MULTI-OBJECTIVE REGRESSION WITH APPLICATION TO THE CLIMATE DOMAIN

By

Zubin Abraham

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Doctor of Philosophy

2013

ABSTRACT

MULTI-OBJECTIVE REGRESSION WITH APPLICATION TO THE CLIMATE DOMAIN

By

Zubin Abraham

Regression-based approaches are widely used in climate research to derive the statistical, spatial, and temporal relationships among climate variables. Despite its extensive literature, existing approaches are insufficient to address the unique challenges arising from the data characteristics and requirements of this domain. For example, climate variables such as precipitation have zero-inflated distributions, which render ineffective any linear regression models constructed from the data. In addition, whereas traditional regression-based approaches emphasize on minimizing the discrepancy between observed and predicted values, there is a growing demand for regression outputs that satisfy other domain-specific criteria. To address these challenges, this thesis presents multi-objective regression frameworks designed to extend current regression-based approaches to meet the needs of climate researchers. First, a framework called Integrated Classification and Regression (ICR) is developed to accurately capture the timing of rain events and the magnitude of rain amount in zero-inflated precipitation data. The second multi-objective regression framework focuses on modeling the extreme values of a distribution without degrading its overall accuracy in predicting non-extreme values. The third framework emphasizes on both minimizing the divergence between the regression output and observed data while maximizing the fit of their cumulative distribution functions. The fourth contribution extends this framework to a multi-output setting, to ensure that the joint distribution of the multiple regression outputs is realistic and consistent with true observations.

ACKNOWLEDGMENTS

This work is partially supported by NSF grant III-0712987 and subcontract for NASA award NNX09AL60G. This work is also partly supported by National Science Foundation Dynamics of Coupled Natural and Human Systems competition Program (CNH Award No.-0909378). The author would like to thank Dr Julie Winkler and Dr Sharon Zhong for valuable discussions on statistical downscaling for climate change projection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Regression is a statistical method for deriving the relationship between a continuous-valued response variable $y \in \Re$ and its predictor variables $\mathbf{x} \in \Re^d$. The relationship is typically expressed as a mathematical function $f : \Re^d \rightarrow \Re$. Numerous regression-based methods have been developed in the past, including discriminative models (such as multiple linear regression [58] and support vector regression [105]) and generative models (such as hidden Markov regression [57]). These methods are primarily designed to minimize a loss function, $\ell[f(\mathbf{x}), y]$, that measures the difference between the observed and predicted values. Although such a loss function is sufficient to ensure a good fit between the predicted and observed values, there are other requirements that must be met when applying such methods to real-world applications. This thesis focuses on developing new regression-based methods that can handle the unique challenges arising from the climate research domain.

1.1 Regression in Climate Research

Given the growing concerns about global warming and its potential influence on human and natural systems [107, 28, 49, 93, 75, 60], there is a pressing need to generate accurate and robust projections of future climate scenarios for researchers, policy-makers and other stakeholders. For example, to aid crop management decision making, the projections can be incorporated into crop models to assess the crop yield response to future climate change.

Towards this end, recent advances in global and regional climate modeling have produced vast amounts of simulation data that can be harnessed to improve our understanding of the climate system and its evolution under different greenhouse gas emission scenarios [31, 46, 116, 85]. These general circulation models (GCMs) and regional climate models (RCMs), as they are called, are physical-based models, developed based on the fundamental laws of physics, chemistry, and fluid dynamics to simulate the response of the Earth system to various external forcings on a three-dimensional spatial grid mesh. However, the scale of these computer-simulated model outputs are often too coarse to be effectively used in climate change impacts, adaptation, and vulnerability (CCIAV) assessment studies. Furthermore, due to the complexity of the climate system and inadequacy of the models in capturing all of its underlying processes, there are inherent biases in the model outputs that must be corrected to enable reproduction of historical climate conditions.

Regression is a popular method to empirically downscale the coarse resolution model outputs to a finer resolution. It can also debias the model outputs to fit the distribution of historical climate data. In addition to generating climate projections, regression can also be applied for the purpose of spatial interpolation, to fill in missing values at locations where observed values are unavailable [64]. However, in spite of extensive number of regression-based approaches that have been proposed to generate downscaled climate projections and bias corrected climate projections, there are still a number of challenges current regression methods have not adequately addressed.

1.2 Challenges

The projections generated by climate models are not an end in themselves but are often integrated into other numerical models (such as crop and hydrological models) to enable assessments of future climate change impacts. These downstream models not only differ in terms of the climate variables needed as input, they may also have distinct expectations regarding the desired characteristics of the climate projections. As an example, the skills of the climate projections in terms of simulating the length of wet and dry runs, i.e., number of consecutive rain and non-rain days, is an important requirement to estimate drought duration and intensity [9]. The least-square loss function employed by multiple linear regression (MLR) or the first-order Markov assumption employed by weather generators are inadequate to simulate the higher order temporal autocorrelation of a precipitation time series. Thus, one of the main challenges of applying regression methods to generate climate projections is that there could be more than one distinct expectation of the projections, which are not always easy to simultaneously achieve.

Furthermore, the projections should be unbiased across all quantiles. An unbiased projection is one whose distributional characteristics are consistent with that of the true values of the response variable, across all the quantiles of its distribution. To illustrate this, consider the histogram of observed daily maximum temperature at a weather station in Michigan, represented by the gray area in Figure 1.1. The red dotted line represents the histogram of daily maximum temperature generated by multiple linear regression. Observe that the MLR outputs have a warm bias for the cooler days and a cold bias during the warmer days. As a result, MLR is not a suitable approach if extreme values are of paramount importance to users of the climate projections.

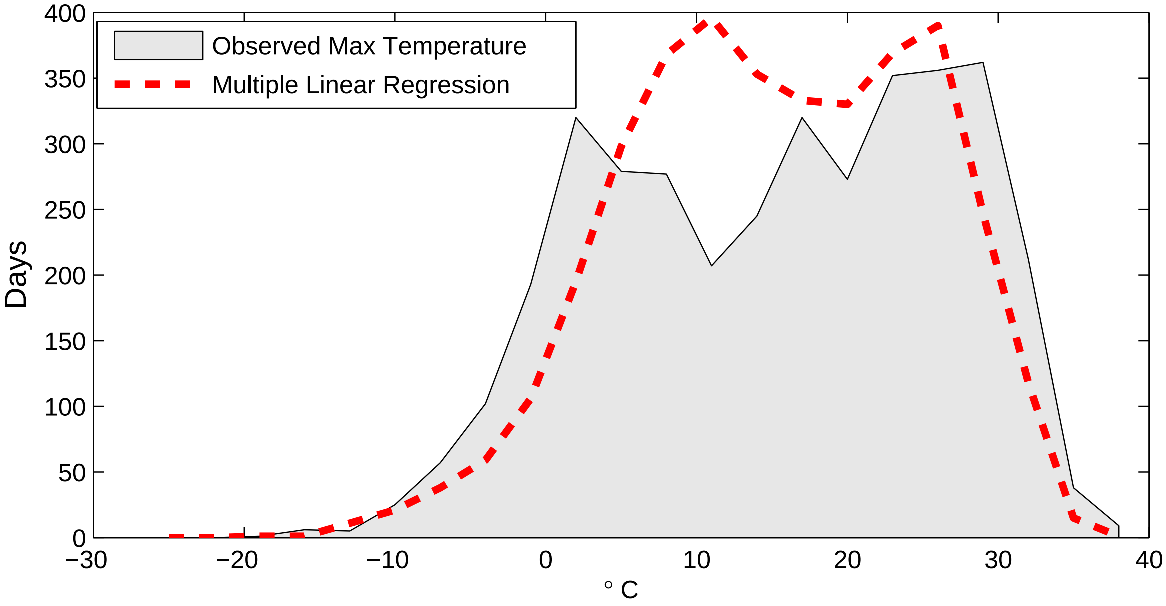


Figure 1.1: Histogram comparing the distribution of predicted daily maximum temperature at a weather station in Michigan to its respective observed values, 1990-1999. [For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation]

To address the requirement for an unbiased distribution, there is a class of bias correction approaches that can be used. Although these approaches can generate climate projections with low bias, their residual errors can be high, which implies a lack of agreement between the observed and predicted values at each time step. Hence, it is imperative to develop regression-based approaches that provide outputs satisfying both requirements of minimal error and unbiased predictions.

In addition, when considering the distribution characteristics of the regression outputs, the end user may be interested in certain quantiles over others. For instance, farmers are often interested in the frequency and magnitude of extreme values in the climate projections, due to the larger economic implications associated with them. However, as most approaches prioritize the conditional mean of the distribution, they tend to underestimate the frequency of extreme-valued data points as shown in Figure 1.2. Thus, another challenge for regression-

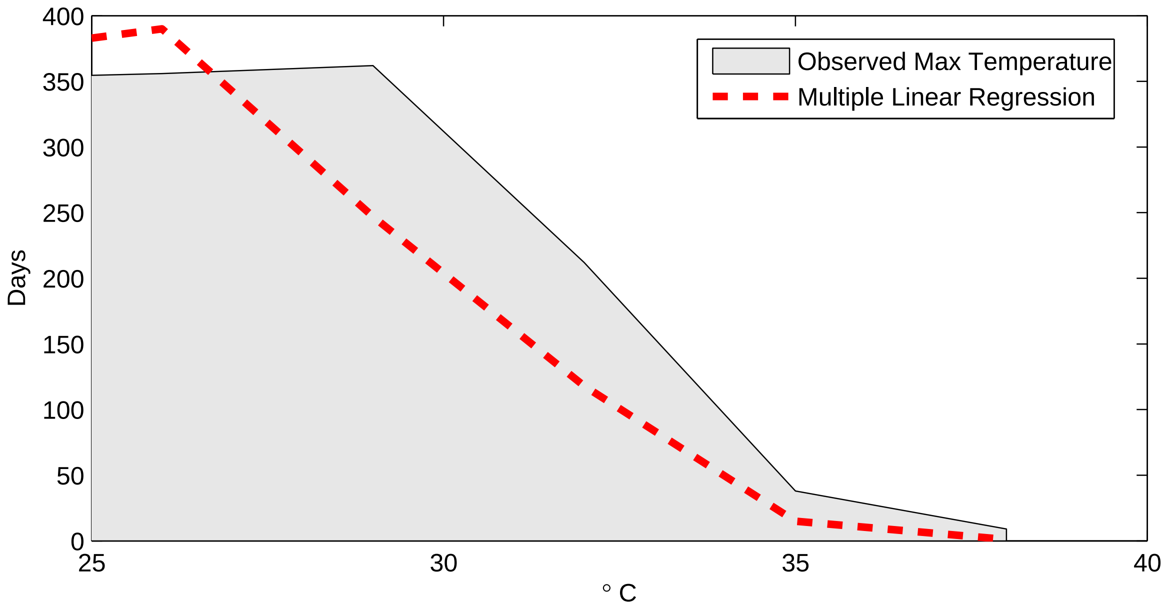


Figure 1.2: Histogram comparing the distribution of extreme values of the predicted daily maximum temperature at a weather station in Michigan, obtained using multiple linear regression, to its respective observed values, 1990-1999.

based approach is to provide a framework that is flexible enough to prioritize the accuracy at the quantiles of the end-user's choice, without significantly degrading the performance at other quantiles.

End users may also require projections of more than one climate variable. For instance, daily minimum and maximum temperature as well as total precipitation are among the common variables needed for CCIAV assessments. Although regression-based approaches can be trained for each variable independently, the resulting projections may not be consistent with each other. Hence, there is a growing demand for multiple output prediction methods capable of capturing the joint distribution of the response variables in a realistic and consistent fashion, while minimizing their residual errors. Unfortunately, current methods are designed to optimize one of the two criteria, but not both, as shown in Figure 1.3. Generating projections for multiple variables while preserving their joint relationships and minimizing the

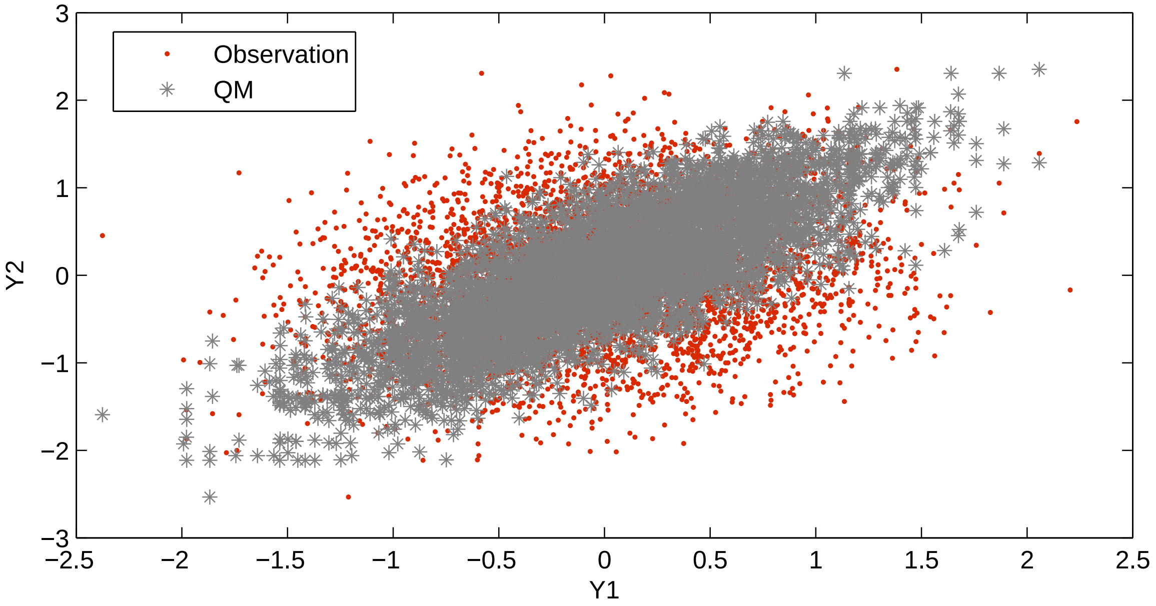


Figure 1.3: Scatter plot comparing the joint distribution between the quantile mapping (QM) predicted output of two response variables and it true values.

residual errors is another challenge that has not been sufficiently addressed.

Some climate variables are also harder to project due to their unique data characteristics. For instance, daily precipitation is notoriously challenging to model due of its zero-inflated distribution, a challenge that conventional regression methods are not well suited to handle. A zero-inflated distribution is a distribution with an abundance of zero values as shown in Figure 1.4. Such distribution can also be found in many other applications that are related to long-term projections, such as ecological modeling, disease monitoring, and traffic monitoring. As conventional regression methods typically prioritize modeling the conditional mean of the distribution, they tend to underestimate the frequency of zero-valued data points as well as the magnitude of the extreme values of a zero-inflated variable. Thus, there is a need to develop models that are geared toward dealing with some of the more uncharacteristic distributions observed among the climate variables.

The source of climate data available for building the regression models may also introduce

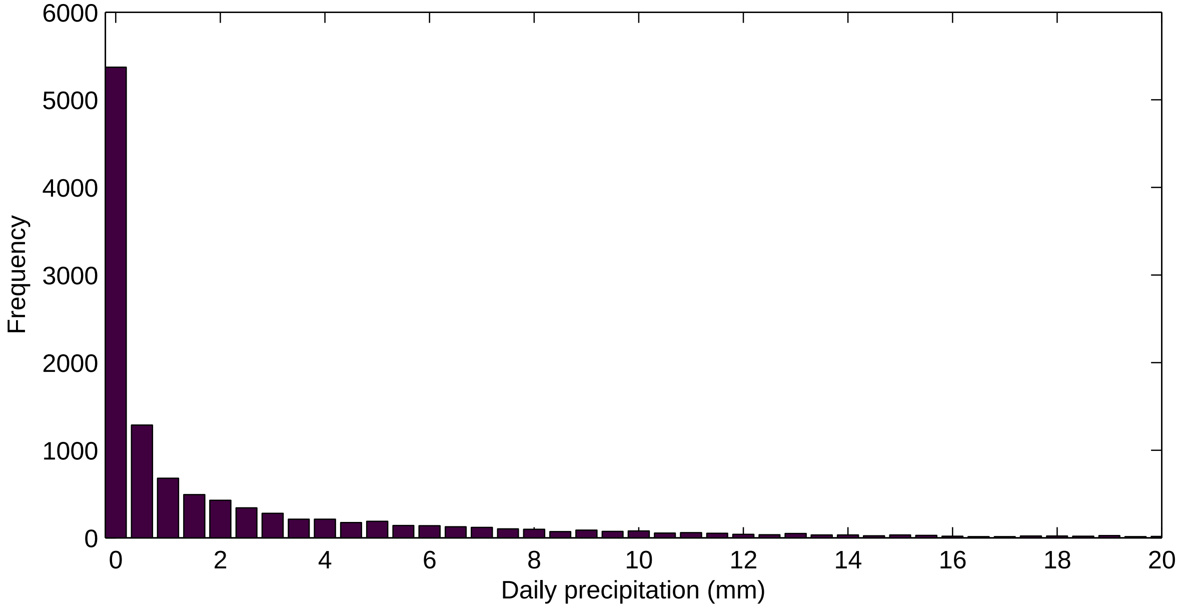


Figure 1.4: Histogram of daily precipitation recorded at a weather station in Canada.

additional complications. For example, the GCM or RCM-simulated precipitation data may not be exactly synchronized with the observed precipitation since each simulated output is only one possible realization of the time series. This affects both the training of regression models as well as model evaluation, which typically assumes there is a one-to-one mapping between the input and output variables for each data point used in the regression model. Fortunately, there are alternative approaches, such as quantile mapping, that cater to modeling data points with asynchronous predictor and response variables. Unfortunately, this flexibility also results in a prediction with relatively higher residual errors, as the models do not utilize the existing mapping information between the predictor and response variables. The challenge here is to develop models that can handle asynchronous data with compromises model accuracy.

For RCM models, the simulations can be driven by either reanalysis data (akin to true observations) or by GCM models as their initial boundary conditions. Reanalysis-driven

RCM runs are typically used to generate hindcast values of the predictor variables whereas GCM-driven RCM runs are needed to generate forecast values of the predictor variables. Most regression-based approaches are trained using reanalysis-driven RCM runs and tested on future data generated by GCM-driven RCM runs. Unfortunately, there are inherent biases in the GCM-driven RCM runs that are not fully accounted for by the regression model. This presents an additional challenge that must be addressed by regression-based approaches, which typically assume that the training and test data have similar distributions.

1.3 Thesis Contributions

This thesis presents multivariate regression-based frameworks that simultaneously addresses multiple objectives pertaining to the individual requirements of an accurate climate projection. By simultaneously optimizing multiple objectives, the frameworks generate a projection that satisfy multiple requirement with minimal degradation of any one objective, unlike existing regression-based approaches that address a single objective at the expense of compromising other requirements. The multiple objectives addressed by each of the frameworks pertain to best replicating the unique distribution characteristics of a response variable, while also ensuring an accurate projection in terms of minimum residual errors.

As mentioned earlier, pragmatic approaches to modeling predictive systems need to take into account any unique distribution characteristics of the response variable that is critical to generating an accurate projection. Chapter 3 presents an integrated multi-objective framework that simultaneously performs classification and regression to accurately predict values of a zero-inflated time series [4]. The regression and classification models are trained to optimize a joint objective function that minimizes both the classification errors (zero and non-zero values) and regression errors for data points that have non-zero values. The framework compensates for the uncertainty in the data by using a smoothing function that prioritizes non-zero valued data point whose response value is consistent with other data points having similar values for their respective predictor variable, during the learning of the regression function. The effectiveness of the framework is demonstrated in the context of its application to a downscaling precipitation climate variable. The semi-supervised extension of the framework in Chapter 3 is elaborated in Chapter 4 and compared with its supervised counterpart [3].

Given that studies and applications that utilize long-term projections for analysis may be interested in certain quantiles of the distribution of the projection over others, Chapter 5 presents a multivariate framework that focuses on the accurate projection of a specific quantile of the distribution (such as those pertaining to extremes values) with minimum deterioration of the accuracy of the projection at the other quantiles of the response variable [8]. Chapter 5 also elaborates on the framework in a semi-supervised setting.

As an extension of the above-mentioned framework, Chapter 6 presents a multi-objective framework called ICR that focuses on reliable prediction of extreme values events for a zero-inflated response variable [7]. This multi-objective framework incorporates the multiple objectives of classification, regression and conditional quantiles. The frameworks in Chapters 5 and 6 were evaluated on climate data and these demonstrated their ability in accurately detecting the frequency, timing and magnitude of extreme temperature and precipitation events effectively compared to several baseline methods.

Chapter 7 shows the limitations of popular regression-based approaches in terms of preserving the distribution characteristics of true response variable across the various quantiles in spite of its minimizing residual errors. Chapter 7 goes on to present a multi-objective regression framework that simultaneously replicates the cumulative distribution properties of the response variable while minimizing the residual errors. The framework is highly flexible and can be applied to linear, nonlinear, and conditional quantile models [5]. 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 is demonstrated along with marked improvement over traditional regression-based approaches, for all climate stations evaluated.

There is a growing demand for a multiple-output prediction that not only is accurate in terms of minimal residual errors but also in terms of accurately capturing the joint distributional characteristic of multiple output variables, so that they are realistic and consistent with each other. Unfortunately, the preservation of these associations is not guaranteed by regular single or multiple output regression approaches. Chapter 8 presents a framework for multiple output regression that preserves the general association patterns among the response variables (including non-linear associations) while minimizing the overall errors of the individual prediction, by coupling regression and geometric quantile mapping. The effectiveness of the framework in modeling temperature and daily precipitation for climate stations in the Great Lakes region is demonstrated [6]. The framework showed significant improvement in reducing residual errors while preserving the joint distribution of the multi-output variables, over the baseline approaches, in all climate stations evaluated.

1.4 Summary

To summarize, the presented frameworks address the challenges pertaining to the application of regression-based approaches to the climate research domain. However, even though the multi-objective frameworks presented in this thesis are originally motivated by the need to generate accurate and unbiased projections of climate variables, they are generic enough to be used for other applications. For example, the ICR framework is applicable to any domains with zero-inflated data distribution whereas the MCR framework is designed for domains that require consistent predictions across multiple response variables.

Additionally, the proposed frameworks leverage ideas from semi-supervised learning, statistical asynchronous regression, and geometric quantiles to address the challenges introduced by the climate research domain. All the experimental results reported in this thesis pertaining to demonstrating the effectiveness of the frameworks are conducted on climate data for a study region involving parts of Canada and the Great Lake region around Michigan. The effectiveness of the frameworks is demonstrated in the context of their applications to bias correcting and downscaling precipitation and temperature data.

Chapter 2

Related Work

Regression is a commonly employed statistical approach for estimating the relationship between a response variable and its respective predictor variables. Popular approaches of regression, such as multiple linear regression (MLR), learn the regression coefficients that estimate the conditional expectation of the response variable, given the predictor variables. Unlike MLR, quantile regression estimates the conditional quantile of the response variables [77]. Similarly, there are numerous other popular variants of regression-based approaches [58, 15, 73], such as ridge regression [62], lasso regression [110], recurrent neural networks [55], Hidden Markov Model Regression [52], and support vector regression [105]. The various regression approaches primarily differ based on the number of predictor variables used, the number of response variables, the type of response variable, the estimation method used to identify the regression coefficients, the bias in the estimated regression coefficients, whether a linear or non-linear regression function is used, etc.

Time series prediction [80] has long been an active area of research with applications in finance [30], network monitoring [81], transportation planning [63][91], weather forecasting [46][31], etc. Regression approaches can be applied to time series data for forecasting purposes. When it comes to using regression for time series analysis, the two most common approaches employed are autoregressive and multivariate regression approaches. While multivariate regression-based approaches are constrained by the requirement of the availability of future values of the predictor variable to make a forecast, autoregressive approaches, such as autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) [23] do not have such constraints. Autoregressive approaches make forecasts by repeatedly invoking a model that makes its prediction, one unit at a time and using the predicted values from the previous iteration to infer future values. However, autoregressive approaches are plagued by the error accumulation problem, on account of the propagation of errors from one prediction step to the next one.

Given the influence of climate on agriculture [28, 107], natural ecosystems [49, 93], human health and natural calamities [60, 75], economic impact [104] etc., considerable effort has been dedicated to generate projections of climate variables, to aid strategic decision making. This effort has also been precipitated by the growth in the number of climate models in the climate science domain [88].

2.1 Forecasting in the Climate Science Domain

In the climate science domain, there has been extensive research on applying time series regression models on climate data obtained from global climate models (GCM) [31, 46, 116]. Global climate models (GCMs) are extensively used for understanding how the global climate may change in the future.

GCMs are computer-generated models for simulating future climate conditions under different greenhouse gas emission scenarios. However, the spatial resolution of GCM outputs are often too coarse to reliably project the future climate scenarios of a local region and do not provide reliable information on scales below about 200 km [Meehl et al., 2007].

Two of the more widely employed approaches to improving the projection of climate variables obtained from climate models such as GCMs and regional climate models (RCM), is downscaling and bias correction. While bias correction is employed to correct the biases in the distributional characteristics of the climate projection, downscaling is employed to refine the granularity of the projection to better represent the climate variability associated with the location of interest of the impact assessment study. In spite of the two distinct objectives, regression is a popular approach employed, to generate bias corrected projections as well as downscaled climate projections.

2.1.1 Downscaling in the Climate Science Domain

Downscaling techniques are used to relate the coarse-scale GCM outputs to the local climate variables such as daily precipitation and temperature [116]. There are two common approaches to downscaling climate variables. The first approach is dynamical downscaling, which nests a regional climate model (RCM) into the GCM to represent the atmospheric physics with a higher grid-box resolution within a limited area of interest. The second approach is statistical downscaling, that statistically links coarse-scale weather with relatively finer resolution observed local-scale weather. Wilby and Wigley [117] classified statistical downscaling into regression methods, weather type approaches, and stochastic weather generators. Multiple linear regression (MLR) is probably the most common regression-based statistical downscaling approach whose objective is to minimize the sum square error. Multiple linear regression with randomization (MLRR) is another regression-based approach that adds a randomization term to compensate for the reduced variability in the prediction of the regression function that is fitted to pass through the centroid of the data. Themeβl et al. [108] applies MLRR to bias correct precipitation data. Analog method (AM) and its variants, such as nearest neighbor analog methods (NNAM), are other common downscaling techniques that are based on the intuition of prototyping. The advantage of approaches such as AM is the minimum training time required. The disadvantage is the need for sufficiently large amount of historical data for accurate prediction as well as the limitation of being unable to predict extreme events beyond the magnitude of extreme events available among the historical data sets. Rummukainen [100] differentiated various statistical downscaling approaches based on the nature of the chosen predictors as either being perfect prog(nosis) (PP) or model output statistics (MOS) (Glahn and Lowry [56]).

2.1.2 Bias Correction in the Climate Science Domain

Often even the results of downscaling, such as RCM simulation data often needs to be bias corrected to accurately reflect the observed distribution of the respective climate variables before being fed to climate change impact models and/or crop growth and yield models so that biases in the simulated data are not propagated. RCM variables, such as temperature, may require bias correction of the mean and standard deviation of the distribution, while variables such as precipitation may require frequency and intensity of the distribution to be additionally calibrated.

The quantile-based bias correction approaches focuses on matching the distribution of the downscaled approaches as closely as possible to that of its observations' distribution. Quantile mapping, modified quantile mapping (EDCDFm) and transfer functions defined by Piani et al.[96][97] are a few examples. Quantile mapping has been extensively used to downscale climate variable across regions, ranging from the smaller regions, such as Japan (Iizumi et al. [67]) to the larger areas such as the European continent as seen by Piani et al [96]. Samuels et al. [102] and Ines et al. [68] use quantile mapping for downscaling and bias correcting climate variables like precipitation. Ceglar et al. [29] has used it to downscale even variables like solar radiation. Quantile mapping can use both as a means to downscale the climate variables as well as a bias correction technique. When used as separate bias correction technique, QM corrects the errors in the shapes of the two distributions either prior or after downscaling the data, using any of the downscaling approaches. Approaches such as linear regression used by Rivington et al. [98] and copula-based approaches used by Favre et al. [47] are also occasionally used for bias correction.

2.1.3 Regression-Based Approaches for Bias Correction and Downscaling

Regression-based approaches are among the most commonly employed approaches for bias correction and downscaling climate projections. In spite of the nuance between the objectives of bias correction and downscaling climate variables for generating projection, there is considerable overlap in terms of the approaches employed by both applications. The various regression-based approaches used for downscaling and bias correcting climate projections can be broadly categorized as either being distribution-driven or accuracy-driven, based on whether their primary objective is minimizing residual errors or ensuring that the cumulative distribution of the projection of the response variable matches, as closely as possible, the distribution of the corresponding observations data at each quantile. Multiple linear regression and QM are respectively, the most popular accuracy-driven and distribution-driven approaches.

Among the various accuracy-driven regression-based approaches, change factor (delta method), linear regression (LR), its multiple variable counterpart and multiple linear regression (MLR) are the most popular regression-based approaches used for downscaling and bias correction [108][98]. Multiple linear regression with randomization used by Themel et al. and copula based approaches used by Favre et al. [47] have also occasionally been used for bias correction and downscaling climate variables.

Unlike accuracy-driven approaches, distribution-based approaches primarily focuses on ensuring that the cumulative distribution of the projection of the response variable matches as closely as possible, the distribution of the corresponding observations data at each quantile. Quantile mapping (QM), modified quantile mapping (EDCDFm), transfer functions defined by Piani et al. and local intensity scaling [108][61][96][97] are few examples of distribution-based regression approaches commonly used for bias correction and downscaling climate projections.

2.2 Multiple-Objective Prediction

Based on the distribution characteristics of the response variable, modeling certain response variables, is more challenging that others. This is often due to the distinct or uncommon distribution characteristics of the response variable. For instance, conventional single-objective regression models that prioritize the conditional mean of the distribution tend to under-estimate the number of zero-valued data points while also under-estimating the values of the extreme values of a zero-inflated response variable. Similarly, the single-objective regression approaches such as MLR, that are commonly used when the emphasis is on minimizing SSR, fare poorly in terms of capturing the shape of the distribution and hence is not well suited in preserving the distribution characteristics of the projection. Thus there is the need to have a framework that caters to simultaneously addressing multiple objectives.

As mentioned earlier, precipitation which is an important driver in a lot of models such as fresh water modeling [78], is considerably more difficult to model than temperature, mostly due to its high spatial and temporal variability and its nonlinear nature. The motivation behind using a multi-objective approach that combines the use of classification and regression models is to address the challenges of zero-inflated distribution observed in precipitation.

Previous studies have shown that additional precautions must be taken to ensure that the excess zeros do not lead to poor fits [11, 13, 42, 20, 114] of the regression models. A typical approach to model a zero-inflated data set is to use a mixture distribution of the form

$$P(y|\mathbf{x}) = \alpha\pi_0(\mathbf{x}) + (1-\alpha)\pi(\mathbf{x})$$

where π_0 and π are functions of the predictor variables $\mathbf{x}$ and α is a mixing coefficient that governs the probability an observation is a zero or non-zero value. This approach assumes that the underlying data are generated from known parametric distributions. For example, π may be Poisson or negative binomial distribution (for discrete data) and lognormal or Gamma (for continuous data). Piani et al. [97] proposed a multi-objective transfer functions, specific to response variables having zero-inflated distribution. Similarly, local intensity scaling (LOCI) has been specialized for bias correcting response variables that have a zero-inflated distribution by accounting for the zero-inflated characteristics of precipitation data [108].

There have been extensive studies on the effect of incorporating unlabeled data to supervised classification problems, including those based on generative models[41], transductive SVM [71], co-training [19], self-training [120] and graph-based methods [18][121]. Some studies concluded that significant improvements in classification performance can be achieved when unlabeled examples are used, while others have indicated otherwise [19, 36, 40, 109, 118]. Blum and Mitchell [19] and Cozman et al. [36] suggested that unlabeled data can help to reduce variance of the estimator as long as the modeling assumptions match the ground truth data. Otherwise, unlabeled data may either improve or degrade the classification performance, depending on the complexity of the classifier compared to the training set size [40]. Tian et al. [109] showed the ill effects of using different distributions of labeled and unlabeled data on semi-supervised learning.

Recently, there have been growing interest on applying semi-supervised learning to regression problems [119][24][39][122]. Some of these approaches are direct extensions of their semi-supervised classification counterparts. Cheng and Tang [33] proposed a semi-supervised learning framework for long-term time series forecasting based on Hidden Markov Model Regression. They also developed a covariance alignment method to deal with the issue of inconsistencies between historical and future data from climate simulation models. None of these semi-supervised learning methods are designed for handling zero-inflated time series data.

2.3 Modeling Extremes

Identifying and modeling extreme events in climatology have recently gained a lot of traction [48]. Unfortunately, the common regression techniques mentioned earlier that may be used for downscaling, focus on predicting the conditional mean of the response variable, while extreme values are better identified by conditional quantiles that corresponds to the extreme values. Hence, unlike the common regression techniques mentioned earlier that focus on predicting the conditional mean, the motivation behind the presented model focuses on the conditional quantile, using an approach similar to quantile regression [77].

Variations of quantile regression, such as non-parametric quantile regression and quantile regression forests, have been used to infer the conditional distribution of the response variable, which may be used to build prediction intervals [106, 87]. Also, variants of quantile regression that estimate the median are used due to their robustness to outliers when compared to traditional mean estimate [82]. Friederichs and Hense [51] presented a statistical downscaling approach to estimate censored conditional quantiles of precipitation that uses QR. The conditional probability of the censored variable is estimated using a generalized linear model (GLM) with a logit function to model the nature of the distribution of precipitation and hence cannot be directly applied to model temperature. Mannshardt-Shamseldin et al. [83] demonstrate another approach to downscaling extremes through the development of a family of regression relationships between the 100 year return value (extremes) of climate modeled precipitation (NCEP and CCSM) and station-observed precipitation values. Generalized extreme value theory based approaches have also be applied to model extreme events like hydrologic and water quality extremes, precipitation, etc [111, 16]. The Pareto distribution [43, 66], Gumbel [22, 14] and Weibull [35] are the more common variants of general extreme value distribution used. But these techniques are probability based that emphasize trends pertaining to the distribution of future extreme events and not the deterministic timing of the occurrence of the extreme event. The drawback of building a model that primarily focuses on only a particular section of the conditional distribution of the response variable is the limited amount of available data. Hence, the motivation for incorporating unlabeled data during model building.

When it comes to accurately predicting extreme values in the presence of zero-inflated data, studies have shown that additional precautions must be taken to ensure that the excess zeros do not lead to poor fits [11, 13, 42, 20, 114] of the regression models. Generally, simple modeling of zero values may not be sufficient, especially in the case of zero-inflated climate data such as precipitation, where extreme value observations need to be accurately modeled. Due to the significance of extreme values in climatology and the increasing trend in extreme precipitation events over past few decades, a lot of work has be done in analyzing the trends in precipitation, temperature, etc., for regions in the United States and Canada among others [79, 99, 44, 37, 101]. Katz [72] introduces the common approaches used in climate change research, especially with regard to extreme values.

The common approaches to modeling extreme events are based on general extreme value theory [53, 89, 50], Pareto distribution [43, 66, 70, 90], generalized linear modeling [35, 34], hierarchical Bayesian approaches [54, 65, 103], etc. Gumbel [22, 14] and Weibull [35] are the more common variants of General extreme value distribution used. There are also Bayesian models such as the model of Cooley et al. [38] that augment the model with spatial information. Watterson et. al. proposed a model that also deals with the skewness of non-zero data/intermittency of precipitation using gamma distribution to interpret changes in precipitation extremes [113]. In contrast, the framework presented in this chapter handles the intermittency of the data by coupling a logistic regression classifier to the quantile regression part of the model.

2.4 Distribution Preserving Modeling

Courtesy of projects such as NARCCAP (North American Regional Climate Change Assessment Program), extensive studies have been done to utilize the long-term future climate projections made available [86, 88]. Many of these studies focus on the impact assessment of climate change on domains, ranging from natural ecosystems [93] [49] to those related to human systems [94]. Since many climate change impact assessment studies are interested in long-term climate projections, the accuracy of the distribution of the projection is often critical. As mentioned earlier, efficient utilization of these projections require the projections to be unbiased across all the quantiles of the distribution [31, 4, 116].

As mentioned earlier, bias correction approaches such as QM [108], Equidistant CDF Matching (EDCDFm), Statistical Asynchronous Regression (SAR)[92], transfer functions proposed by Piani et al.(2010b), etc, have been applied, to address these biases in the projection of climate data. However, these approaches are best suited when there is no day-to-day mapping available between the predictor and the response variable, as is the case of downscaling from GCMs or data from RCMs driven by GCMs. QM is very well equipped to generate a projection of the response variable with an unbiased distribution. However, these bias correction approaches under-perform in terms of accuracy of prediction of individual data points. This is because these bias correction approaches do not leverage the original mapping information between the response and predictor variables during training. This drawback is all the more impeding, 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.

Thus, common distribution driven single output regression approaches are best suited when the predictor and output variables are asynchronous and there is less emphasis on low sum squared residual error (SSR). Accuracy-driven regression approaches, such as MLR, Ridge, Lasso and analog methods [108], are commonly used when the emphasis is on minimizing SSR but fare poorly in terms of capturing the shape of the distribution (Figures 1.1). Thus, commonly used regression approaches are not well suited in preserving the distribution characteristics of the projection. Given the drawbacks of regression and quantile-based approaches, approaches such as Contour Regression (CR) [5] have been proposed that try to simultaneously minimize error and preserve the shape of the forecast distribution. CR uses a regression function that regularizes the area between the CDF of the target response variable and the regression result.

To address the limitation of single output regression (SOR), numerous multiple output regression (MOR) models have been proposed including the commonly used multi-output regression [59] and structured output regression [17]. A number of regression-based multiple output models focus on penalizing of the regression matrix using low rank penalization methods such as reduced rank regression [69]. However, these approaches do not model correlation in output dimensions. Another common approach to multiple output prediction is to penalize input space shared, for co-linearity, such as partial least square regression discriminant analysis (PLSDA) [95]. However these models, too, do not capture the association among various response variables. ”Curds and whey” is an example of regression based approach that models output correlation [25]. However, modeling output correlation assumes the relation among the response variables is linear. Multiple output SVR is another approach that takes advantage of correlation among response variables and extends SVR to multi-output systems by considering Cokriging (a multi-variable version of Kriging) [112]. Cokriging models multiple output variables by computing cross covariances between the different outputs. Group lasso [74], LL-MIMO [21], gaussian process MOR [10] are other examples of MOR.

However, none of the above-mentioned approaches preserve the full range of variability of the joint distribution of the response variables. He et al.[61] proposed bivariate quantile mapping to address the limitations of QM in bivariate space and use the intuition proposed by Buja et al. regarding geometric quantiles [26]. However, in spite of the approach faring very well in terms of capturing the requisite marginal distribution characteristics of the two response variables, due to its asynchronous nature, it suffers from poor SSR.

In the following chapters, approaches that address the challenges of modeling zero-inflated response variables, prioritizing extremes in a distribution of the response variable, preserving the overall distribution characteristics of the response variable, in both a single output and a multi-output setting are proposed.

Chapter 3

Modeling Zero-Inflated Data

This chapter presents a multi-objective approach for predicting future values of a time series data that are inherently zero-inflated. The proposed framework decouples the prediction task into two objectives—a classification step to predict whether the value of the time series is zero and a regression step to estimate the magnitude of the non-zero time series value.

3.1 Introduction

Predictive models for time series data are commonly employed in the fields of economics, finance, epidemiology, ecology, and meteorology, among others. The prediction accuracy is subject to the choice of model used, which in turn, may be limited by characteristics of the time series observations. For example, studies have shown that the performance of classical regression models is degraded when applied to data sets with excess zero values [11, 13, 42, 20, 114]. Such data are typically encountered in applications such as climate and ecological modeling, disease monitoring, manufacturing defect detection, and traffic monitoring.

Figure 3.1 shows the histogram of daily precipitation (in log scale) at a weather station in Canada for the period between January 1, 1961 and December 31, 2000. Nearly half of the observations have precipitation values equal to zero. Such zero-inflated data, as they are commonly known, often lead to poor model fitting using standard regression methods as

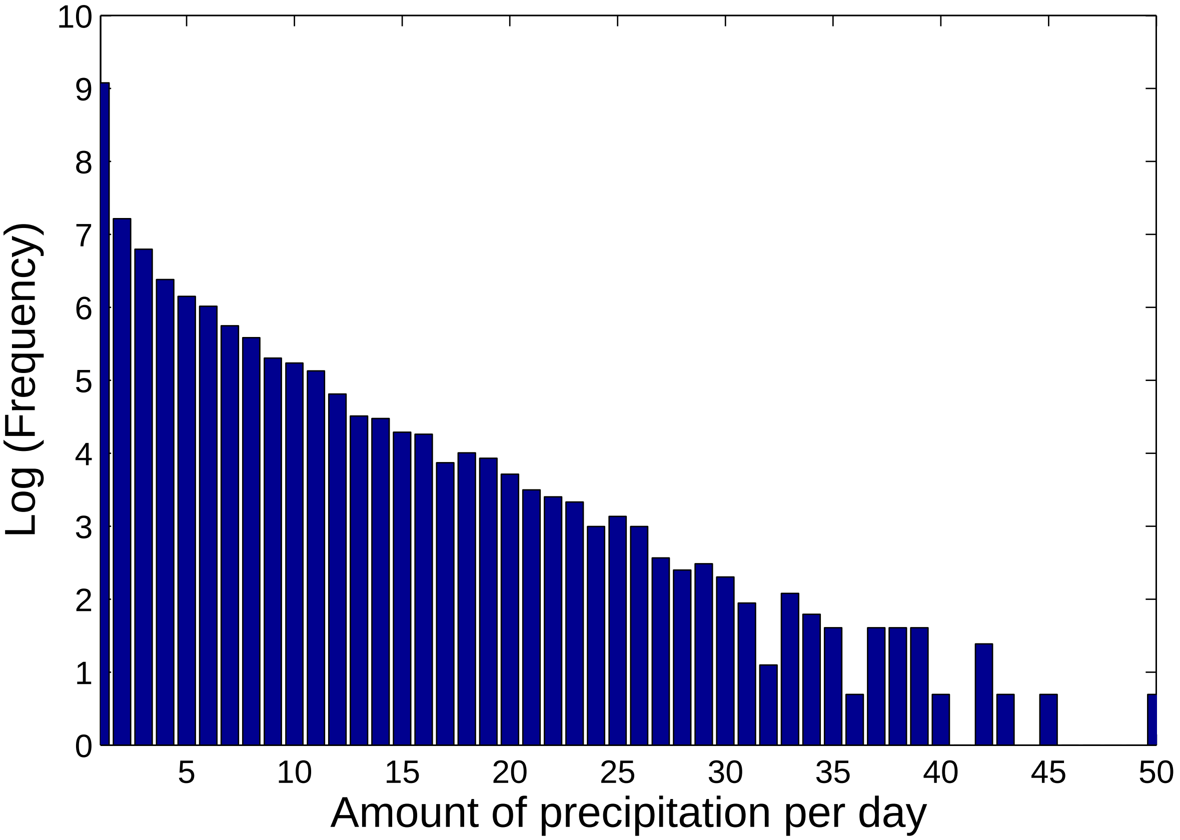
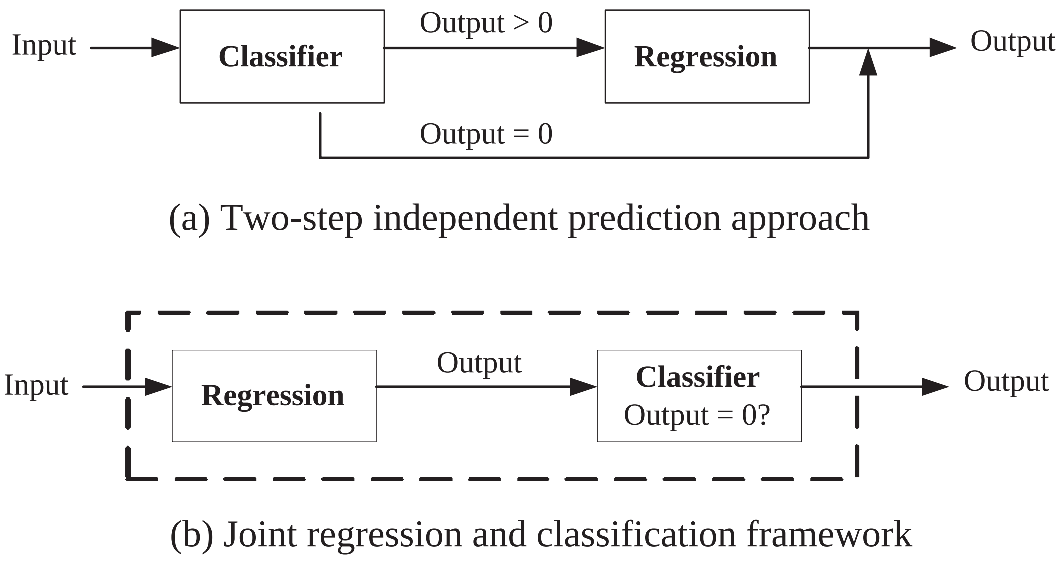


Figure 3.1: A zero-inflated frequency distribution of daily precipitation at a weather station in Canada

they tend to underestimate the frequency of zeros and the magnitude of non-zero values of the data. A typical strategy for handling such type of data is to first invoke a classification model to predict whether the output value is zero. A regression model, which has been trained on the non-zero data points, is then applied to estimate its magnitude only if the classifier predicts a non-zero output. Such an approach is commonly used for statistical downscaling of precipitation [115], in which the occurrence of rain or wet days is initially predicted prior to applying a regression model to estimate the amount of rainfall for the predicted wet days. The limitation of this approach is that the classification and regressions models are often built independent of each other. As a result, neither models can glean



(a) Two-step independent prediction approach

(b) Joint regression and classification framework

Figure 3.2: Comparison between independent modeling approach and proposed framework for predicting zero-inflated data

information from the other to potentially improve their prediction accuracy.

