} \xrightarrow{F_Y^{-1}} \hat{y}_i = y_j$$

[1] The regression function cannot be trained with GCM-driven RCM outputs because there is no day-to-day correspondence between the model outputs and observation data.

Every value $y_j \in \mathcal{Y}$ must have a corresponding $\hat{y}_i$ and $x_i$ it was mapped from using QM. Therefore, the CDFs for $F_{\hat{Y}}$ and $F_Y$ must be identical. ■

**Proposition 2** *The sum-of-square residual error (SSE) for a QM function that perfectly replicates the distribution of the response variable can be larger than zero.*

**Proof:** Consider a labeled training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. Let $R$ and $O$ be the multisets containing the ranks of all the data points in $\{x_i\}$ and $\{y_i\}$, respectively. The sum-of-square residual error for QM is given by

$$\begin{aligned}\text{SSE} &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i \left[ F_Y^{-1}(F_Y(y_i)) - F_Y^{-1}(F_X(x_i)) \right]^2 \\ &= \sum_i \left[ F_Y^{-1}(O_i) - F_Y^{-1}(R_i) \right]^2,\end{aligned}$$

where $O_i = F_Y(y_i)$ and $R_i = F_X(x_i)$. To prove this, consider the case where the ranks are unique and $R_i = n - O_i + 1$ (i.e., the ordering in $R$ and $O$ are completely opposite of each other). Since there is a one-to-one correspondence between $O_i$'s and $R_i$'s, QM will perfectly transform $X$ to replicate the distribution of $Y$ (see Proposition 1). However, $F_Y^{-1}(O_i) - F_Y^{-1}(R_i) \neq 0$ since $O_i \neq R_i$. Thus, SSE $> 0$. ■

**Proposition 3** *If the rank correlation between the predictor variable X and response variable Y is 1, then QM on $\mathcal{X}$ will perfectly replicate the distribution of Y and its SSE is equal to zero.*

**Proof:** Let $R$ and $O$ be the multisets containing the ranks of all data points in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Since their rank correlation is 1, $\forall i : O_i = R_i$. QM will perfectly replicate the distribution of $Y$ because there is a one-to-one correspondence between the unique elements of $R$ and $O$ (see Proposition 1). Furthermore, SSE $= \sum_i [F_Y^{-1}(O_i) - F_Y^{-1}(R_i)]^2 = 0$. ■

## 4. CONTOUR REGRESSION FRAMEWORK

As previously noted, regression methods seek to maximize the prediction accuracy of individual data points but are limited by the potential bias in their predicted distribution, as depicted by the area between the two CDFs shown in Fig. 1. Our proposed framework, called contour regression, aims to minimize the bias by adding a distribution regularization term, which measures the nonoverlapping area between the empirical CDF of the target function $f(\mathbf{x}; \boldsymbol{\beta})$ and the actual distribution $y$, directly into the objective function:

$$\min_{\beta} \sum_{i=1}^n \left\{ \gamma \mathcal{L}\Big[ y_i, f(\mathbf{x}_i; \boldsymbol{\beta}) \Big] + (1-\gamma) \mathcal{L}\Big[ y_{(i)}, \hat{y}_{(i)} \Big] \right\}, \quad (2)$$

where $y_{(i)}$ is the $i$-th largest value of the response variable $y$ and $\hat{y}_{(i)}$ is the $i$-th largest value of $f(\mathbf{x}, \boldsymbol{\beta})$. $0 \leq \gamma \leq 1$ is a user-defined parameter that controls the trade-off between maximizing accuracy and minimizing bias in the predicted distribution. $\mathcal{L}$ can be any generic loss function, including sum of squared error and quantile loss. Section 4.1 presents multivariate linear contour regression (MLCR), which uses the least-square loss function. Section 4.2 introduces kernel contour regression (KCR), which is an extension of MLCR to the nonlinear case. Section 4.3 incorporates the distribution regularization term into the loss function for quantile regression.