The objective of this chapter is to develop an integrated framework that accurately estimates the future values of a zero-inflated time series by simultaneous training the classification and regression models. Specifically, the models are trained to optimize a joint objective function that penalizes errors in classifying a data point and errors in predicting the magnitude of non-zero data points. Given a test point, the regression model is applied to estimate the magnitude of the predicted value. The output from the regression model along with the values of other predictor variables of the test point are then fed into a classification model to determine whether the predicted value should be adjusted to zero. The distinction between the traditional two-step independent modeling approach and the proposed framework is illustrated in Figure 3.2.

The effectiveness of the learning framework is demostrated in the context of precipitation prediction, using climate data from the Canadian Climate Change Scenarios Network Web site [1]. Specifically, the performance of the integrated framework was compared against two baseline methods. The first baseline corresponds to applying standard multiple linear regression (MLR) method on the entire training data, which includes both dry and rain days. The second baseline method (SVM-MLR) uses a combination of support vector machine classifier to predict dry/wet days and multiple linear regression to predict rainfall amount on wet days. Both the models are trained independently. Empirical results showed that the proposed framework outperforms both MLR and SVM-MLR on the majority of the weather stations investigated in this study.

In summary, the main contributions of this chapter are as follows:

- An integrated framework for simultaneously learning classification and regression models.

- The proposed framework was found to be more effective at predicting zero-inflated time series than building a single regression model or building independent classification and regression models to fit the time series data.

- The framework was successfully applied to the real-world problem of downscaling precipitation time series for climate impact assessment studies.

3.2 Preliminaries

Consider a multivariate time series $\mathbf{L} = (\mathbf{x}_t, \mathbf{c}'_t)$, where $t \in \{1, 2, \cdots, n\}$ denote the elapsed time, $\mathbf{x}_t$ is a d-dimensional vector of predictor variables at time t, and c_t is the corresponding value for the response (target) variable. Given an unlabeled sequence of multivariate observations $\mathbf{x}_\tau$, where $\tau \in \{n+1, \cdots, n+m\}$, the goal was to learn a target function $f(\mathbf{x}_\tau, \mathbf{w})$ that best estimates the future values of the response variable at each time τ. The set of weights $\mathbf{w} = [w_1, w_2, ..., w_d]^T$ are the regression coefficients to be estimated from the training data $\mathbf{L}$. For applications such as statistical downscaling, the predictor variables $\mathbf{x}_\tau$ correspond to climate variables at large spatial scales generated from computer-driven general circulation models (GCMs).

For zero-inflated data, the frequency of zero values in the time series is relatively larger than the frequency of each non-zero values, as shown in Figure 3.1. The response variable c'_t can be mapped into a binary class c_t, where

$$c_t = \begin{cases} 1, & \text{if } c'_t > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

For brevity, the notation $y' \equiv f(\mathbf{x}, \mathbf{w})$ was used as the predicted value of the response variable and y as its corresponding predicted class.

3.3 Framework for Simultaneous Classification and Regression

In this chapter, a framework is presented for predicting future values of a time series with the following unique characteristics:

1. The framework simultaneously performs classification and regression to improve the accuracy of predicting the magnitude of non-zero values in a zero-inflated time series.
2. The framework can be easily extended to a semi-supervised learning setting via graph regularization.

This chapter considers a framework for modeling zero-inflated variables using a combination of classification and regression models. The models in the framework are trained to optimize a joint objective function that considers both the classification errors on the time series and regression errors for the non-zero values. The framework presented in this chapter trains an SVM classifier only once after the parameters of the regression model have been determined. Proofs of convergence of our algorithm are also presented in this section.

Multiple linear regression (MLR) was considered as the underlying regression model in this study, in which $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. Extending the approach to nonlinear models will be a subject for future research.

3.3.1 Objective Function

The classification and regression models developed in this study are designed to minimize the following objective function:

$$\begin{aligned} \arg\min_{\mathbf{w},\mathbf{y}} L(\mathbf{w}, \mathbf{y}) &= \sum_{i=1}^{n} c_i(c_i' - y_i y_i')^2 + T_1 \sum_{i=1}^{n} (y_i - c_i)^2 \\ &+ T_2 \sum_{i,j=1}^{n} s_{i,j}[c_i y_i' - c_j y_j']^2 + T_3 ||w||^2 \end{aligned}$$

where,

$$y_i' = \sum_d w_d x_{i,d}, \quad y_i \in \{0, 1\}$$

and s_{ij} is the similarity between the values of the predictor variables at t_i and t_j

The rationale for the design of our objective function is as follows. The first term is somewhat similar to the standard least-square formulation of multiple linear regression, except the estimation of $\mathbf{w}$ is based on the non-zero values in the time series. The regression model is therefore biased towards estimating the non-zero values more accurately instead of being influenced by the over-abundance of zeros in the time series. The product $y_i y_i'$ in the first term corresponds to the predicted output of our joint classification and regression models. The second term in the objective function is equivalent to misclassification error in training data. The third term corresponds to a graph regularization constraint to ensure smoothness and consistency in the model predictions. Specifically, for two highly similar data points $\mathbf{x}_p$ and $\mathbf{x}_q$, i.e., s_{pq} is large, the model is penalized if the predicted values of the response variables are inconsistent. Finally, the last term in the objective function is equivalent to the L_2 norm used in ridge regression models to shrink the coefficients in $\mathbf{w}$.

Each data point was considered to be a given elapsed time $t \in \{1, 2, \cdots, n\}$ in the time series. An $n \times n$ similarity matrix $\mathbf{S} = [s_{ij}]$ is computed between every pair of data points based on the similarities of their predictor variables. Prior to computing the similarity matrix, each variable is standardized by subtracting its mean value and then dividing by its corresponding standard deviation. The standardization of the variables is needed to account for their varying scales. Pearson correlation coefficient was used to compute the similarity between each pair of data points and then transform the value to a range between 0 and 1 to ensure ensure all the terms in the objective function are non-negative. The choice of Pearson correlation as the similarity measure is due to the popularity of the measure in the Earth science domain.

3.3.2 Parameter Estimation

The objective function can be further expanded as follows:

$$
\begin{aligned}
L(\mathbf{w},\mathbf{y}) &= \sum_{i=1}^{n} c_i(c_i' - y_i \sum_d w_d x_{i,d})^2 + T_1 \sum_{i=1}^{n} (y_i - c_i)^2 \\
&+ T_2 \sum_{i,j=1}^{n} s_{i,j} \Big(\sum_d c_i w_d x_{i,d} - \sum_d c_j w_d x_{j,d} \Big)^2 \\
&+ T_3 ||w||^2
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
L(\mathbf{w},\mathbf{y}) &= \sum_{i=1}^{n} c_i(c_i' - y_i \sum_d w_d x_{i,d})^2 \\
&+ T_1 \sum_{i=1}^{n} (y_i - c_i)^2 + T_3 ||w||^2 \\
&+ T_2 \sum_{i,j=1}^{n} s_{i,j} \bigg(\Big(\sum_d c_i w_d x_{i,d} \Big)^2 + \Big(\sum_d c_j w_d x_{j,d} \Big)^2 \\
&- 2 \sum_{d,d'} c_i c_j w_d w_{d'} x_{i,d} x_{j,d'} \bigg)
\end{aligned}
$$

To estimate the regression parameter $\mathbf{w}$ and class labels $\mathbf{y}$, the following iterative procedure was employed. First, the partial derivative of $L(\mathbf{w},\mathbf{y})$ is computed with respect to each of the w's and set it to zero (assuming $\mathbf{y}$ is fixed):

$$\begin{aligned}\frac{\partial L}{\partial w_k} = \Bigg[& -2\sum_{i=1}^{n} c_i\Big(c_i' - y_i \sum_d w_d x_{i,d}\Big)\Big(y_i x_{i,k}\Big) \\ & + 2T_2 \sum_{i,j=1}^{n} s_{i,j}\Big(\Big(\sum_d c_i w_d x_{i,d}\Big)\Big(c_i x_{i,k}\Big)\Big) \\ & + 2T_2 \sum_{i,j=1}^{n} s_{i,j}\Big(\Big(\sum_d c_j w_d x_{j,d}\Big)\Big(c_j x_{j,k}\Big)\Big) \\ & - 2T_2 \sum_{i,j=1}^{n} s_{i,j}\Big(\sum_d c_i c_j w_d (x_{i,d}x_{j,k} + x_{i,k}x_{j,d})\Big) \\ & + 2T_3 w_k \Bigg] = 0\end{aligned}$$

This reduces to a system of linear equations of the form $\mathbf{Aw} = \mathbf{b}$ where

$$b_k = \sum_{i=1}^{n} c_i y_i c_i' x_{i,k}$$

and A is a square matrix of dimension $d \times d$ whose non-diagonal elements is given by,

$$\begin{aligned}\mathbf{A}_{k,l} = & \; 2T_2 \sum_{i,j=1}^{n} s_{i,j} c_i x_{i,l} x_{i,k} \\ & - 2T_2 \sum_{i,j=1}^{n} s_{i,j} c_i c_j x_{i,l} x_{j,k} \\ & + \sum_{i=1}^{n} c_i y_i x_{i,l} x_{i,k}\end{aligned}$$

and diagonal elements

$$
\begin{aligned}
\mathbf{A}_{k,k} = 2T_2 & \sum_{i,j=1}^{n} s_{i,j} c_i x_{i,k}^2 \\
& - 2T_2 \sum_{i,j=1}^{n} s_{i,j} c_i c_j x_{i,k} x_{j,k} \\
& + \sum_{i=1}^{n} c_i y_i x_{i,k}^2 + T_3
\end{aligned}
$$

To estimate $\mathbf{y}$, the following part of the objective function that depends on $\mathbf{y}$ is minimized:

$$
L_c(\mathbf{y}) \quad = \quad \sum_{i=1}^{n} c_i (c_i' - y_i y_i')^2 + T_1 \sum_{i=1}^{n} (y_i - c_i)^2
$$

subject to the constraint $y_i \in \{0, 1\}$. It is straightforward to show that L_c is minimized according to the following rule:

$$
y_i = \begin{cases} 1, & \text{if } c_i = 1 \text{ and } (c_i' - y_i')^2 > c_i'^2 + T_1; \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}
$$

The predicted class labels $\mathbf{y}$ are then used to re-estimate the regression coefficients $\mathbf{w}$. This procedure is repeated until the regression coefficients and class labels converge.

3.3.3 Proof of Convergence

This section presents the proof of convergence of our iterative update algorithm. Let $(w_t, \mathbf{y}_t)$ be the regression coefficients and class labels estimated after the t-th iteration and $(w_{t+1}, \mathbf{y}_{t+1})$ be the regression coefficients and class labels estimated after the $(t+1)$-th iteration.

Proposition 3.3.1. *Assuming that the class labels $\mathbf{y}_t$ are fixed, $L(w_{t+1}, \mathbf{y}_t) \leq L(w_t, \mathbf{y}_t)$.*

Proof. For a fixed $\mathbf{y}_t$, let $L_r(\mathbf{w})$ be part of the objective function that depends on the regression coefficients $\mathbf{w}$:

$$\begin{aligned} L_r(\mathbf{w}) &= \sum_{i=1}^{n} c_i(c_i' - y_i \sum_d w_d x_{i,d})^2 + T_3||w||^2 \\ &+ T_2 \sum_{i,j=1}^{n} s_{i,j} \Big(\sum_d c_i w_d x_{i,d} - \sum_d c_j w_d x_{j,d} \Big)^2 \end{aligned}$$

The Hessian matrix $\mathbf{H}$ of $L_r(\mathbf{w})$ is given by:

$$\begin{aligned} \frac{\partial^2 L_r}{\partial w_k \partial w_l} &= 2 \sum_{i=1}^{n} c_i {y_i}^2 x_{i,k} x_{i,l} + 2T_3 \delta_{kl} \\ &+ 2T_2 \sum_{i,j=1}^{n} s_{i,j} (c_i x_{i,k} - c_j x_{j,k})(c_i x_{i,l} - c_j x_{j,l}) \end{aligned}$$

where $\delta_{kl} = 1$ if $k = l$ and zero otherwise. Since the parameters T_2 and T_3 are non-negative, it can be shown that, for any non-zero vector $\mathbf{z}$ with real values, $\mathbf{z}^T \mathbf{H} \mathbf{z} \geq 0$, i.e., the Hessian matrix is positive semi-definite. Thus, the stationary point $\mathbf{w}_{t+1}$ minimizes $L(\mathbf{w}_{t+1})$ and $L_r(\mathbf{w}_{t+1}) \leq L_r(\mathbf{w}_t)$.

□

Proposition 3.3.2. *Assuming that the regression coefficients are fixed, $L(\mathbf{w}_{t+1}, \mathbf{y}_{t+1}) \leq L(\mathbf{w}_{t+1}, y_t)$.*

Proof. For a fixed $\mathbf{w}_{t+1}$, let $L(\mathbf{w}_{t+1}, \mathbf{y}_t) = L_c(\mathbf{y}_t) + T_2 \sum_{i,j=1}^{n} s_{i,j} [c_i y_i' - c_j y_j']^2 + T_3||w||^2$. Note that last two terms are independent of $\mathbf{y}_t$. Since the update formula for $\mathbf{y}_t$ minimizes

$L_c(\mathbf{y})$, it follows that $L(\mathbf{w}_{t+1}, \mathbf{y}_{t+1}) \leq L(\mathbf{w}_{t+1}, \mathbf{y}_t)$.

□

Theorem 3.3.1. *The objective function $L(w)$ is monotonically non-increasing given the update formula for $\mathbf{w}$ and $\mathbf{y}$.*

Proof. The update formula iteratively modifies the objective function as follows: $L(\mathbf{w}_t, \mathbf{y}_t) \Rightarrow L(\mathbf{w}_{t+1}, \mathbf{y}_t) \Rightarrow L(\mathbf{w}_{t+1}, \mathbf{y}_{t+1})$. Using the above propositions we have $L(\mathbf{w}_{t+1}, \mathbf{y}_t) \leq L(\mathbf{w}_t, \mathbf{y}_t)$ and $L(\mathbf{w}_{t+1}, \mathbf{y}_{t+1}) \leq L(\mathbf{w}_{t+1}, \mathbf{y}_t)$. Therefore, $L(\mathbf{w}_{t+1}, \mathbf{y}_{t+1}) \leq L(\mathbf{w}_t, \mathbf{y}_t)$

□

Lemma 3.3.1. *The objective function will eventually converge, as the value of the loss function is always non-negative and since we know $L(w)$ is monotonically decreasing.*

3.3.4 Classification of Test Data

The update formula presented in the previous subsections compute the regression coefficients $\mathbf{w}$ and class labels $\mathbf{y}$ of the training examples in such a way that minimizes the objective function. For a given test example $\mathbf{x}_\tau$, where $\tau \in \{n+1, \cdots, n+m\}$, the predicted value of the regression model can be computed as follows: $y'_\tau = \mathbf{w}^T \mathbf{x}_\tau$. However, the classification output cannot be determined since the update formula for $\mathbf{y}$ depends on the true class labels $\mathbf{c}$, as shown in Equation (3.2). Therefore, to predict the class label $\mathbf{y}$, an SVM classifier on $(\mathbf{x}_t, \mathbf{y}'_t)$ as the $d+1$-dimensional feature vector and the estimated $(\mathbf{y}_t)$ as the class labels using only examples from training data. Once the classifier has been constructed, it can be applied to predict the class label of a test example. The final output of the joint classification and regression model is the product $y_\tau y'_\tau$ (see Figure 3.2).

Empirically, it was found that SVM may be used as an alternate classifier to predict y at each iteration, instead of the update formula described above. But since the objective function of the generic classifier does not necessarily minimize both the first and second term of $L_c(y)$ simultaneously, convergence cannot be guaranteed.

3.4 Integrated Classification and Regression Algorithm

The Integrated Classification and Regression (ICR) framework takes as input $(\mathbf{x}_t, \mathbf{c}'_t)$ (a multivariate time series with d-dimensional predictor variables $\mathbf{x}_t$ and response variable $\mathbf{c}'_t$) and a sequence of unlabeled observations $(\mathbf{x}_\tau)$. The output returned by the framework are the regression coefficients $(\mathbf{w})$ and the predicted values of the unlabeled sequence $(\mathbf{z}_\tau)$. For the training phase set $\mathbf{c} = (\mathbf{c}' > 0)$ and initialize $\mathbf{y} = \mathbf{c}$. Then until convergence update $\mathbf{w}$ by solving $\mathbf{A}\mathbf{w} = \mathbf{b}$ followed by updating $\mathbf{y}$ using Equation (3.2). Then train an SVM classifier $g : (\mathbf{x}_t, \mathbf{y}'_t) \rightarrow \mathbf{y}_t$. During the testing phase, set $\forall\tau : y'_\tau = \mathbf{w}^T\mathbf{x}_\tau$, $\forall\tau : y_\tau = g(\mathbf{x}_\tau, y'_\tau)$ and $\forall\tau : z_\tau = y_\tau y'_\tau$.

It is assumed that the time series data has been partitioned into a training set, a validation set (for model selection), and a test set. Model selection is needed to estimate the parameters T_1,T_2,T_3 of our objective function $L(\mathbf{w}, \mathbf{y})$.

The class labels $\mathbf{c}$ of the training examples are obtained based on the response variable $\mathbf{c}'$. The training phase of the algorithm starts by setting $\mathbf{y} = \mathbf{c}$ for all the n-training examples. It then iteratively updates the regression coefficients $\mathbf{w}$ and class labels $\mathbf{y}$ according to the methodology presented in the previous section. At this stage, the value of the objective function is computed and saved for testing convergence of the objective function. Upon convergence, an SVM classifier g is constructed to learn the mapping between the input features $\mathbf{x}, \mathbf{y}'$ and output class $\mathbf{y}$.

Once the training phase is completed, the Testing phase begins. Testing is performed by first applying the multiple linear regression model to the predictor variables $\mathbf{x}_\mathcal{T}$. This is followed by invoking the SVM classifier to predict the class label $y_\mathcal{T}$ for the m test examples. The classifier takes $\mathbf{x}_\mathcal{T}$ and $y'_\mathcal{T}$ as input and returns class labels $y_\mathcal{T}$. Finally, the prediction output is obtained by setting $z_\mathcal{T} = y_\mathcal{T} y'_\mathcal{T}$.

The time complexity of the training phase of the algorithm is $O(k(n^2d + d^3))$, where n is the number of training examples, d the number of predictor variables and k is the maximum number of iterations required for convergence. The computational complexity of the training phase is composed of two major parts: the first that requires computing the similarity matrix and the second that requires iteratively solving $\mathbf{w}$ and $\mathbf{y}$. The time needed to compute the similarity matrix is $(O(n^2d))$. The time complexity of each iteration refers to the time needed to compute $\mathbf{w}$ $(O(n^2d^2 + d^3))$ plus time needed to compute $\mathbf{y}$ $(O(n))$. Hence, for maximum iterations set to k, the time complexity for the training phase is $O(k(n^2d + d^3))$, where $d \ll n$.

3.5 Experimental Evaluation

This section presents the experimental results to demonstrate the effectiveness of the proposed framework.

3.5.1 Experimental Setup

The ICR algorithm was run on climate data obtained for 37 weather stations in Canada, from the Canadian Climate Change Scenarios Network Web site [1]. The response variable to be regressed corresponds to daily precipitation values measured at each weather station. The predictor variables correspond to 26 coarse-scale climate variables derived from the NCEP Reanalysis data set, which include measurements of airflow strenght, sea-level pressure, wind direction, vorticity, and humidity, as shown in Table 7.1. The data span a 40-year period, 1961 to 2001. The time series was truncated for each weather station to exclude days for which the precipitation values are missing.

Table 3.1: List of predictor variables for precipitation prediction.

Predictor Variables	
Mean sea level pressure	Surface zonal velocity
Surface airflow strength	Surface meridional velocity
Surface vorticity	Surface wind direction
Surface divergence	Mean temp at 2m
500 hPa airflow strength	850 hPa airflow strength
500 hPa zonal velocity	850 hPa zonal velocity
500 hPa meridional velocity	850 hPa meridional velocity
500 hPa vorticity	850 hPa vorticity
500 hPa geopotential height	850 hPa geopotential height
500 hPa wind direction	850 hPa wind direction
500 hPa divergence	850 hPa divergence
Relative humidity at 500 hPa	Relative humidity at 850 hPa
Near surface relative humidity	Surface specific humidity

A comparison of the performance of the algorithm(ICR) was made against the multiple linear regression (MLR) model and an approach that combined SVM and MLR (SVM-MLR). MLR uses the least square criterion to estimate the weight vector $\mathbf{w}$ of the model. In SVM-MLR, SVM was used to learn a classifier model to differentiate between `Rain` and `NoRain` days, and MLR was learnt on rain days only. Finally, for the given test set MLR is applied only to those days classified as a `Rain` day. As far as choice of SVM is concerned, during the evaluation phase a choice of the kernel (Linear or RBF) and its respective parameter is made. The choice of the SVM kernel for ICR was limited to a linear kernel. Future experiments will include a wider selection during the evaluation phase.

The following criteria was used to evaluate the performance of the models:

- Root Mean Square Error (RMSE), which measures the difference between the actual and predicted values of the response variable, i.e.: RMSE $= \sqrt{\frac{\sum_1^n (c_i' - y_i')^2}{n}}$.

- Accuracy, which measures the number of Rain and NoRain days predicted correctly by the model.

- F-measure, which is the harmonic mean between recall and precision values for rain days.

3.5.2 Experimental Results

The purpose of the experiment is to demonstrate the following:

1. Limitations of classical regression models in terms of handling zero-inflated time series data.

2. Performance comparison between classical regression models and the proposed framework.

3.5.2.1 Effect of Zero-Inflated Time Series Data

The objective of this experiment is to demonstrate the effect of increasing number of zeros in a time series on the performance of a regression model. Specifically, given the precipitation time series of a randomly selected weather station, each day was classified as NoRain or Rain, depending on the amount of precipitation it receives is equal to or greater than zero. Several training sets of different sizes and varying percentage of NoRain and Rain days by randomly

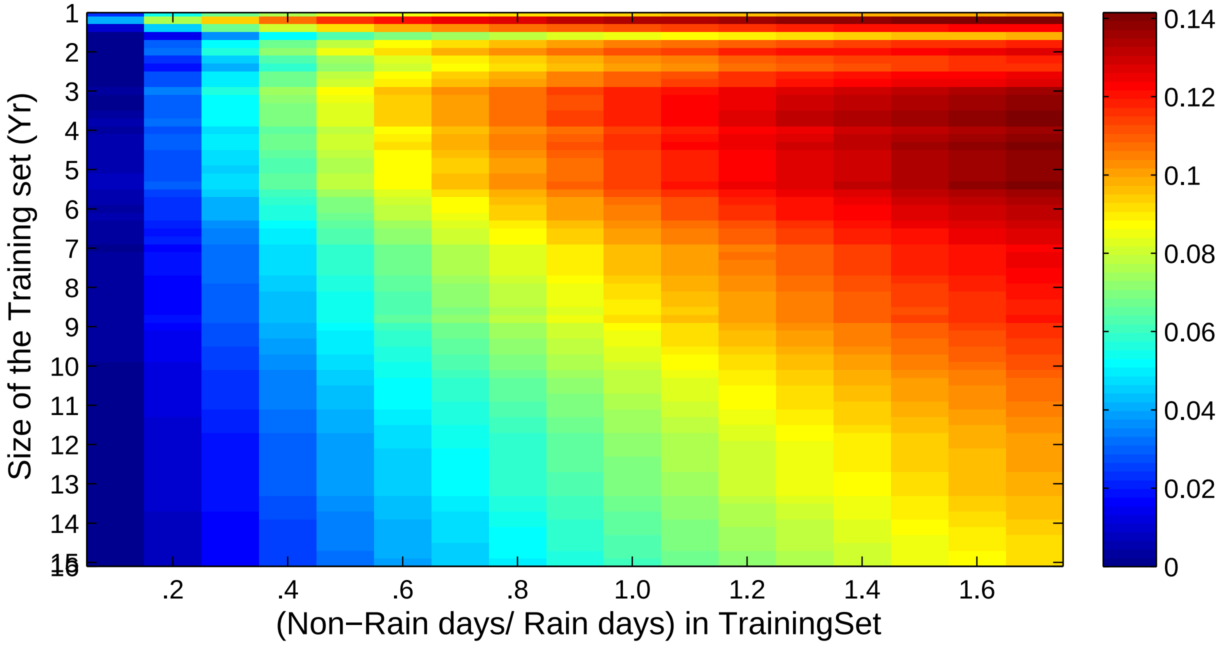


Figure 3.3: Effect of increasing the number of NoRain days on performance of regression model (best viewed in color).

sampling the original time series, were created. A disjoint test set of size ten years, is used for all the experiments in this subsection.

The performance of two multiple linear regression (MLR) models was evaluated: (1) MLR_1, which is trained on both Rain and NoRain days and (2) MLR_2, which is trained on Rain days only. Figure 3.3 compares the RMSE values of both models for Rain days in the test set. The horizontal axis corresponds to the ratio of NoRain to Rain days in the training set. The larger the ratio, the more inflated the number of zeros in the training data. The vertical axis corresponds to the training set size, where each unit on the scale represents a period of three months. The value of each cell indicates the performance improvement when using MLR_2 to predict the Rain days:

$$\%\text{Improvement} = \frac{\text{RMSE}(\text{MLR}_1) - \text{RMSE}(\text{MLR}_2)}{\text{RMSE}(\text{MLR}_1)} \tag{3.3}$$

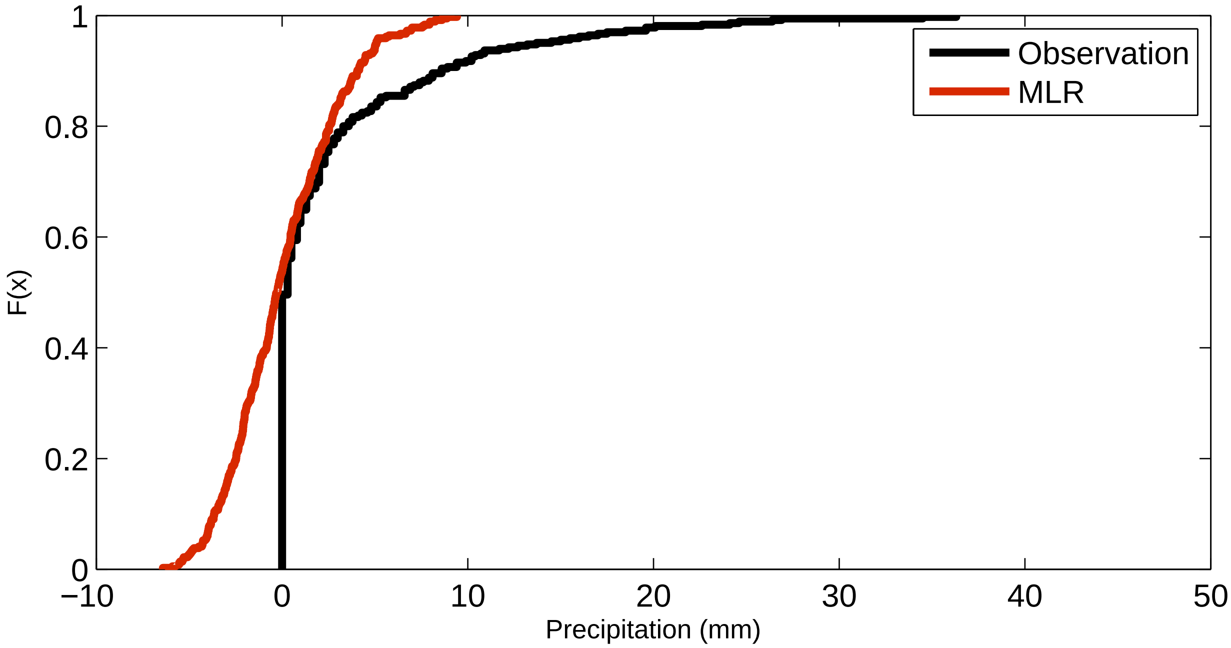


Figure 3.4: The cumulative distribution function of multiple linear regression output for a zero-inflated precipitation response variable.

Since the % Improvement is greater than or equal to zero, this indicates that MLR_2 consistently outperforms MLR_1 in terms of predicting future `Rain` days irrespective of the training set size. The amount of improvement becomes even more pronounced when the percentage of `NoRain` days in the training data increases. A similar improvement pattern is observed for all the weather stations investigated in this study, as shown in Figure 3.5. In contrast, MLR_1, which is trained on both `Rain` and `NoRain` days, has a lower RMSE compared to MLR_2 when applied to all the days in the test set, as shown in Figure 3.6. This is because MLR_2 tends to overestimate the amount of precipitation for the `NoRain` days.

In summary, the experiment given in this section clearly justifies the rationale for applying a combination of classification and regression models to better estimate the precipitation amount of `Rain` days.

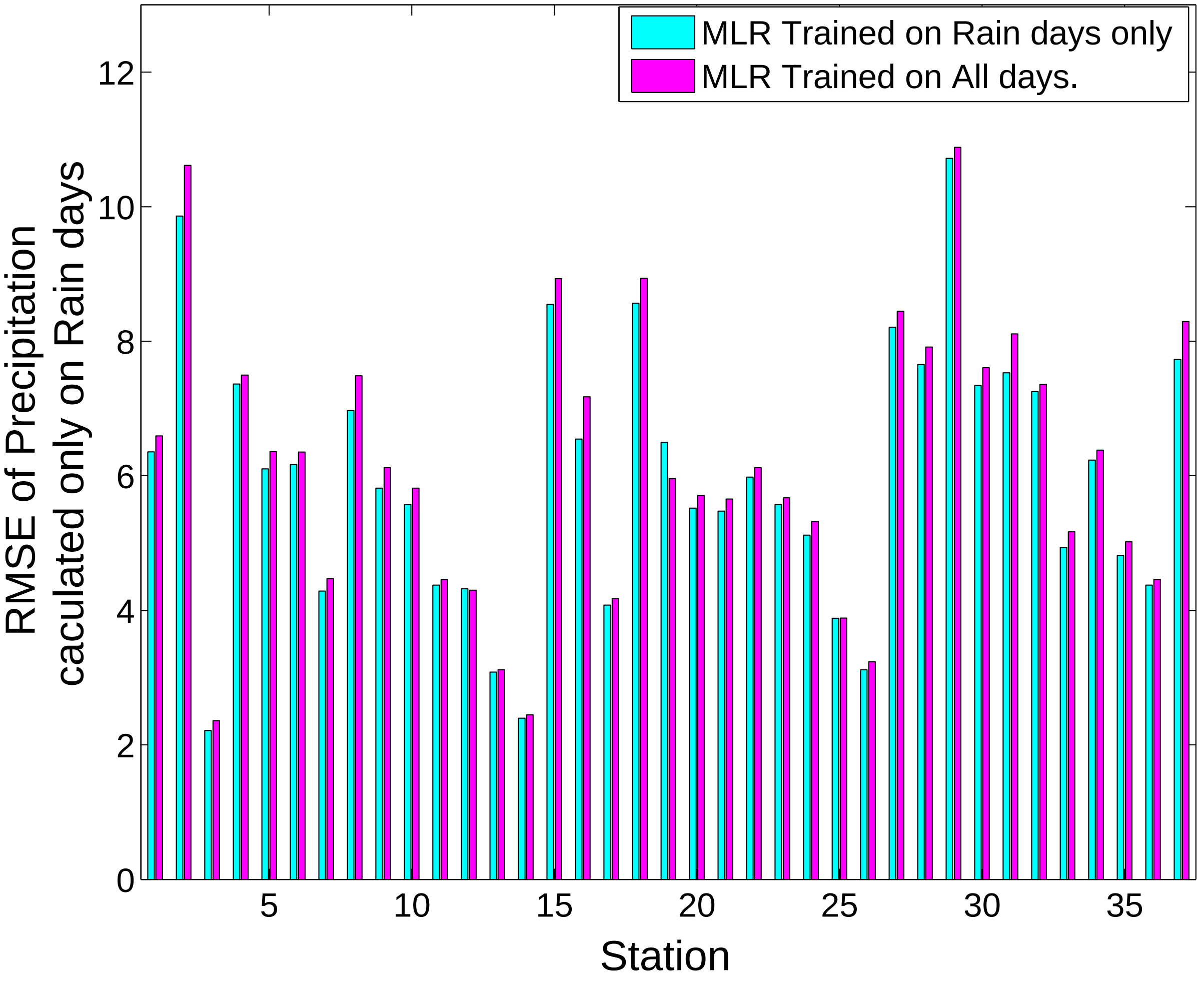


Figure 3.5: Comparison of RMSE values (tested on Rain days only) for MLR models trained on all days compared with models trained only on Rain days.

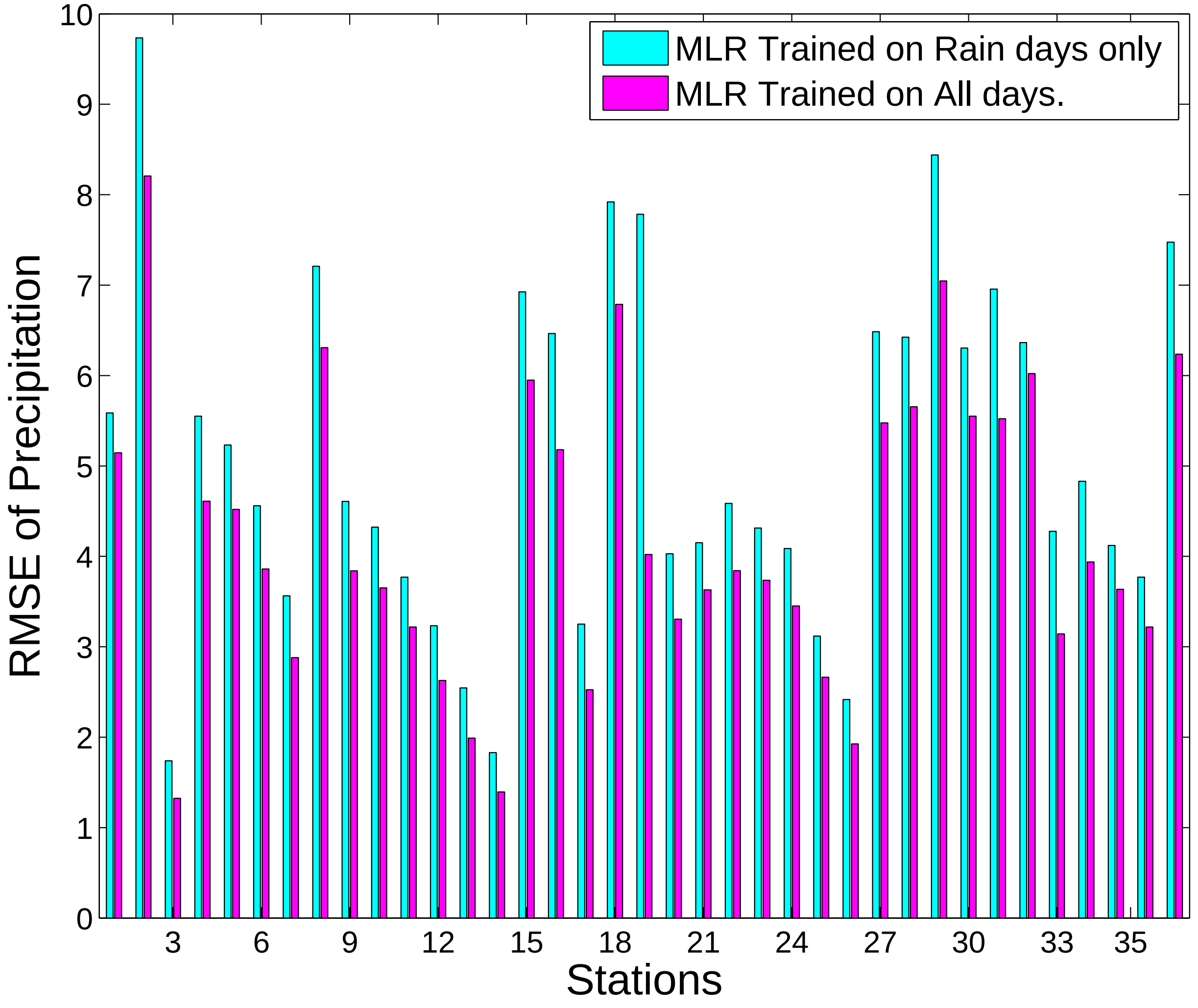


Figure 3.6: Comparison of RMSE values (tested on All days) for MLR models trained on all days compared with models trained only on Rain days.

3.5.2.2 Impact of Coupling the Classifier and Regression Model Creation

The objective of this experiment is to demonstrate the advantage of building a classifier and regression model in conjunction with each other, as against building them independent of the other for zero-inflated time-series data. Specifically, empirical results demonstrating improvement in the classification accuracy, F-measure of classification as well as RMSE of the predictors are provided.

The performance of two multiple linear regression models was evaluated and compared. In the first model, MLR is trained on all days and a quadratic discriminant analysis (QDA) trained on ground truth response variable. In the second model, again MLR is trained on all days but the QDA trained on the predicted response values $\mathbf{y'} = \mathbf{w}^T\mathbf{x}$. The results of the experiment show that the model trained on the predicted response values outperformed the model trained on ground truth response variable for all 37 stations, when it came to RMSE, Classification Accuracy and F-Measure. In particular, the average improvements were 13.4% and 19.3% when it came to RMSE and classification accuracy.

In summary, these empirical results provide motivation to try and integrate the classifier and regression models to take into consideration the accuracy of the other's prediction for each individual data point.

3.5.2.3 Performance Comparison

This section compares the RMSE, accuracy, and F-measure values of the predicted response variable (Precipitation) for our proposed supervised (ICR) framework against that of multiple linear regression (MLR), SVM-MLR (A model that combines MLR and SVM) and classification and regression tree (CART). All the experiments were performed using a training size (n) of 3 years starting from the first observation in the time series. The test set size (m) was also fixed at 1 year. After calculating the RMSE on the test set, the training set was shifted by 3 years, such that it now occupied the data set used for testing in the previous iteration. The experiment is repeated 5 times for each station. The RMSE values reported in this section is the mean value of all 5 iterations. The same approach is used to compute the RMSE values for `Rain` days, accuracy (for all days), F-measure for `Rain` days only and F-measure for `NoRain` days only. The results for 37 weather stations when ICR is compared with both MLR and SVM-MLR, is presented. Classification accuracy, and F-measures related to classification accuracy of MLR is not plotted on account of MLR not having an explicit classifier. CART fared comparatively poorly in terms of residual errors and classification accuracy and F-measure. However, CART fared well in replicating the cumulative distribution function of the response variable as shown later in the experiment section.

As shown in Figures 3.7 and 3.8, our supervised model, ICR significantly outperformed the MLR model (trained on all days) and the SVM-MLR model in terms of their RMSE values for predicting both `Rain` and `NoRain` days.

ICR outperformed MLR in 36 out of 37 stations and outperformed SVM-MLR in 30 out of the 37 stations. In terms of percentage improvement in RMSE for all days, ICR indicated an average 8% improvement over MLR and 5.8% improvement when compared to SVM-MLR.

In terms of the RMSE values for `Rain` days only, as shown in Figures 3.9 and 3.10, ICR consistently outperformed both the MLR and SVM-MLR model with ICR outperforming MLR in 35 stations and ICR outperforming SVM-MLR in 33 stations. When evaluating average RMSE value for `Rain` days only, ICR had an improvement of 5.3% over MLR and 8.6% over SVM-MLR.

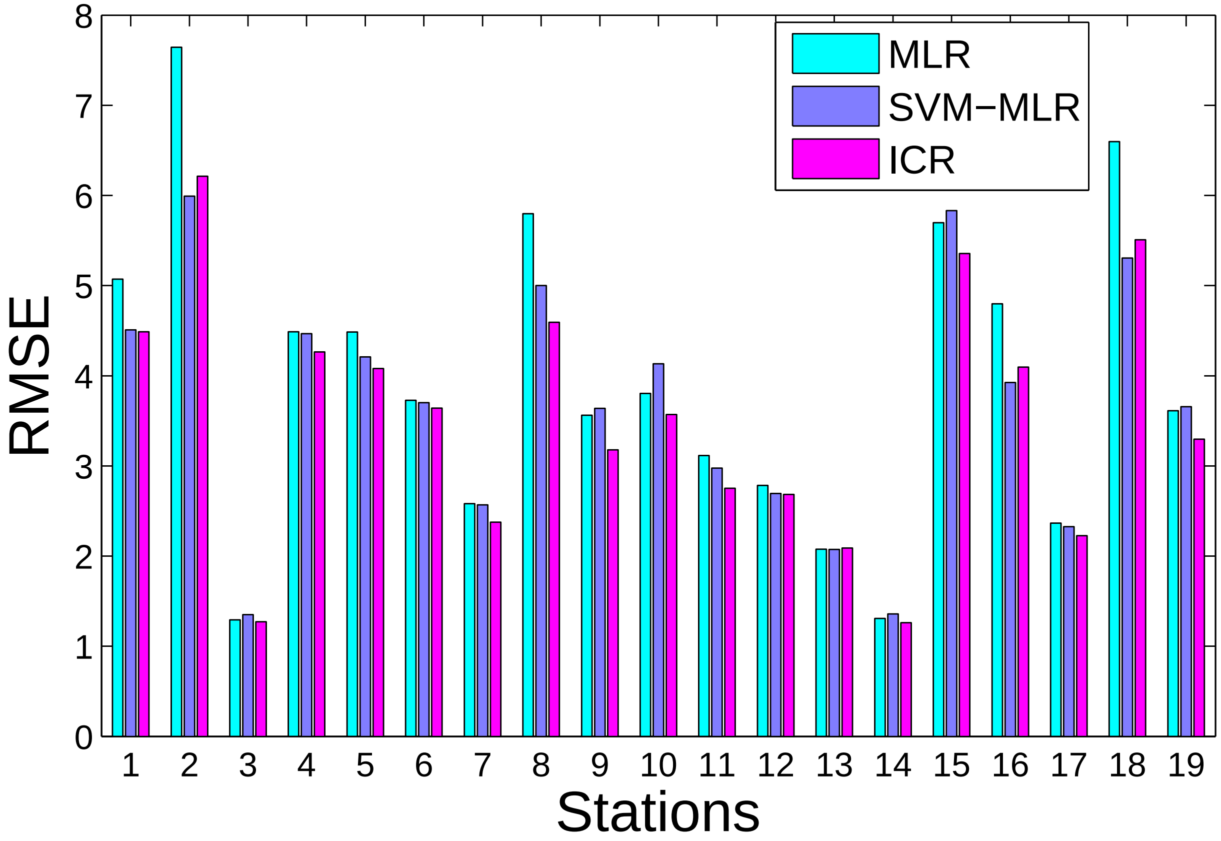


Figure 3.7: Comparison of RMSE values (for all days) among MLR, SVM-MLR and ICR.