### 4.1. Multiple Linear Contour Regression (MLCR)

This section describes a variant of CR using a linear model as its target function and a squared error loss function.

$$\sum_{i=1}^n \left\{ \gamma \Big[ y_i - f(\mathbf{x}_i, \boldsymbol{\beta}) \Big]^2 + (1-\gamma) \Big[ z_i - \hat{y}_{(i)} \Big]^2 \right\}$$

where, $z_i = y_{(i)}$ and $\hat{y}_{(i)}$ is the $i$-th largest value of $f(\mathbf{x}, \boldsymbol{\beta})$. To estimate the parameter $\beta$, we employ the following iterative approach. First, $\beta$ is initialized to be the solution of MLR. A constant vector $\mathbf{z}$ that contains the ordered values of $\{y_i\}$ is also constructed. Since the values of $\{\hat{y}_i\}$ depend on the current estimate of $\boldsymbol{\beta}$, we need to sort the values of $f(\mathbf{x}_i, \boldsymbol{\beta})$ in order to compute the distribution bias (i.e., the second term of the objective function). Therefore, at each iteration, we reorder the rows (data points) in both matrix $\mathbf{X}$ and vector $\mathbf{y}$ based on their current estimated values of $\hat{y}_i$, i.e. if the index of row $i$ is less than row $j$, then $f(\mathbf{x}_i, \boldsymbol{\beta}) < f(\mathbf{x}_j, \boldsymbol{\beta})$. Once the data points in $\mathbf{X}$ and $\mathbf{y}$ have been sorted[2], we can update $\boldsymbol{\beta}$ as follows:

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

Convergence is obtained when the ordering of the data points remains the same after updating $\boldsymbol{\beta}$.

The user-specified parameter $\gamma$ controls the trade-off between model accuracy and distribution bias. As $\gamma \to 1$,

[2] Note that the vector $\mathbf{z}$ remains fixed for all iterations.

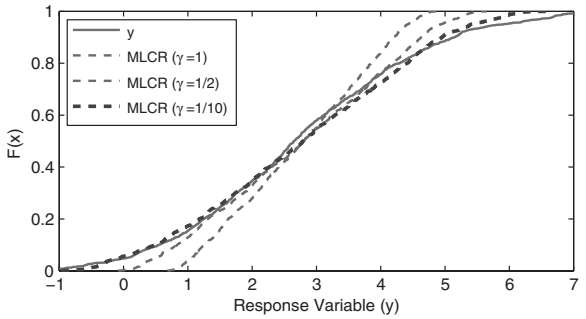


Fig. 3 Influence of the parameter $\gamma$ on the distribution bias of MLCR. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

MLCR reduces to the MLR approach. Figure 3 shows the different CDFs obtained by MLCR when applied to the data set described in Fig. 1.

**Algorithm 1** Parameter estimation for MLCR.

1) Set $z_i = y_{(i)}$
2) Initialize $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$
3) Reorder the rows of $\mathbf{X}$ and corresponding $\mathbf{y}$, such that the new indices conform to the following condition: if $i < j$, then $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \leq f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})$. (Note: $\mathbf{z}$ remains invariant.)
4) Update $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\gamma\mathbf{X}^T\mathbf{y} + (1-\gamma)\mathbf{X}^T\mathbf{z})$.
5) Repeat steps 3 and 4 until convergence.
6) Return $\hat{\boldsymbol{\beta}}$.

#### 4.1.1. *Proof of convergence*

This section presents the proof of convergence for the iterative update algorithm described in the previous section. Let $\boldsymbol{\beta}_t, \hat{\mathbf{y}}_t, \mathbf{X}_t$ be the regression parameters, predicted values, and input data matrix at the $t$-th iteration, while $\boldsymbol{\beta}_{t+1}, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1}$ represent their respective values after the $(t+1)$-th iteration. Note that $\mathbf{X}_{t+1}$ is a reordering of the rows in $\mathbf{X}_t$ (step 3 of the MLCR algorithm).

LEMMA 1: Consider an ordered vector of real numbers, $\mathbf{a} = [a_1 a_2 \cdots a_n]$, where $a_1 \leq a_2 \cdots \leq a_n$. If $\mathbf{b} = [b_1 b_2 \cdots b_n]$ is a vector of length $n$ whose elements are chosen (without replacement) from a given set of $n$ real numbers, then

$$\max_{\mathbf{b}} \ \mathbf{a}^T\mathbf{b} = \mathbf{a}^T\mathbf{b}_o = \sum_i a_{(i)}b_{(i)},$$

where $\mathbf{b}_o = [b_{(1)}b_{(2)} \cdots b_{(n)}]$, $b_{(1)} \leq b_{(2)} \cdots \leq b_{(n)}$.

**Proof:** By contradiction, assume the dot product between $a$ and $\mathbf{b}_o$ is not maximized. Assume there exists another vector $\hat{\mathbf{b}}$, which has the same values as $\mathbf{b}_o$ except its $i$th and $j$th elements are swapped, i.e. $\hat{b}_i = b_{(j)} > \hat{b}_j = b_{(i)}$, but $\mathbf{a}^T\mathbf{b}_o < \mathbf{a}^T\hat{\mathbf{b}}$. Since the two vectors differ only in terms of their $i$th and $j$th elements, therefore

$$a_{(i)}b_{(i)} + a_{(j)}b_{(j)} < a_{(i)}\hat{b}_i + a_{(j)}\hat{b}_j.$$

But

$$\begin{aligned} &a_{(i)}b_{(i)} + a_{(j)}b_{(j)} - a_{(i)}\hat{b}_i - a_{(j)}\hat{b}_j \\ &\quad = a_{(i)}b_{(i)} + a_{(j)}b_{(j)} - a_{(i)}b_{(j)} - a_{(j)}b_{(i)} \\ &\quad = (a_{(i)} - a_{(j)})(b_{(i)} - b_{(j)}) \\ &\quad \geq 0, \end{aligned}$$

which is a contradiction. Therefore $\mathbf{a}^T\mathbf{b}_o \geq \mathbf{a}^T\hat{\mathbf{b}}$. ■

LEMMA 2: If $\boldsymbol{\beta}$ is fixed and we reorder the rows of $\mathbf{X}$ and $\mathbf{y}$ in increasing order of $\hat{\mathbf{y}}$ (step 3 of MLCR algorithm), then $\mathcal{L}(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_t, \mathbf{X}_t) \geq \mathcal{L}(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1})$

**Proof:** Let $\mathcal{L}(\boldsymbol{\beta}, \hat{\mathbf{y}}, \mathbf{X}) = \gamma\mathcal{L}^{\text{SSE}} + (1-\gamma)\mathcal{L}^{\text{bias}}$, where $\mathcal{L}^{\text{SSE}}$ refers to the first term of the objective function that minimizes the prediction error and $\mathcal{L}^{\text{bias}}$ refers to the second term of the objective function that minimizes the distribution bias. Since exchanging the rows in $\mathbf{X}$ and $\mathbf{y}$ does not affect $\mathcal{L}^{\text{SSE}}$, we focus only on $\mathcal{L}^{\text{bias}}$.

$$L^{\text{bias}} = \sum_{i=1}^{n}(z_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n}(z_i^2 + \hat{y}_{(i)}^2 - 2\hat{y}_{(i)}z_i)$$

Reordering of the rows will not affect the first two terms of the sum. Hence, $L(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_t, \mathbf{X}_t) - L(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1}) = \sum_{i=1}^{n}(\hat{y}_{(i)}^{(t+1)}z_i^{(t+1)} - \hat{y}_i^{(t)}z_i^{(t)})$, where $\hat{y}_{(i)}^{(t+1)}$ are the reordered values of $f(\mathbf{x}, \boldsymbol{\beta})$ and $\hat{y}_i^{(t)}$ are the values of $f(\mathbf{x}, \boldsymbol{\beta})$ before reordering. Since $z_i^{(t+1)} = z_i^{(t)} = y_{(i)}$ and using Lemma 1, we have $\sum_{i=1}^{n}(\hat{y}_{(i)}^{(t+1)}y_{(i)} - \hat{y}_i^{(t)}y_{(i)}) \geq 0$, thus completing the proof. ■

LEMMA 3: If the rows in $\mathbf{X}$ and $\mathbf{y}$ are fixed, then $\mathcal{L}(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1}) \geq \mathcal{L}(\boldsymbol{\beta}_{t+1}, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1})$ after $\boldsymbol{\beta}$ is updated (step 4 of MLCR algorithm).

**Proof:** Since $\boldsymbol{\beta}_{t+1}$ is obtained by minimizing the objective function (assuming $\mathbf{X}_t$ and $\mathbf{y}_t$ are fixed), therefore $\mathcal{L}(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1}) \geq \mathcal{L}(\boldsymbol{\beta}_{t+1}, \hat{\mathbf{y}}_{t+1}, \mathbf{X}_{t+1})$. ■

THEOREM 1: The objective function is monotonically nonincreasing upon applying the iterative MLCR algorithm.

**Proof:** Let $\mathcal{L}(\boldsymbol{\beta}_t, \hat{\mathbf{y}}_t, \mathbf{X}_t)$ be the value of the objective function after $t$ iterations. $\boldsymbol{\beta}$ is updated after applying steps

3 and 4 of the MLCR algorithm. The proof follows directly from Lemmas 2 and 3. ■

Based on Theorem 1 and the fact that the objective function is nonnegative, the algorithm will eventually converge to a local minimum.