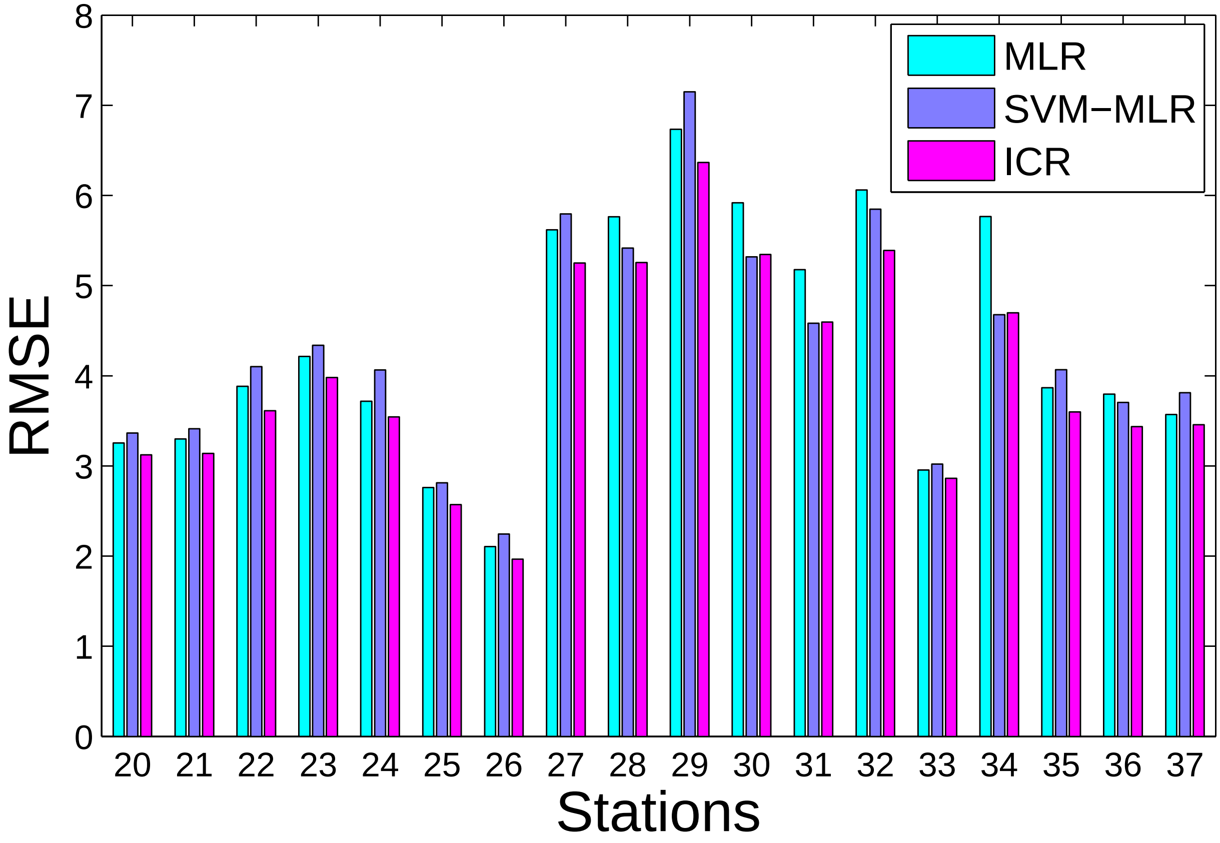


Figure 3.8: Comparison of RMSE values (for all days) among MLR, SVM-MLR and ICR.

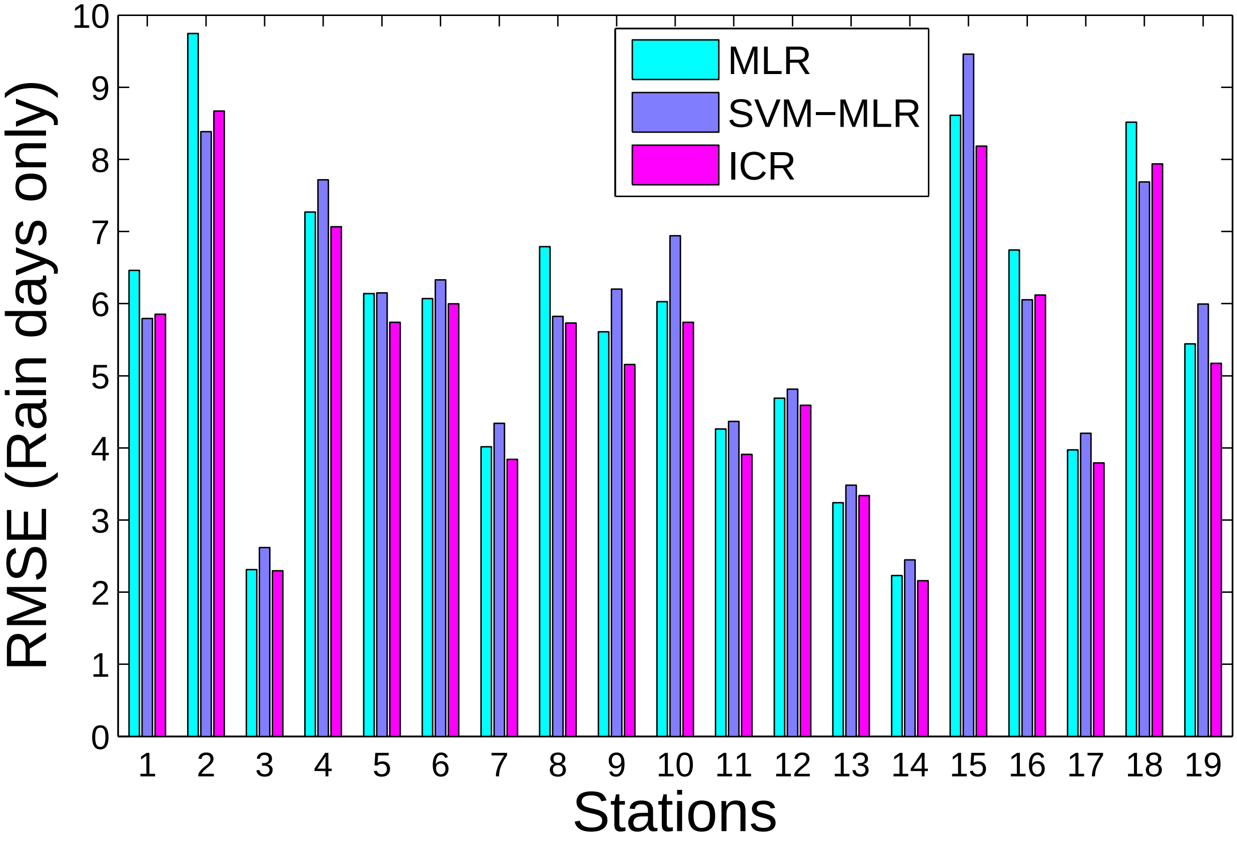


Figure 3.9: Comparison of RMSE values (for `Rain` days) among MLR, SVM-MLR and ICR.

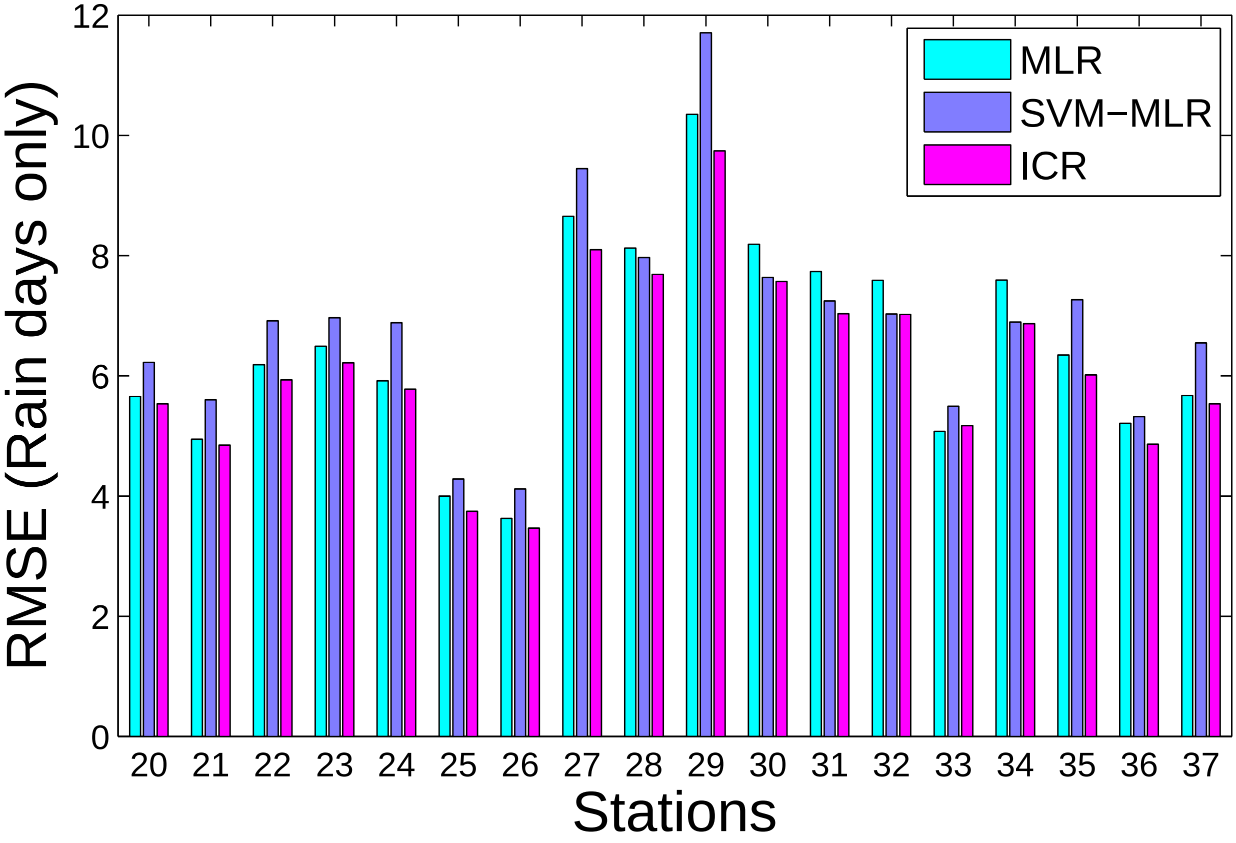


Figure 3.10: Comparison of RMSE values (for `Rain` days) among MLR, SVM-MLR and ICR.

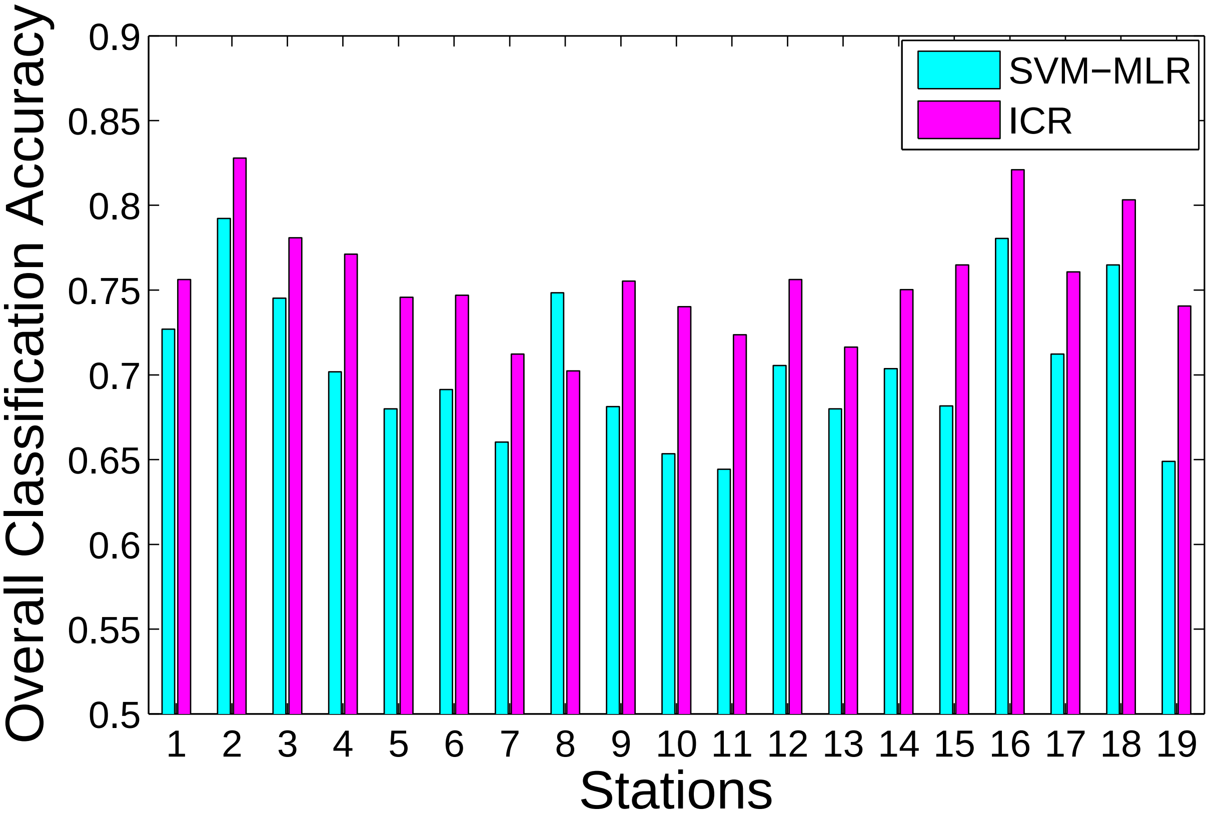


Figure 3.11: Comparison of classification accuracy (for all days) between SVM-MLR and ICR.

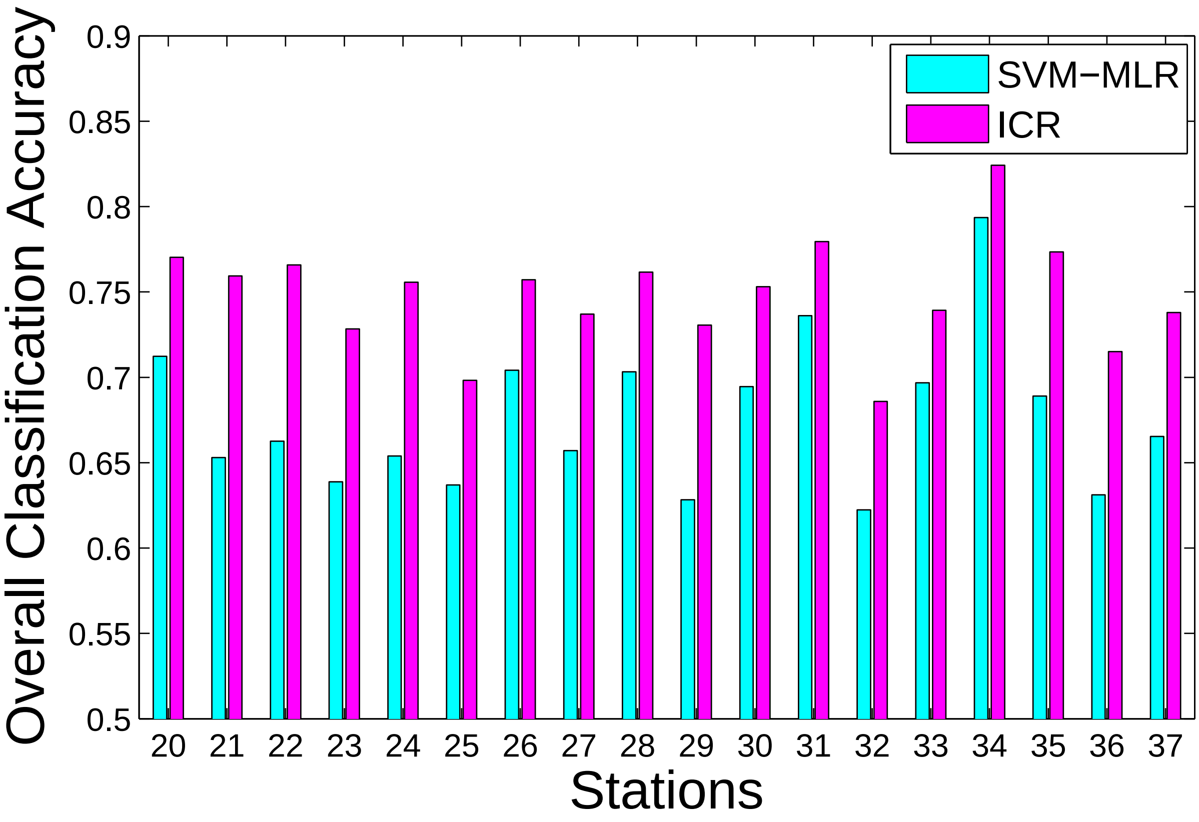


Figure 3.12: Comparison of classification accuracy (for all days) between SVM-MLR and ICR.

MLR does not inherently classify any days as Rain or NoRain. Hence, a comparison between ICR and MLR with regards to classification accuracy and F-measure, is not plotted.

As shown in Figures 3.11 and 3.12, ICR outperformed SVM-MLR in 36 of the 37 stations and showed a 9.1% improvement in classification accuracy. At the same time, in terms of F-measure for Rain days, the model outperformed SVM-MLR, as shown in Figures 3.13, 3.14. ICR outperformed SVM-MLR in 35 out of the 37 stations.

Although, MLR does not inherently classify any days as Rain or NoRain, a Quadratic Discriminant Analysis(QDA) classifier mentioned earlier was trained on the MLR output. ICR witnessed a 21.2% improvement in overall classification accuracy.

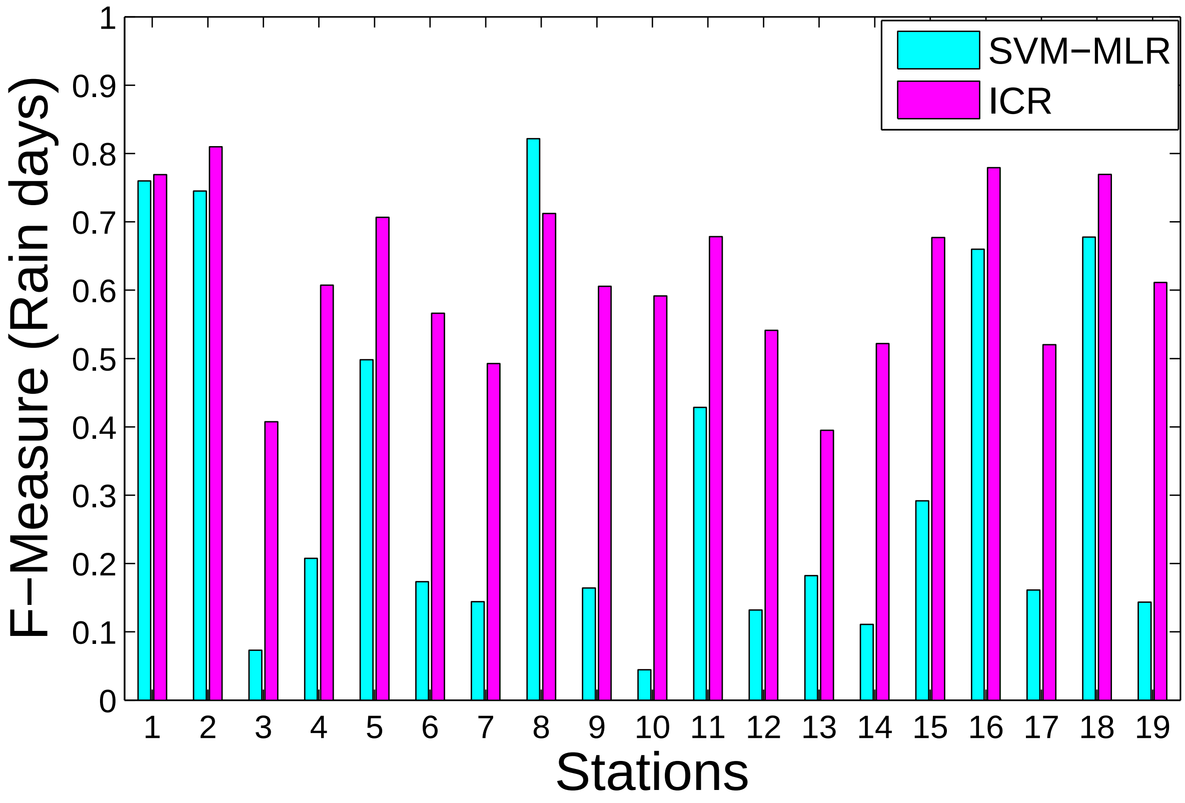


Figure 3.13: Comparison of F-measure (for Rain days) between SVM-MLR and ICR.

With regard to F-measure for NoRain days, ICR outperformed SVM-MLR, in 36 stations.

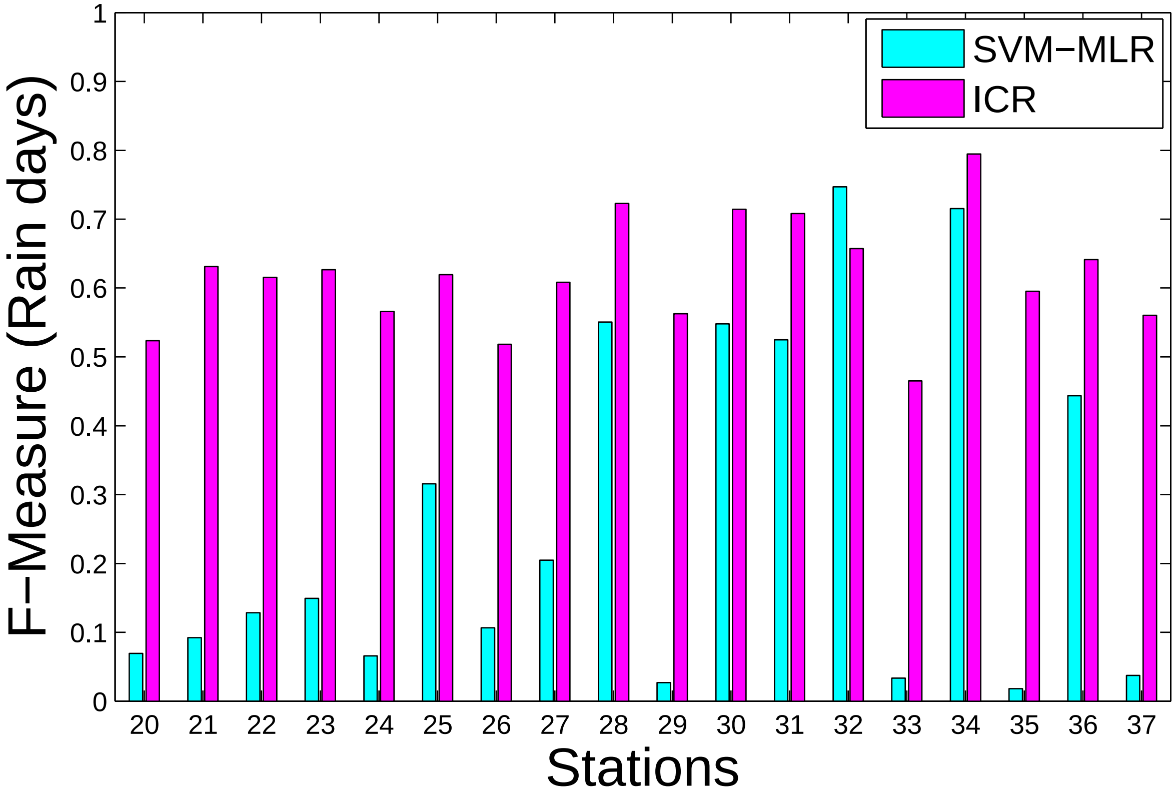


Figure 3.14: Comparison of F-Measure (for `Rain` days) between SVM-MLR and ICR.

As shown in Figures 3.15,3.16 that shows the comparison of F-Measure for `NoRain` days between SVM-MLR and ICR, ICR outperformed SVM-MLR in all but one station and witnessed an 8.1% improvement in F-measure results.

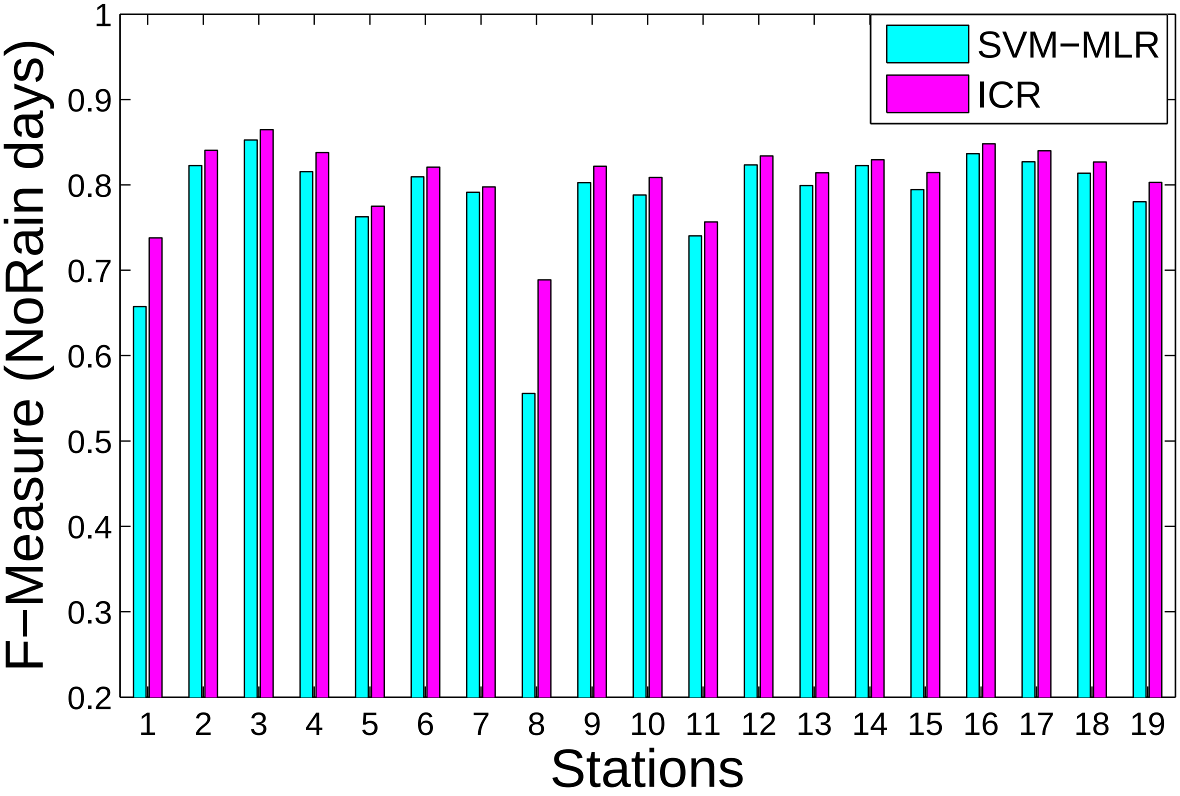


Figure 3.15: Comparison of F-Measure (for `NoRain` days) between SVM-MLR and ICR.

As shown in the following figure, ICR as well as SVM-MLR was able to capture the frequency of zero-valued data points in the distribution as well as improve the shape of the cumulative distribution function when compared to that of multiple linear regression. CART fared the best in terms of replicating the distribution of the zero inflated precipitation observed, especially for the higher quantiles. However, CART fared less favorably in terms of capturing the frequency of zero-valued data points as shown in Figure 3.17. Also, ICR prediction showed a 28.9% and 24.2% improvement over the CART output in terms of

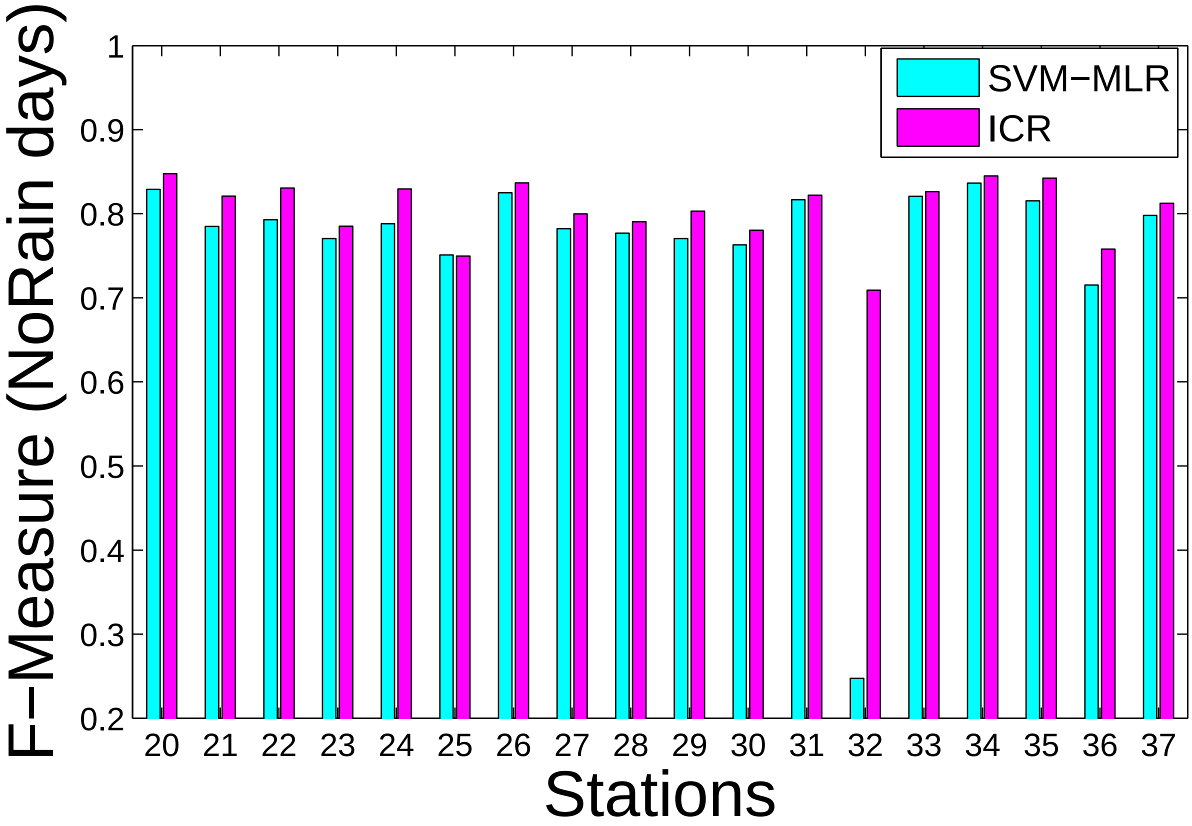


Figure 3.16: Comparison of F-Measure (for `NoRain` days) between SVM-MLR and ICR.

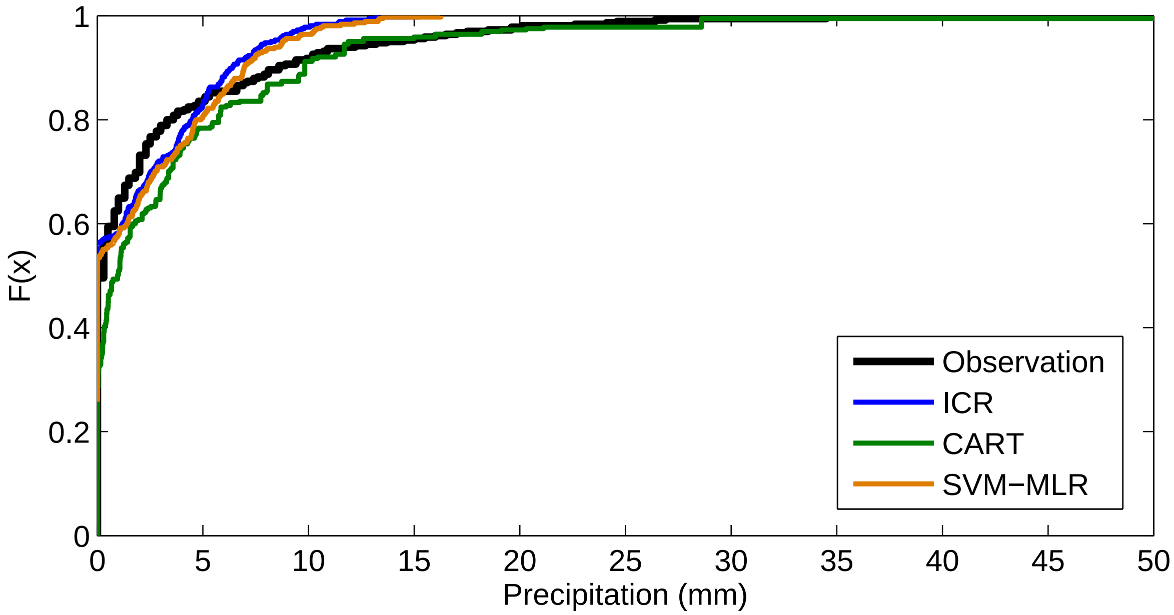


Figure 3.17: Comparing the cumulative distribution function of the predicted values of precipitation according to the various models.

RMSE for all-days and rain-days respectively. Similarly, the classification accuracy of ICR was 12.9% better than CART. The F-measure of ICR for rain days and non-rain days 5.3% and 25.2% better.

3.6 Conclusions

This chapter presents a novel approach for predicting future values of a time series data that are inherently zero-inflated. The proposed framework decouples the prediction task into two steps—a classification step to predict whether the value of the time series is zero and a regression step to estimate the magnitude of the non-zero time series value. The effectiveness of the model was demonstrated on climate data to predict the amount of precipitation at a given station.

The framework presented in this chapter assumes a linear relationship between the predictor and response variables. The framework can also be extended to a semi-supervised learning setting as shown in the Chapter 3.

Chapter 4

Semi-Supervised Modeling of Zero-Inflated Data

This chapter demonstrates a semi-supervised extension of the ICR framework, presented in Chapter 3 . The purpose of the framework is to utilize both labeled and unlabeled data to accurately estimate the future values of a zero-inflated variable, by simultaneously performing classification and regression. The regression and classification models are simultaneously learned by optimizing a unified objective function that includes a graph regularization term to ensure smoothness of their target functions and consistency between the labeled and unlabeled examples. The effectiveness of the semi-supervised learning framework is also demonstrated in the context of precipitation prediction using climate data obtained from the Canadian Climate Change Scenarios Network website [1]. The proposed framework significantly outperforms regression models trained on both zero and non-zero parts of the time series for the majority of the weather stations investigated in this study.

4.1 Preliminaries

Let $\mathbf{L} = (\mathbf{X}_l, \mathbf{c}'_l)$ be a multivariate time series of length l, where the predictor variables $\mathbf{X}_l = [\mathbf{x}_{l1}, \mathbf{x}_{l2}, ..., \mathbf{x}_{ln}]^T$ is a d-dimensional sequence of values and $\mathbf{c}'_l = \left[c'_{l1}, c'_{l2}, ..., c'_{ln}\right]^T$ is the corresponding ground truth values for the response variable. The objective of time series prediction is to learn a target function $f(\mathbf{x}, \mathbf{w})$ that best estimates the future values of the response variable, $\mathbf{c}'_u = \left[c'_{u1}, c'_{u2}, ..., c'_{um}\right]^T$, given the historical data $\mathbf{L}$ and the unlabeled data, $\mathbf{X}_u = [\mathbf{x}_{u1}, \mathbf{x}_{u2}, ..., \mathbf{x}_{um}]^T$, where $\mathbf{w} = [w_1, w_2, ..., w_d]^T$ is the set of weights associated with the target function. $\mathbf{X}_u$ may be obtained, for example, using computer-driven simulation models. In the semi-supervised framework proposed in this study, let n represent the number of labeled training points and m the number of unlabeled training points. In the supervised framework proposed, m represents the number of unlabeled testing points.

In this study, the relative frequency of zero values in $\mathbf{c}'_l$ and $\mathbf{c}'_u$ is assumed to be larger than the frequency of non-zero values. Furthermore, the response variable c' can be mapped into a binary class c, where $c = 1$ if $c' > 0$, and $c = 0$ otherwise. For brevity, the notation $y' \equiv f(\mathbf{x}, \mathbf{w})$ is used as the predicted value of the response variable and y as the predicted class. Let, $\mathbf{y}'_u = \left[y'_{l1}, y'_{l2}, ..., y'_{lm}\right]^T$ and $\mathbf{y}_u = [y_{l1}, y_{l2}, ..., y_{lm}]^T$.

In the semi-supervised framework proposed, let $\tilde{\mathbf{y}}$ be a vector of length $n+m$ whose first n elements are initialized with the vector $\mathbf{c}_l$ and whose remaining m elements are initialized with the vector $\mathbf{y}_u$. Hence, in the supervised framework proposed, as there are no unlabeled training points, $\tilde{\mathbf{y}}$ is a vector of length n and is initialized with the vector $\mathbf{c}_l$.

4.2 Semi-Supervised Framework for Simultaneous Classification and Regression

Unlike ICR, the semi-supervised extension modifies the graph regularization term to be summed over $n + m$ data points, where the m refers to the unlabeled data points. In the remainder of this chapter, the supervised and semi-supervised versions of the algorithm are denoted as ZICR-S and ZICR-SS, respectively (where ZICR stands for Zero-Inflated Classification-Regression method).

For brevity, only linear regression models were considered, where $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. Extending the approach to nonlinear models will be a subject for future research.

The goal was to simultaneously estimate the values of the weight parameters $\mathbf{w}$ and the class labels $\mathbf{y}$ to minimize the following objective function:

$$\begin{aligned}\arg\min_{\mathbf{w},\mathbf{y}} f(\mathbf{w}) &= \sum_{i=1}^{n} c_i(c_i' - y_i y_i')^2 + T_1 \sum_{i=1}^{n} (y_i - c_i)^2 \\ &+ T_2 \sum_{i=1}^{n} \sum_{j=1}^{n+m} s_{i,j}[c_i y_i' - \tilde{y}_j y_j']^2 + T_3 ||w||^2\end{aligned}$$

where,

$$\sum_d x_{i,d} w_d = y_i'.$$

Intuitively the first term of the objective function is equivalent to the least square formulation of multiple linear regression, except the estimation of $\mathbf{w}$ is performed based on the rain days only. The second term of the objective function measures the classification accuracy on the training data. The third term in the objective function computes the sum of squared difference in the predicted response values for every pair of data points, weighted by the similarity value of their predictor variables. This represents a graph regularization constraint to ensure smoothness of the objective function and can be used to extend the framework to a semi-supervised learning setting. Unlike ICR, ZICR the observed class label term in the graph regularizer component of the equation is replaced by its expected value, to that it could be extended to include unlabeled data points. Finally, the last term of the objective function is equivalent to the L_2 norm used in ridge regression models to penalize models that have many large non-zero weights.

Note that each data point corresponds to a given time period in the time series. The similarity matrix S is computed according to the Pearson correlation coefficient between every pair of data points in $\mathbf{X}$. Prior to computing the similarity matrix, each attribute value of the data set is standardized by subtracting the mean value of the attribute and then dividing by its corresponding standard deviation. The standardization of each column is done to account for differences in the variance of the various attributes in the data set. The Pearson correlation value is then transformed to range between 0 and 1. The choice of Pearson correlation as our similarity measure is due to the popularity of the measure in the Earth science domain.

The purpose of the similarity function is to identify how closely related two data points are to one another, and to use this information in creating the regression model which gives more credence to closeness in the predicted amount of precipitation for data points that are similar as against to data points that are dissimilar. As the similarity function has values ranging between 0 to 1, dissimilar data points have limited impact on the error function while similar data points that differ significantly on the amount of predicted precipitation have the largest impact on the error function. The model further emphasizes on using data points that are categorized as rain events by using '0' and '1' as class labels. Such that '0' is assigned to days that are categorized as '`NoRain`' days and '1' to '`Rain`' days.

The supervised version of the framework is obtained by considering only the labeled training examples for the third term in the objective function.

An iterative procedure was employed to solve the objective function. First, the partial derivative of $f(\mathbf{w})$ with respect to each of the w's is computed, and set to zero:

$$\begin{aligned}\frac{\partial f}{\partial w_k} = \Bigg[&2T_2 \sum_{i,j=1}^{n+m} s_{i,j}\bigg(\Big(\sum_d \tilde{y}_i w_d x_{i,d}\Big)\Big(\tilde{y}_i x_{i,k}\Big)\bigg) \\ &+ 2T_2 \sum_{i,j=1}^{n+m} s_{i,j}\bigg(\Big(\sum_d \tilde{y}_j w_d x_{j,d}\Big)\Big(\tilde{y}_j x_{j,k}\Big)\bigg) \\ &- 2T_2 \sum_{i,j=1}^{n+m} s_{i,j}\bigg(\sum_d \tilde{y}_i \tilde{y}_j w_d (x_{i,d} x_{j,k} + x_{i,k} x_{j,d})\bigg) \\ &- 2\sum_{i=1}^{n} c_i \Big(c_i' - y_i \sum_d w_d x_{i,d}\Big)\Big(x_{i,k}\Big) \\ &+ 2T_3 w_k\Bigg] = 0\end{aligned}$$

This reduces to a system of linear equations of the form $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{x} = [w_1 w_2 .... w_d]^T$ and

$$b_k = \sum_{i=1}^{n} c_i c_i' x_{i,k}$$

A is a square matrix of dimension $d \times d$ where the non-diagonal elements,

$$\begin{aligned}\mathbf{A}_{k,l} = &2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i x_{i,l} x_{i,k} - 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i \tilde{y}_j x_{i,l} x_{j,k} \\ &+ \sum_{i=1}^{n} c_i y_i x_{i,l} x_{i,k}\end{aligned}$$

and the diagonal elements

$$\mathbf{A}_{k,k} = 2T_2 \sum_{i,j=1}^{n+m} s_{i,j}\tilde{y}_i x_{i,k}^2 - 2T_2 \sum_{i,j=1}^{n+m} s_{i,j}\tilde{y}_i\tilde{y}_j x_{i,k}x_{j,k} + \sum_{i=1}^{n} c_i y_i x_{i,k}^2 + T_3$$

Next quadratic discriminant analysis (QDA) was applied on the predicted response values $\mathbf{y'} = \mathbf{w}^T\mathbf{x}$ to estimate the class labels of the unlabeled data points. The updated class labels $\mathbf{y}$ are then used to re-estimate the regression weights $\mathbf{w}$. This procedure is repeated until convergence. A summary of the framework is presented in Algorithm 1. In the remainder of this paper, the supervised and semi-supervised versions of our algorithm are denoted as ZICR-S and ZICR-SS, respectively (where ZICR stands for Zero-Inflated Classification-Regression method).