### 4.2. Kernel Contour Regression (KCR)

Ridge regression can be used to avoid overfitting by providing a sparse solution for $\boldsymbol{\beta}$. It can also provide a way to overcome the hurdle of taking the inverse of a covariance matrix with correlated features. By writing $\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{\alpha}$, the dual form of the ridge regression solution can be written as follows

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

Using the kernel trick, we can project $\mathbf{X}$ to a higher dimensional feature space and replace $\mathbf{X}\mathbf{X}^T$ by the Gram (kernel) matrix $\mathbf{K}$. This extends ridge regression to deal with nonlinear relationships between $X$ and $Y$. A similar approach can be used to extend CR to a nonlinear setting. Specifically, the objective function for KCR can be written as follows:

$$\mathcal{L}_{\text{KCR}} = \mathcal{L}_{\text{MLCR}} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta},$$

where $\lambda$ is the ridge parameter. Replacing $\mathbf{X}$ by $\boldsymbol{\Phi}$, we have

$$\boldsymbol{\beta} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I})^{-1}\left[\gamma\boldsymbol{\Phi}^T\mathbf{y} + (1-\gamma)\boldsymbol{\Phi}^T\mathbf{z}\right]$$

Let $\boldsymbol{\beta} = \boldsymbol{\Phi}^T\alpha$. After some manipulation, we obtain

$$\alpha = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\mathbf{I})^{-1}(\gamma\mathbf{y} + (1-\gamma)\mathbf{z})$$

### 4.3. Quantile Contour Regression (QCR)

MLR focuses on accurately predicting the conditional mean of the response variable. As illustrated in Figs 1 and 2, it is not well suited for predicting extreme values. To overcome this limitation, an alternative approach known as quantile regression (QR) [14] can be used to improve the estimate of the $\tau^{th}$ conditional quantiles of the prediction. This is achieved by minimizing the following asymmetrical loss function:

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

where $\rho_\tau(u) = \begin{cases}\tau u & u > 0\\ (\tau-1)u & u \le 0\end{cases}$ and the $\tau^{th}$ quantile of a random variable $Y$ is given by:

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

We extend CR to quantile regression by modifying the loss function as follows:

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

where $z_i = y_{(i)}$ and $\hat{y}_i$ is the $i$-th lowest value of $f(\mathbf{x}, \boldsymbol{\beta})$. Algorithm 1 can be modified to solve the quantile contour regression (QCR) problem. Instead of solving $\boldsymbol{\beta}$ in closed form in step 4, we apply linear programming to minimize the objective function for QCR (assuming $\mathbf{z}$, $\mathbf{X}$, and $\mathbf{y}$ are fixed).

## 5. CONTOUR REGRESSION FOR MULTISOURCE DATA

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\tilde{\mathcal{X}} = \{\tilde{x}_i\}_{i=1}^n$ be two different sources of predictor variables and $\mathcal{Y} = \{y_i\}_{i=1}^n$ be the values of response variables. Suppose there is a one-to-one correspondence between each instance in $\mathcal{X}$ and $\mathcal{Y}$ but an instance in $\tilde{\mathcal{X}}$ cannot be matched to an instance in $\mathcal{Y}$. For climate scenario downscaling, $\mathcal{X}$ may represent the predictor variables obtained from NCEP-driven simulations whereas $\tilde{\mathcal{X}}$ may represent the predictor variables from GCM-driven simulations. Our goal is to learn a regression function $f(\mathbf{x}, \boldsymbol{\beta})$ that can predict the response value of a previously unseen instance from the domain of $\tilde{\mathcal{X}}$.

The contour regression framework described in the previous section can be automatically applied to the preceding regression problem. The first term of the objective function in Eq. (2) is computed using $\mathcal{X}$ whereas the second term of the objective function is computed using $\tilde{\mathcal{X}}$. This is possible since the second term matches the quantiles of the distributions, and thus, does not require one-to-one correspondence between the predictor and response variables. We consider two variations of the CR framework for multisource data. The first approach incorporates the predictor variables from $\tilde{\mathcal{X}}$ directly into the second term of the objective function. We termed this approach as $\text{MLCR}_{\text{HET}}$. The second approach adjusts the distribution of $\tilde{\mathcal{X}}$ to match the distribution of $\mathcal{X}$ before applying the adjusted value to the second term of the objective function. We termed this approach as $\text{MLCR}_{\text{GQ}}$.

### 5.1. $\text{MLCR}_{\text{HET}}$

For MLCR, the second term of the objective function given in Eq. (2) measures the squared difference of the predicted and observed distribution, i.e. $\sum_i[y_{(i)} - \hat{y}_{(i)}]^2$, where $\hat{y}_{(i)}$ is the $i$-th lowest predicted value $f(\mathbf{x}, \boldsymbol{\beta})$. This approach can be easily extended to heterogeneous data by replacing $\mathbf{x} \in \mathcal{X}$ with $\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}$ and sorting the predicted values to obtain the $\hat{y}_i$'s. This approach essentially tries to

learn $\boldsymbol{\beta}$ in such a way that simultaneously minimizes the prediction error on $\mathcal{X}$ and the distribution bias on $\tilde{\mathcal{X}}$.