4.3 Experimental Evaluation

This section presents the experimental results to demonstrate the effectiveness of our proposed framework.

4.3.1 Experimental Setup

The set up of the experiments discussed in this chapter is similar to the experiment setup described in Chapter 3. The performance of the algorithm was compared against the multiple linear regression (MLR) model. MLR uses the least square criterion to estimate the weight

Algorithm 1 Concurrent Semi-supervised Regression and classification using simultaneous equations iteratively.

```
Input:
$\mathbf{X}$ (An $(n+m) \times d$ matrix of NCEP weather data)
$\mathbf{c}$ (A $n$-dimension vector of class labels (1-Rain/0-NoRain))
$\mathbf{c'}$ (A $n$-dimension vector of precipitation values for each day.)
Output:
$\mathbf{w}$ (A $d$-dimensional vector of weights)
$\mathbf{y}$ (A $(n+m)$-dimensional vector containing class labels.)
$\mathbf{y'}$ (A $(n+m)$-dimensional vector containing regressional values of amount of precipitation for each day)
Method:
Partition data 3 ways (training, evaluation and test)
1) Perform MLR on the training set (Size-$n$) to get $\mathbf{w}$.
2) Use the $\mathbf{w}$ on the testing set (Size-$m$) to get $\mathbf{y}'_i$.
3) Calculate the objective function error using the present $\mathbf{w}$ and save the value
4) Quadratic Discriminant Analysis (QDA) is performed on $\mathbf{y}'_i$ to get $\mathbf{y}_i$
5) In the semi-supervised approach (ZICR-SS) initialize $\tilde{\mathbf{y}}$ to $\mathbf{c}$ for the first $n$ datapoints and initialize the remaining $m$ points of $\tilde{\mathbf{y}}$ with $\mathbf{y}$ from step-3.
  In the supervised approach (ZICR-SS), $\tilde{\mathbf{y}}$ is initialized to $\mathbf{c}$ only.
6) Solve $\mathbf{w}$, using the $d$ equations got after differentiating the objective function $f(w)$
7) After having solved $\mathbf{w}$, solve for $\mathbf{y'}$ using the linear equation $\mathbf{y'} = \mathbf{xw}$
8) Apply QDA to find class labels for the training data points $\mathbf{y}$.
10) Calculate the objective function error using the present $\mathbf{w}$
11) For a fixed number of iterations (e.g., 10) or based on the convergence of the objective function, repeat steps 4 to 10
12) Evaluate the model by testing the RMSE error on the test data set.
```

vector **w** of the model. The following criteria was used to evaluate the performance of the models:

- Root Mean square error (RMSE), which measures the difference between the actual and predicted values of the response variable, i.e.: RMSE $= \sqrt{\frac{\sum_1^n (c'_i - y'_i)^2}{n}}$.
- Accuracy, which measures the number of `Rain` and `NoRain` days predicted correctly by the model.
- F-measure, which is the harmonic mean between recall and precision values for rain days.

4.3.2 Experimental Results

The purpose of the experiment was to demonstrate the following:

1. Limitations of classical regression models in terms of handling zero-inflated time series data.
2. Rationale of incorporating unlabeled data for precipitation prediction.
3. Performance comparison between classical regression models and our proposed framework.

4.3.2.1 Rationale for Incorporating Unlabeled Data

The objective of this section is to demonstrate the utility of incorporating unlabeled data for semi-supervised learning in precipitation prediction. Previous studies have shown that unlabeled data are helpful as long as their distribution is similar to those in the labeled training data [19, 36]. For climate data, the study showed that the natural periodic behavior of the predictor variables and the response variable (precipitation) provides an opportunity to leverage the unlabeled data to improve precipitation prediction.

Figure 4.1 shows the average similarity values of the predictor variables over time. The horizontal axis corresponds to the width of two time periods in the time series while the vertical axis corresponds to the average similarity for all pairs of time periods with the given width. For example, consider a time series of length 10,000 days. To compute the average similarity of width 3 months, we compare the similarity of the predictor variables on days 1 and 91, days 2 and 92, and so on. We use Pearson correlation as the similarity measure. The plot shows there are clear cycles in the average similarity values demarcated by years, i.e., the predictor variables for a given day is more similar to another observation that is 1yr, 2yr, or 3yrs apart when compared to observations that are 1.5yr, 2.5yr and 3.5yr apart. Figure 4.1 shows that this trend in similarity of observations is valid even for differences as large as 30 years. More subtle trends of cycles of a decade and a half were also observed. One of the encouraging observations is that the similarity of the predictors showed very slow decay with time. This observation encourages the notion that the predictor variables even if separated by large time differences still contain useful information that can be exploited for predicting future precipitation events.

One caveat is that though the similarity of predictor variables may not differ much over time, the similarity of the relationship between the predictor and response variables over time tend to decrease at a much faster rate, as shown in Figure 4.3. The product of similarity between predictor variables and similarity between response variables for two time periods of a given width was used to represent the vertical axis. This was the case because, as shown in Figure 4.2, though the similarity of precipitation was periodic in nature, it was fluctuating more rapidly compared to similarity of the predictor variables over time.

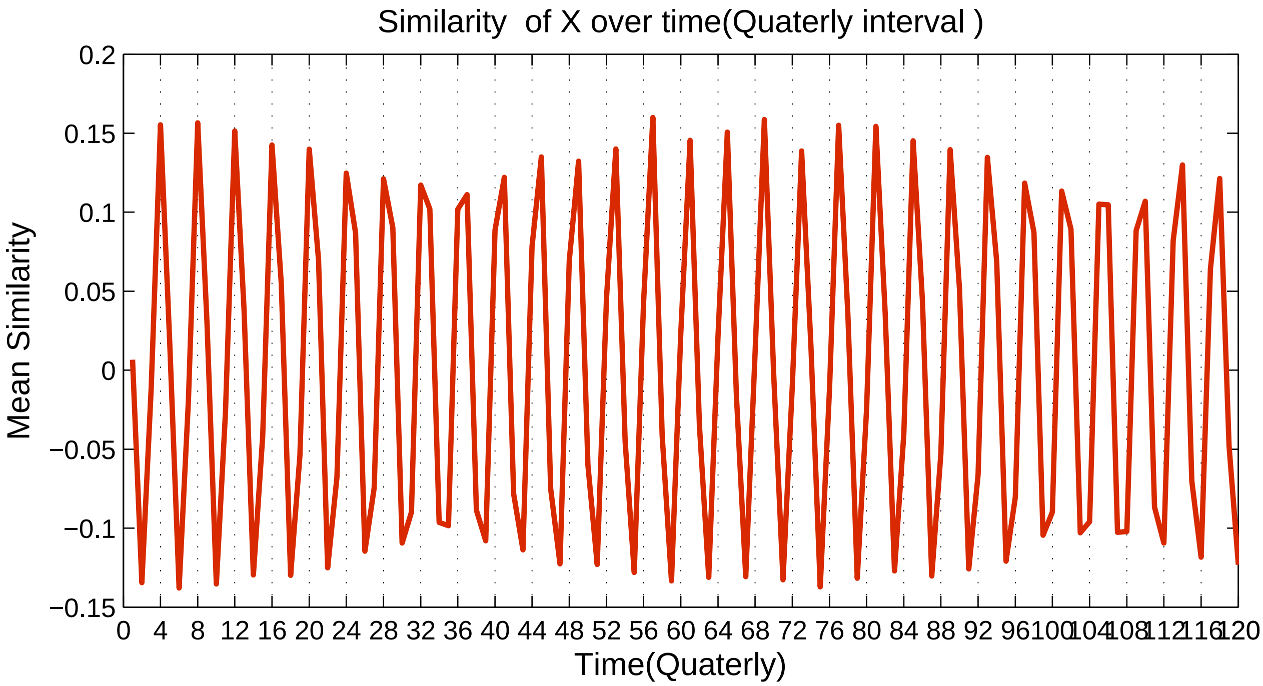


Figure 4.1: Similarity of predictor variables for all pairs of time periods of a given width.

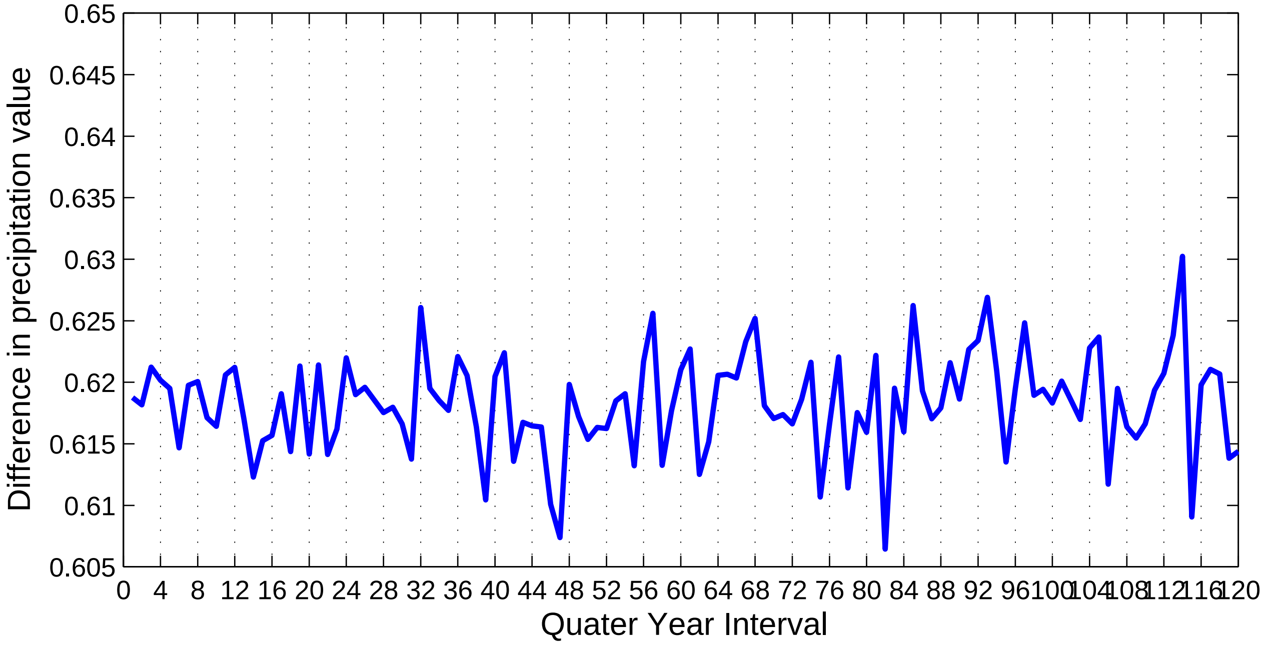


Figure 4.2: Similarity of response variable for all pairs of time periods of a given width.

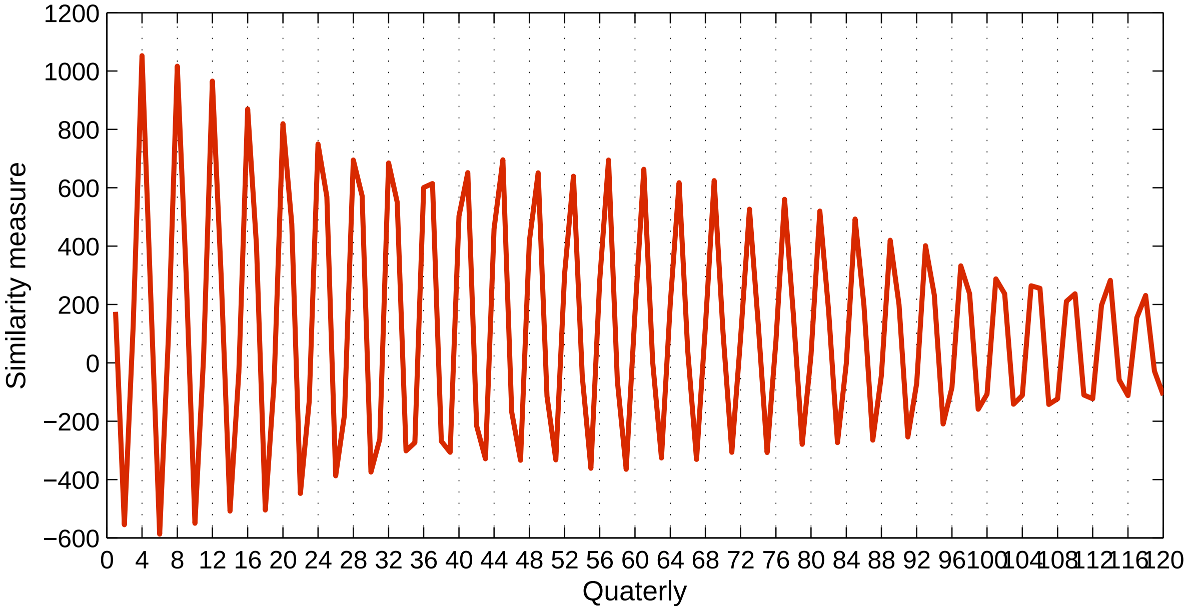


Figure 4.3: Similarity of the relation between Predictors and Predicants over time

4.3.2.2 Performance Comparison

This section compares the RMSE, accuracy, and F-measure values for our proposed supervised (ZICR-S) and semi-supervised (ZICR-SS) framework against the precipitation prediction results of multiple linear regression (MLR). All the experiments were performed using a training size (n) of 3 years starting from the first observation in the time series. The test set size (m) was also fixed at 3 years. After calculating the RMSE on the test set, the training set was shifted by 3 years, such that it now occupied the data set used for testing in the previous iteration. The experiment is repeated 7 times for each station. The RMSE values reported in this section is the mean value of all 7 iterations. The same approach is used to compute the RMSE values for `Rain` days, accuracy (for all days), F-measure for `Rain` days only and F-measure for `NoRain` days only. Due to space restriction, we show the results for 20 weather stations.

As shown in Figure 4.4, both our models, ZICR-S and ZICR-SS, significantly outperformed the MLR model (trained on all days) in terms of their RMSE values for predicting both `Rain` and `NoRain` days.

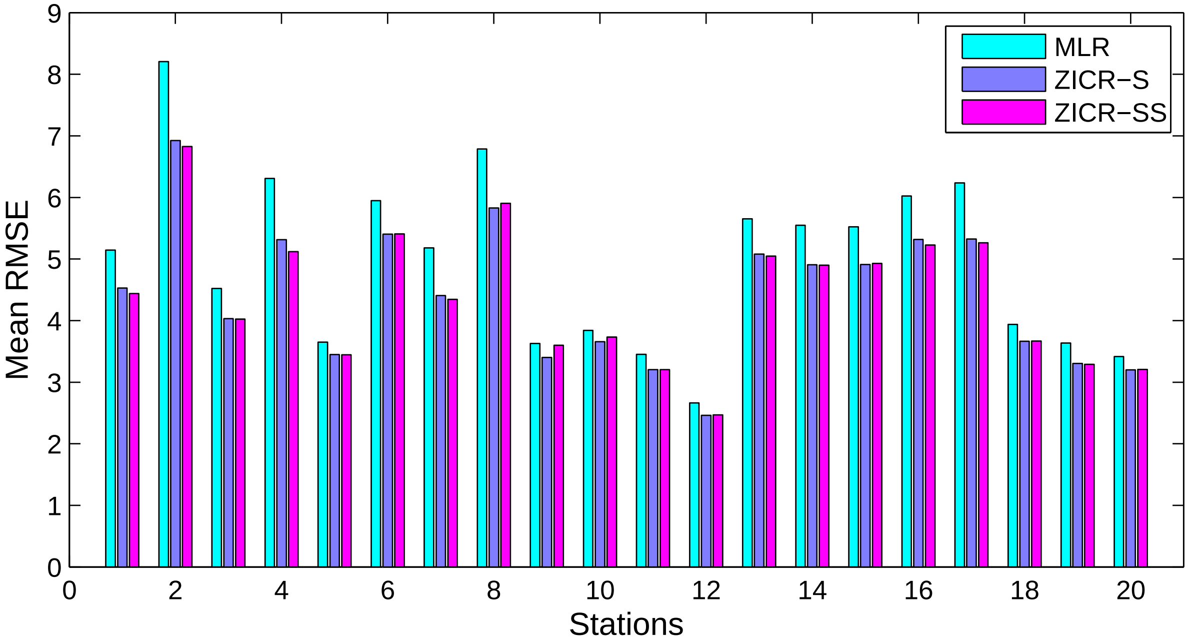


Figure 4.4: Comparison of RMSE values (for all days) among MLR, ZICR-S, and ZICR-SS.

The supervised version of the approach outperformed MLR for all 37 stations, while the semi-supervised approach outperformed MLR in 34 out of the 37 stations. In terms of percentage improvement in RMSE, the RMSE for MLR was at an average 8.8% and 8.4% worse than ZICR-S and ZICR-SS respectively. ZICR-S outperformed ZICR-SS in 22 out of the 37 stations.

However, in terms of the RMSE values for `Rain` days only, Figure 4.5 MLR had an average RMSE value for `Rain` days only that was 4.9% and 5.2% higher than ZICR-S and ZICR-SS respectively. Both ZICR-S and ZICR-SS consistently outperform the MLR model with ZICR-S outperforming in 34 and ZICR-SS outperforming in 32 stations. ZICR-S outperformed ZICR-SS in 21 out of the 37 stations.

Although MLR does not inherently classify any days as `Rain` or `NoRain`, the Quadratic

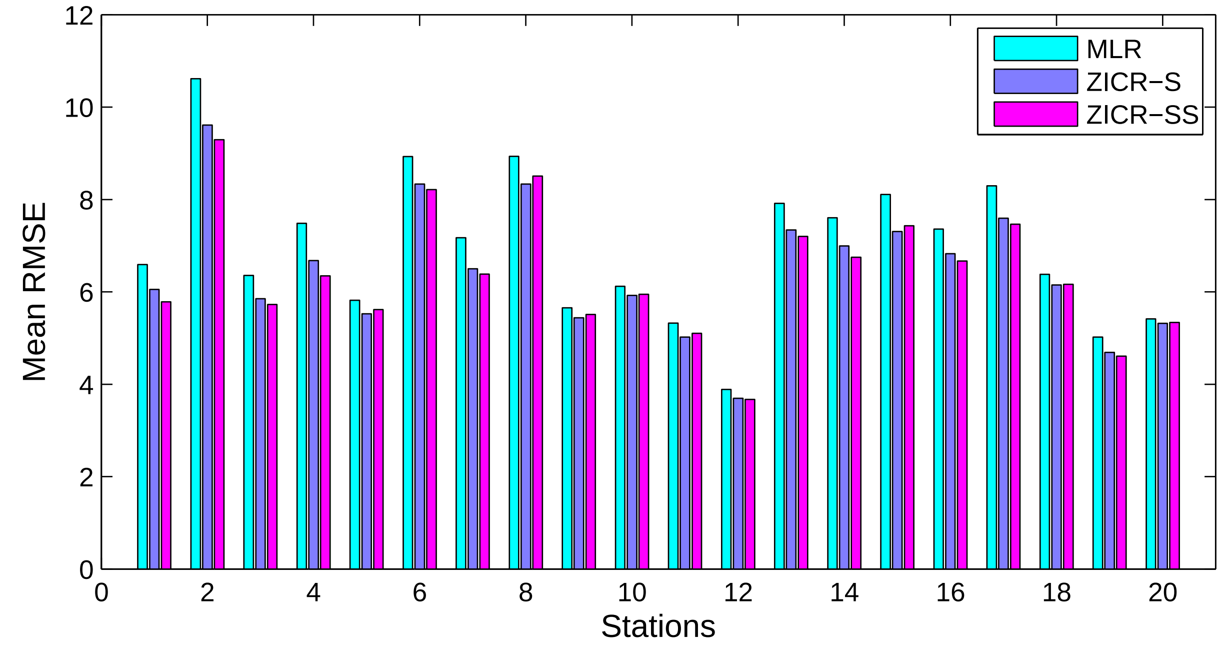


Figure 4.5: Comparison of RMSE values (for `Rain` days) among MLR, ZICR-S, and ZICR-SS.

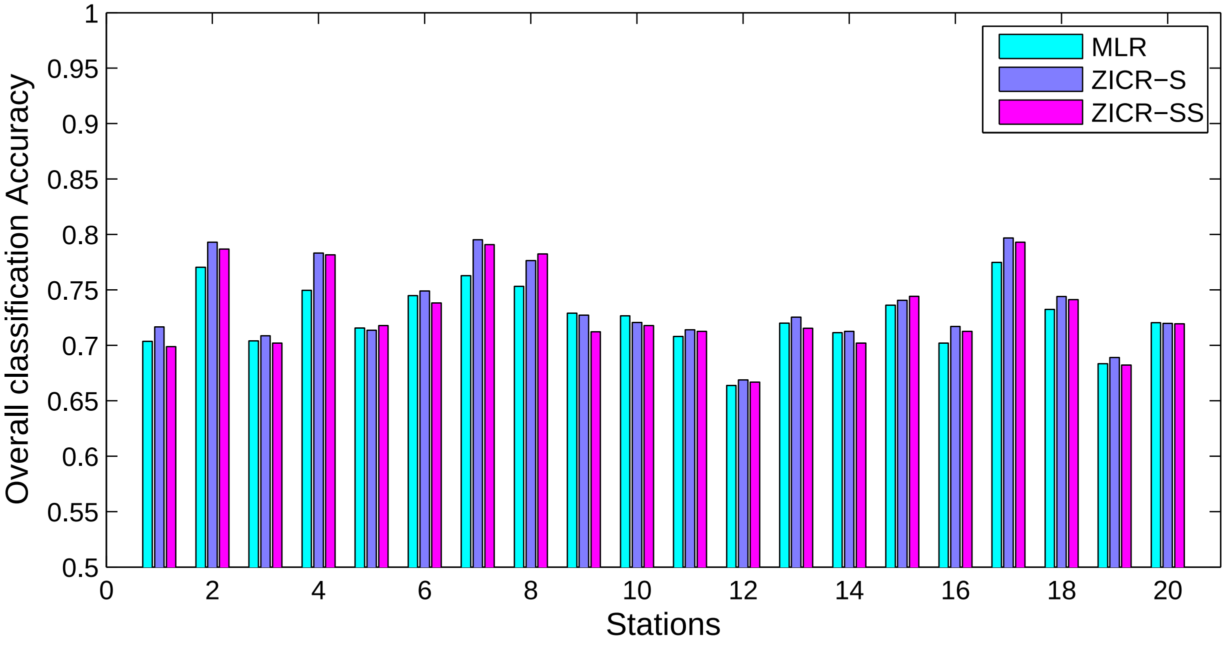


Figure 4.6: Comparison of classification accuracy (for all days) among MLR, ZICR-S, and ZICR-SS.

Discriminant Analysis (QDA) classifier used in our framework, was trained on the MLR outputs to compare its classification accuracy and F-Measure against those of ZICR-S and ZICR-SS. As shown in Figure 4.6, all 3 ZICR-S ZICR-SS and MLR were comparable in terms of classification accuracy with ZICR-SS outperforming MLR in approx 60% of the stations. Nevertheless, in terms of F-measure for `Rain` days, both the models consistently outperformed MLR as shown in Figure 4.7 with ZICR-S outperforming MLR in 32 stations while ZICR-SS outperformed MLR in 33 stations.

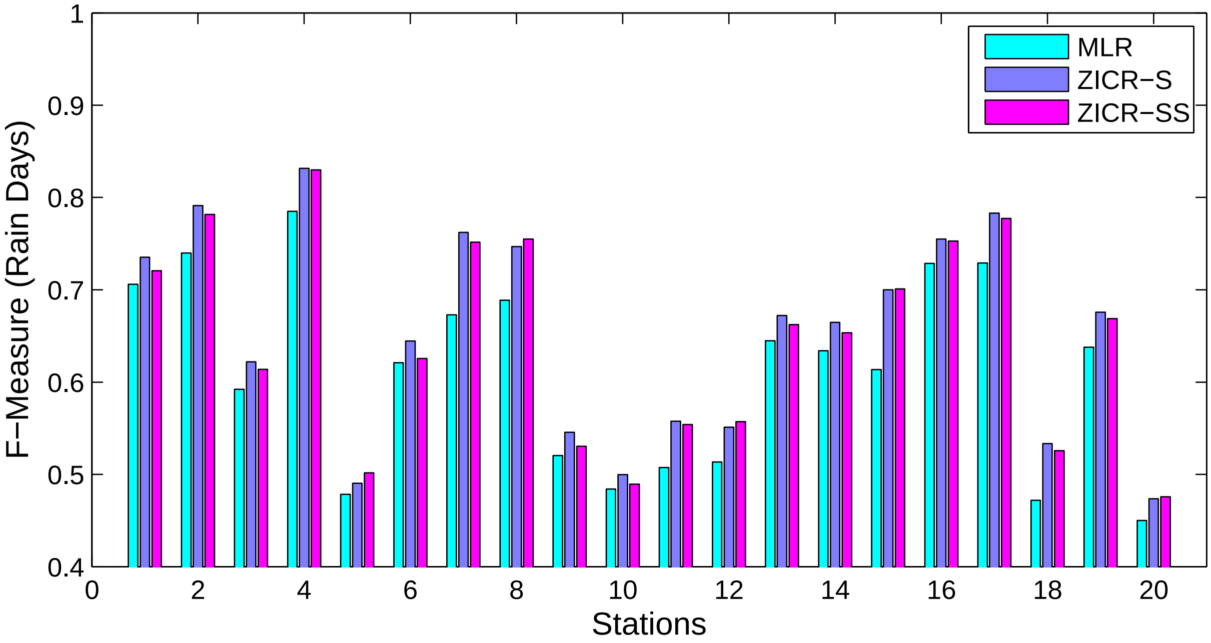


Figure 4.7: Comparison of F-Measure (for `Rain` days) among MLR, ZICR-S, and ZICR-SS.

With regard to the number of stations that MLR was outperformed in F-measure for `Rain` days, ZICR-S outperformed MLR in 32 and ZICR-SS in 33 stations. Figure 4.8 shows the comparison of F-measure for `NoRain` days between MLR, ZICR-S, and ZICR-SS.

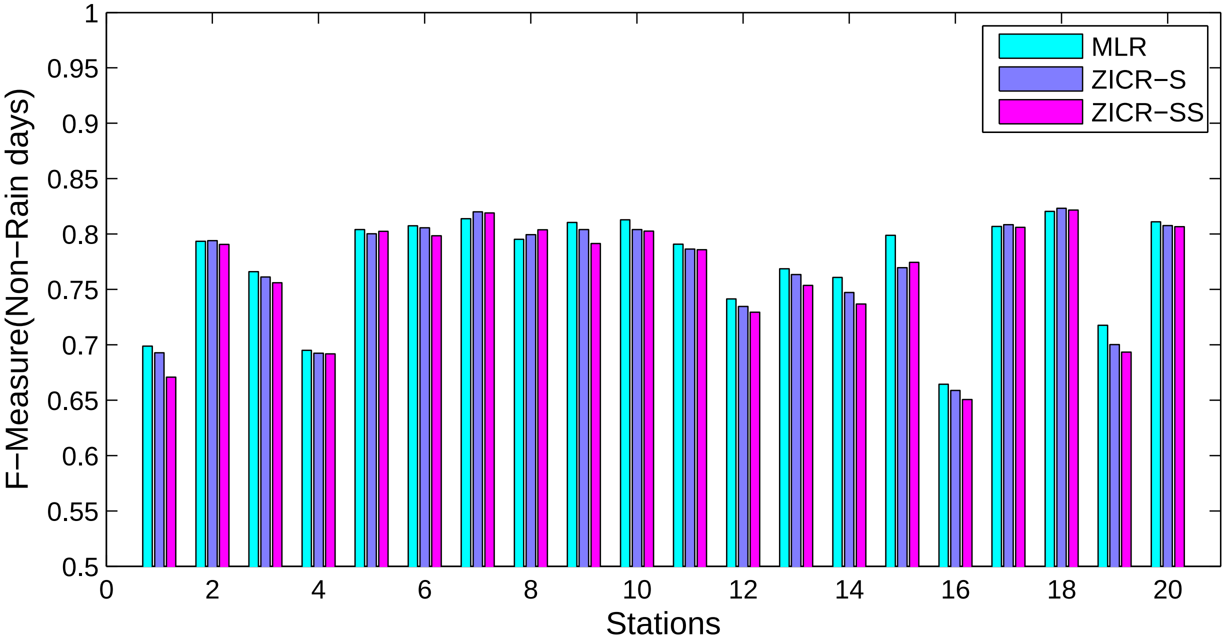


Figure 4.8: Comparison of F-Measure (for `NoRain` days) among MLR, ZICR-S, and ZICR-SS.

4.4 Conclusions

This chapter elaborates on extending the ICR framework detailed in Chapter 3, to a semi-supervised learning setting. This chapter compares the performance of the framework when it utilizes both unlabeled data and labeled data instead of using only labeled data during training of the model.

Chapter 5

Modeling Conditional Quantiles

This chapter as well as the following chapter elaborates the importance of accurately predicting the frequency, timing and magnitude of extreme values in the distribution of the response variable. Specifically, a semi-supervised framework for smoothed quantile regression (LSSQR) is presented that focuses on accurate prediction of extreme values without significantly degrading the sum-of-square residual errors.

5.1 Introduction

An integral part of climate modeling is downscaling, which seeks to project future scenarios of the local climate based on the coarse resolution outputs produced by global climate models (GCMs). Two of the more common approaches to downscaling are dynamic downscaling and statistical downscaling. Dynamic downscaling uses a numerical meteorological model to simulate the physical dynamics of the local climate while utilizing the climate projections from GCMs as initial boundary conditions. Though it captures the geographic details of a region unresolved by GCMs, the simulation is computationally demanding while its spatial resolution remains too coarse for many climate impact assessment studies. Statistical downscaling establishes the mathematical relationship between the coarse-scale GCM outputs and the fine-scale local climate variables based on observation data. Unlike dynamic downscaling, it is flexible enough to incorporate any predictor variable and is relatively inexpensive.

Most of the statistical downscaling approaches employ regression methods such as multiple linear regression, ridge regression, and neural networks to estimate the conditional mean of the future climate conditions. These methods are ill-suited for predicting extreme values of the climate variables.

An alternative approach is to use techniques such as quantile regression, which aims to minimize an asymmetrically weighted sum of absolute errors, to estimate the particular quantile that corresponds to extreme values [77]. Unfortunately, quantile regression tends to overestimate the response variable resulting in a large number of data points being falsely predicted to be extreme. Figure 5.1 represents the histogram of the distribution of observed temperature at a weather station in Canada. The lines represent the distribution of the predicted values for temperature obtained using multiple linear regression (MLR) and quantile regression. An observation is considered an extreme data point if its response variable is in the top 5 percentile of observations. The shape of the tail of the distribution that represents extreme data points (observed and projected) is shown in Figure 5.2. It is clear from the figures that methods such as multiple linear regression (green line) that estimate the conditional mean tend to underestimate the tail of observed probability distribution, while quantile linear regression (red line) overestimates the tail part of the probability distribution. As elaborated in Section 5.4, it was found that for the 37 stations evaluated, at an average, quantile regression predicted a datapoint to be an extreme point more than twice as frequently as the actual frequency of observed extreme data points.

To address this overestimation, a method known as smoothed quantile regression (LSQR) is proposed, that reduces the absolute error of extreme data points by introducing a smoothing term that brings the predicted response value of extreme points closer to the value corresponding to the percentile of extreme data points. This smoothing term also provides a

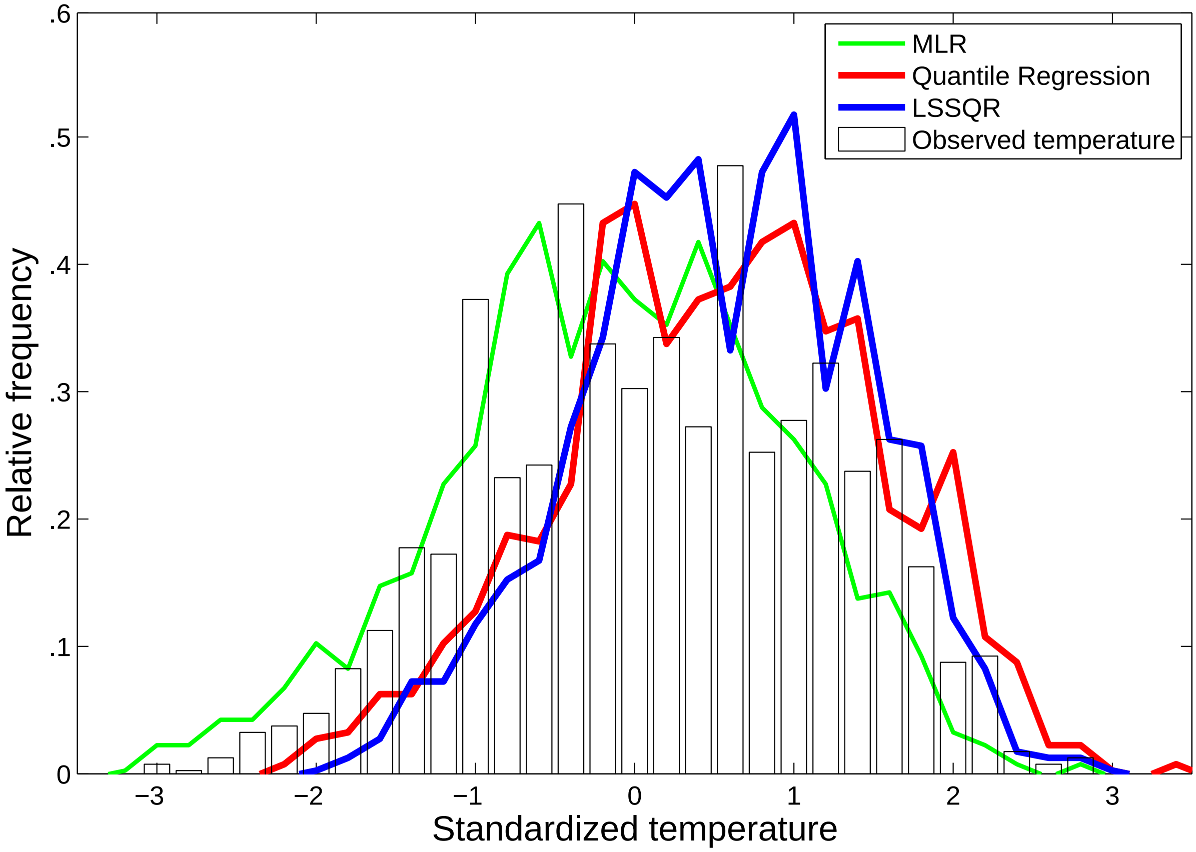


Figure 5.1: Histogram of observed temperature.

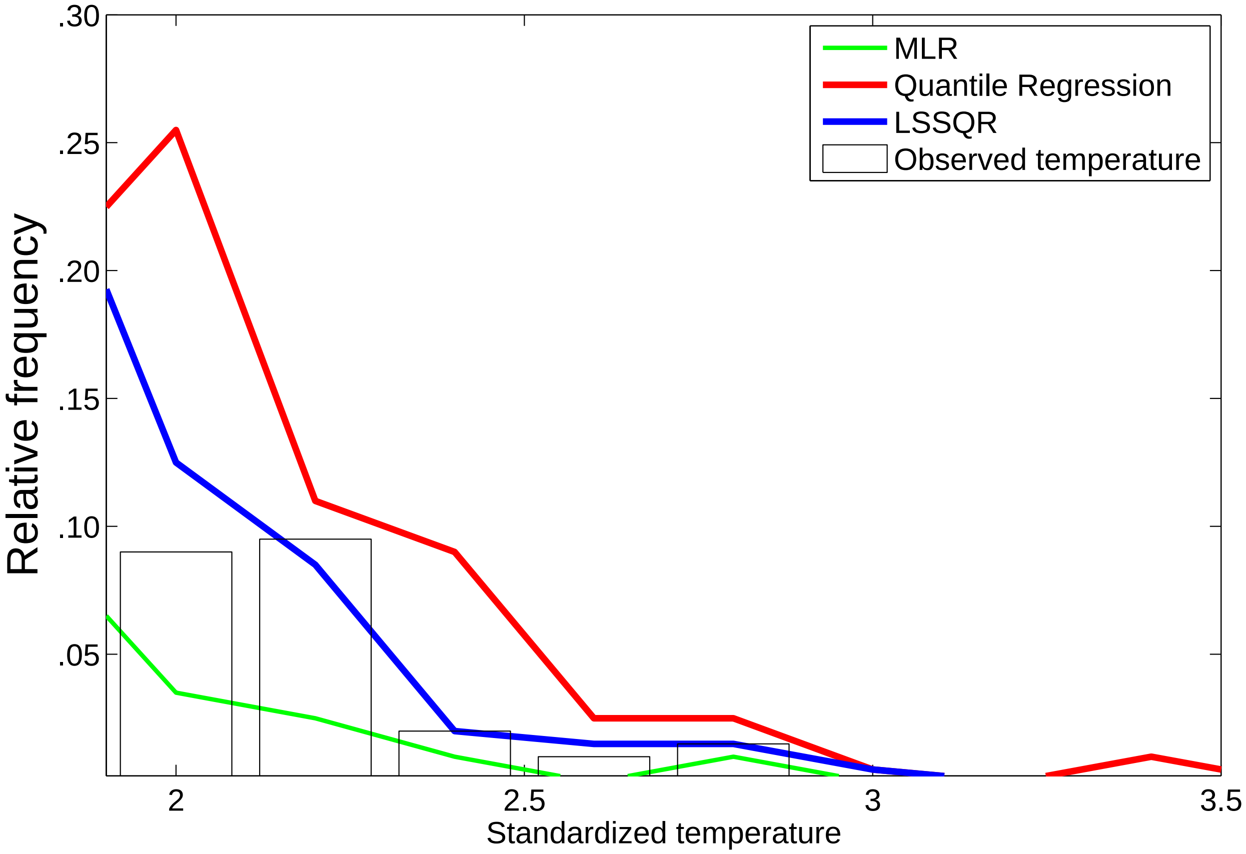


Figure 5.2: Tail of the histogram.

means to easily extend the objective function to a semi-supervised learning setting (LSSQR). Semi-supervised learning, in addition to using the training data, can also use the distribution characteristics of the predictor variables of the test set to glean a better estimate of the distribution of data upon which the model will be applied.

In summary, the main contributions of this chapter are as follows:

- Demonstrating the limitation of MLR, ridge regression and quantile regression in predicting extreme values.
- Presenting a smoothed quantile regression framework for extreme values prediction.
- Extending the framework to a semi-supervised setting.
- Demonstrating the efficacy of our learning framework on climate data (temperature) obtained from the Canadian Climate Change Scenarios Network website [1]. Both the supervised and the semi-supervised proposed frameworks outperformed the baseline methods in 85% of the 37 stations evaluated, in terms of magnitude, frequency and the timing of the extreme events.

5.2 Preliminaries

Let $D_l = \{(x_i, y_i)\}_{i=1}^{n}$ 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. Similarly, $D_u = \{(x_i, y_i)\}_{i=n+1}^{n+m}$ corresponds to the unlabeled dataset. 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 of the target function. n represents the number of labeled training points and m represents the number of unlabeled testing points.

5.2.1 Multiple Linear Regression (MLR) and Ridge Regression

One of most widely used forms of regression is multiple linear regression. It solves 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 of squared residuals

$$(y - X\beta)^T (y - X\beta)$$

which leads to a closed-form expression for the solution

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

A variant of MLR, called ridge regression or Tikhonov regularization is often used to mitigate overfitting. Ridge regression also provides a formulation to overcome the hurdle of a singular covariance matrix $X^T X$ 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

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

where, the ridge coefficient $\lambda > 0$ results in a non-singular matrix $X^T X + \lambda I$ always being invertible. The problem with both MLR and ridge regression is that they try to model the conditional mean, which is not best suited for predicting extremes.

5.2.2 Quantile Linear Regression(QR)

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 $\tau \in [0, 1]$.

Unlike MLR that estimates the conditional mean, quantile regression estimates the quantile (e.g., median) of Y.To estimate the τ^{th} conditional quantile $Q_{Y|X}(\tau)$, quantile regression minimizes an asymmetrically weighted sum of absolute errors. To be more specific, the loss function for quantile linear regression is:

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

where,

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

Unlike MLR and ridge regression that have a closed-formed solution, quantile regression is often solved using optimization methods such as linear programming. 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, $u_i \geq 0$ and $v_i \geq 0$. But as shown in Figures 5.1 and 5.2, quantile regression often overestimates data points resulting in too many false positive extreme events predicted.

5.3 Framework for Smoothed Quantile Regression

Given that the primary objective of the model is to accurately regress extreme valued data points and quantile regression has been shown to perform relatively better that its least square counterparts that tend to underestimate the frequency and magnitude of extreme data points, the proposed objective approach of the proposed frameworks is modeled around linear quantile regression. Section 5.3.1 describes smoothed quantile regression (LSQR) and its objective function. Section 5.3.2 proposes a semi-supervised extension to LSQR which is then followed by mathematical properties of the behavior of the objective function.

5.3.1 Smoothed Quantile Regression (LSQR)

A quantile-based linear regression model was proposed, based on the assumption of smoothness, i.e., data points whose predictor variables are similar, should have a similar response. The notion of smoothness as an integral part of the framework, as experiments provided in Section 5.4 demonstrate this characteristic in the dataset used. The smoothness assumption could be described as the constraint

$$\sum_{i,j}^{n} w_{ij}(f_i - f_j)^2 < c$$

where w_{ij} is a measure of similarity between data point i and j, f the predicted value of the response variable and c is a constant.

Also, since the framework doesn't restrict the training set only to extreme data points, the smoothing component of the objective function tends to implicitly cluster data points resulting in better distinction of the response variables of an extreme valued data point and a non-extreme valued data point. Empirical results comparing supervised quantile regression to the proposed semi-supervised model illustrate this point as shown in Section 5.4. The term

$$w_{ij} = \exp(-\frac{||x_i - x_j||^2}{\sigma}) \quad i, j \in [1, 2, \ldots, n]$$

is equivalent to the radial basis function and is used to capture the similarity between the predictor variables of data point i and data point j. σ is a scale parameter used to control the distance above which two data points are not considered as being highly coupled.

Assuming linear regression, $f(x_i, \beta) = x_i\beta$, the smoothing term can be reformulated as

$$\sum_{i,j}^{n} w_{ij}(f(x_i, \beta) - f(x_j, \beta))^2 = f^T \mathbf{\Delta} f = \beta^T \mathbf{\Sigma} \beta$$

where,

$$\mathbf{\Sigma} = X^T \mathbf{\Delta} X$$

$$\mathbf{\Delta} = D - W$$

and D is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{n} w_{ij}$ and $W = \{w_{ij}\}|_{i,j=1}^{n}$.

Coupling smoothing with the objective function of linear qunatile regression, we end up with the following optimization problem.

$$\min_{\beta} \sum_{i=1}^{n} \rho_{\tau}(y_i - x_i^T \beta) + \lambda \beta^T \mathbf{\Sigma} \beta$$

As can be clearly observed from the objective functions of $LSQR$, $\lambda \rightarrow 0$ results in an estimate similar to quantile linear regression while, $\lambda \rightarrow \infty$ results in the estimate of the response variable converging towards the target quantile of data. This is because a large λ would penalize any non-zero difference between f_i and f_j very harshly thereby minimizing the error by setting $f_i = \alpha, \forall i \in [1, 2, \ldots, n]$, thereby reducing the error from the second component of the equation to 0. This reduces the loss function to the following

$$f(\beta) = \sum_{i=1}^{n} \rho_{\tau}(y_i - \alpha), \quad \beta = (\alpha, 0, 0, \ldots, 0)^T$$

The formal proof of this is provided in the following theorem.