### 5.2. $\text{MLCR}_{\text{GQ}}$

One potential limitation of $\text{MLCR}_{\text{HET}}$ is that it does not explicitly correct the difference between $\mathcal{X}$ and $\tilde{\mathcal{X}}$. An alternative approach is to transform the latter to have the same distribution characteristics as the former. The transformation can be performed using various methods including geometric quantile mapping (GQM) [15], covariance alignment [16], and change factors. In this paper, we consider the MLCR approach using GQM to transform $\tilde{\mathcal{X}}$ into its corresponding value in $\mathcal{X}$. The transformed value is then incorporated into $\hat{\mathbf{y}}$ in the second term of the objective function given in Eq. (2). The transformation only needs to be performed once (during preprocessing) and remains unchanged as Algorithm 1 progresses from one iteration to the next.

To perform the transformation, we need to define a notion of multidimensional quantiles and positions of data points. Geometric quantile is the multidimensional equivalent of a univariate quantile function [17]. The position [18] of a data point $\mathbf{z}$ relative to a set of points $\mathbf{Z} = (\mathbf{z}_1, .., \mathbf{z}_m)^T$ is iteratively computed as follows [6,15]:

$$\mathbf{p}_Y^k(\mathbf{z}) = \frac{1}{\kappa n} \sum_{i=1}^{n} \frac{\mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i)}{\parallel \mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i) \parallel}$$

$$\mathbf{p}_Y^1(\mathbf{z}) = \frac{1}{\kappa n} \sum_{i=1}^{n} \frac{\mathbf{z} - \mathbf{y}_i}{\parallel \mathbf{z} - \mathbf{y_i} \parallel} \tag{3}$$

Thus, for each data point $\tilde{\mathbf{x}}_i \in \tilde{\mathcal{X}}$, we apply the transformation given in Eq. (3) to identify its corresponding position and look for the image point $\mathbf{x}_i \in \mathcal{X}$ that is closest to its position. The second term of the objective function given in Eq. (2) is subsequently computed using the image points of the data points in $\tilde{\mathcal{X}}$.

## 6. EXPERIMENTAL EVALUATION

This section describes the data set used for evaluation as well as the experimental setup and results obtained from applying the proposed CR framework.

### 6.1. Data Sources

We evaluated the performance of our framework on 20-year climate data from 14 stations in Michigan. The response variables correspond to daily maximum temperature, minimum temperature, and precipitation observed at each station, spanning the period of 1980–1999. The predictor variables used in this study correspond to RCM simulations obtained from the North American Regional Climate Change Assessment Program (NARCCAP) [8]. Nine predictor variables were chosen—zonal and meridional surface wind speed, maximum, minimum, and average surface air temperature, surface pressure, humidity, precipitation, and 500 hPa geo-potential height. The three RCMs used in this study are the Canadian regional climate model (CRCM), the Weather Research and Forecasting Model (WRFG), and the regional climate model Version-3 (RCM3). The gridded RCM data have a spatial resolution of 50 km. The nearest land grid cell to each station is chosen to provide the RCM data for that station.

The RCM simulations can be driven either by NCEP/DOE AMIP-II Reanalysis (NCEP) or GCMs as their lateral boundary conditions. The NCEP-driven runs span the years 1979–2004 and is a retrospective model of the atmosphere based on observed data. Since there is a day-to-day mapping between the predictor variables provided by the NCEP-driven runs and the response variables provided by station data, regression methods can be applied to their combined data. By contrast, the GCM-driven simulations are available for the years 1971–2000 (historical period) and 2041–2070 (future period). Since each GCM-driven simulation provides one of many possible realizations of the climate scenarios, it is not meaningful to associate the simulated predictor variables for a given day with the response variables for that day. Without a day-to-day mapping between the predictor and response variables, regression methods are inapplicable to such data. Nevertheless, QM and distribution-regularized contour regression can be used here because they require the matching of quantiles between two distributions, instead of matching the predictor and response values by their dates.

### 6.2. Experimental Setup

Our experiments were conducted using NCEP-driven RCM simulations and station data from 1980-1999. The regression functions and CDFs for QM were trained using the first 10 years of data and evaluated on the remaining 10 years. Nine data sets were created, each corresponding to a combination of RCM (WRFG, CRCM, or RCM3) and the response variable ($T$ for maximum temperature, $t$ for minimum temperature, and $P$ for precipitation). For example, the data set named 'WRFG-T' uses predictor variables from WRFG and maximum temperature as its response variable. The parameter $\gamma$ for contour regression was fixed across all stations and the maximum number of iterations was set to 10. For quantile regression, the parameter $\tau$ is chosen to be the upper or lower 5th percentile of the distribution. Specifically, $\tau = 0.95$ for modeling

extreme precipitation and maximum temperature, whereas $\tau = 0.05$ for modeling extreme minimum temperature. For experiments with heterogeneous data sources, GCM-driven RCM runs from 1980 to 1989 were used (in addition to the NCEP-driven runs) to provide the distribution information needed to calibrate the parameters of the CR function. The CR function is then applied to GCM-driven RCM simulations from 1990 to 1999 and compared against the observed distribution for the given time period.

MLR, QM, lasso, and quantile regression are chosen as the baseline algorithms for comparison. The performance of the different methods is evaluated in terms of its root mean square error (RMSE) and distribution bias (RMSE-CDF):

$$\text{RMSE} = \sqrt{\sum_i \frac{(y_i - \hat{y}_i)^2}{m}},$$

$$\text{RMSE-CDF} = \sqrt{\sum_i \frac{(y_{(i)} - \hat{y}_{(i)})^2}{m}},$$

where $m$ is the number of test examples and $\hat{y}$ corresponds to the predicted values.

### 6.3. Experimental Results

First, we compare the performance of MLCR against MLR and lasso regression. The results are shown in Table 1. Barring possible overfitting, MLR should, in principle, have a lower RMSE given its emphasis on minimizing residual errors. Lasso is more robust to overfitting and produces sparser models. For maximum and minimum temperatures (the first six data sets in Table 1), MLCR is slightly worse than MLR, with an increase in RMSE of about $< 3\%$. However, MLCR improves its distribution bias (RMSE-CDF) over MLR considerably, with an average improvement around 40%. This improvement is observed across all climate stations in each data set, as shown by the 100% win−loss percentage in Table 1. A similar result was observed with Lasso. although not shown in the table, MLCR also has a lower RMSE than QM on all the data sets.

For precipitation, MLR and Lasso fared considerably worse due to the zero-inflated nature of the data (see Fig. 4). Since MLR struggled to capture the shape of the distribution, we chose a smaller value for the $\gamma$ parameter of MLCR than was used for temperature data to obtain a smaller bias in the predicted distribution. Consequently, its RMSE is considerably worse than MLR or Lasso, but it achieves an impressive average RMSE−CDF improvement of more than 70%. We use two metrics to evaluate the distribution bias. First, we compute the difference in standard deviation ($\sigma$) between the predicted distribution ($\hat{y}$) and observed distribution ($y$) on the test data (from 1990 to 1999). The results given in Table 2 suggest that MLCR was able to capture variance of the observed distribution better than both MLR and Lasso. Nevertheless, the variance of the predicted distribution using QM is still closer to observation data than MLCR, which is not surprising considering the former is a distribution-driven method and the latter is a hybrid of distribution-driven and accuracy-driven methods. Second, to measure the similarity between their empirical CDFs, we also compute the correlation ($\rho_{CDF}$) between the sorted values of the predicted and observed CDFs. MLCR once again consistently outperforms MLR and Lasso for every station (see Table 2). MLCR also produces a higher $\rho_{CDF}$ than QR for some of the stations (especially for precipitation data) because its $\gamma$ parameter is chosen to give more emphasis on minimizing the distribution bias. Even for stations where the performance of MLCR does not fit as well as QM, its overall distribution is still closer to observation data than MLR (see Fig. 4).

**Table 1.** 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 |

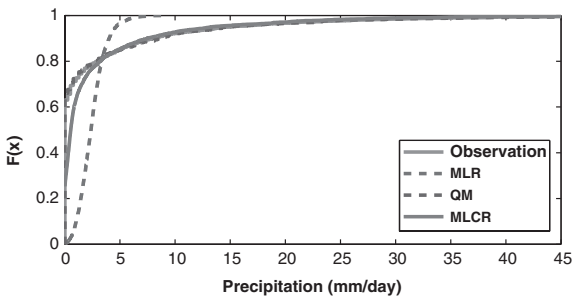


Fig. 4 CDF of predicted daily precipitation at a weather station in Michigan over the years 1990−1999. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