Theorem 1: $f(x_i, \beta) \rightarrow y_{(n\tau)}$ as $\lambda \rightarrow \infty$, $\forall i \in [1, 2, \ldots, n]$.

Proof : Let $y_{(i)}$ be the i^{th} smallest element among $y_k|_{k=1}^{n}$ and $y_{(i)} < \alpha_i <= y_{(i+1)}$. When $\lambda \rightarrow \infty$, the loss function can be rewritten in terms of α_i as follows

$$\sum_{k=1}^{i} (1-\tau)(\alpha_i - y_{(k)}) + \sum_{k=i+1}^{n} \tau(y_{(k)} - \alpha_i) + \sum_{i,j=1}^{n} W_{ij}(\alpha_i - \alpha_i)$$

which is equivalent to minimizing

$$\tau \sum_{k=1}^{n} y_{(k)} - \sum_{k=1}^{i} y_{(k)} - (n\tau - i)\alpha_i$$

or maximizing

$$\sum_{k=1}^{i} y_{(k)} + (n\tau - i)\alpha_i = l_i$$

Therefore,

$$l_j - l_{j-1} = y_j - \alpha_{j-1} + (n\tau - j)(\alpha_{j-1} - \alpha_j)$$

Hence, $\forall j : j \leq n\tau, \quad l_j - l_{j-1} >= 0$, since $(y_j - \alpha_{j-1})$, $(n\tau - j)$ and $(\alpha_{j-1} - \alpha_j)$ are all ≥ 0. Similarly, $\forall j : j \geq n\tau$,

$$l_j - l_{j+1} = \alpha_{j+1} - y_{j+1} + (n\tau - j)(\alpha_j - \alpha_{j+1}) \geq 0$$

Hence, if $\exists i : i = n\tau$, then $\alpha = y_{(n\tau)}$. But if, $i < n\tau < (i+1)$, then α is in the interval $[y_{(i)}, y_{(i+1)}]$ □

Figure 5.3 is a plot that tracks the values of β for different λ values. The figure shows that the regression parameter vector $\boldsymbol{\beta}$ will converge to $(\alpha, 0, 0, \ldots, 0)^T$ as λ increases. β_0 is the regression parameter that corresponds to the column of 1's in the design matrix.

Figures 5.4 and 5.5 plots the influence of λ on the predicted values returned from LSSQR. i.e., as the value of λ increases, LSSQR shrinks the prediction range to the quantile τ. Figure 5.5 is a zoomed-in image, capturing the tail of Figure 5.4.

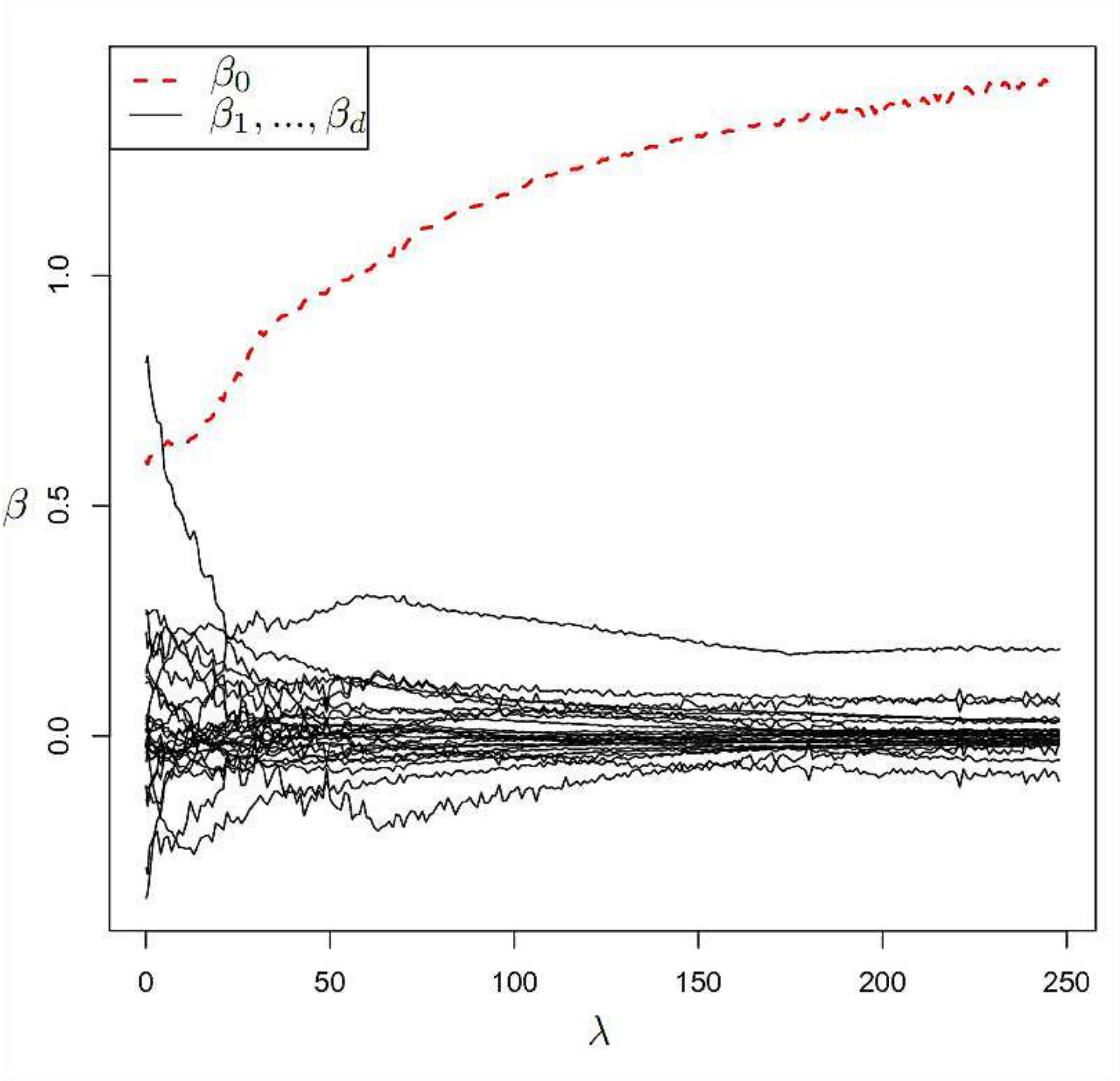


Figure 5.3: Influence of parameter λ on the regression coefficients β in LSQR.

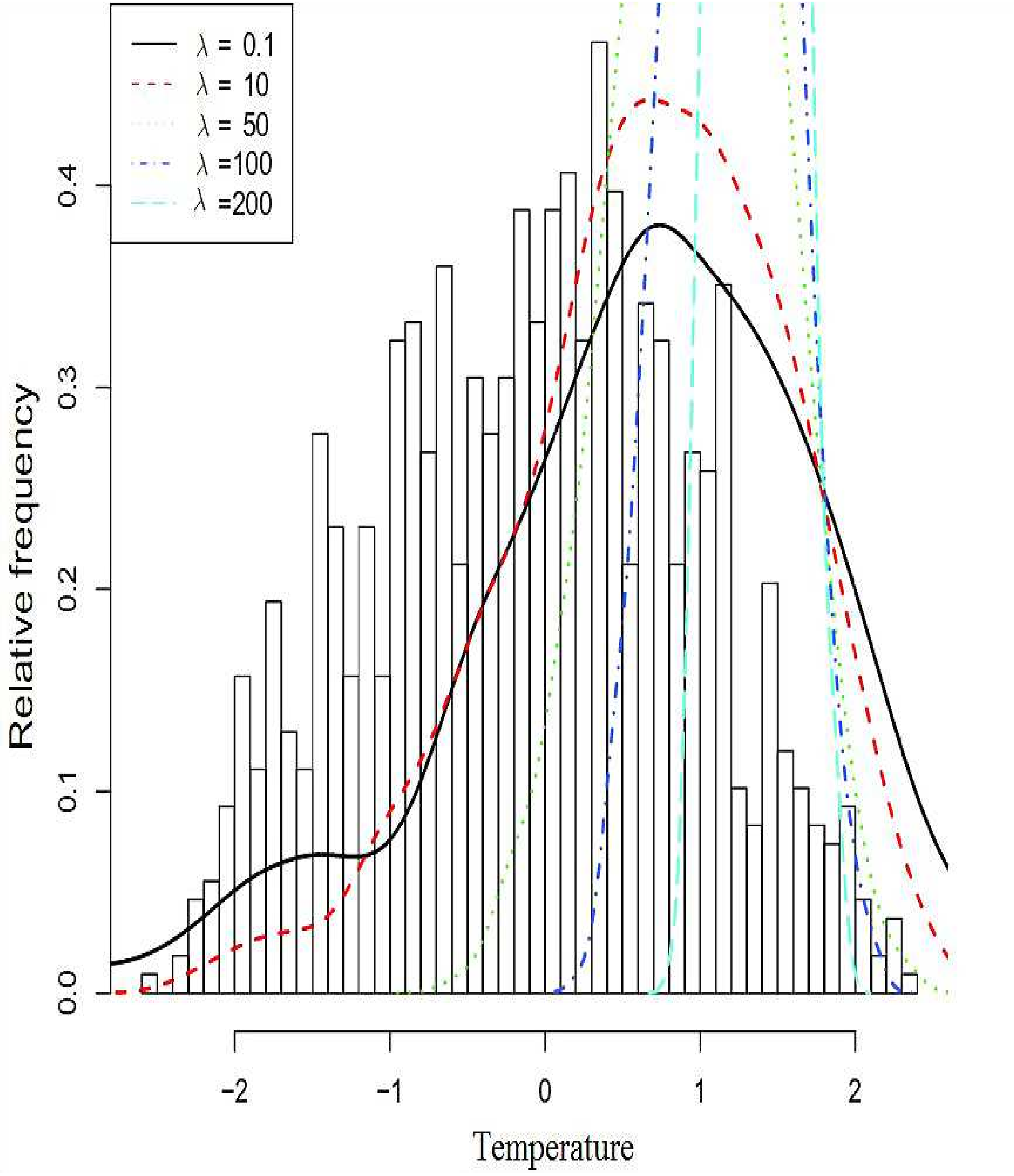


Figure 5.4: Influence of λ on the probability distribution of the predicted values obtained from LSSQR.

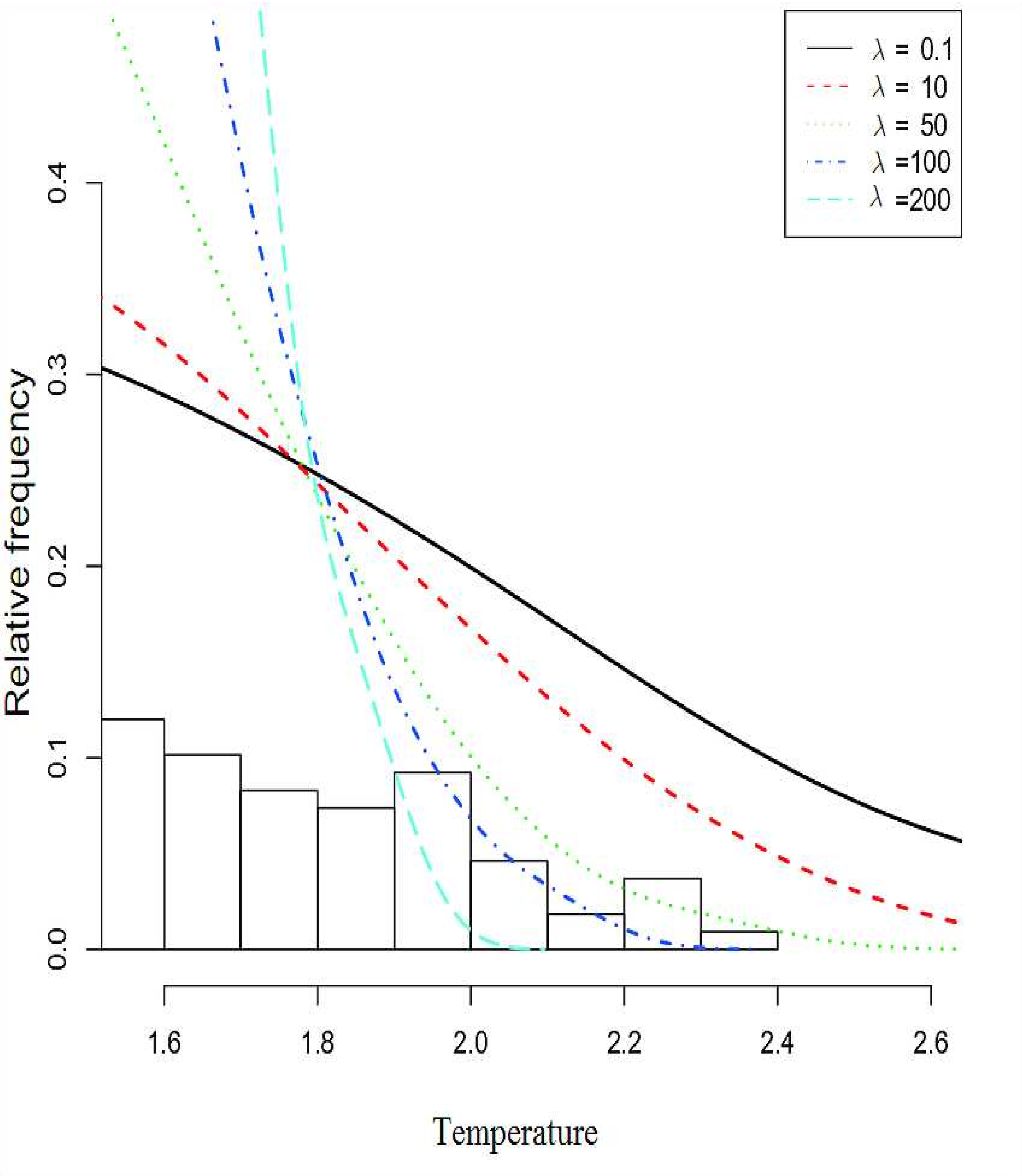


Figure 5.5: Influence of λ on the probability distribution of the predicted extreme values obtained from LSSQR.

5.3.2 Linear Semi-Supervised Quantile Regression (LSSQR)

The objective function of LSQR can be easily extended to a semi-supervised learning setting since the smoothing factor (the second term in the equation) is independent of y. Therefore, by extending the range of the indices i and j of the smoothing term to span 1 to $n+m$, the predictor variables of the unlabeled data $X_u = [x_{u1}, ..., x_{um}]^T$ can be harvested.

The objective function of the LSSQR is

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho_{\mathcal{T}}(y_i - x_i^T \beta) + \lambda \sum_{i,j}^{n+m} w_{ij}(x_i^T \beta - x_j^T \beta)^2$$

5.4 Experimental Results

In this section, the climate dataset that is used for statistical downscaling is described. This is followed by the experimental setup, which address the inherent properties of the dataset, such as its periodic nature. Once the dataset is introduced, we analyze the behavior of baseline models developed using MLR, ridge regression and quantile regression and contrast them with LSQR and LSSQR. The efficacy of the models in accurately measuring the magnitude, the relative frequency and timing of forecasting a data point as an extreme event is measured.

5.4.1 Data

All the algorithms were run on climate data obtained at 37 weather stations in Canada, from the Canadian Climate Change Scenarios Network website [1]. The response variable to be regressed (downscaled) corresponds to daily temperature values measured at each weather station. The predictor variables for each of the 37 stations correspond to 26 coarse-scale climate variables derived from the NCEP re-analysis data set, which include measurements of airflow strength, sea-level pressure, wind direction, vorticity, and humidity, as shown in Table 5.1. The predictor variables used for training were obtained from the NCEP re-analysis data set that span a 40-year period (1961 to 2001). The time series was truncated for each weather station to exclude days for which temperature or any of the predictor values are missing.

Table 5.1: List of predictor variables for temperature prediction.

Predictor Variables	
500 hPa airflow strength	850 hPa airflow strength
500 hPa zonal velocity	850 hPa zonal velocity
500 hPa meridional velocity	850 hPa meridional velocity
500 hPa vorticity	850 hPa vorticity
500 hPa geopotential height	850 hPa geopotential height
500 hPa wind direction	850 hPa wind direction
500 hPa divergence	850 hPa divergence
Relative humidity at 500 hPa	Relative humidity at 850 hPa
Near surface relative humidity	Surface specific humidity
Mean sea level pressure	Surface zonal velocity
Surface airflow strength	Surface meridional velocity
Surface vorticity	Surface wind direction
Surface divergence	Mean temp at 2 m

5.4.2 Experimental Setup

As is well known, temperature, which is the response variable in our experiments, has seasonal cycles. To efficiently capture the various cycles, de-seasonalization is performed prior to running the experiments. As is common practice in the field of climatology, a common approach to de-seasonalization is to split the data into 4 seasons (DJF, MAM, JJA, SON) where 'DJF' refers to the months of December-January-February in the temperature time-series. Similarly, 'MAM' refers to March-April-May, and 'JJA' refers to June-July-August and 'SON', September-October-November. In effect, for each station, 4 different models, corresponding to the 4 seasons were built. The training size used spanned 6 years of data and the test size, 12 years. During validation, the parameter λ was selected using the score returned by RMSE for extreme data points. A data point is considered extreme if its response variable is greater than .95 percentile (Threshold-1) of the whole dataset corresponding to the station. QR was implemented using the interior point algorithm as detailed in [76]. Broyden Fletcher Goldfarb Shanno (BFGS) method was used to solve the LSQR and LSSQR optimization problem.

5.4.3 Evaluation Criteria

The motivation behind the selection of the evaluation metrics was the intent to evaluate the different algorithms in terms of accuracy of the prediction of extreme values, the timing of the extreme events as well as the frequency with which a data point is predicted to be an extreme data point. The following metrics are used to capture the above evaluation criteria for the various models:

- Root Mean Square Error (RMSE), which measures the difference in magnitude between the actual and predicted values of the response variable, i.e.:
  RMSE $= \sqrt{\frac{\sum_{i=1}^{n}(y_i' - f_i')^2}{n}}$. RMSE was computed on those days that were observed to be extreme data points.

- Precision and recall of extreme events are computed to measure the timing accuracy of the prediction. F-measure, which is the harmonic mean between recall and precision values, will be used as a score that summarizes the precision and recall results.
  F-measure $= \frac{2 \times Recall \times Precision}{Recall + Precision}$

- The frequency of predicting extreme data point for the various methods was measured by computing the ratio of the number of data points that were predicted to be extreme to the number of observed extreme data points.

To summarize, RMSE is used for measuring the accuracy of the predicted magnitude of the response variable, whereas F-measure can be thought of as measuring the correctness of the timing of the extreme events.

5.4.4 Baseline

We compared the performance of LSQR and LSSQR with baseline models created using multiple linear regression (MLR), ridge regression (Ridge), and quantile regression (QR). All the baselines were run for the same 37 stations and for all the 4 seasons. Also, a comparison of the performance of the proposed supervised framework (LSQR) is made with its semi-supervised counterpart (LSSQR), where LSSQR demonstrated an improved performance over LSQR for the 37 stations evaluated upon as shown in Table 5.2. Table 5.2 summarizes the tally of percentage of times LSSQR outperformed LSQR over the 4 seasons for the given 37 stations. As seen in the table, LSSQR showed an improved performance in terms of both RMSE and F-measure.

Table 5.2: The relative performance of LSSQR compared with LSQR with regard to the extreme data points.

	Win	Loss	Tie
RMSE	68.25%	31.75%	0%
F-measure	60.14%	37.16%	2.7%

5.4.5 Results

As mentioned earlier, experiments were run separately using each of the baseline approaches and LSQR and LSSQR for the 4 seasons (DJF, MAM, JJA, SON) of the year for each of the 37 stations' data. The results over all the seasons and stations are summarized in Tables 5.3 and 5.4 while the individual results of each season in Figures 5.6 and 5.8. Table 5.3 summarizes the relative performance of LSQR with respect to the baseline methods in terms of RMSE of extreme data points and F-measure of identification of extreme data points. During testing, a data point is considered extreme, if its response variable is greater than .95 percentile (Threshold-1) of the whole dataset corresponding to the station. For the purpose of analysis, results of using the .95 percentile of the response variable in the training set (Threshold-2) to identify extreme data points are also summarized. The fact that the results obtained by using the two different baselines is an indicator that the training data did capture the distribution of the response variable reasonably well. LSQR consistently outperformed the baselines both in terms of RMSE and F-measure. It must also be noted that LSQR did outperform MLR and Ridge in terms of recall of extreme events comprehensively across each of the 37 stations and seasons.

Table 5.3: The percentage of stations LSQR outperformed the respective baselines, with regard to the extreme data points.

		MLR	Ridge	QR
RMSE	Threshold-1	88.51%	87.84%	80.40%
	Threshold-2	89.19%	87.84%	79.05%
F-measure	Threshold-1	59.45%	60.13%	72.97%
	Threshold-2	56.08%	58.10%	79.05%

Similarly, Table 5.4 summarizes the relative performance of LSSQR with respect to the baseline methods in terms of RMSE of extreme data points and F-measure of identification of extreme data points. Like LSQR, LSSQR consistently outperformed the baselines both

Table 5.4: The percentage of stations LSSQR outperformed the respective baselines, with regard to the extreme data points.

		MLR	Ridge	QR
RMSE	Threshold-1	87.16%	85.14%	85.13%
	Threshold-2	87.84%	86.49%	81.76%
F-measure	Threshold-1	60.13%	58.78%	75.67%
	Threshold-2	56.75%	59.45%	81.75%

in terms of RMSE and F-measure. It must be noted that LSSQR outperform MLR and Ridge in terms of recall of extreme events comprehensively across each of the 37 stations and seasons.

Figure 5.6 gives a breakdown of the performance of the LSSQR over each of the 4 seasons of the 37 stations using Threshold-1 for the purpose of marking a data point as extreme. The figure is a bar chart of percentage of stations that LSSQR outperformed MLR, ridge regression and QR in prediction accuracy for only extreme data points in the test set. RMSE was used to compute the accuracy of each model in predicting extreme value data points, at the 37 stations. As seen in the plot, LSSQR outperforms MLR, ridge regression and QR in each of the four seasons across the 37 stations.

Figure 5.7 shows a graph that depicts the percentage of stations LSSQR outperformed MLR, ridge regression and QR in terms of identifying extreme data points over 37 stations. Again, LSSQR comprehensively outperforms MLR and ridge regression over all the 37 stations and 4 seasons. But as expected, QR outperforms LSSQR in terms of recall performance for each of the 4 seasons due to the overestimating nature of QR, which consequently resulted in poor precision and which is reflected in its F-measure score. At an average, quantile regression, predicted a datapoint to be an extreme point more than twice as frequently as the actual frequency of observed extreme data points. In fact, QR lost out to LSSQR in 91% of 37 stations across 4 seasons in terms of precision of identifying extreme data points.

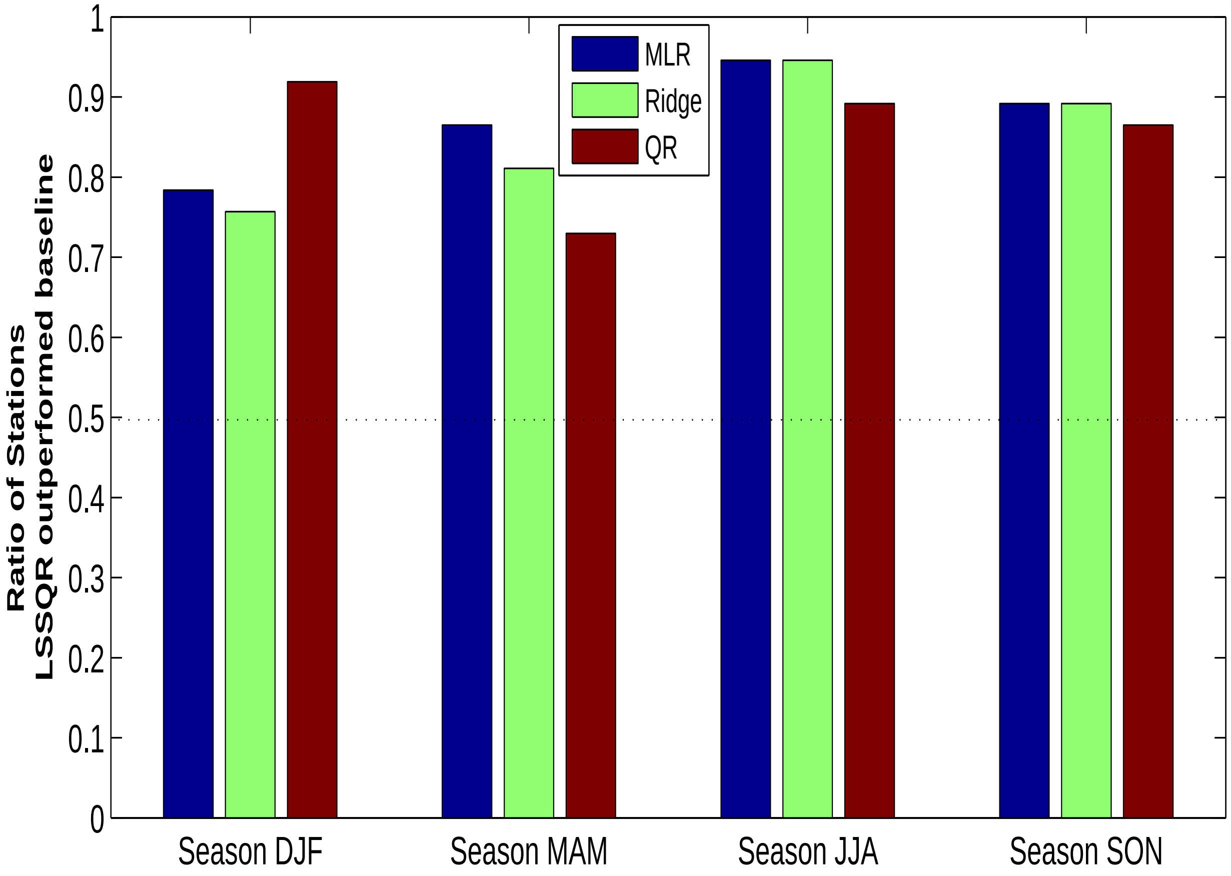


Figure 5.6: Ratio of stations LSSQR outperforming baseline in terms of RMSE of extreme data points.

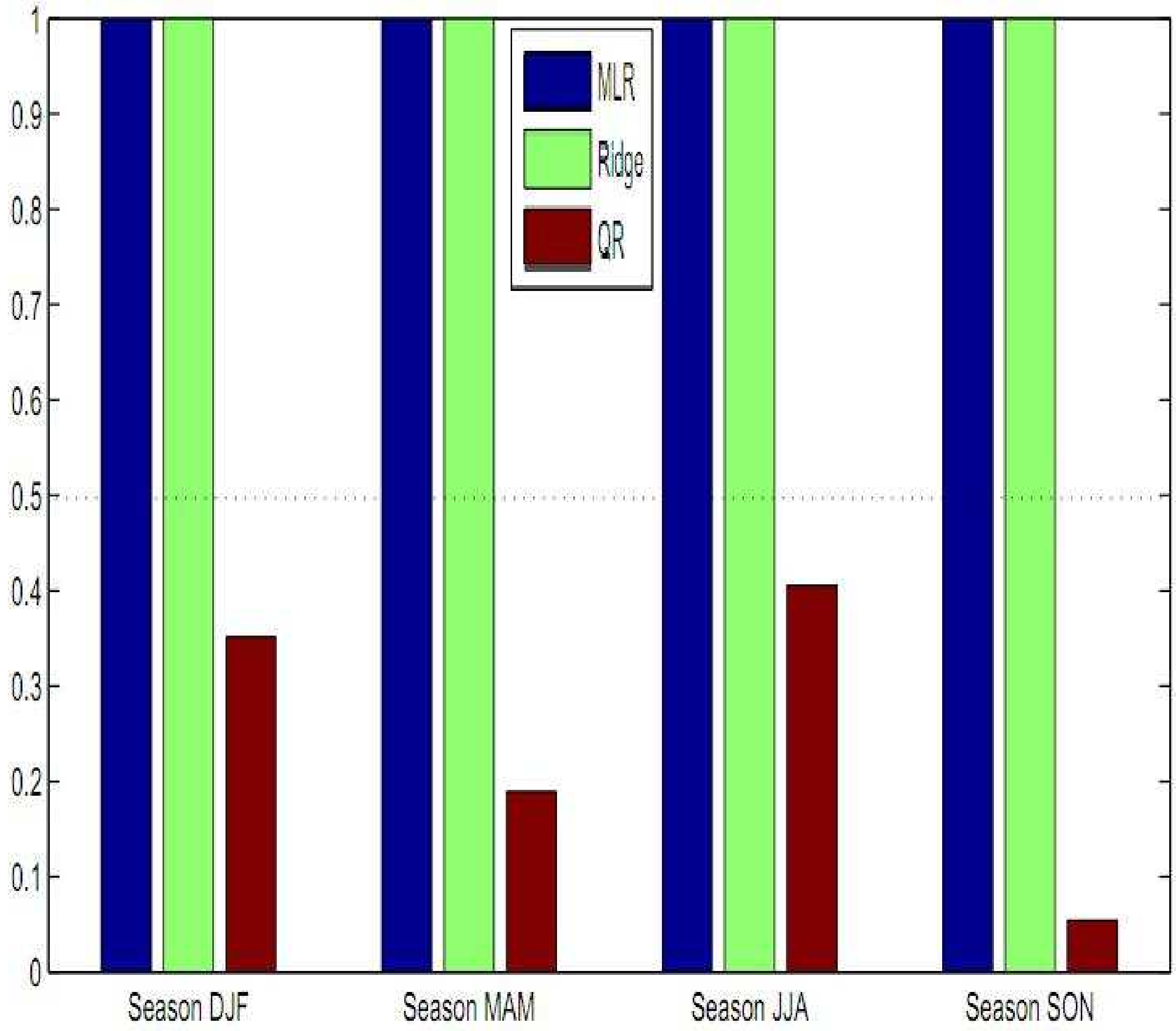


Figure 5.7: Ratio of stations LSSQR outperforming baseline in terms of recall of extreme data points.

Figure 5.8 shows a graph that depicts the percentage of stations where LSSQR outperformed MLR, ridge regression and QR in prediction accuracy based on F-measure of the identifying extreme data points over 37 stations. Again, LSSQR outperforms MLR, ridge regression and QR for all the 4 seasons.

The performance improvement obtained by LSSQR in terms of predicting the extreme values can be easily visualized in Figure 5.9. Figure 5.9 is a plot comparing the predicted response variable of the various methods. The plot is restricted to only extreme data points

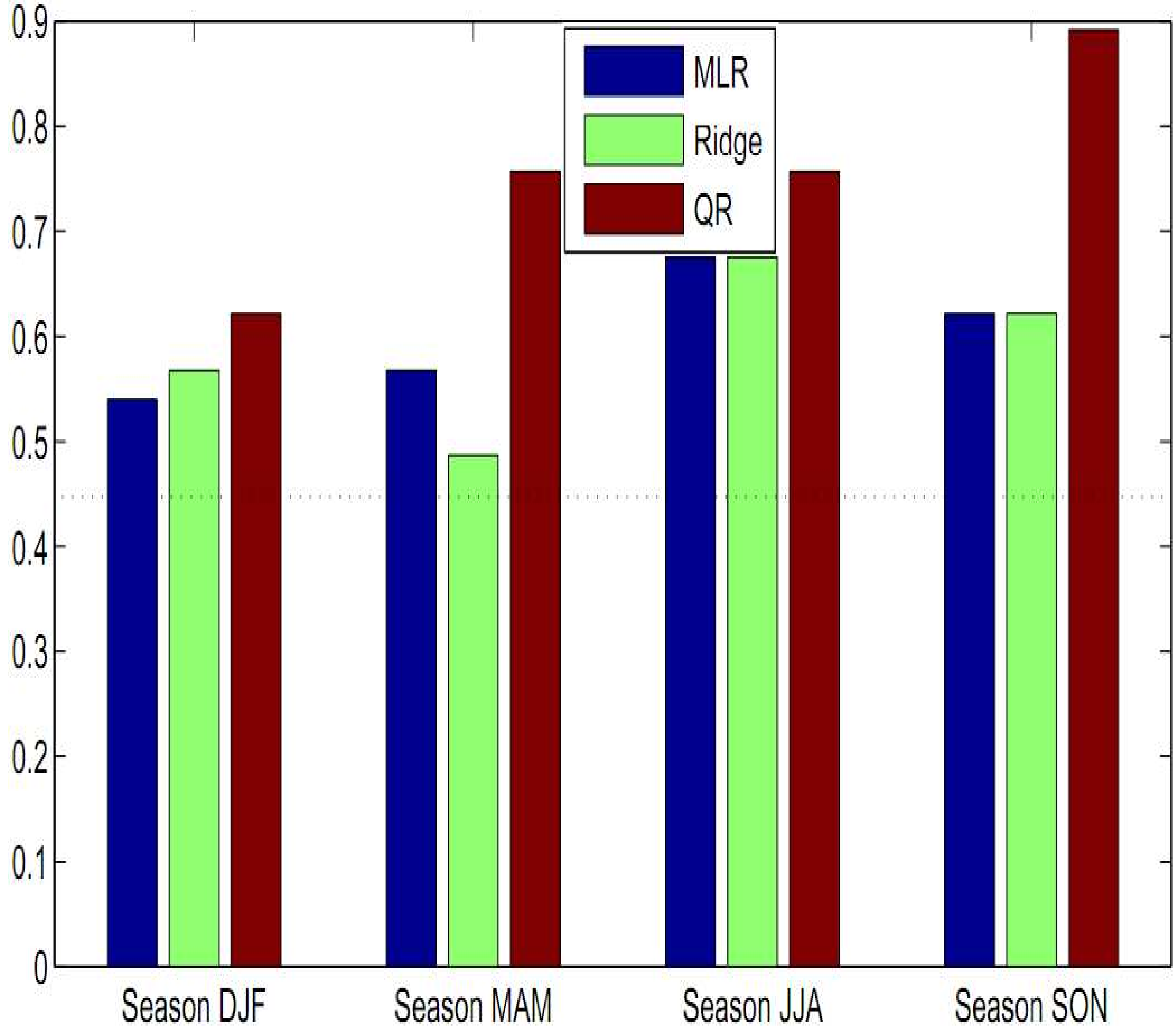


Figure 5.8: Ratio of stations LSSQR outperformin baseline in terms of F-measure of extreme data points.

for a station. As expected, the predicted value of the response variable using multiple linear regression is often underestimating the observed temperature, while quantile regression regularly overestimates the prediction of temperature and LSSQR lies in between MLR and QR and closer to the observed temperature.

5.5 Conclusions

This chapter presents a semi-supervised framework (LSSQR) for accurately predicting values of extreme data points. The proposed approach was applied to real world climate data spanning 37 stations and was compared against MLR, ridge regression and quantile regression in terms of the effectiveness the model demonstrated in identifying and predicting extreme temperatures for the given stations. The next chapter merges the intuition of the framework presented in this chapter, related to extreme values, with the integrated classification and regression framework presented in Chapter 3.

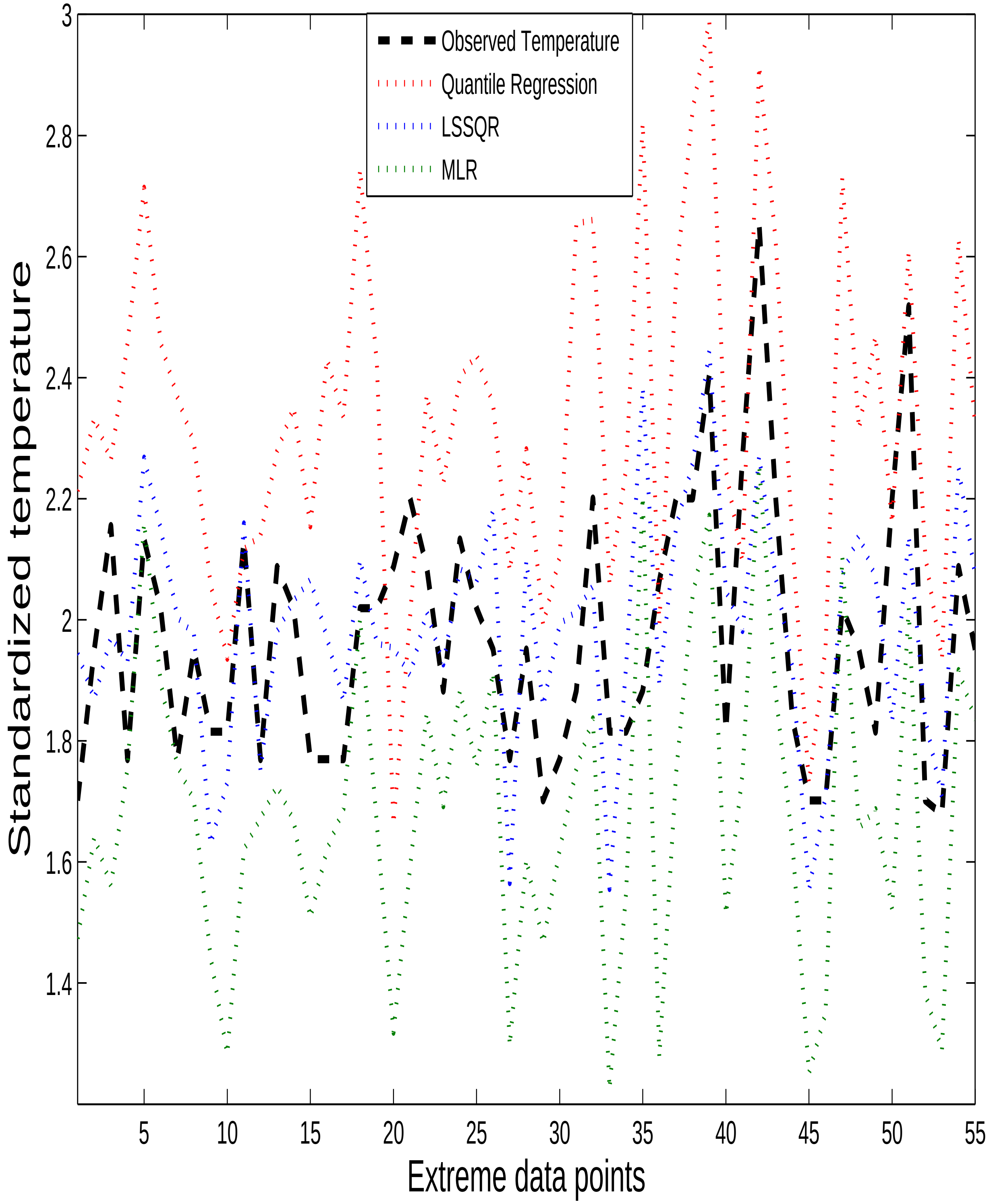


Figure 5.9: Prediction performance of extreme data points using MLR, Ridge, QR, LSSQR.

Chapter 6

Modeling Extremes in Zero-Inflated Data

This chapter extends the *LSQR* framework presented in Chapter 5, that emphasizes the accurate predicting of the frequency, timing and magnitude of extreme values in a the distribution of the response variable, to handle zero-inflated response variables such as daily precipitation.

6.1 Introduction

The notion behind being able to foretell the occurrence of an extreme event in a time series is very appealing, especially in domains with significant ramifications associated with the occurrence of an extreme events. Predicting pandemics in an epidemiological domain or forecasting natural disasters in a geological and climatic environment are examples of applications that give importance to detection of extreme events. Unfortunately, the accurate prediction of the timing and magnitude of such events is a challenge given their low occurrence rate. More so, the prediction accuracy depends on the regression method used as well as characteristics of the data. On the one hand, standard regression methods such as generalized linear model (GLM) emphasize estimating the conditional expected value, and thus, are not best suited for inferring extremal values. On the other hand, methods such as quantile regression are focused towards estimating the confidence limits of the prediction, and thus, may overestimate the frequency and magnitude of the extreme events. Though methods for inferring extreme value distributions do exist, combining them with other predictor variables for prediction purposes remains a challenging research problem.

Standard regression methods typically assume that the data conform to certain parametric distributions (e.g., from an exponential family). Such methods are ineffective if the assumed distribution does not adequately model characteristics of the real data. For example, a common problem encountered especially in modeling climate and ecological data is the excess probability mass at zero. Such zero-inflated data, as they are commonly known, often lead to poor model fitting using standard regression methods as they tend to underestimate the frequency of zeros and the magnitude of extreme values in the data. One way for handling such type of data is to identify and remove the excess zeros and then fit a regression model to the non-zero values. Such an approach, can be used, for example, to predict future values of a precipitation time series [115], in which the occurrence of wet or dry days is initially predicted using a classification model prior to applying the regression model to estimate the amount of rainfall for the predicted wet days. A potential drawback of this approach is that the classification and regressions models are often built independent of each other, preventing the models from gleaning information from each other to potentially improve their predictive accuracy. Furthermore, the regression methods used in modeling the zero-inflated data do not emphasize accurate prediction of extreme values.

The chapter presents an integrated framework that simultaneously classifies data points as zero-valued or not, and apply quantile regression to accurately predict extreme values or the tail end of the non-zero values of the distribution by focussing on particular quantiles.

We demonstrate the efficiency of the proposed approach on modeling climate data (precipitation) obtained from the Canadian Climate Change Scenarios Network website [1]. The performance of the approach is compared with four baseline methods. The first baseline is the general linear model (GLM) with a Poisson distribution. The second baseline used is the general linear model using an exponential distribution coupled with a binomial distribution classifier (GLM-C). A zero-inflated Poisson was used as the third baseline method (ZIP). The fourth basesline was quantile regression. Empirical results showed that the proposed framework outperforms the baselines for majority of the weather stations investigated in this study.

In summary, the main contributions of this chapter are as follows:

- Comparison and analysis of the performance of models created using variants of GLM, quantile regression and ZIP approaches to accurately predict values for extreme data points that belong to a zero-inflated distribution.
- Presenting an approach optimized for modeling zero-inflated data that outperforms the baseline methods in predicting the value of extreme data points.
- Successfully demonstrating the proposed approach to the real-world problem of downscaling precipitation climate data with application to climate impact assessment studies.

6.2 Preliminaries

Consider a multivariate time series $\mathbf{L} = (\mathbf{x}_t, y_t)$, where $t \in \{1, 2, \cdots, n\}$ is a discrete-valued index for time, $\mathbf{x}_t$ is a d-dimensional vector of predictor variables at time t, and y_t is the corresponding value for the response (target) variable. Given an unlabeled sequence of multivariate observations $\mathbf{x}_\tau$, where $\tau \in \{n+1, \cdots, n+m\}$, the goal is to learn a target function $f(\mathbf{x}, \boldsymbol{\beta})$ that best estimates the values of the response variable by minimizing the expected loss $\mathcal{E}_{\mathbf{x},y}[\mathcal{L}(\mathbf{y}, f(\mathbf{x}, \boldsymbol{\beta}))]$. The weight vector $\boldsymbol{\beta}$ denotes the regression coefficients to be estimated from the training data $\mathbf{L}$.

Multiple linear regression (MLR) is one of most widely used regression methods due to its simplicity. It assumes $f(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T\mathbf{x}$ (where $\mathbf{x}$ is a $(d+1)$-dimensional vector whose first element $x_0 = 1$ and $\boldsymbol{\beta} \in \Re^{d+1}$ is the weight vector) and the response variable y is related to $f(\mathbf{x}, \boldsymbol{\beta})$ via the following equation:

$$y = \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2).$$

As a result, $P(y|\mathbf{x}) \sim N(\boldsymbol{\beta}^T\mathbf{x}, \sigma^2)$ and $\mathcal{E}_{y|\mathbf{x}}[y] = \int yP(y|\mathbf{x})dy = \boldsymbol{\beta}^T\mathbf{x}$. Since the predicted value of the response variable for a test data point $\mathbf{x}_\tau$ is $\boldsymbol{\beta}^T\mathbf{x}_\tau$, this implies that the predictions made by MLR focus primarily on the average value of y given $\mathbf{x}_\tau$. This explains the limitation of MLR in terms of inferring extreme values in a given time series. The parameter vector $\boldsymbol{\beta}$ in MLR can be estimated using the maximum likelihood (ML) approach to obtain

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

where $\mathbf{X}$ is the $n \times (d+1)$ design matrix and $\mathbf{y}$ is an $n \times 1$ column vector for the observed values of the response variable.

The drawback of simple linear regression is that it is built on a strong assumption -namely, normality. Unfortunately, real world data may not always have a normal distribution and may be skewed to one side or may not cover the whole range of real numbers or may have a heavier tail than the normal distribution, etc. Hence, alternative approaches that are not constrained by such assumptions such as GLM may be used.

6.2.1 Generalized Linear Model and 2-Step GLM (GLM-C)

The generalized linear model is one of most widely used regression methods due to its simplicity. Generally, a GLM consists of three elements:

1. The response variable $\mathbf{Y}$, which has a probability distribution from the exponential family.

2. A linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$

3. A link function $g(\cdot)$ such that $E(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\eta)$

where, $\mathbf{Y} \in \mathcal{R}^{n\times 1}$ is the response variables vector, $\mathbf{X} \in \mathcal{R}^{n\times d}$ is the design matrix with all 1 in the last column. $\boldsymbol{\beta} \in \mathcal{R}^{p\times 1}$ is the parameter vector. Since the link function shows the relationship between the linear predictor and the mean of the distribution, it is very important to understand the detail about the data before arbitrarily using the canonical link function. In this case, since the precipitation data are always non-negative and values represented using a millimeter scale, the non-zero data may be treated as count data allowing the use Poisson distribution or an exponential distribution to describe the data. Hence, in these experiments $\log(\cdot)$ is chosen as the link function and Poisson distribution chosen. We scale the Y used in the regression model to be $10 \times Y$:

$$(10 \times Y_i)|X_i \sim Poi(\lambda_i)$$

$$E((10 \times Y_i)|X_i) = \lambda_i = g^{-1}(\eta_i) = g^{-1}(X_i\beta);$$

Considering the large number of zeros, one is motivated to perform classification first to eliminate the zero values before any regression. There are many classification methods available. But for the purpose of these experiments, logistic regression (which is also a variation of GLM) was chosen to do the classification. The response variable Y^* of logistic regression is a binary variable defined as:

$$Y^* = \begin{cases} 1 & Y > 0, \\ 0 & Y = 0 \end{cases}$$

The detail of the model is as follows: The link function is a logit link $g(p) = \log(\frac{p}{1-p})$, such that,

$$Y_i^* | X_i \sim Bin(p_i)$$

$$E(Y_i^* | X_i) = p_i = g^{-1}(\eta_i) = g^{-1}(X_i \beta);$$

When the fitted values are derived, they will be transferred to be binary:

$$f^* = \begin{cases} 1 & 1 \geq \hat{Y}^* > 0.5, \\ 0 & 0.5 \geq \hat{Y}^* \geq 0 \end{cases}$$

The second part is a GLM with exponential distribution, the response variable Y' is just those non-zero data, and the link function is $g(\cdot) = \log(\cdot)$:

$$Y_i' | X_i \sim Exp(\lambda_i)$$

$$E(Y_i' | X_i) = \lambda_i = g^{-1}(\eta_i) = g^{-1}(X_i \beta);$$

Then, the fitted-value $\mathbf{f'}$ was found for all X_i

Finally, the product of those two fitted-values $\hat{\mathbf{Y}} = \mathbf{f}^* \times \mathbf{f'}$ was reported.

To fit the GLM model, iteratively reweighted least squares(IRLS) method was used for maximum likelihood estimation of the model parameters.

6.2.2 Zero Inflated Poisson Regression (ZIP)

Differing from the methods above, zero inflated poisson regression treats the zero as a mixture of two distributions: a Bernoulli distribution with probability π_i to get 0, and a Poisson distribution with parameter μ (let $Pr(\cdot;\mu)$ denote the probability density function). In fact, the ZIP regression model is defined as:

$$Pr(Y = y_i|x_i) = \begin{cases} \pi_i + (1 - \pi_i)Pr(Y_i = 0; \lambda_i) & y_i = 0, \\ \\ (1 - \pi_i)Pr(Y = y_i; \lambda_i) & y_i > 0 \end{cases}$$

where $0 < \pi_i < 1$, and

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}) = x_i\boldsymbol{\beta}_1$$

$$\log(\mu_i) = x_i\boldsymbol{\beta}_2$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ are all regression parameter. Both of them could be found by maximizing the likelihood function. For the purpose of the experiments, the R package 'pscl' was used to fit the model.

6.2.3 Quantile Linear Regression(QR) and 2-step QR (QR-C)

Quantile regression was used to estimate the specified quantile of a population. Hence, if the objective of the regression is to estimate the conditional quantile(e.g., median) of $\mathbf{Y}$ instead of a conditional mean like MLR and Ridge regression, one may use quantile regression. Its loss function for the linear regression model is:

$$f(\mathbf{b}) = \sum_{i=1}^{N} \rho_\mathcal{T}(Y_i - \mathbf{X}_i^T \mathbf{b}), \text{ and } \hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} f(\mathbf{b}),$$

where

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

Let $F_Y(y) = P(Y \leq y)$ be the distribution function of a real valued random variable Y. The τ^{th} quantile of Y is given by:

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

It can be proved that the $\hat{y}$ which minimizes $E\rho_\mathcal{T}(y - \hat{y})$ should satisfy that $F_Y(\hat{y}) = \tau$. Thus, quantile regression will find the τ^{th} quantile of a random variable, for example:

$$\text{Median}(\mathbf{Y}|\mathbf{X}) = X\hat{\boldsymbol{\beta}}^{qr}; \hat{\boldsymbol{\beta}}^{qr} = \arg\min_{\mathbf{b}} \sum \rho_{0.5}(y_i - \mathbf{X_i^T b})$$

For the purpose of the experiments conducted, $\tau = 0.95$ was used to represent extreme high value. Unlike the least squares methods mentioned above, which could be solved by numerical linear algebra, the solution to quantile regression is relatively non-trivial. Linear programming is used to solve the loss function by converting the problem to the following form.

$$\min_{\mathbf{u},\mathbf{v},\mathbf{b}} \{\tau \mathbf{e}_{\mathbf{N}}^{\mathbf{T}} \mathbf{u} + (1-\tau)\mathbf{e}_{\mathbf{N}}^{\mathbf{T}} \mathbf{v} | \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{u} - \mathbf{v}; \mathbf{b} \in \mathcal{R}^{p}; \mathbf{u}, \mathbf{v} \in \mathcal{R}_{+}^{N}\}$$

For the same reason as mentioned in the Section 6.2.1, a classification method should be incorporated along with the regression model. Logistic regression was used for classification, and quantile regression on those nonzero Y. Finally, the product of those two fitted values is reported. Quantile regression may return a negative value, which we force to 0. We do this because precipitation is always non-negative.

6.3 Framework for Integrated Classification and Regression

With the introduction of quantile regression, which is an integral part of the objective function, the motivation behind the various components of the proposed objective function needs to be elaborated. Since zero-inflated data is best described with the help of a classifier that help identify non-zero values and a regression component to address non-zero values, this framework consists of both components. For the classifier component, a least square support vector machine is used and for the regression component, the intuition of quantile regression is used to help focus the regression of extreme values. Since the final prediction of the data point using this framework is a product of the regression and classification component, the quantile regression component is built to work on the eventual predicted return value, thereby integrating both the classifier and regression components.

6.3.1 Integrated Classifier and Regression for Extreme Values (ICRE)

The classification and regression models developed in this study are designed to minimize the following objective function:

$$\begin{aligned} \arg\min_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} L(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) &= \frac{1}{n}\sum_{i=1}^{n}(1-(2y_i-1)f_i)^2 \\ &+ \frac{1}{n^*}\sum_{i=1}^{n} y_i\rho_{\mathcal{T}}(y_i'-f_i'\times(f_i+1)/2)+\lambda(||\boldsymbol{\omega}_1||^2+||\boldsymbol{\omega}_2||^2) \end{aligned} \tag{6.1}$$

where n^* is the number of nonzero y_i. Then it can be expanded as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} L(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) &= \frac{1}{n}\sum_{i=1}^{n}(1-(2y_i-1)(\mathbf{x}_i^T\boldsymbol{\omega}_2))^2 \\ &+ \frac{1}{n^*}\sum_{i=1}^{n} y_i\rho_{\mathcal{T}}(y_i'-(\mathbf{x}_i^T\boldsymbol{\omega}_1)\times(\text{sign}((\mathbf{x}_i^T\boldsymbol{\omega}_2+1)/2))) \\ &+ \lambda(||\boldsymbol{\omega}_1||^2+||\boldsymbol{\omega}_2||^2) \end{aligned} \tag{6.2}$$

The rationale for the design of our objective function is as follows. The first term which corresponds to the regression part of the equation represents quantile regression performed for only the observed non-zero values in the time series. The regression model is therefore biased towards estimating the non-zero extreme values more accurately and not be adversely influenced by the over-abundance of zeros in the time series. The product $f_i'\times(f_i+1)/2$ in the first term, corresponds to the predicted output of our joint classification and regression model. The second term in the objective function, which is the main classification component, is equivalent to the least square support vector machine. And the last two terms in the objective function are equivalent to the L_2 norm used in ridge regression models to shrink the coefficients in $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$.

Each data point is considered to be a representative reading at an instance of time $t \in \{1, 2, \cdots, n\}$ in the time series. Each predictor variable is standardized by subtracting its mean value and then dividing by its corresponding standard deviation. The standardization of the variables is needed to account for the varying scales.

The optimization method used while performing experiments is 'L-BFGS-B', described by Byrd et. al. [27]. It is a limited memory version of BFGS methods. This method does not store a Hessian matrix, just a limited number of update steps for it, and then it uses derivative information. Since this model includes a quantile regression component, which is not differentiable, this method of optimization is well suited to the objective function.

To solve the objective function, the inverse logistic function of $\mathbf{x}_i^T \boldsymbol{\omega}_2$instead of sign$((\mathbf{x}_i^T \boldsymbol{\omega}_2 + 1)/2))$ was used. The decision was motivated by the fact that the optimizer tries to do a line search along the steepest descent direction and finds the positive derivative along this line, which would result in a nearly flat surface for the binary component. Hence, conversion of the binary report to an inverse logistic function of $\mathbf{x}_i^T \boldsymbol{\omega}_2$ was used to address this issue. During the prediction stage, the binary-fitted values from the SVM component was used.

6.4 Experimental Evaluation

In this section, the climate data that are used to downscale precipitation is described. This is followed by the experiment setup. Once the dataset is introduced, the behavior of baseline models was analyzed and contrasted with ICRE, in terms of relative performance of the various models when applied to this real world dataset to forecast future values of precipitation.

6.4.1 Data

All the algorithms were run on climate data obtained for 29 weather stations in Canada, from the Canadian Climate Change Scenarios Network website [1]. The response variable to be regressed (downscaled), corresponds to daily precipitation values measured at each weather station. The predictor variables correspond to 26 coarse-scale climate variables derived from the NCEP Reanalysis data set and the H3A2a data set(computer generated simulations), which include measurements of airflow strength, sea-level pressure, wind direction, vorticity, and humidity. The predictor variables used for training were obtained from the NCEP Reanalysis data set while the predictor variables used for the testing were obtained from the H3A2a data set. The data span a 40-year period, 1961 to 2001. The time series was truncated for each weather station to exclude days for which temperature or any of the predictor values are missing.

6.4.2 Experimental Setup

The first step was to standardize the predictor variables by subtracting its mean value and then dividing by its corresponding standard deviation to account for their varying scales. The training size used was 10yrs worth of data and the test size, 25yrs. During the validation process, the selection of the parameter λ was done using the score returned by RMSE-95. Also, to ensure the experiments replicated the real world scenario where the prediction for a future timeseries needs to be performed using simulated values of the predictor variables for the future time series, simulated values for the corresponding predictor variables obtained from H3A2a climate scenario was used as $\mathbf{X}_U$, while $\mathbf{X}_L$ are values obtained from NCEP. All the experiments were run for 37 stations.

6.4.3 Baseline Algorithm

We compare the performance of ICRE with baseline models created using general linear model(GLM), general linear model with classification (GLM-C), quantile regression(QR), quantile regression with classification and zero-inflated Poisson(ZIP). Further details about the baselines are provided below.

6.4.3.1 General Linear Model (GLM)

The baseline GLM refers to the generalized linear model that uses a Poisson distribution as a link function, resulting in the regression function $\log(\lambda) = X\beta$, where $E(Y|X) = \lambda$

6.4.3.2 General Linear Model with Classification (GLM-C)

Unlike the previous baseline (GLM), GLM-C refers to a two step generalized linear model that uses a Binomial distribution, for the classifier with the model described as $logit(p) = X\beta$, and $E(Y' = 1|X) = p$ which $Y' = 1$ when $Y > 0$ and $Y' = 0$ when $Y = 0$ and a second step that uses a generalized linear model with an exponential distribution that is built only on non-zero response data points. The regression function is $\log(\lambda) = X\beta$, which $E(Y|X) = \lambda$. The eventual predicted value for each data point is the product of the two respective fitted values.

6.4.3.3 Quantile Regression (QR)

The baseline QR refers to the regular quantile regression described earlier in the preliminary section 6.2

6.4.3.4 Quantile Regression with Classification (QR-C)

The baseline QR-C refers to a two step model that has a GLM that uses a binomial distribution that acts as a classifier and a regular quantile regression model that is built on non-zero valued data points as described earlier in the preliminary section. These two models that comprise QR-C are built independent of each other and the eventual predicted value for each data point is the product of the two respective fitted values.

6.4.3.5 Zero Inflated Poisson (ZIP)

Zero Inflation Poisson model used as a baseline and is similar to the ZIP model described in Section 6.2.

6.4.4 Evaluation Criteria

The motivation behind the selection of the various evaluation metrics was to evaluate the different algorithms in terms of predicting the magnitude and the timing of the extreme events.The following criteria to evaluate the performance of the models are used:

- Root Mean Square Error (RMSE), which measures the difference between the actual and predicted values of the response variable, i.e.:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y'_i - f'_i f_i)^2}{n}}$$

- RMSE-95, was used to measure the difference between the actual and predicted value of the response variable for only the extreme data points(j). Extreme data points refer to the points whose actual value were 95 percentile and above. The equation is with respect to 95 percentile, as throughout this chapter, we associate data points that are

95 percentile and above as extreme values, i.e.:

$$\text{RMSE-95} = \sqrt{\frac{\sum_{j=1}^{n/20}(y'_j - f_i f'_j)^2}{n/20}}$$

- Confusion matrices will be computed to visualize the precision and recall of extreme and non-extreme events. F-measure, which is the harmonic mean between recall and precision values was used as a score that evaluates the precision and recall results.

$$\text{F-measure} = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

To summarize, RMSE-95 is used for measuring magnitude and F-measure measures the correctness of the timing of the extreme events.

6.4.5 Experimental Results

The results section consists of two main sets of experiments. The first set of experiments evaluates the impact of zero-inflated data on modeling extreme values. The second section compares the performance of ICRE with the baseline methods which are followed .

6.4.5.1 Impact of Zero-Inflated Data on Extreme Value Prediction

Unlike regular data which may be modeled using regression, modeling zero-inflated data usually involves a classifier and a regression component. The classifier is used to identify zero and non-zero values, which is followed by regression for the non-zero values. But since the focus of the chapter is on extreme data points within zero-inflated data, the impact of the classifier is unclear. In this section, the impact of including the classifier in modeling extreme values of zero-inflated data is compared. QR is compared with QR-C and GCM with GCM-C and show the results in Table 6.1. Note that the percentage of wins for F-measure, recall, precision may not total to 100 in the case of a tie.

Table 6.1: Percentage of stations won

	QR-C	QR	GLM-C	GLM
RMSE-95	0	100	67.57	32.43
F-Measure	81.08	18.92	18.92	35.13

As shown in the Table 6.1, it isn't clear that using an independent classifier along with regression for modeling extreme values among zero inflated data is preferred. But the results do indicate that the inclusion or exclusion of a classifier with the regression model built independent of each other may compromise either RMSE-95 (by overestimating the magnitude) or F-measure (mistiming predicting an extreme value), without necessarily compromising both together.

6.4.5.2 Comparison of ICRE to Baseline Methods

Table 6.2 shows the relative performance of ICRE to all the baseline methods in terms of percentage of stations outperformed against the baseline method in terms of RMSE-95 values calculated on extreme rain days. In terms of RMSE of extreme rain days, as shown in Table 6.2, ICRE outperformed the baselines (except QR) in almost every one of the 37 stations. But QR was the best across all methods for RMSE-95 of extreme days. In terms of F-measure that was computed based on recall and precision of identifying extreme events,

Table 6.2: Percentage of stations ICRE outperformed the baseline

	QR-C	QR	GLM-C	GLM	ZIP
RMSE-95	91.89	0	97.3	97.3	97.3
F-Measure	43.24	62.16	89.19	89.19	91.9

ICRE again outperformed the baselines(except QR-C) in majority of the 37 stations. But ICRE was only able to outperform QR-C in 16 or the 37 stations in terms of F-measure. Although QR performed the best in terms of estimating magnitude for those extreme events, it over-estimated the timing of the events as seen by the relatively lower F-measure score. QR-C did the reverse, it did reasonably well in terms of modeling the timing, but performed very poorly in terms of the magnitude of the events by overestimating.

6.5 Conclusions

This chapter compares and analyzes the performance of models created using variants of GLM, quantile regression and ZIP approaches to accurately predict values for extreme data points that belong to a zero-inflated distribution. An alternate framework(ICRE) was present that outperforms the baseline methods and the effectiveness of the model was demonstrated on climate data to predict the amount of precipitation at a given station.

Chapter 7

Contour Regression

The *LSSQR* and *ICRE* frameworks presented in Chapter 5 and Chapter 6 prioritize the accuracy of the prediction of a response variable at a user specified quantile. However, there are climate model applications that are interested in capturing the accurate distribution characteristics of the response variable across all quantiles. In this chapter, the limitations of current regression-based approaches in terms of preserving the distribution of observed climate data is shown and a multi-objective regression framework that simultaneously fits the distribution properties and minimizes the prediction error is presented. The framework is highly flexible and can be applied to linear, nonlinear, and conditional quantile models. The chapter 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.

7.1 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 [107, 60, 104]. In addition to minimizing the residuals of the predicted outputs, some of

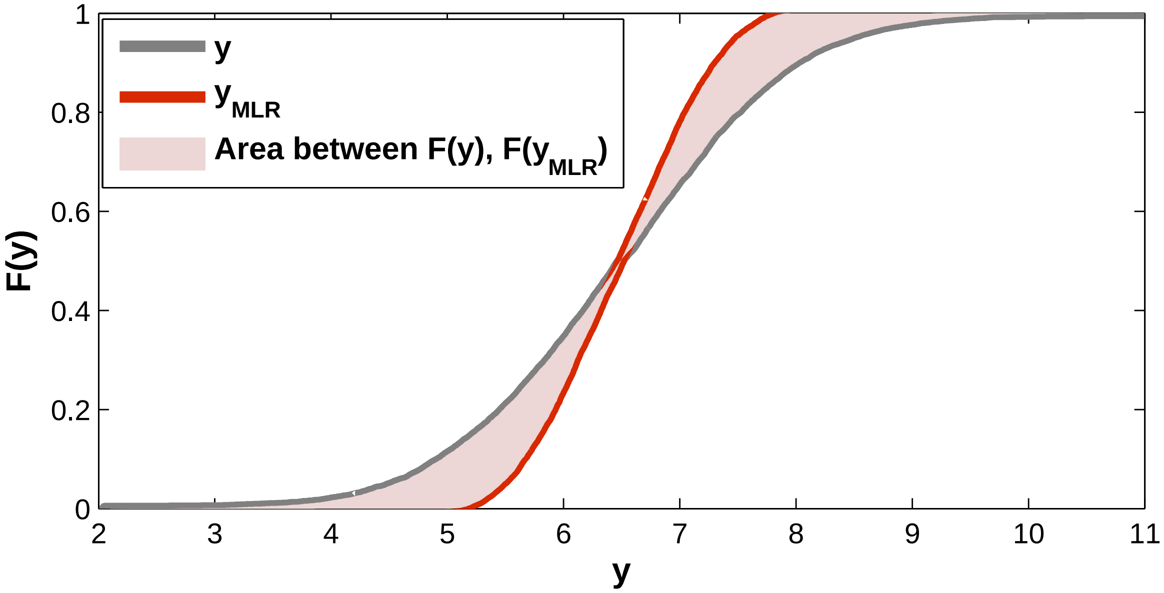


Figure 7.1: Area between the CDF of y and y_{MLR}.

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.

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 7.1. Even though the regression model was trained using ten thousand data points, it is clear from Figure 7.1 that MLR fails to replicate the shape of the cumulative distribution for y, particularly the tails of the distribution.

As another example, Figure 7.2 compares the histograms of daily maximum temperature observed at a climate station in Michigan and the predicted outputs of MLR. In this case,

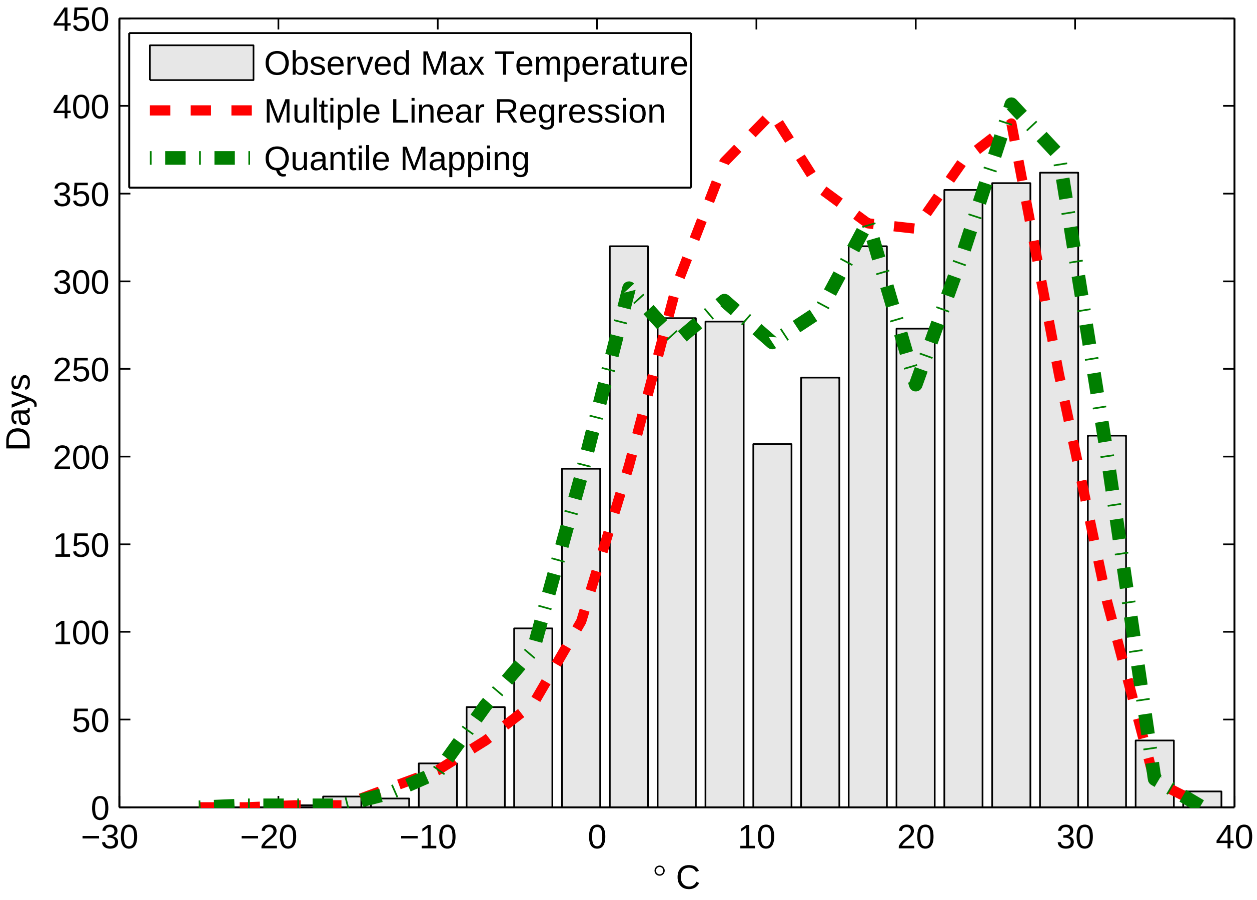


Figure 7.2: Histogram of predicted daily maximum temperature at a weather station in Michigan, 1990-1999.

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-driven approaches [108, 97] 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

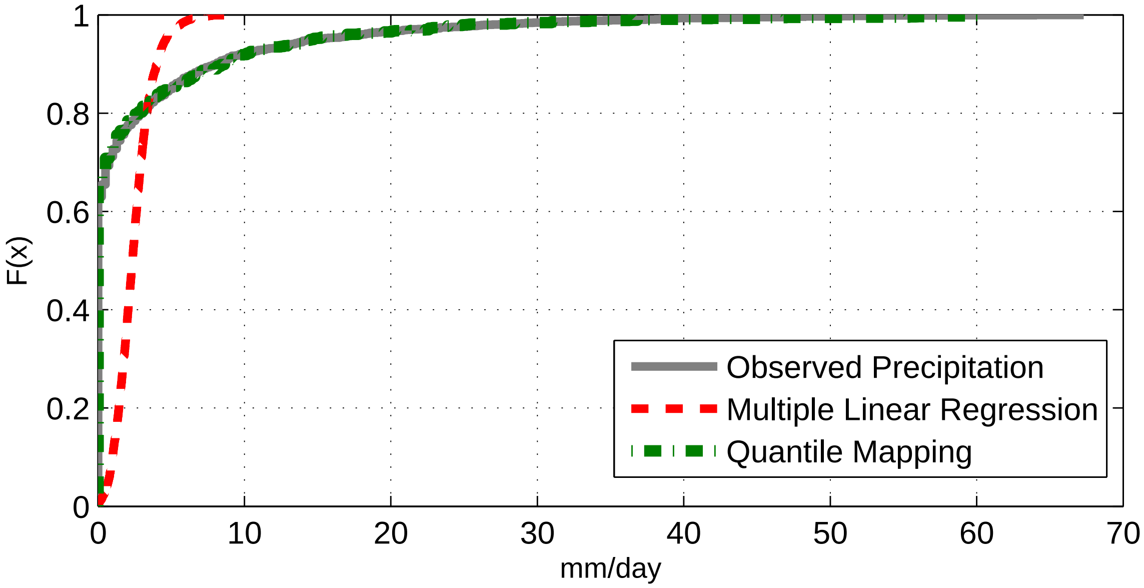


Figure 7.3: CDF of predicted daily precipitation at a weather station in Michigan, 1990-1999.

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 [45].

Since the fidelity of both the distribution characteristics and the accuracy of projections are important, a framework for multivariate regression is proposed 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 chapter, the effectiveness of the proposed framework is demonstrated by downscaling and bias correcting daily temperature and precipitation to match their corresponding observations.

In summary, the main contributions of this study are:

- Identification of the limitations of existing least squared error regression techniques.
- Presentation of 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.
- Demonstration of the feasibility of adapting the framework to fit various objective functions such as multivariate ordinary least squares, QR and non-linear kernel ridge regression.
- Evaluation of 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.

7.2 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.

7.2.1 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

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

7.2.2 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.

It can be shown that a QM function that accurately replicates the distribution characteristics of the response variable may not guarantee $RMSE = 0$, nor a relatively small $RMSE$.

Proposition 7.2.1. *A QM function that accurately replicates distribution of the observation may have* $RMSE > 0$

Proof. Given $QM : y = F_{\mathbf{y}}^{-1}(F_{\mathbf{x}}(x))$, where $F_X \in [0,1]$ and $F_Y \in [0,1]$ are the empirical cumulative distribution function of $\mathbf{x}$ and $\mathbf{y}$ respectively. Let R and O be the multiset containing the quantile values of $\mathbf{x}$ in $F_{\mathbf{x}}$ and $\mathbf{y}$ in $F_{\mathbf{y}}$ respectively. i.e., $F_{\mathbf{x}}(\mathbf{x}) = R$ and $F_{\mathbf{y}}(\mathbf{y}) = O$. Let $\varepsilon(i) = |F_{\mathbf{y}}(O(i)) - F_{\mathbf{y}}(R(i))|$ be the QM prediction error of data point x_i.

$\Rightarrow RMSE = \sqrt{\sum_i \varepsilon^2(i)/n}$. A necessary and sufficient condition for the quantile function to replicate the distribution of observation is that the cardinality, member and multiplicity of the multiset O equals the multiset R, i.e.,

$$|O \bigcap R| = |O| = |R|$$

The above requirement does not eliminate that possibility that $\exists i$, s.t., $O(i) \neq R(i)$, where $R(i)$ corresponds to the quantile value of data point x_i.

$\Rightarrow$ if $\exists i$, s.t., $O(i) \neq R(i)$, and $F_Y(O(i)) \neq F_Y(R(i))$. $\Rightarrow RMSE > 0$. Hence, quantile mapping function that accurately replicates the distribution characteristics of the response variable may not guarantee $RMSE = 0$, nor a relatively small $RMSE$. $\diamondsuit$ $\square$

Proposition 7.2.2. *A QM function accurately returns the distribution characteristics of the response variable as well as $RMSE = 0$ when rank correlation $\Gamma = 1$ between the predictor and response variables.*

Proof. Let R and O be the multiset quantiles $F_{\mathbf{x}}(\mathbf{x})$ and $F_{\mathbf{y}}(\mathbf{y})$ of $\mathbf{x}$ and $\mathbf{y}$ respectively. Let $\varepsilon(i) = |F_{\mathbf{y}}(O(i)) - F_{\mathbf{y}}(R(i))|$ be the ith error of the predicted values from QM. $\Rightarrow RMSE = \sqrt{\varepsilon^2(i)/n}$. Given $(\Gamma = 1) \equiv (\forall i, R(i) = O(i))$, we have $\varepsilon(i) = |F_{\mathbf{y}}(O(i)) - F_{\mathbf{y}}(O(i))|$. $\Rightarrow RMSE = \sqrt{\sum_i \varepsilon^2(i)/n} = 0$ $\diamondsuit$ $\square$

7.3 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 7.1, 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 7.3.1 elaborates on CR and describes multivariate linear contour regression (MLCR) which has an objective function that is based on ordinary least squares. Section 7.3.2 proposes kernel contour regression (KCR) that is a kernel-based interpretation of the CR framework. Section 7.3.3 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.

7.3.1 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 $\gamma\rightarrow 1$, MLCR converges to the solution of MLR as seen in Figure 7.4, 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

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

Since it is often not possible to guarantee that X is pre-sorted correctly according to $f(x_i,\beta)$, one 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 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.

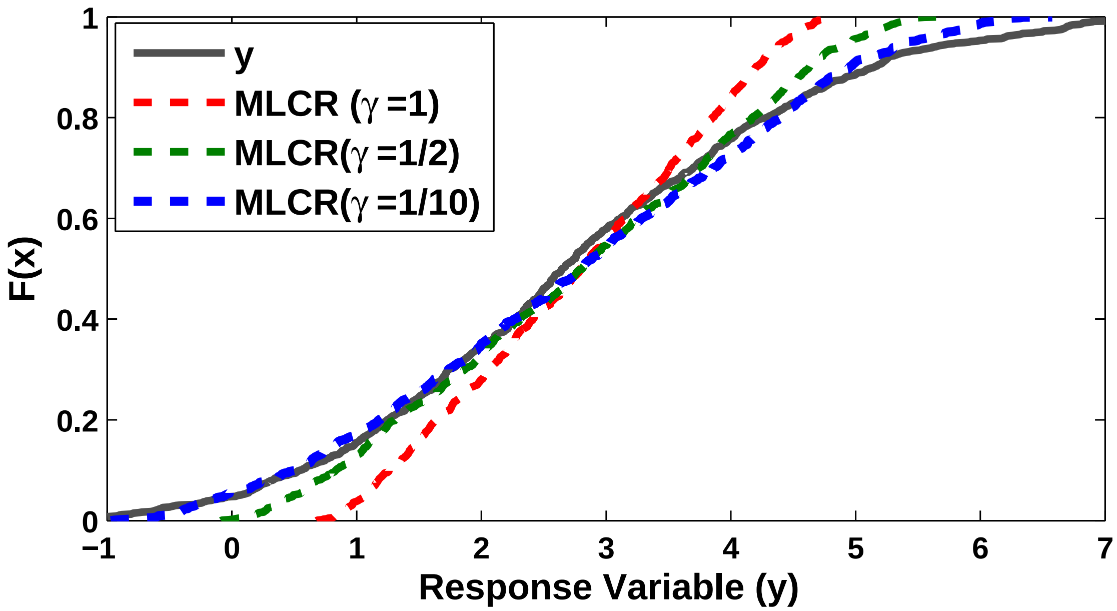


Figure 7.4: Influence of gamma parameter on fidelity of the response variable's cumulative distributive function.

7.3.1.1 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 7.3.1. *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 7.3.2. *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

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

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

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

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 7.3.1. *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 7.3.1. *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.*

7.3.2 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 $X^T X$ 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

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

where, the ridge coefficient $\lambda > 0$ results in a non-singular matrix $X^T X + \lambda I$ 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.

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

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

$$\Rightarrow \alpha = (G + \lambda I)^{-1} (\gamma y + (1-\gamma) z)$$

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

7.3.3 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) [77] 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 $\beta \in R^d$.

7.3.3.1 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 7.3.3. *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 7.3.4. *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 $L(\beta_{t+1}, f_{t+1}, X_t) = L^y_{t+1} + L^z_t$ where, L^y_{t+1} refers to the first half of the loss function that performs QR on y and L^z_t 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, $L^z_t = 0.5 \sum_{i=1}^{n}(f_i - z_i)$ and $L^z_{t+1} = 0.5 \sum_{i=1}^{n}(f_{(i)} - z_i)$

$$\Rightarrow L^z_t = L^z_{t+1}$$

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

Theorem 7.3.2. *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 $L(\beta_{t+1}, f_{t+1}, X_t) = L(\beta_{t+1}, f_{t+1}, X_{t+1})$.

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

□

7.4 Experimental Results

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

7.4.1 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) [2] (Table 7.1). 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.

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

Table 7.1: 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

$kg/m^2/s$ 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.

7.4.2 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. [45], and the relative performances being consistent across the two set-ups, the results of 10-fold cross validation are not included in this chapter.

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 chapter, the extreme 5 percentile of a distribution is defined 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.

7.4.3 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. The performance of MLCR was compared 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, MLCR was compared first with 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, MLR was used as a baseline to evaluate possible 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

Table 7.2: 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

maximum and minimum temperature) (Table 7.2) while improving the average error in terms of empirical cumulative distribution frequency (RMSE-CDF), around 40% (Figure 7.6).

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 7.2). 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 7.5). Since, MLR struggled to capture the shape of the precipitation distribution, a smaller value for the γ parameter for MLCR was chosen, 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, the Kolmogorov-Smirnov statistic (K) is used, 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

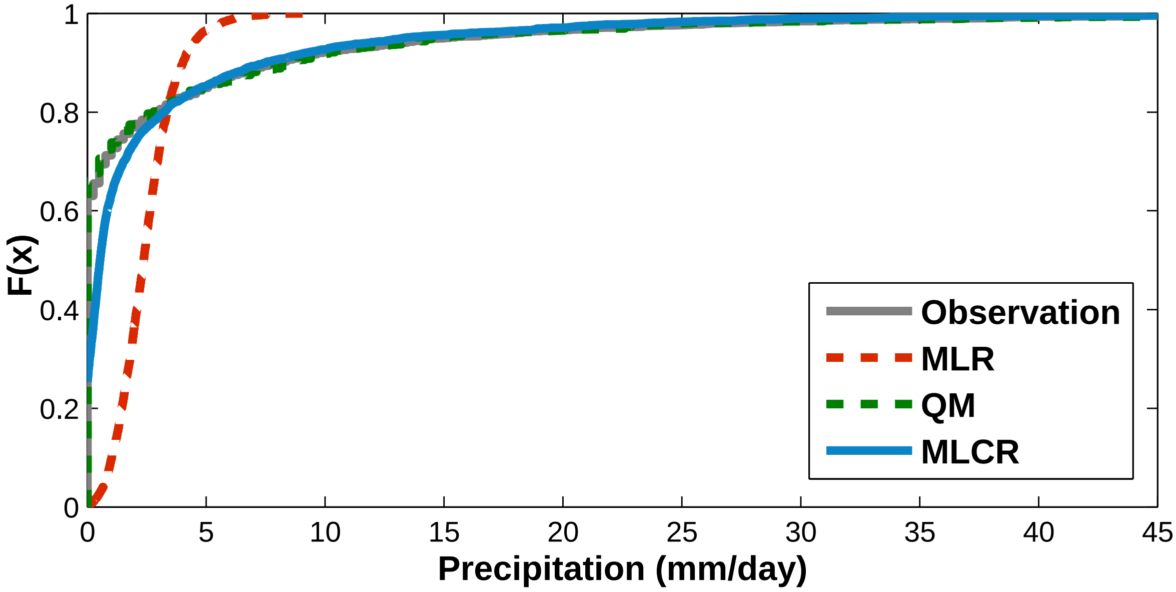


Figure 7.5: CDF of predicted daily precipitation at a weather station in Michigan over the years 1990-99.

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 7.3), 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 7.3: 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

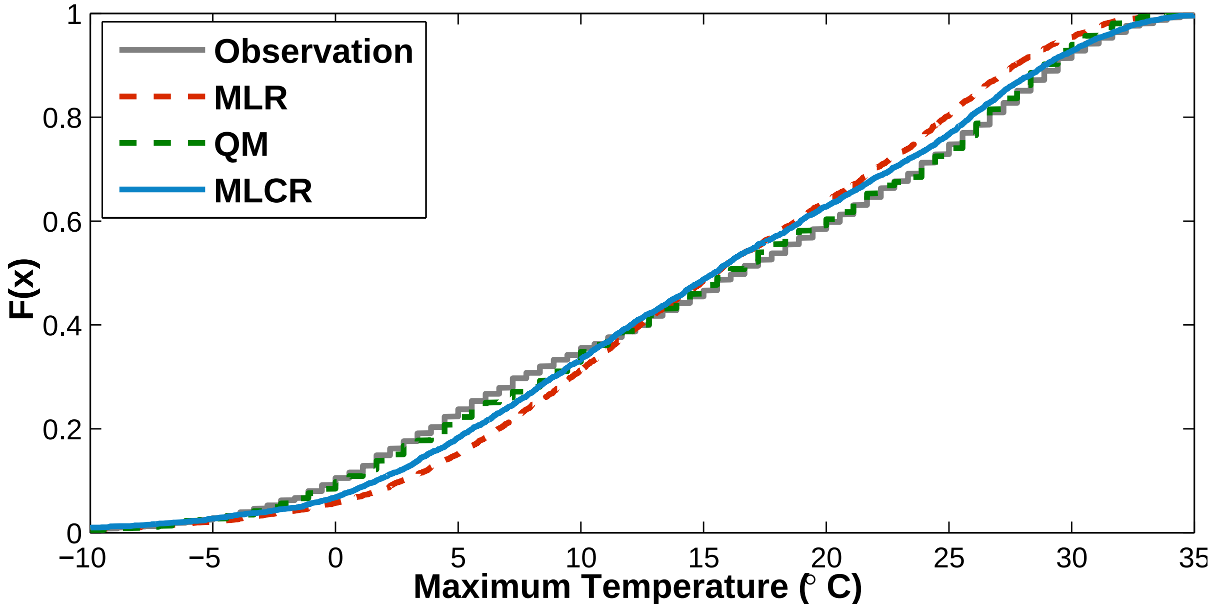


Figure 7.6: CDF of predicted daily maximum temperature at a weather station in Michigan, 1990-99.

7.4.3.1 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 7.8), 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

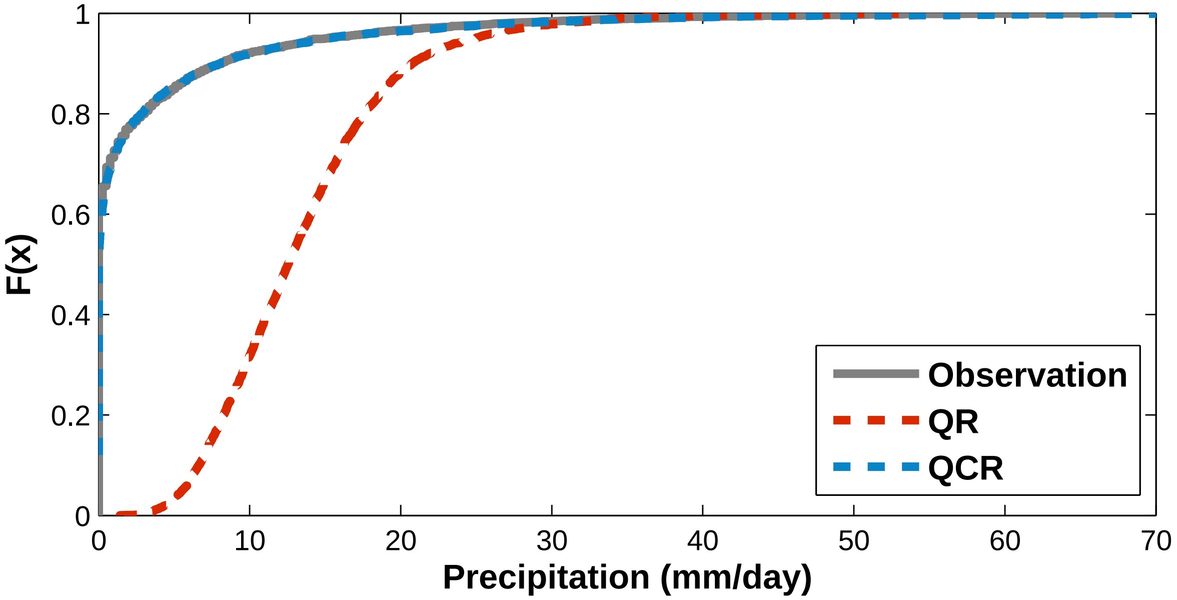


Figure 7.7: CDF of predicted daily precipitation at a weather station in Michigan, 1990-99.

based approaches in terms of correctly identifying extreme values across all the stations. Figure 7.7 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 7.4, QCR regularly outperformed QR in terms of the other metrics such as correlation.