#### 6.3.1. QCR results

We compared the performance of QCR against QR in terms of its ability to capture the shape as well as extreme values of the distribution. The following evaluation

**Table 2.** Percentage of stations that MLCR outperformed baseline in terms of $\sigma$ and $\rho_{\text{CDF}}$.

| | $\sigma$ win−loss% | | | $\rho_{\text{CDF}}$ 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 |

**Table 3.** Percentage of stations that QCR outperformed QR in terms of RMSE, $F$-measure, Kolmogorov-Smirnov statistic ($K$) and correlation for data points considered extreme value.

| Dataset | RMSE | $F$-measure | $k$ | $\rho$ |
|---|---|---|---|---|
| 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 |

metrics were used to compare the two approaches. First, we compute RMSE using only observation data from the extreme values (top 5th percentile for $T$ and $P$ or bottom 5th percentile for $t$) of the distribution. The second metric, $F$-measure, determines how well QR and QCR can detect the timing of extreme days during the test period of 1990–1999. A true positive (TP) event is when the method predicts the value to be extreme and the actual observed value is also extreme. A false positive (FP) event is when the method predicts an extreme value day but the observed value is nonextreme. A false negative (FN) event is when the method predicts a nonextreme value for an extreme value day. $F$-measure is computed as follows:

$$F\text{-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

The third metric, Kolmogorov-Smirnov (K) statistic, measures the similarity of two distributions $F_1$ and $F_2$ by taking the maximum absolute value of their difference, i.e. $\max_x |F_1(x) - F_2(x)|$. The last metric $\rho$ computes the correlation ($\rho$) between the observed and predicted values only for the days identified as extremes (i.e., in top-5th or bottom-5th percentiles).

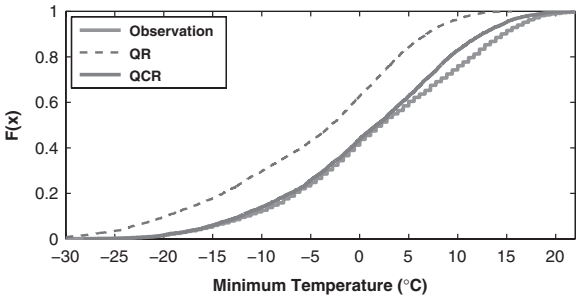


Fig. 5 CDF of predicted daily minimum temperature at a weather station in Michigan, 1990–1999. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 3 summarizes the results. In terms of RMSE, $F$-measure, and $K$, QCR outperforms QR for all stations in all nine data sets evaluated. This suggests that QCR can effectively detect the magnitude and timing of extreme valued days better than QR. However, for correlation of extreme precipitation values ($\rho$), QCR does not perform as well as QR for many stations because its $\gamma$ parameter has been tuned to aggressively focus on fitting the distribution well (see Fig. 5 for example of CDFs obtained using QR and QCR at one of the stations in Michigan).

### 6.3.2. *Comparison of* $\text{MLCR}_{\text{GQ}}$ *and* $\text{MLCR}_{\text{HET}}$

This section presents our experimental results applying the CR framework to heterogeneous sources of predictor variables. The first approach, $\text{MLCR}_{HET}$ incorporates predictor variables from GCM-driven RCM simulations directly into the distribution bias term of our objective function, while the second approach, $\text{MLCR}_{\text{GQ}}$ uses GQM to adjust the distribution of GCM-driven RCM simulations before incorporating them into the objective function. Figure 6 compares the CDFs of predicted daily minimum temperature at a weather station located in Eau Claire, Michigan. Both $\text{MLCR}_{\text{GQ}}$ and $\text{MLCR}_{\text{HET}}$ were able to match the observed distribution quite well, though $MLCR_{GQ}$ shows a slightly lower bias especially for minimum temperature above 0°C. For precipitation, both methods appear to underestimate the frequency of nonrain days. This is a known artifact of most regression-based methods when applied to zero-inflated time series data [19]. $\text{MLCR}_{\text{GQ}}$ shows marginal improvement over $\text{MLCR}_{\text{HET}}$, particularly for precipitation between 5 and 15 mm (see Fig. 7).

## 7. CONCLUSIONS

We propose a framework called contour regression that simultaneously maximizes the prediction accuracy of regression methods as well as reduces biases in the overall shape of the predicted distribution. The framework can

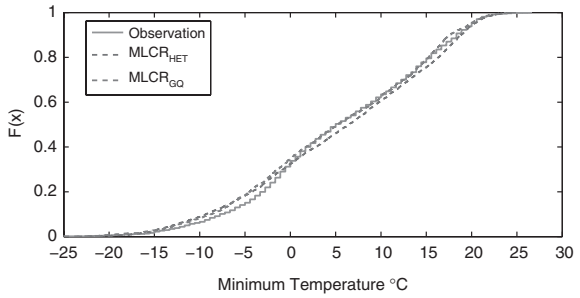


Fig. 6 Comparing the cumulative distribution function of $MLCR_{HET}$ and $MLCR_{GQ}$ for daily minimum temperature at a weather station in Michigan. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

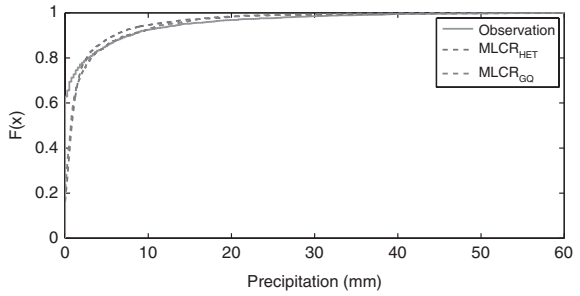


Fig. 7 Comparing the cumulative distribution function of $MLCR_{HET}$ and $MLCR_{GQ}$ for daily precipitation at a weather station in Michigan. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

incorporate different loss functions as well as nonlinear relationships between the predictor and response variables. Our experimental results showed that the CR framework can improve the fit of predicted distribution without degrading its prediction accuracy significantly.

## ACKNOWLEDGMENT

This work was supported by NSF Award CNH 0909378. The views and conclusions contained in this paper are those of the authors and do not reflect the views or policies of the funding agency.

## REFERENCES

[1] C. Tebaldi and D. Lobell, Towards probabilistic projections of climate change impacts on global crop yields, Geophys Res Lett 35(8) (2008), L08705, doi: 10.1029/2008GL033423.

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

[3] D. Scott and G. McBoyle, Climate change adaptation in the ski industry, Mitig Adapt Strateg Glob Change 12(8) (2007), 1411–1431.

[4] J. M. Themeßl, A. Gobiet, and A. Leuprecht, Empirical-statistical downscaling and error correction of daily precipitation from regional climate models, Int J Climatol, 31(10) (2011), 1530–1544.

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

[6] Z. Abraham, P.-N. Tan, P. Perdinan, J. Winkler, S. Zhong, and M. Liszewska, Position preserving multi-output prediction, In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), 2013.

[7] C. Monteleoni, G. A. Schmidt, S. Saroha, and E. Asplund, Tracking climate models, Stat Anal Data Mining 4(4) (2011), 372–392.

[8] North American Regional Climate Change Assessment Program, http://www.narccap.ucar.edu/. Accessed on January, 2012.

[9] S. P. Charles, B. C. Bates, I. N. Smith, and J. P. Hughes, Statistical downscaling of daily precipitation from observed and modelled atmospheric fields, Hydrol Process 18(8) (2004), 1373–1394.

[10] R. Wilby, S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods, 2004.

[11] J. A. Winkler, G. S. Guentchev, Perdinan, P.-N. Tan, S. Zhong, M. Liszewska, Z. Abraham, T. Niedzwiedz, and Z. Ustrnul, Climate scenario development and applications for local/regional climate change impact assessments: an overview for the non-climate scientist. Part II: Considerations when using climate change scenarios, Geogr Compass 5/6 (2011), 301–328.

[12] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich, Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user, Rev Geophys 48 (2010), 134.

[13] J. A. Winkler, G. S. Guentchev, M. Liszewska, Perdinan, and P.-N. Tan, Climate scenario development and applications for local/regional climate change impact assessments: an overview for the non-climate scientist. Part II: Considerations when using climate change scenarios, Geogr Compass 5/6 (2011), 301–328.

[14] R. Koenker, Quantile Regresssion, Wiley Online Library, Cambridge University Press, 2005.

[15] X. He, Y. Yang, and J. Zhang, Bivariate downscaling with asynchronous measurements, J Agric Biol Environ Stat 17(3) (2012), 476–489.

[16] H. Cheng and P.-N. Tan, Semi-supervised learning with data calibration for long-term time series forecasting, In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, 133–141.

[17] P. Chaudhuri, On a geometric notion of quantiles for multivariate data, J Am Stat Assoc 91(434) (1996), 862–872.

[18] J. I. Marden, Positions and qq plots, Stat Sci 19(4) (2004), 606–614.

[19] Z. Abraham and P.-N. Tan, An integrated framework for simultaneous classification and regression of time-series data, In SIAM International Conference on Data Mining (SDM), 2010, 653–664.