7.5 MCR Using Heterogeneous Data

The MCR framework can also incorporate heterogeneous data sources of predictor variables. This extension is referred to as MCR_{HET}. An example of incorporating the heterogeneous data sources is utilizing a reanalysis values for the predictor variables as well as asynchronous data obtained from GCM driven runs of RCM. The asynchronous predictor

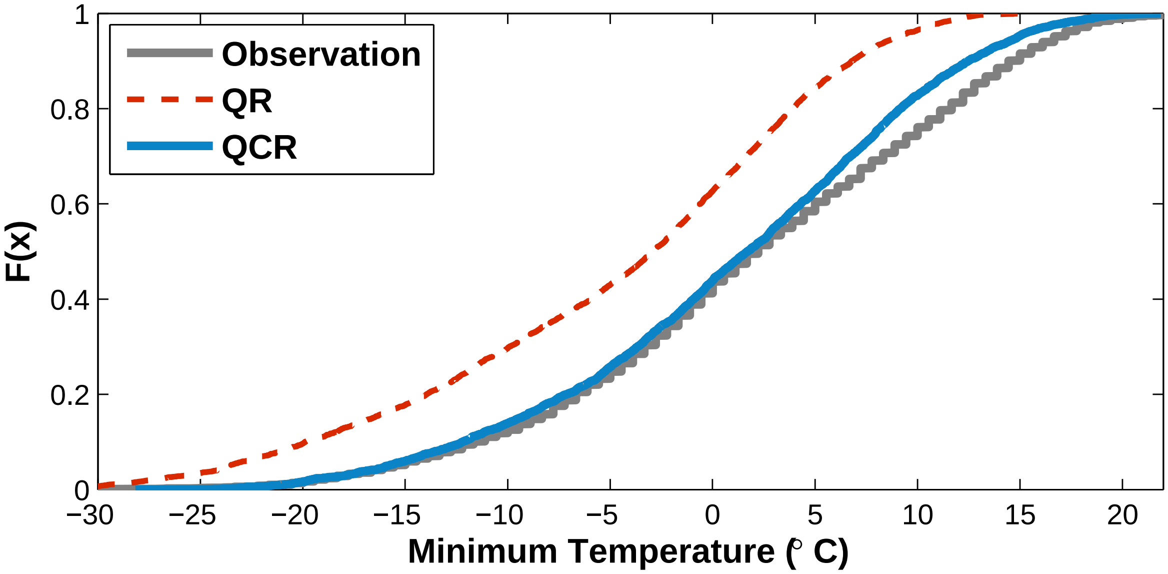


Figure 7.8: CDF of predicted daily minimum temperature at a weather station in Michigan, 1990-99.

variables obtained from GCM driven runs of RCM is incorporated into the second term of MCR. Figure 7.9 shows the CDF of predicted daily minimum temperature at a weather station (Eau Claire)in Michigan, 1990-99, using asynchronous regional climate model data. Similarly, Figure 7.10 shows the CDF of predicted daily precipitation at the same weather station in Michigan, 1990-99, using asynchronous regional climate model data.

7.5.1 Geometric Quantile Mapping

The multi-dimensional equivalent of quantile function is geometric quantile [32].

For a univariate random variable $X \in \Re$, let $F_X(x)$ be its cumulative distribution function (CDF), i.e., $F_X(x) = P(X \leq x)$. The corresponding α-quantile of X is given by $\inf\{x \in \Re : F_X(x) \geq \alpha\}$. More generally, the position [84] of data point $\mathbf{z}$ relative to a set of points $\mathbf{Z} = (\mathbf{z}_1, .., \mathbf{z}_m)^T$ is given by

Table 7.4: 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

$$\mathbf{p_Z}(\mathbf{z}) = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \eta(\mathbf{z}-\mathbf{z}_i) \qquad \text{where} \quad \eta(\mathbf{w}) = \begin{cases} \frac{\mathbf{w}}{\|\mathbf{w}\|}, & \text{if } \mathbf{w} \neq \mathbf{0} \\ 0, & \text{if } \mathbf{w} = \mathbf{0} \end{cases}$$

For univariate data, the position $p_Z(z)$ is equal to $2F_Z(z) - 1$, where $F_Z(z)$ is the cumulative distribution function of Z. The multi-dimensional equivalent of quantile function is geometric quantile [32].

Distribution correction methods such as quantile mapping is only applicable if one can match the position of a data point in one univariate distribution (say for x) to its corresponding position in another univariate distribution (say for y). This is possible using the preceding definition of position for univariate data since the values of p_Z are always fixed in the range between $[-1, +1]$ irrespective of the values in Z. Unfortunately, when extended to multivariate positions, the range of values for p_Z may vary depending on the values in Z. To overcome this problem, He et al. [61] introduce the notion of a stationary position by iteratively applying the following position transformation function until convergence:

$$\mathbf{p}_Y^k(\mathbf{z}) = \frac{1}{\kappa n}\sum_{i=1}^{n} \frac{\mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i)}{\| \mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i) \|}, \qquad \mathbf{p}_Y^1(\mathbf{z}) = \frac{1}{\kappa n}\sum_{i=1}^{n} \frac{\mathbf{z} - \mathbf{y}_i}{\| \mathbf{z} - \mathbf{y_i} \|} \tag{7.1}$$

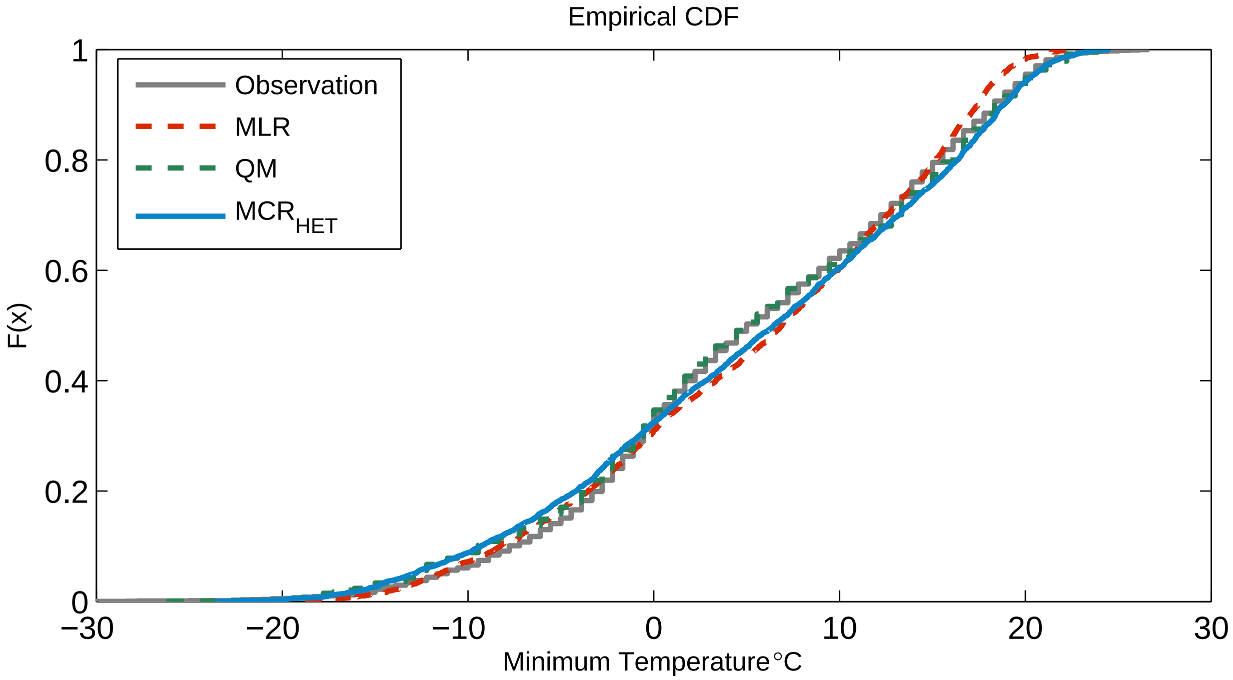


Figure 7.9: Cumulative distribution function of predicted daily minimum temperature at a weather station in Michigan, 1990-99, using asynchronous regional climate model data.

Here each component in $\mathbf{y}_i$ must be converted to its marginal rank first before applying the position transformation function. Marginal rank refers to the rank of the data point divided by the largest rank and then normalized to the range $[-1, 1]$. The normalization is done to negate the effect of variables having values that correspond to different ranges. Data points with normalized marginal rank close to ± 1 correspond to extreme values for the particular variable, while those close to $\mathbf{0}$ are located near the median of the distribution. In practice, the number of iterations needed to reach a stationary distribution is quite small, typically $K > 5$ [61]. For univariate data, it can be shown that $\mathbf{P}^k$ reaches a stationary distribution at $k = 1$. The term κ in Equation (7.1) is a normalization factor to ensure the distribution of the geometric positions is supported in a q-dimensional unit hypersphere. In the case of bivariate response variable $\mathbf{Y}$, the stationary geometric quantile distribution is circularly symmetric around the origin, with the radial density of $r/\sqrt{1-r^2}$ for $r \in (0, 1)$

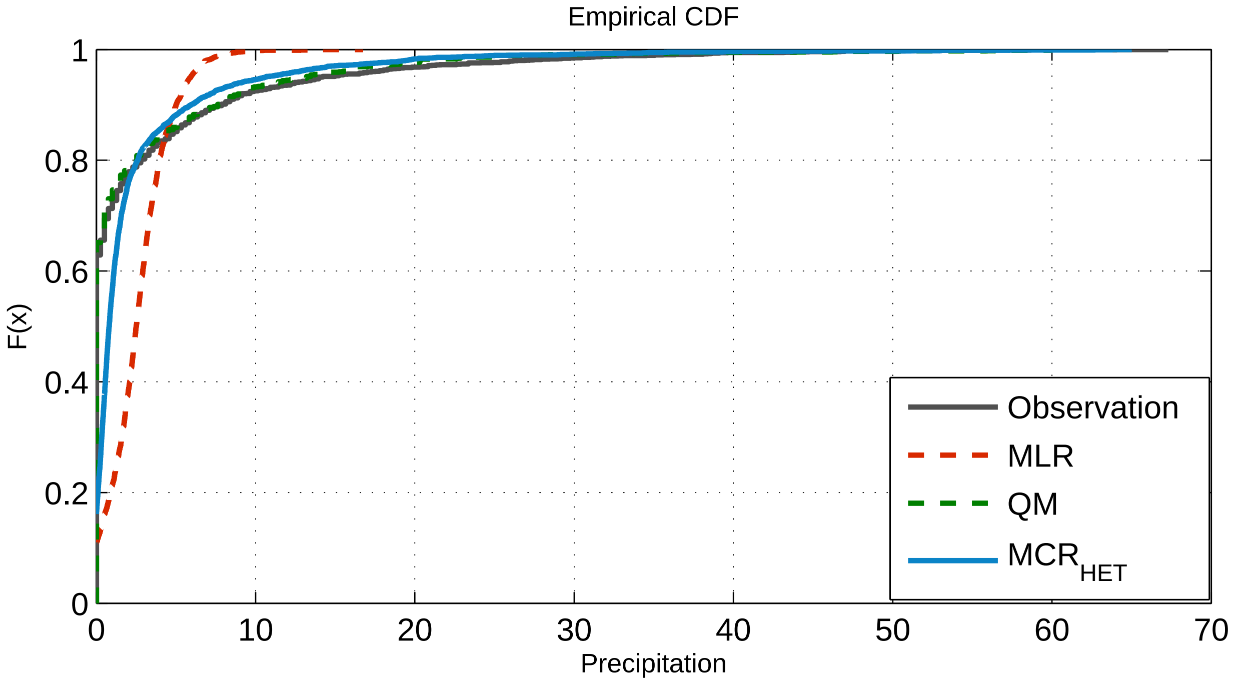


Figure 7.10: Cumulative distribution function of predicted daily precipitation at a weather station in Michigan, 1990-99, using asynchronous regional climate model data.

[61].

7.5.2 MCR Using Geometric Quantile Mapped Heterogeneous Data

The predictor variables from the asynchronous data can be transformed to the have the same geometric distribution characteristics of the given synchronous predictor variables data using methods such as a geometric quantile mapping (GQM) and covariance alignment. The MCR approach that used Geometric quantile mapping on the predictor variables having asynchronous data, is referred to as MCR_{GQ}.

Figure 7.11 compares the CDF of predicted daily minimum temperature at a weather station (Eau Claire)in Michigan, 1990-99, with and without geometric quantile mapping the asynchronous predictor variables to match the synchronous predictor variables. Using geometric quantile mapping the asynchronous predictor variables to match the synchronous

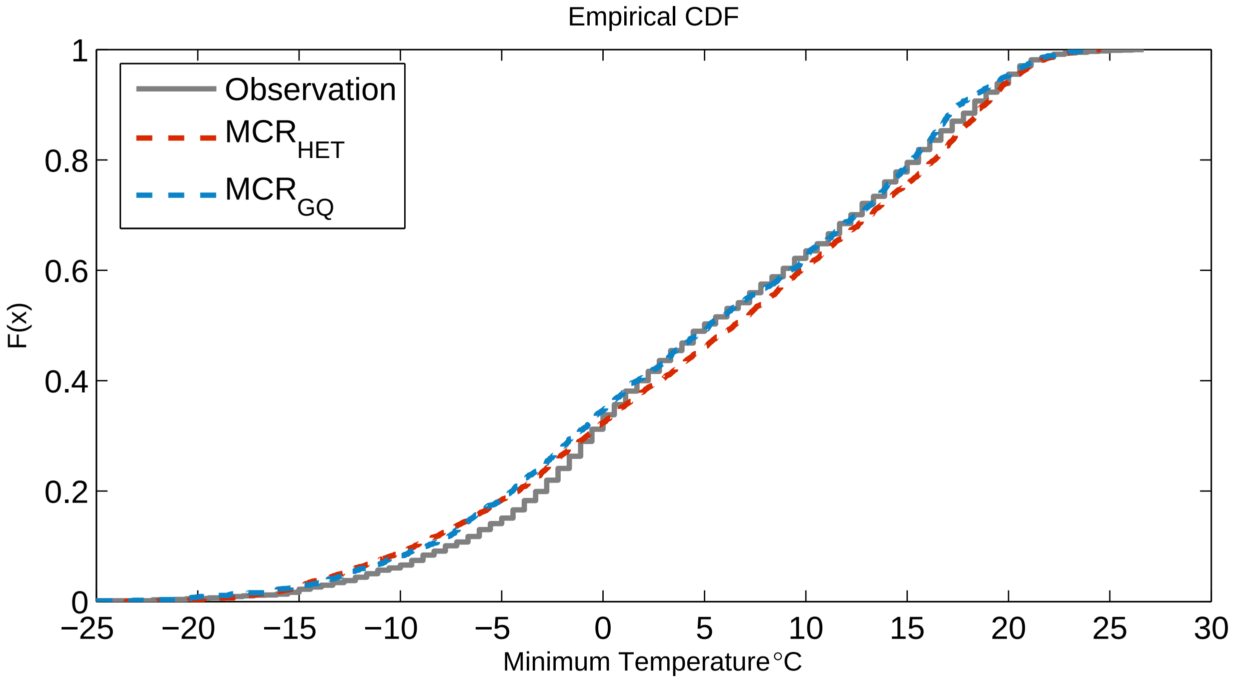


Figure 7.11: Comparing the cumulative distribution function of MCR_{HET} and MCR_{GQ} output of predicted daily minimum temperature at a weather station in Michigan, 1990-99, using asynchronous regional climate model data.

predictor variables prior to applying GQM showed marginal improvement. Similarly, results were also seen in the case of the CDF of predicted daily precipitation at the same weather station in Michigan, 1990-99, with and without geometric quantile mapping the asynchronous predictor variables to match the synchronous predictor variables (Figure 7.12).

7.5.3 Projections For The Years 2040-2049

Figure 7.13 and Figure 7.14 shows the projected distribution of the climate variables for the years 2040-2049, for the same weather station in Michigan.

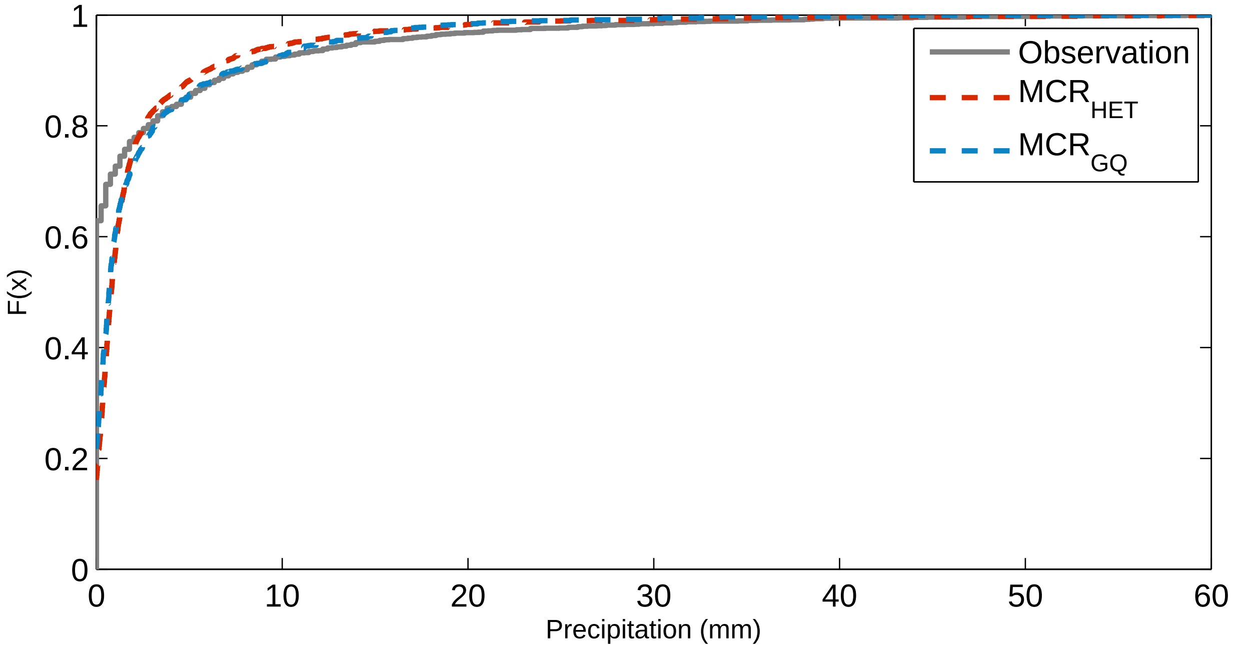


Figure 7.12: Comparing the cumulative distribution function of MCR_{HET} and MCR_{GQ} output of predicted daily precipitation at a weather station in Michigan, 1990-99, using asynchronous regional climate model data.

7.6 Conclusions

This chapter presents 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. 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, is demonstrated. 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.

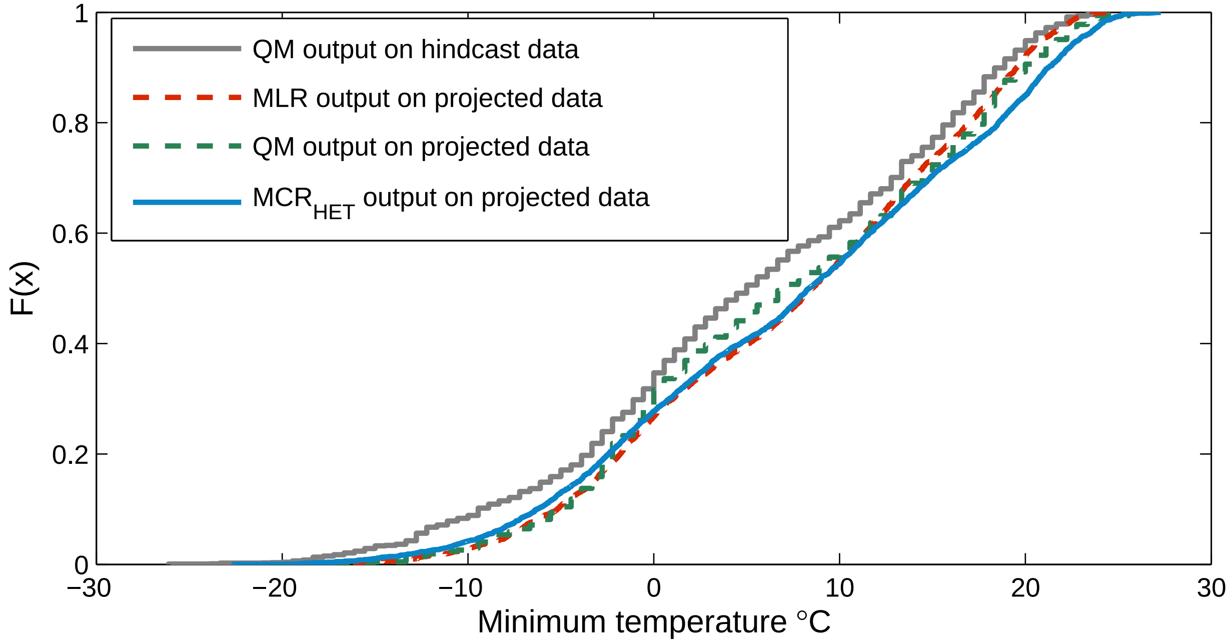


Figure 7.13: Comparing the cumulative distribution function of projected daily minimum temperature at a weather station in Michigan, for the period 2040-2049 to that of QM model output for the years 1990-99

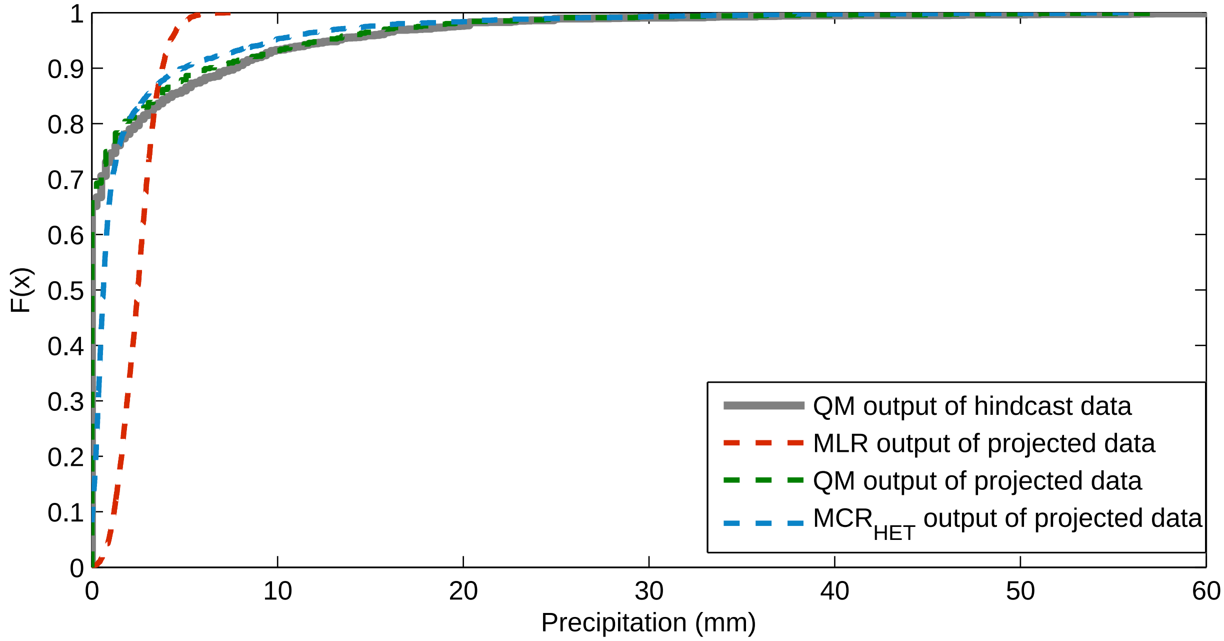


Figure 7.14: Comparing the cumulative distribution function of projected daily precipitation at a weather station in Michigan, for the period 2040-2049 to that of QM model output for the years 1990-99

Chapter 8

Multivariate Contour Regression

This chapter presents a framework for multiple output regression that extends the single output regression framework presented in Chapter 7. The framework is motivated by the growing demand for multiple output prediction methods capable of both minimizing residual errors and capturing the joint distribution of the response variables in a realistic and consistent fashion. The multiple output regression presented in this chapter, preserves the relationships among the response variables (including possible non-linear associations) while minimizing the residual errors of prediction by coupling regression methods with geometric quantile mapping.

8.1 Introduction

Multiple output regression (MOR) is the task of inferring the joint values of multiple response variables from a set of common predictor variables. The response variables are often related, though their true relationships are generally unknown *a priori*. An example application of multiple output regression is to simultaneously estimate the projected future values of temperature, precipitation, and other climate variables needed for climate change impact, adaptation and vulnerability (CCIAV) assessments. The projected values are used as the driving input variables for phenological and hydrological models to simulate the responses of the ecological system to future climate change scenarios. To ensure the projected values

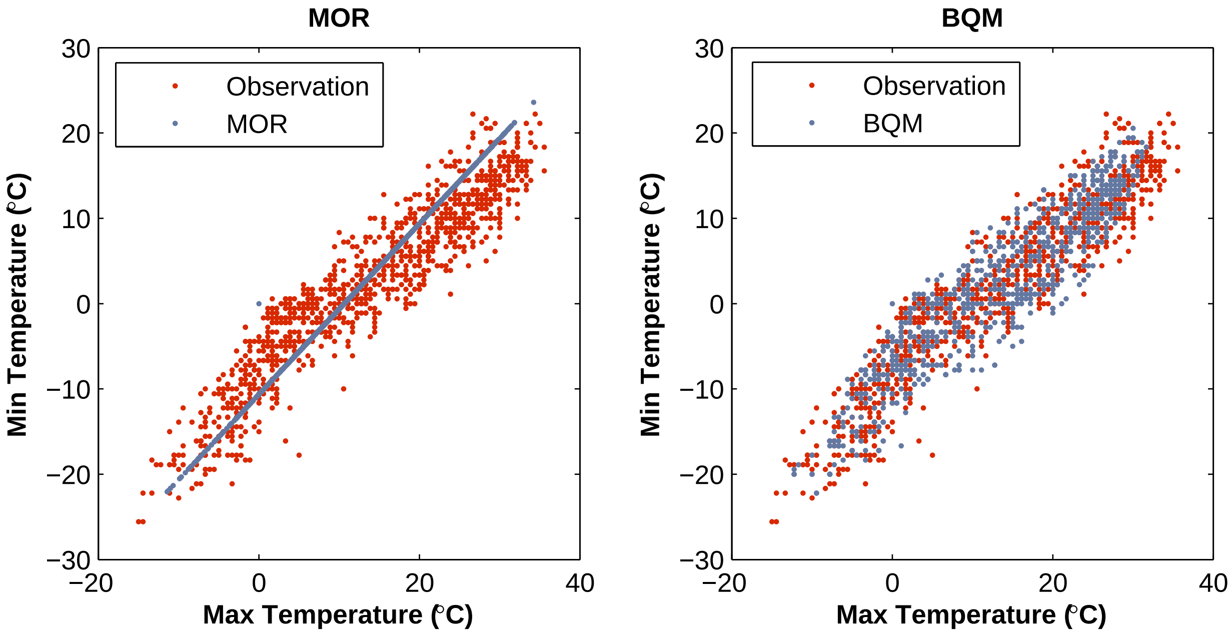


Figure 8.1: Scatter plot of observed daily maximum and minimum temperature at a climate station in Michigan, USA.

are realistic, there are certain constraints on the relationship among the response variables that must be preserved; e.g., minimum temperature must not exceed maximum temperature or liquid precipitation should be zero when temperature is below freezing. While there have been numerous multiple output regression methods developed in recent years [25, 112, 12, 95, 69], most of them are focused on fitting the conditional mean or preserving covariance structure of the outputs. Such methods do not adequately capture the full range of variability in the joint output distribution, as illustrated in Figure 8.1(a).

The inability of standard regression-based approaches to reproduce the shape of the true distribution of output variables, even for univariate response variables, is well-documented [5]. Univariate *distribution-driven approaches* such as quantile mapping (QM) [108] and statistical asynchronous regression (SAR) [92] have been developed to address this limitation, but the accuracy of these approaches is generally poor since they are not designed to minimize residual errors. Quantile mapping approaches map a univariate predictor variable x

to its corresponding response variable y by transforming the cumulative distribution function (CDF) of x to match that of y. More recently, a bivariate quantile mapping approach (BQM) (see Figure 8.1(b)) has been developed to generate bivariate response values that mimic the joint distribution of the observed response data [61]. However, as will be shown in this chapter, the residual error is significantly worse when compared to regression-based methods because the position and rank correlation between the predictor and response variables remain invariant under QM-based transformation, which in turn, hinders its ability to minimize residual errors. Thus, unless the predictor variable has a high rank correlation with the response variable, the residual error upon applying QM-based approaches is likely to be large.

This suggests a possible hybrid approach to improve both the residual errors and distribution fitting is by first applying a regression-based method to transform the predictor variables so that their rank correlation with respect to the response variable is high, before applying quantile mapping to adjust for the fit in distribution. However, maximizing the rank correlation of the data points is necessary but not sufficient condition for improvement in the residuals for QM, unless the response values of the data points are uniformly spaced. Hence, the need for position regularization, that would prioritize the prediction accuracy of data points whose position, when incorrectly estimated, results in high residual. The term 'position' here refers to the geometric quantile of a data point with respect to a multivariate distribution, which is analogous to the quantile of a data point in the case of univariate distribution. In this chapter, a position-regularized, multi-output prediction framework called Multi-Output Contour Regression (MCR) is presented. MCR addresses the dual objective of preserving the associations among the multiple output variables as well as minimizing residuals. MCR is able to achieve the dual objective by applying a novel, position-regularized regression method, followed by geometric quantile mapping (GQM) to improve the fit in distribution. The position-regularized regression helps to alleviate the limitation associated with the rank invariant property of QM, which contributes to the high residuals of QM-based approaches. MCR additionally addresses the challenge of ensuring that its prediction of the response variables will always abide by the constraints of the actual response data. MCR is also not limited by the number of predictor variables that may be used nor does it require them to have high correlation with the response variables, unlike quantile mapping. The flexible nature of our framework allows for the incorporation of other loss functions such as the L_1 loss used in quantile regression.

8.2 Preliminaries

Let $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_n]^T$ be an $(n \times d)$ data matrix and $\mathbf{Y} = [\mathbf{y}_1, .., \mathbf{y}_n]^T$ be the corresponding $(n \times q)$ response matrix, such that $\mathbf{x}_i \in \Re^d$ and $\mathbf{y}_i \in \Re^q$ are column vectors representing the respective values of predictor and response variables for the i^{th} data point. The objective of multi-output regression (MOR) is to learn a target function $h(\mathbf{x}, \Omega)$ that best estimates the multi-output response $\mathbf{y}$, where $\Omega = (\omega_1, .., \omega_q)$ is the parameter set of the target function.

For a univariate random variable $X \in \Re$, let $F_X(x)$ be its cumulative distribution function (CDF), i.e., $F_X(x) = P(X \leq x)$. The corresponding α-quantile of X is given by $\inf \{x \in \Re : F_X(x) \geq \alpha\}$. Intuitively, each quantile indicates the value in which a certain fraction of the data points are below it, and thus, provides a measure of its position in the data. For example, the median, which is equivalent to the 0.5-quantile, is the central location of the distribution. More generally, the position [84] of data point $\mathbf{z}$ relative to a set of points $\mathbf{Z} = (\mathbf{z}_1, .., \mathbf{z}_m)^T$ is given by

$$\mathbf{p_Z}(\mathbf{z}) = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \eta(\mathbf{z}-\mathbf{z}_i) \qquad \text{where} \quad \eta(\mathbf{w}) = \begin{cases} \frac{\mathbf{w}}{\|\mathbf{w}\|}, & \text{if } \mathbf{w} \neq \mathbf{0} \\ 0, & \text{if } \mathbf{w} = \mathbf{0} \end{cases}$$

For univariate data, the position $p_Z(z)$ is equal to $2F_Z(z)-1$, where $F_Z(z)$ is the cumulative distribution function of Z. The multi-dimensional equivalent of quantile function is geometric quantile [32].

Distribution correction methods such as quantile mapping is only applicable if one can match the position of a data point in one univariate distribution (say for x) to its corresponding position in another univariate distribution (say for y). This is possible using the preceding definition of position for univariate data since the values of p_Z are always fixed in the range between $[-1,+1]$ irrespective of the values in Z. Unfortunately, when extended to multivariate positions, the range of values for p_Z may vary depending on the values in Z. To overcome this problem, He et al. [61] introduce the notion of a stationary position by iteratively applying the following position transformation function until convergence:

$$\mathbf{p}_Y^k(\mathbf{z}) = \frac{1}{\kappa n}\sum_{i=1}^{n} \frac{\mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i)}{\parallel \mathbf{p}_Y^{k-1}(\mathbf{z}) - \mathbf{p}_Y^{k-1}(\mathbf{y}_i) \parallel}, \qquad \mathbf{p}_Y^1(\mathbf{z}) = \frac{1}{\kappa n}\sum_{i=1}^{n} \frac{\mathbf{z}-\mathbf{y}_i}{\parallel \mathbf{z}-\mathbf{y_i} \parallel} \tag{8.1}$$

Here each component in $\mathbf{y}_i$ must be converted to its marginal rank first before applying the position transformation function. Marginal rank refers to the rank of the data point divided by the largest rank and then normalized to the range $[-1,1]$. The normalization is done to negate the effect of variables having values that correspond to different ranges. Data points with normalized marginal rank close to ±1 correspond to extreme values for the particular variable, while those close to $\mathbf{0}$ are located near the median of the distribution. In practice, the number of iterations needed to reach a stationary distribution is quite small, typically $K > 5$ [61]. For univariate data, it can be shown that $\mathbf{P}^k$ reaches a stationary distribution at $k = 1$.

The term κ in Equation (8.1) is a normalization factor to ensure the distribution of the geometric positions is supported in a q-dimensional unit hypersphere. In the case of bivariate response variable $\mathbf{Y}$, the stationary geometric quantile distribution is circularly symmetric around the origin, with the radial density of $r/\sqrt{1-r^2}$ for $r \in (0,1)$ [61]. Therefore,

$$\kappa = \int_0^1 \frac{r}{\sqrt{1-r^2}} dr \Rightarrow \kappa = \frac{\pi}{4}$$

In this chapter, the position of the multivariate data points in $\mathbf{Y}$ is denoted as $\mathbf{P}_Y = [\mathbf{p}_Y(\mathbf{y}_1), .., \mathbf{p}_Y(\mathbf{y}_n)]^T$, where $\mathbf{p}_Y(\mathbf{y}_i) \in [-1,1]^q$. The notation $\mathbf{z}_{XY}(\mathbf{y}) = \mathbf{p}_X^{-1}(\mathbf{p}_Y(\mathbf{y}))$ is used to represent a point in the domain of $\mathbf{X}$ that has the same geometric quantile position as the data point $\mathbf{y}$ in $\mathbf{Y}$, i.e., $\mathbf{p}_X(\mathbf{z}_{XY}(\mathbf{y})) = \mathbf{p}_Y(\mathbf{y})$. Consequently, $\mathbf{z}_{YY}(\mathbf{y}_i) = \mathbf{y}_i$. Finally, let $\mathbf{Z}_{XY}(\mathbf{y}) = [\mathbf{z}_{XY}(\mathbf{y}_1)^T, .., \mathbf{z}_{XY}(\mathbf{y}_n)^T]^T$ be the geometric quantiles in X that correspond to the data points in Y.

8.2.1 Properties of the Geometric Quantiles

Proposition 8.2.1. *For $q = 1$, $\mathbf{P}^k$ reaches a stationary distribution at $k = 1$.*

Proof. Based on Equation 8.1, for $q = 1$, $\mathbf{P}_X^k(x)$ computes rank of the univariate variable x, scaled to the range $(-1,1)$. Since, each iteration of $\mathbf{P}_X^k(x)$ re scales the marginal rank to range $(-1,1)$, the rank is preserved. Hence, for $q = 1$, $\mathbf{P}^k$ reaches a stationary distribution at $k = 1$. ♢ □

Proposition 8.2.2. *Multivariate distributions that are movement transformations of each other have the same P^k distribution.*

Proof. Geometric quantile distribution is invariant under movement transformation, if, given $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, .., \mathbf{y}_n]$, having geometric quantile distribution $\mathbf{P}_X$ and $\mathbf{P}_Y$, $\mathbf{P}_X = \mathbf{P}_Y$. Since, $\mathbf{X}$ and $\mathbf{Y}$ are scalar transformations of each other, by definition $\exists \Delta \in \mathcal{R}^q$ s.t., $\mathbf{X} = \mathbf{Y} + \Delta$. Given, $\mathbf{z}_Y \in Y$ such that

$$\mathbf{p}^1_Y(\mathbf{z}_Y) = \frac{1}{\kappa n} \sum_{i=1}^{N} \frac{\mathbf{z}_Y - \mathbf{y}_i}{\| \mathbf{z}_Y - \mathbf{y}_i \|} \quad \Rightarrow \mathbf{p}^1_Y(\mathbf{z}_Y) = \frac{1}{\kappa n} \sum_{i=1}^{N} \frac{\mathbf{z}_X - \Delta - \mathbf{x}_i + \Delta}{\| \mathbf{z}_X - \Delta - \mathbf{x}_i + \Delta \|}$$

$\Rightarrow P^k_X = P^k_Y \forall$ k. Hence P^k is invariant under movement transformation. $\diamondsuit$ □

Proposition 8.2.3. *Multivariate distributions that are scale transformations of each other have the same P^K distribution.*

Proof. Geometric quantile distribution is invariant under scale transformation, if, given $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, .., \mathbf{y}_n]$, having geometric quantile distribution P_X and P_Y, $P_X = P_Y$. Since, X and Y are scale transformations of each other, by definition $\exists \alpha \in \mathcal{R}$ s.t., $X = \alpha Y$. Given,

$$P^1_X(z_x) = \frac{1}{\kappa n} \sum_{t \in X} \frac{z_x - X(t)}{\| z_x - X(t) \|}$$

$$\Rightarrow P^1_X(z_x) = \frac{1}{\kappa n} \sum_{t \in X} \frac{\alpha z_Y - \alpha Y(t)}{\| \alpha z_Y - \alpha Y(t) \|} \quad \Rightarrow P^1_X(z_x) = \frac{1}{\kappa n} \sum_{t \in X} \frac{\alpha(z_Y - Y(t))}{\alpha \| z_Y - Y(t) \|}$$

$\Rightarrow P_X = P_Y$, which means that P^K is invariant under scale transformation. $\diamondsuit$ □

8.2.2 Quantile Mapping-Based Approaches

Quantile mapping transforms a univariate predictor variable X to its corresponding response variable Y by adjusting the cumulative distribution function F_X to match that of F_Y:

$$QM : \hat{y} = F_Y^{-1}(F_X(x)) \tag{8.2}$$

It can be shown that QM preserves the rank correlation[1] between the variables. For instance, consider the example in Table 8.1 where $\mathbf{y}$ is the response variable and $\mathbf{x}_1$, $\mathbf{x}_2$ are two independent predictor variables. Let QM($\mathbf{x}_1$) and QM($\mathbf{x}_2$) be the corresponding QM outputs for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. If we sort the vectors in ascending order, it is easy to see that the resulting rank vectors are invariant under QM transformation. As a result, the rank correlation between $\mathbf{x}_1$ (or $\mathbf{x}_2$) and $\mathbf{y}$ is identical to the rank correlation between QM($\mathbf{x}_1$) (or QM($\mathbf{x}_2$)) and $\mathbf{y}$. Furthermore, the empirical CDF for QM($\mathbf{x}_1$) as well as QM($\mathbf{x}_2$) are identical to that for $\mathbf{y}$, i.e., $F_Y = F_{\text{QM}(\mathbf{x}_1)} = F_{\text{QM}(\mathbf{x}_2)}$.

Even though quantile mapping was able to replicate the empirical distribution of $\mathbf{y}$ perfectly, QM($\mathbf{x}_1$) has a higher residual error than QM($\mathbf{x}_2$). This can be explained by the lower rank correlation between $\mathbf{x}_1$ and $\mathbf{y}$ compared to the rank correlation between $\mathbf{x}_2$ and $\mathbf{y}$. Note that the inverse relationship between rank correlation and residual error holds only if the values of the response variable are uniformly spaced. For example, if the response value y for the fourth data point changes from 0.4 to 0.7, the residual error for QM($\mathbf{x}_2$) increases from 0.02 to 0.32, and is larger than the residual error for QM($\mathbf{x}_1$), which remains at 0.06. In this case, a high rank correlation for $\mathbf{x}_2$ does not translate to lower residual error when applying quantile mapping. A formal proof showing the relationship between rank correlation and

[1] Examples of rank correlation measures include Kendall τ and Spearman's ρ coefficients.

Table 8.1: Quantile Mapping Example

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}$	QM($\mathbf{x}_1$)	QM($\mathbf{x}_2$)
0.6	0.7	0.2	0.1	0.2
0.8	0.6	0.1	0.3	0.1
0.7	0.9	0.3	0.2	0.4
0.9	0.8	0.4	0.4	0.3
		SSR=	0.06	0.02

$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{y}$	QM($\mathbf{x}_3$)	QM($\mathbf{x}_4$)
0.7	0.6	0.2	0.2	0.1
0.6	0.7	0.1	0.1	0.2
0.9	0.8	0.3	0.7	0.3
0.8	0.9	0.7	0.3	0.7
		SSR=	0.32	0.02

residual error for uniformly spaced data is given in the next section.

Since most data sets are non-uniform, maximizing rank correlation is not a sufficient condition to ensure a low residual error. Nevertheless, the data points were observed to be associated with quantiles that are located in sparse regions (i.e., far from their next closest quantiles) will contribute to higher residual error when incorrectly ranked compared to data points associated with quantiles located in dense regions. This is demonstrated by the example shown in Table 8.1, where both $\mathbf{x}_3$ and $\mathbf{x}_4$ have the same rank correlation with respect to the response variable $\mathbf{y}$, yet have different SSR. The response values for the first three data points (0.2, 0.1, and 0.3) are closer to each other than the last data point (0.7). An incorrect ranking of the fourth data point will lead to much higher residual error compared to the first three data points. Since $\mathbf{x}_3$ ranked the fourth data point incorrectly, its residual error is larger than $\mathbf{x}_4$ even though they both have the same rank correlation. This suggests a possible heuristic for improving both rank correlation and residual error by emphasizing on data points that contribute to high residual errors in prediction if ranked incorrectly.

8.2.3 Rank Correlation and Residual Errors of Quantile Mapping

This section presents several properties of the QM approach with respect to the rank correlation and residual error of its output. First, quantile mapping was shown to preserve the rank correlation between the predictor and response variables.

Proposition 8.2.4. *Rank correlation is invariant under QM transformation if the values of the predictor and response variables in a data set are unique.*

Proof. Consider a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ that contains n points. Let $\hat{y}_i$ be the quantile mapped value for the data point with predictor variable x_i. To prove that rank correlation is invariant under QM transformation, it is sufficient to show that the rank for x_i is identical to the rank of $\hat{y}_i$ after quantile mapping. Without loss of generality, assume the data points in $\mathcal{D}$ are sorted in increasing order of their x values. Thus, the rank for data point x_i is i (since the x values are unique). Equation (8.2) can be rewritten as follows

$$F_Y(\hat{y}_i) = F_X(x_i)$$

Since $F_X(x_i) = i/n$, therefore $F_Y(\hat{y}_i) = F_X(x_i) = i/n$. Given that the response values y_i are distinct, the rank for $\hat{y}_i$ is also i. ◇ □

Next, the relationship between rank correlation and residual error of QM, for data sets with uniformly spaced response values is illustrated.

Proposition 8.2.5. *The residual error of QM is negatively proportional to the rank correlation given a data set with uniformly spaced response variable.*

Proof. Consider a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ that contains n points. Let r_i be the rank of x_i and s_i be the rank of the response value y_i. To simplify the discussion, it is assumed that the ranks in $\mathbf{r}$ and $\mathbf{s}$ are unique. Since $\mathbf{y}$ is uniformly spaced, it can be easily shown that $y_i = s_i c_1 + c_0$, where c_0 and c_1 depend only on the minimum and maximum values in $\mathbf{y}$.

Following Proposition 1, since QM preserves the rank of the data point x_i, $\hat{y}_i = r_i c_1 + c_0$. The Spearman rank correlation between $\mathbf{x}$ and $\mathbf{y}$ can be written as

$$\rho = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2 \sum_i (s_i - \bar{s})^2}}$$

which can be further simplified as $\rho = (\sum_i r_i s_i + c_3)/c_2$ where, c_2 and c_3 are constants for a fixed n. Since the QM output $\hat{\mathbf{y}}$ is a reordering of $\mathbf{y}$, its residual error can be computed as $SSR = \sum_i (y_i - \hat{y}_i)^2 = 2(\sum_i y_i^2 - \sum_i y_i \hat{y}_i)$. Furthermore, $\sum_i y_i \hat{y}_i = c_1^2 \sum_i r_i s_i + c_4$. where, c_4 depends on c_0, c_1, and n. Therefore, $SSR = 2(\sum_i y_i^2 - c_1^2 c_2 \rho - c_3 - c_4)$. Since, c_2 is a non-negative constant, $c_1^2 c_2$ will always be non-negative. Hence, SSR is negatively proportional to ρ for a given data set with uniformly spaced $\mathbf{y}$. $\diamondsuit$ $\square$

Although Proposition 2 is applicable only to uniformly spaced response variable, there are other situations where the residual error of QM output can be reduced if its rank correlation increases, as will be shown in the next proposition.

Proposition 8.2.6. *Correcting the rank of x_i to match the rank of its corresponding response variable y_i, maintains, if not, improves the residual error of QM output, as long as it does not deteriorate the ranks of other data points in* $\mathbf{x}$.

Proof. Given the response variable $\mathbf{y}$, let $\hat{\mathbf{y}}$ be the QM output of predictor variable $\mathbf{x}$. The sum squared residual of QM($\mathbf{x}$) is $SSR_{\mathbf{x}} = \sum_i (y_i - \hat{y}_i)^2$ where, the residual error of the ith data point $\varepsilon_{\mathbf{x}_i} = y_i - \hat{y}_i = F_{\mathbf{Y}}^{-1}(F_{\mathbf{Y}}(y_i)) - F_{\mathbf{Y}}^{-1}(F_{\mathbf{Y}}(\hat{y}_i))$. Following Equation (8.2), we have $F_{\mathbf{Y}}(\hat{y}_i) = F_{\mathbf{X}}(x_i)$. Consequently, the residual error $\varepsilon_{\mathbf{x}_i}$ can be rewritten as $\varepsilon_{\mathbf{x}_i} = F_{\mathbf{Y}}^{-1}(F_{\mathbf{Y}}(y_i)) - F_{\mathbf{Y}}^{-1}(F_{\mathbf{X}}(x_i))$. Assuming the residual error of the QM output for $\mathbf{x}$ is non-zero, there must exist a data point x_j such that $F_{\mathbf{Y}}(y_j) \neq F_{\mathbf{X}}(x_j)$. Next, consider an *improved* vector $\mathbf{x'}$, which is a reordering of the values in vector $\mathbf{x}$ subject to the following two conditions: (1) x'_i is equal to x_i if the i^{th} data point of $\mathbf{x}$ is ranked correctly. i.e., $x'_i = x_i$ if $F_{\mathbf{X}}(x_i) = F_{\mathbf{Y}}(y_i)$. (2) If i^{th} data point of $\mathbf{x}$ is not ranked correctly, then either $x'_i = x_i$ or $x'_i = F^{-1}_{\mathbf{X}}(F_{\mathbf{Y}}(y_i))$. Note that there must be at least one data point ranked incorrectly in $\mathbf{x}$ but correctly $\mathbf{x'}$. The second condition also implies that any data point that has been reordered, it must be ranked correctly. Thus, for all the data points in condition (1), $\varepsilon^2_{\mathbf{x}_i} = \varepsilon^2_{\mathbf{x'}_i}$ since their ranks remain unchanged. On the other hand, for all the data points in condition (2), $\varepsilon^2_{\mathbf{x}_i} \geq \varepsilon^2_{\mathbf{x'}_i}$ since $F_{\mathbf{X}}(x'_i)$ is either the same as $F_{\mathbf{X}}(x_i)$ or $F_{\mathbf{Y}}(y_i)$. Therefore, $\forall i, \varepsilon^2_{\mathbf{x}_i} \geq \varepsilon^2_{\mathbf{x'}_i}$. Hence, $SSR_{\mathbf{x}} \geq SSR_{\mathbf{x'}}$. Thus, by correcting the ranks of those data points that do not have the same rank as its corresponding response value, while ensuring the ranks of all other data points remain the same, the SSR of QM output can be improved. ♢ □

Even though the above proposition suggests that one can maximize rank correlation and improve residuals simultaneously, this requires a flexible target function that allows all possible orderings of $\mathbf{x}$. For linear functions, it might not be possible to produce a reordering of values in $\mathbf{x}$ without affecting the ranks of other data points. Thus, an alternate scheme was proposed that focuses on correcting the ranks of data points associated with high residuals, which is explained in the next section.

8.3 Multi-Output Contour Regression Framework (MCR)

Since QM and regression-based approaches have their own distinct advantages which have been successfully exploited in a hybrid manner by approaches such as CR, we propose a framework that extends the intuition behind hybrid approaches that exploits the unique advantages of both QM and regression, to work in a multi-output setting. The approach uses a position regularized regression function $h(\mathbf{x}, \hat{\Omega})$ that prioritizes matching the positions of output to best match the positions of the observed response data. This step is followed by correcting the geometric quantiles of the output from the previous step to match the observed response data using the intuition of QM. This hybrid approach addresses the limitation of QM regarding the number of predictor variables that may be used as well as requirement of the predictor variables being highly correlated to the response variable. The hybrid approach was further exposed to be flexible enough to work in a multi-output setting so as to be able to capture the multi-output associations that are often ignored.

To prioritize improving the positions of the output, the proposed multi-output contour regression (MCR) framework learns the regression function $h(\mathbf{x}, \hat{\Omega})$. The regression function $h(\mathbf{x}, \hat{\Omega})$ consists of two components. The first component is similar to conventional regression loss function where the data matrix is made to regress with respect to the observed response variable. This component emphasizes minimizing residual error of the regression function.

The second component of $h(\mathbf{x}, \hat{\Omega})$ is the position regularizer that helps improve rank correlation of $h(\mathbf{x}, \Omega)$ and $\mathbf{y}$. At a first glance, one would expect the second term to be regressing on the position of the data points. Instead of regressing on the position of the data points, we regress on the geometric quantiles of the data points obtained by inverse mapping their positions to the output response space. This is done so that the position regularizer assigns a larger penalty to those data points whose position when incorrectly estimated, results in a larger minimum residual errors. To accomplish this, the data matrix is made to regress on $\mathbf{z}_{\hat{Y}Y}$, where,

$$\hat{\mathbf{z}}_{\hat{Y}Y}(y) = \mathbf{p}_{\hat{Y}}^{-k}(\mathbf{p}_{Y}^{k}(\mathbf{y})) \tag{8.3}$$

is the geometric quantile value in the $h(\mathbf{x}, \hat{\Omega})$ regression output space that corresponds to the position of the observed response variable y.

The regression function of MCR is shown in Equation (8.4),

$$\min_{\Omega} \sum_{i=1}^{n} (\gamma \mathcal{L}(h(\mathbf{x}_i, \Omega), \mathbf{y}_i) + (1-\gamma)\mathcal{L}(h(\mathbf{x}_i, \Omega), \mathbf{z}_{\hat{Y}Y})) \tag{8.4}$$

where $0 \leq \gamma \leq 1$ is a user defined parameter that may be used for either prioritizing fidelity of regression accuracy or its position correlation.

$\mathcal{L}$ can be any generic loss function such as ordinary least square (that multiple linear regression adopts), or quantile mapping (if certain quantiles are to be prioritized overs others, such as in the case of a heavy tail distribution).For instance, when the loss function $\mathcal{L}$ is ordinary least square, Equation 8.4 takes the form

$$\min_{\Omega} \sum_{j=1}^{q} \sum_{i=1}^{n} (\gamma(\mathbf{x}_i^T \Omega_j - \mathbf{y}_i)^2 + (1-\gamma)(\mathbf{x}_i^T \Omega_j - \mathbf{z}_{\hat{Y}Y})^2)$$

which corresponds to the following matrix form

$$\hat{\Omega} = \arg\min_{\Omega} \; tr(\gamma(\mathbf{X}\Omega - \mathbf{Y})^T(\mathbf{X}\Omega - \mathbf{Y}) + (\mathbf{X}\Omega - \mathbf{Z}_{\hat{Y}Y})^T(\mathbf{X}\Omega - \mathbf{Z}_{\hat{Y}Y}))$$

The regression parameters $\hat{\Omega}$ is learnt in an iterative manner. At each iteration, the regression output space from the previous iteration is used to compute $\mathbf{z}_{\hat{Y}Y}$ in the second component of the regression function $h(\mathbf{x}, \hat{\Omega})$. For the very first iteration, the regression output space is that of regular multiple linear regression.

Once $h(\mathbf{x}, \hat{\Omega})$ is learnt, the MCR prediction for a given data point $\mathbf{x}$ having corresponding observed multi-output response $\mathbf{y}$ and a regression estimation of $\hat{\mathbf{y}} = h(\mathbf{x}, \hat{\Omega})$ is obtained by inverse geometrically quantile mapping $\mathbf{p}^k_{\hat{Y}}(\hat{\mathbf{y}})$ to its corresponding value in the observed response variable space, to give the MCR prediction $\hat{\mathbf{z}}_{Y\hat{Y}}$,

$$MCR : \hat{\mathbf{z}}_{Y\hat{Y}} = \mathbf{p}_Y^{-k}(\mathbf{p}^k_{\hat{Y}}(h(\mathbf{x}, \hat{\Omega}))) \tag{8.5}$$

where, $\mathbf{p}_Y^{-k}(\mathbf{p}^k_{\hat{Y}}(\hat{y}))$ maps the stationary geometric quantile position of $h(\mathbf{x}, \hat{\Omega})$ to its corresponding data point in $\mathbf{Y}$.

To summarize, multi-output contour regression (MCR) performs multi-output regression of the predictor variables such that the position of its output is highly correlated with respect to position of the observed response variable, thereby reducing position errors of the multi-output regression results. This multivariate regression output is then mapped to its corresponding geometric quantile counterpart in the observed multi-output response space using geometric quantiles. The rationale behind using the regularized regression results, prior to performing multi-output geometric quantile mapping in MCR, is to improve on SSR by increasing the correlation among the multivariate ranks of the predictors and response variable.

8.3.1 Estimating Inverse Geometric Quantile Position

The value $\hat{\mathbf{z}}(\mathbf{p})$ that corresponds to a given geometric quantile position $\mathbf{p}$, in a multivariate distribution F_Y i.e., $\mathbf{p}_Y(\mathbf{p})$, is empirically computed by minimizing the generalized multivariate quantile loss function [32]

$$\hat{\mathbf{z}}(\mathbf{p}) = \arg\min_{\mathbf{z}\in\Re^q} \sum_{i=1}^{n}(\|\mathbf{y}_i - \mathbf{z}| + <\mathbf{p}, \mathbf{y}_i - \mathbf{z}>) \tag{8.6}$$

where, $\mathbf{p} \in \Re^q$ and $< .,. >$ denotes the Euclidean inner product. So long all the values of y_i does not fall on the same line, $\hat{\mathbf{z}}(\mathbf{p})$ will be unique for a given $\mathbf{p}$ for $q \geq 2$ [32]. Algorithms such as Newton-Raphson's method can be used to solve the above loss function geometric quantile $\hat{\mathbf{z}}(\mathbf{p})$ using the following update $\hat{\mathbf{z}} \leftarrow \hat{\mathbf{z}} - \frac{\delta}{\delta'}$ where, $\delta = \sum_{i=1}^{n}((n\kappa)\mathbf{p} - \|\mathbf{z}-\mathbf{y}_i\|^{-1}(\mathbf{z}-\mathbf{y}_i))$

$$\delta' = \textstyle\sum_{i=1}^{n} \|\mathbf{z}-\mathbf{y}_i\|^{-1}(I_q - \|\mathbf{z}-\mathbf{y}_i\|^{-2} \times (\mathbf{z}-\mathbf{y}_i)(\mathbf{z}-\mathbf{y}_i)^T)$$

For a univariate distribution, F_Y, it can be easily shown that equation (8.6) boils down to the same loss function used to identify the αth regression quantile in a linear regression setup for quantile regression [77], where $0 < \alpha < 1$ and $p = 2\alpha - 1$. i.e, $\sum_1^n(|y_i - z| + p(y_i - z))$ is minimized for z that corresponds to the αth quantile of Y.

8.3.2 Alternative Approximation-Based Approach for MCR

If one can make the assumption that given the position $(\mathbf{p})$ of a test data point $(\mathbf{y}^{test})$ that belongs to the distribution F_Y, and $\exists \mathbf{y}_i \in \mathbf{Y}$ such that $\mathbf{y}^{test} \simeq \mathbf{y}_i$, then the search space for $\hat{\mathbf{z}} = \mathbf{y}^{test}$ can be limited to data points in $\mathbf{Y}$.

Given that the search space for $\hat{\mathbf{z}}$ is finite it will not always possible to find the exact same point in F_Y using the loss function δ, as it returns a vector. Alternatively, the following range bound approximation that is equivalent to Equation 8.6, can be used to find the best

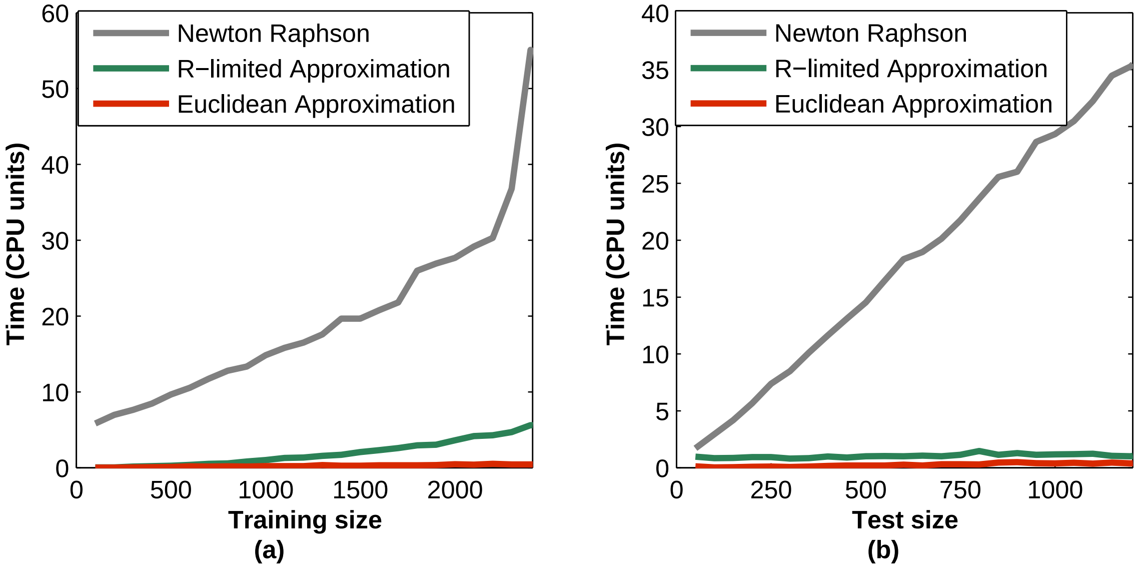


Figure 8.2: Relative computation time of the various approximation-based approaches for estimating inverse geometric quantile positions.

solution [61, 32].

$$\arg\min_{\mathbf{z}} \sum_{i=1}^{n} \{\parallel \mathbf{y}_i - \mathbf{z} \parallel + \frac{1}{\kappa}(\mathbf{y}_i - \mathbf{z})^T \mathbf{p}\} \tag{8.7}$$

where κ in the scaling factor chosen in Equation (8.4)

As shown in the experiment section, there was only a marginal performance deterioration in the solution obtained from the above approximation, due to sufficient amount of training data points. Another approximation approach with even less tighter bounds than Equation 8.7, having $O(n)$ time complexity is to use the following Euclidean approximation.

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{y}_i}((\mathbf{p} - \mathbf{p}_Y(\mathbf{y}_i))(\mathbf{p} - \mathbf{p}_Y(\mathbf{y}_i))^T) \tag{8.8}$$

The R-limited approximation approach (Equations 8.7) as well as the Euclidean approximation approach (8.8) show considerable improvement in the computation time across varying training size (Figure 8.2.a) and test size (Figure 8.2.b), with minimum deterioration in terms of accuracy of the inverse geometric quantile positions estimated.

8.4 Variations of MCR

As mentioned above, when the loss function $\mathcal{L}$ is ordinary least square, equation 8.5 takes the form

$$\min_{\Omega}\sum_{j=1}^{q}\sum_{i=1}^{N}(\gamma(\mathbf{x}_i^T\Omega_j - \mathbf{y}_i)^2 + (1-\gamma)(\mathbf{x}_i^T\Omega_j - \mathbf{z}_{\hat{Y}Y})^2)$$

which corresponds to the following matrix form

$$\hat{\Omega} = \arg\min_{\Omega} \; tr(\gamma(\mathbf{X}\Omega - \mathbf{Y})^T(\mathbf{X}\Omega - \mathbf{Y}) + (\mathbf{X}\Omega - \mathbf{Z}_{\hat{Y}Y})^T(\mathbf{X}\Omega - \mathbf{Z}_{\hat{Y}Y}))$$

The following subsection demonstrates the use of alternative loss functions in the MCR framework.

8.4.1 Quantile Multi-Output Contour Regression (MCR_Q)

An alternative to using ordinary least square loss function, which is well suited when the response variable has a non-uniform distribution such as a heavy tail, is to use quantile regression (QR) loss function for $\mathcal{L}$ so as to prioritize the rank correlation for the ranks that have the highest variance and corresponding highest impact on residual errors while quantile mapping, if incorrectly ranked [77]. Additionally, MCR_Q may also be suited to estimate the extremes of Y, by prioritizing the correct estimation of ranking extreme values.

The objective function of QR can be adopted by MCR 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,

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

8.4.2 Non-linear Multi-Output Contour Regression (MCR_{NL})

Unlike the above mentioned linear interpretations of MCR, MCR_{NL} uses a non-linear approach. By mapping the predictor variable X to a higher dimension feature space F to give ϕ , 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 MCR can be mapped to a higher dimension feature space F by using the ridge counterpart of the loss function of MCR.

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

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

$$\Rightarrow \alpha = (G + \lambda I)^{-1}(\gamma y + (1-\gamma)z)$$

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

8.5 Experimental Results

The objective of the experiments was to evaluate the ability of MCR in replicating the associations among multiple climate response variables while minimizing sum square residuals.

All the algorithms were run using climate data obtained at fourteen weather stations in Michigan, USA. The response variables used were maximum temperature, minimum temperature, and the total precipitation for each day spanning twenty years. The predictor variables used in this study are simulated climate data obtained from regional climate models (RCM) that best correspond to the observed response variables at each of the fourteen weather stations. Three different RCM data sets for each of the climate stations were obtained from North American Regional Climate Change Assessment Program (NARCCAP) [2]. 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). For the purpose of the experiments, there were a total of 126 data sets with univariate response variables, 126 data sets with bivariate responses and 42 data sets with trivariate responses.

8.5.1 Experimental Setup

Twenty year of predictor and response data, spanning the years 1980-1999 was split into two parts for training and testing. For the purpose of the evaluation of the relative skill in preserving associations among the multi-output responses, popular regression and quantile mapping approaches such as MLR, Ridge regression (Ridge), QM, EDCDFm, MOR, CR, BQM as well as ad-hoc approaches that sequently combine regression and quantile mapping approaches were used as baseline. An example of the ad-hoc baseline approach used is MOR in combination with BQM (RBQM) and MLR and QM (RQM). γ was set to 0.5 for all experiments. For CR and MCR based experiments, the maximum number of iterations was set to ten.

After discarding the missing values, each experiment run for each of the stations, across all the data sets, had a minimum of one thousand training and test data points. All the results provided in the following section are on test data (out-of-sample results). Kendall τ rank correlation and Spearman ρ rank correlation were the two rank correlation metrics used for evaluation univariate rank correlation. In the following experiment section, results of only one of the two rank correlation metrics were included, when their results were very similar. Root mean square error (RMSE), was used as a metric to compare the performance of the various approaches evaluated in terms of its output residual errors. Two dimensional and three dimensional scatter plots were used to visualize the relative skill of the various approaches in preserving the associations among the multi-output responses.

8.5.2 Results

8.5.2.1 Univariate MCR

The rank correlation of the various response variables were computed in a single output MCR setting and it was found that across all the different data sets and stations and response variables (i.e, 126 datasets), MCR consistently improved the rank correlation across both rank correlation metrics. The 126 individual data sets that corresponded to univariate response data were grouped into nine larger data sets, where each of the larger data sets were a grouping of data sets that shared the same response variable as well as the same RCM source for the predictor variables.

Figure 8.3 is a box plot representing the percentage of stations in each of the nine data

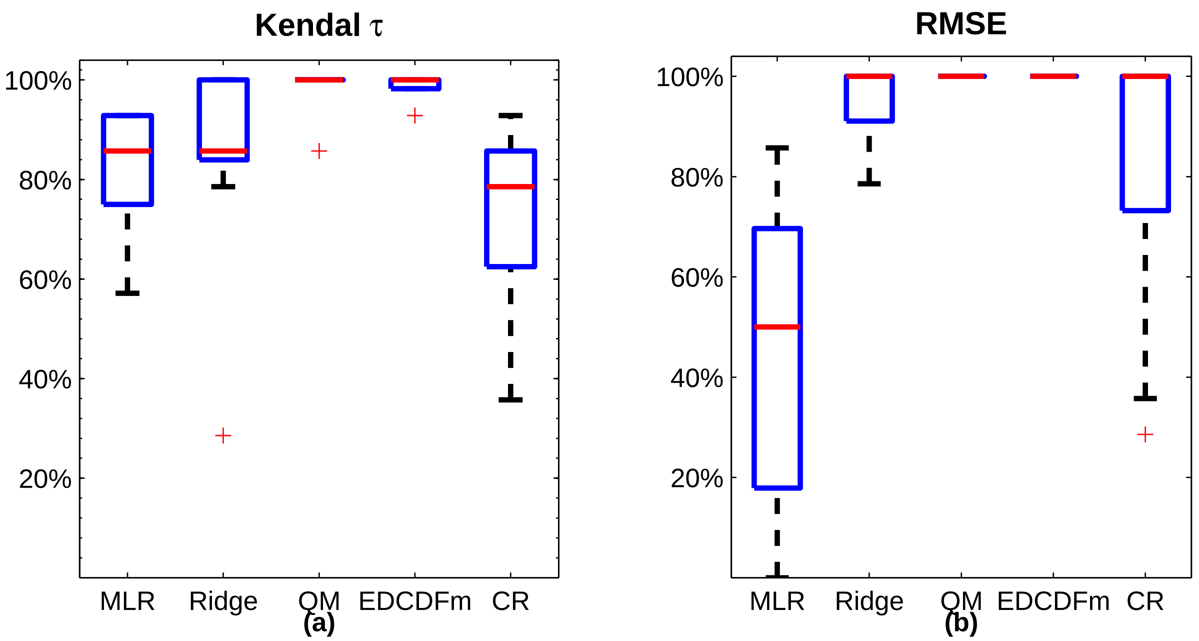


Figure 8.3: Box plot of the percentage stations where MCR showed improvement over single output baselines, in terms of Kendall τ rank correlation and RMSE, across all RCM's and variables.

sets where the rank correlation regularizer used in Equation 8.4, improved rank correlation and reduced residuals when compared to baselines approaches.

The box plot in Figure 8.4 shows that in spite of MCR's reported improvement across majority of stations in terms of τ and RMSE, for both regression and quantile mapping based approaches, the improvement was not significant when compared to the regression based approaches. However, the rank correlation regularizer showed a significant improvement in terms of RMSE at each station when compared to the corresponding quantile mapping based approaches.

8.5.2.2 Bivariate MCR

Bivariate modeling for all the combinations of bivariate response variables were evaluated. As shown in Figure 8.5, MCR performed best in replicating both the bivariate associations

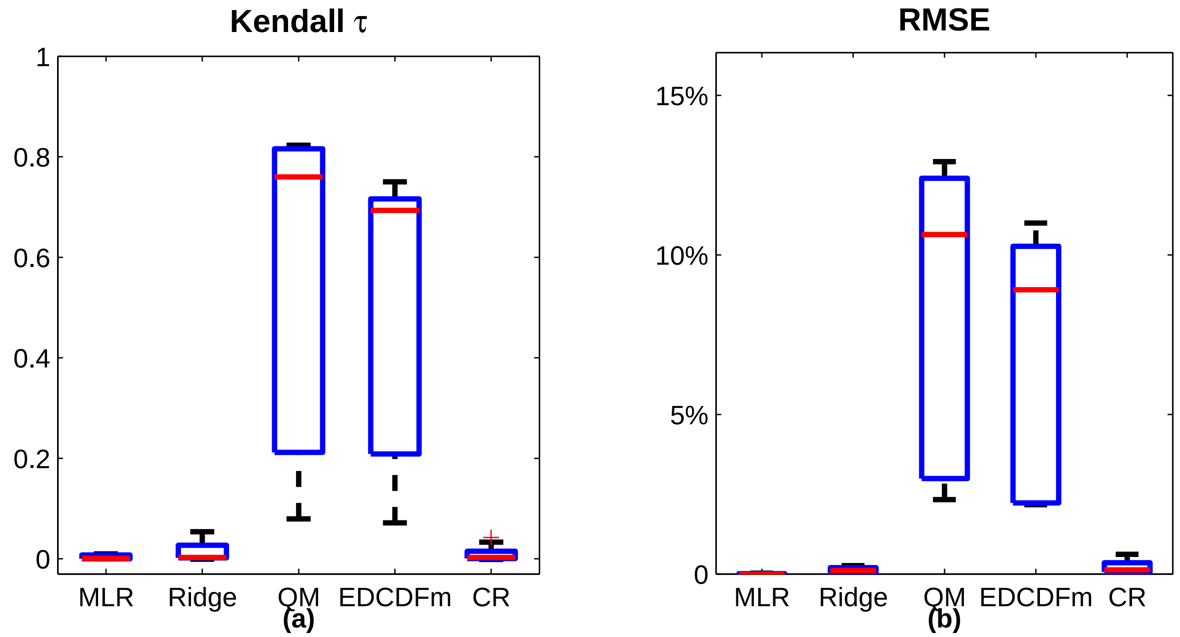


Figure 8.4: Box plot of MCR's improvement over baseline approaches in terms of Kendall τ rank correlation and RMSE, across all RCM's and variables.

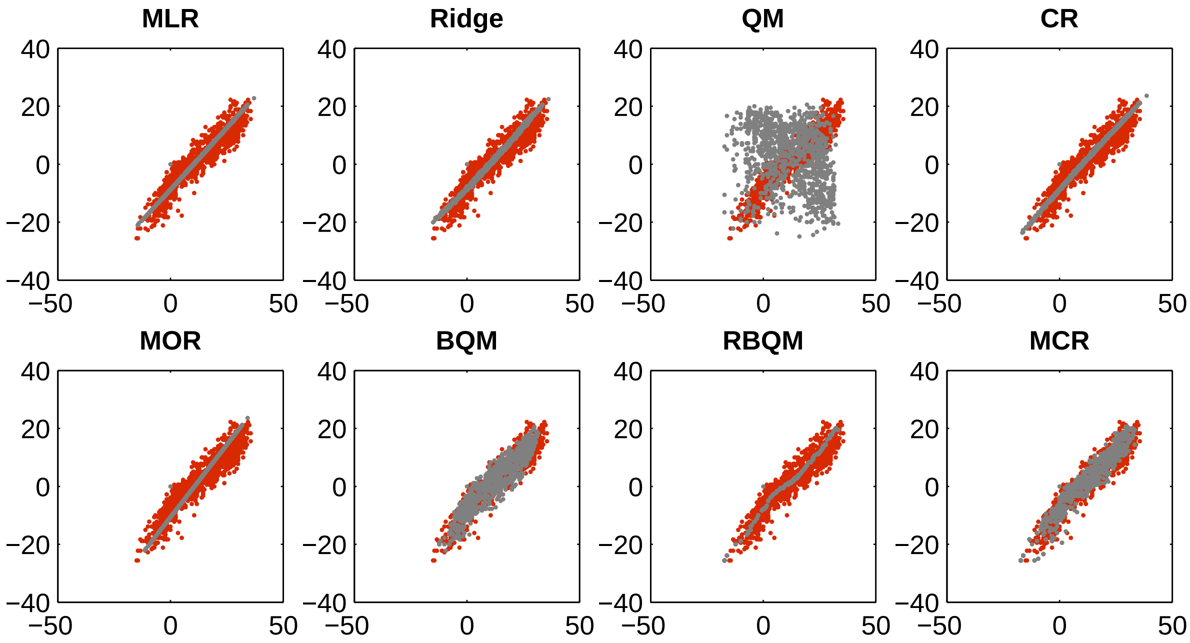


Figure 8.5: Scatter plot portraying the fidelity of forecast values of various approaches replicating the observed associations among the bivariate temperature response variables.

Table 8.2: Performance (RMSE) of bivariate MCR over baseline approaches

Data set	RMSE					
	% of stations outperformed baseline			Avg.improvement across stations over baseline		
	MOR	QM	BQM	MOR	QM	BQM
$WRFG_1$	29	100	100	-0.06	0.18	0.17
$WRFG_2$	07	100	100	-0.08	0.16	0.16
$WRFG_3$	00	100	100	-0.07	0.31	0.30
$CRCM_1$	93	100	100	0.06	0.25	0.25
$CRCM_2$	71	100	100	0.03	0.23	0.23
$CRCM_3$	07	100	100	-0.02	0.35	0.34
$RCM3_1$	43	100	100	-0.02	0.20	0.20
$RCM3_2$	36	100	100	-0.03	0.19	0.18
$RCM3_3$	00	100	100	-0.07	0.31	0.30

and minimizing SSR, although BQM performed as well in terms of replicating the bivariate associations. Regression based approaches (both SOR and OMR) fared poorly in preserving associations in the 2D space, while single output quantile mapping based approaches, were very sensitive to correlation of the predictor variables with response resulting in poor bivariate associations in spite of replicating the marginal distributions of the individual responses very well.

In terms of residuals, MCR had considerably lower residuals when compared of the various quantile mapping baseline approaches as shown in Table 8.2. But as expected, MCR showed marginal increase in residuals when compared to the respective SOR and MOR based approaches.

8.5.2.3 Trivariate MCR

The performance of modeling the association among three response variables was also evaluated and is shown in Figure 8.6. The performance is compared against single output, and multiple output models. We also use as a baseline, an trivariate extension of the bivariate

Table 8.3: Performance (Kendall τ) of bivariate MCR over baseline approaches

Data set	Kendall τ					
	% of stations outperformed baseline			Avg.improvement across stations over baseline		
	MOR	QM	BQM	MOR	QM	BQM
$WRFG_1$	64	100	100	0.03	0.40	0.41
$WRFG_2$	79	100	100	0.04	0.38	0.39
$WRFG_3$	0	100	100	-0.01	0.75	0.67
$CRCM_1$	100	100	100	0.13	0.52	0.53
$CRCM_2$	100	100	100	0.12	0.49	0.52
$CRCM_3$	14	100	100	-0.01	0.78	0.73
$RCM3_1$	79	100	100	0.06	0.46	0.46
$RCM3_2$	79	100	100	0.06	0.47	0.45
$RCM3_3$	0	100	100	-0.01	0.81	0.78

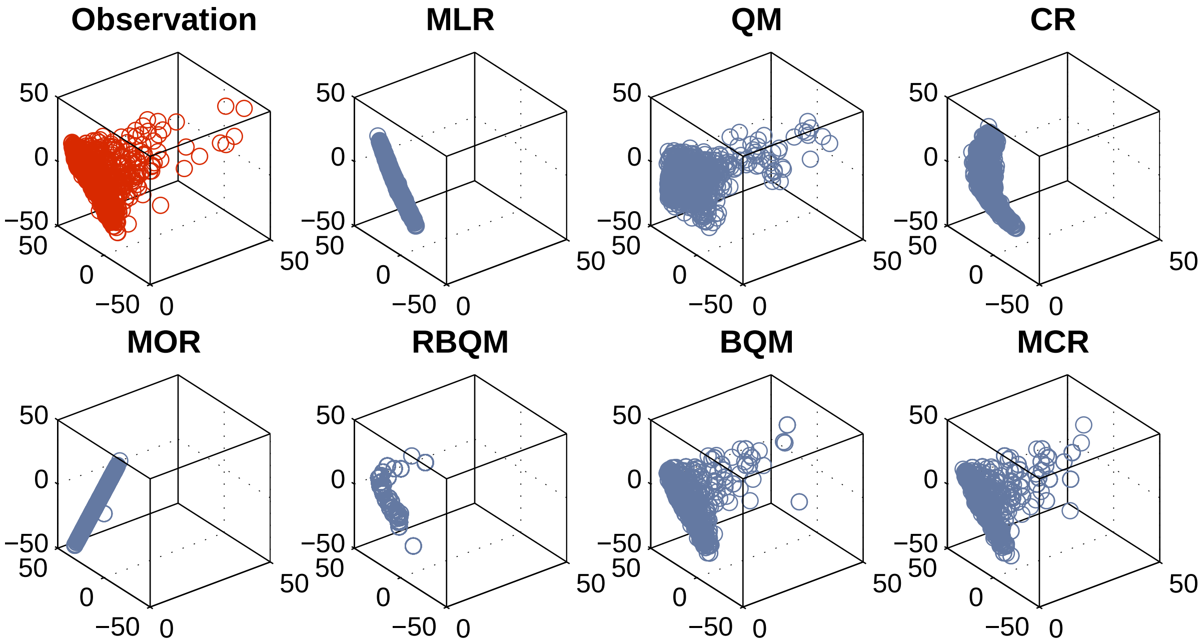


Figure 8.6: Three dimensional scatter plot of the observed associations among maximum temperature, minimum temperature and precipitation as well as the respective forecasts made by the various single output and multiple output approaches.

BQM approach, as an additional baseline. Along with MCR, the trivariate extension of BQR fared best in replicating the observed associations among three variables when compared to the baseline approaches.

Additionally, MCR was also able to improve upon its BQM counterpart in terms of reduction of residuals. MCR produced lower RMSE for all the station across all the tri-variate datasets with an average reduction of RMSE in excess of 10%. The average improvement of the three variables in terms of rank correlation τ was found to be 0.41.

8.6 Conclusions

This chapter present a multi-output regression framework extension of the single output regression framework presented in Chapter 7. The multi-output regression framework preserves the general association patterns among multiple response variables while minimizing the overall residual errors by coupling regression and geometric quantile mapping. The chapter demonstrates the effectiveness of the framework in significantly reducing residuals while preserving the joint distribution of the multi-output variables, over the baseline approaches in all the climate stations evaluated.

Chapter 9

Conclusions and Future Work

In this thesis, a number of multivariate frameworks are presented that improve upon existing regression based approaches used for generating projections, by integrating multiple objectives pertaining to the unique characteristics of response variable, as well as the expectations of a long-term projection.

The four primary multivariate frameworks, as well as its logical extensions, address the following four primary challenges. The first framework addresses the challenge of modeling response variables with irregular distribution characteristics, in particular, zero-inflated response variables. The second framework addresses the challenge of extremes in a distribution, by prioritizing the conditional quantile associated with extreme values. The third framework addresses the challenge of building a regression framework that preserves the distribution characteristics of the response variable, so as to provide an unbiased projection across all the quantiles of the distribution. The fourth framework extends the intuition behind the above-mentioned single output distribution preserving framework to an multi-output setting, such that not only is each projection unbiased, but it also maintains the relationships among multiple outputs.

Given that most of the emphasis of the evaluation of the frameworks presented in this thesis assumed a linear relationship between the predictor and response variables, a detailed evaluation of the performance, while assuming a non-linear relationship needs to be explored as well. In the frameworks presented in Chapter 3 and Chapter 4, Pearson correlation coefficient was chosen as the default similarity measure. As future work, the impact of the choice of kernel/similarity measures chosen, needs to be evaluated. Exploration of non-linear approaches to further improve the precision and recall of zero and non-zero valued data points, would provide new insights. Given the availability of numerous sources of available data, there is extensive scope for further exploiting the available heterogeneous datasets in modeling zero-inflated data.

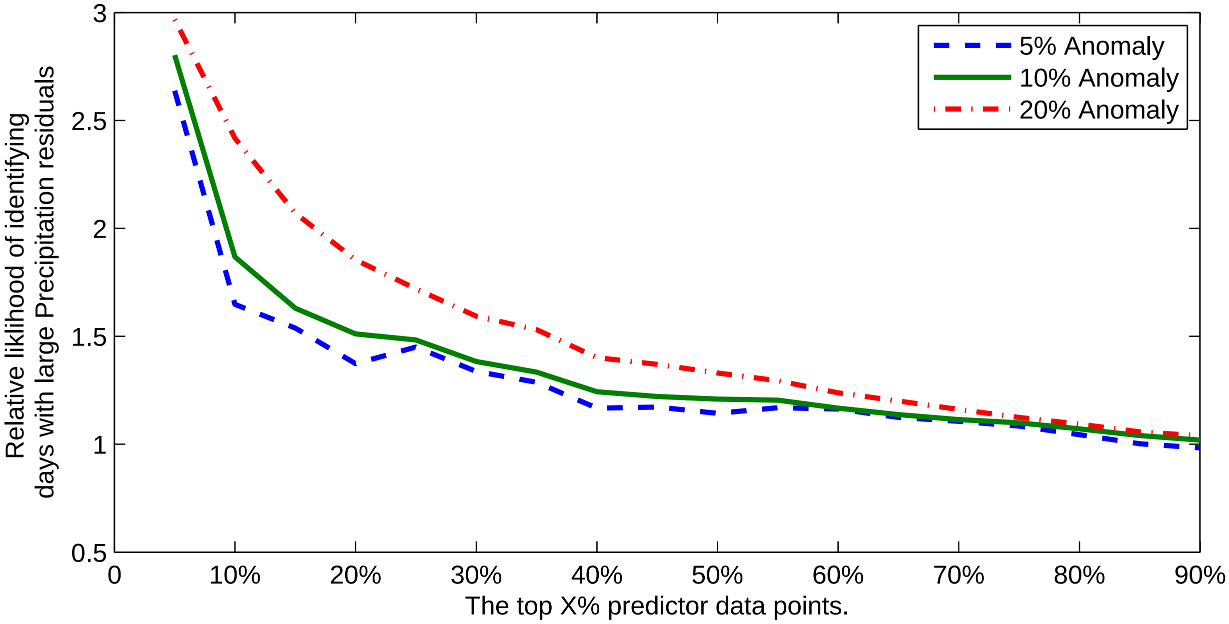


Figure 9.1: Relative likelihood of identifying large precipitation residuals

The main challenge in identifying erroneous values within the training dataset is being able to differentiate it from valid anomalies. Also, a model that cleans spurious data during model building should also be able to do so for out-of-sample data upon which the model is to be applied. Discarding data points due to a faulty value in one of the predictor variables may result in a very small training data set. This is all the more the case when the occurrence of errors across predictor variables are independent of each other and there are a large number of predictor variables– even if the percentage of error values for any individual predictor variable is relatively small. Often in many practical applications due to the scarcity of available data needed for training a model, discarding large amount of data could be unfeasible. The other drawback of dropping values that may be erroneous is that the model's response value for such data points in the test set may be extremely poor on account of not having been trained on similar data points.

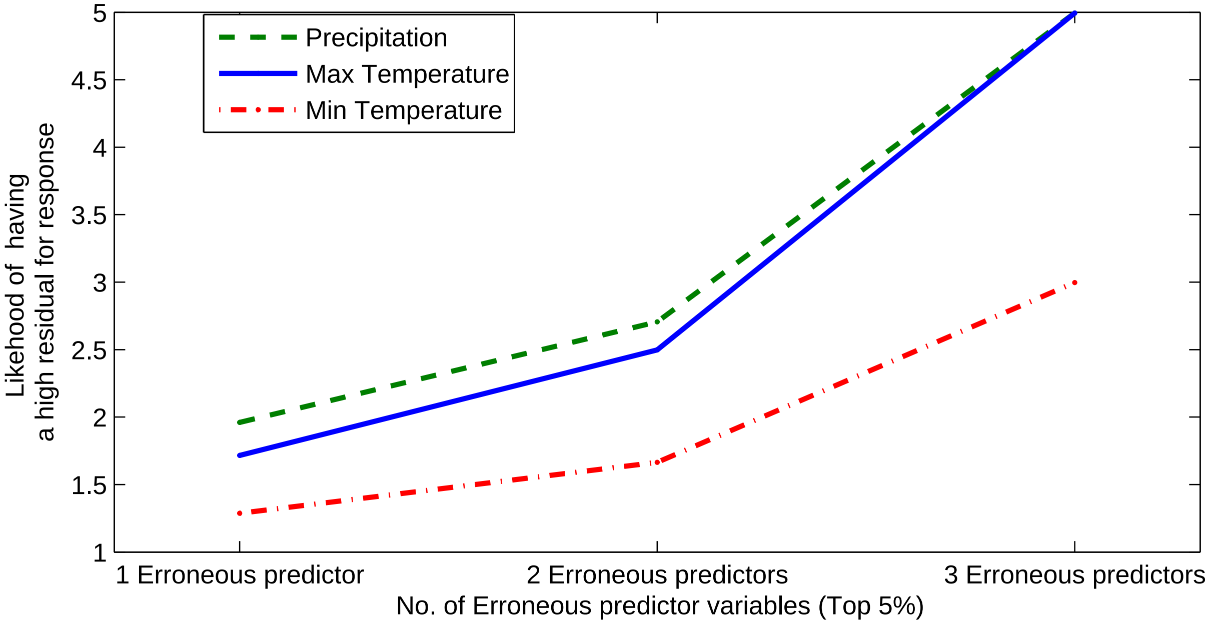


Figure 9.2: Relative likelihood of identifying days with large residuals

A possible approach for future work is therefore to build regression models in the presence of uncertain data to identify data points that have a high likelihood of being erroneous, based on its relationship with other corresponding variables, and assign weights for each data point that is a function of this uncertainty similar to weighted regression where weights are the inverse of the variance observed for the respective response variable. Unlike weighted regression, the weights chosen can be a function of the fidelity of the values for the data points.

As shown in figure 9.1, it was observed that data points whose predictor variables seemed anomalous with respect to the rest of its corresponding predictor variables were far more likely to have large residuals for the response variable. In the case of modeling precipitation using multiple linear regression, it was found that the data points that belonged to the top 5% of those likely to have an erroneous value for the precipitation predictor variable were thrice as likely to have large residuals for the response variable.

Figure 9.2 shows that the relative likelihood of a data point being one with a high residual, increases as a function of the number of predictor variables for the data point being erroneous increases.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Canadian Climate Change Scenarios Network, Environment Canada. http://www.ccsn.ca/.

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

[3] Z. Abraham and P.-N. Tan. A semi-supervised framework for simultaneous classification and regression of zero-inflated time series data with application to precipitation prediction. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 644–649. IEEE, 2009.

[4] 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)*, pages 653–664, 2010.

[5] Z. Abraham, P.-N. Tan, P. Perdinan, J. Winkler, S. Zhong, and M. Liszewska. Distribution regularized regression framework for climate modeling. In *SIAM International Conference on Data Mining (SDM)*, 2013.

[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] Z. Abraham and F. Xin. Extreme value prediction for zero-inflated data. In *Advances in Knowledge Discovery and Data Mining*, pages 318–329. Springer, 2012.

[8] Z. Abraham, F. Xin, and P.-N. Tan. Smoothed quantile regression for statistical downscaling of extreme events in climate modeling. pages 92–106, 2011.

[9] D. E. Akyuz, M. Bayazit, and B. Onoz. Markov chain models for hydrological drought characteristics. *J. Hydrometeor.*, 13:298–309, 2012.

[10] M. A. Alvarez and N. D. Lawrence. Sparse convolved multiple output gaussian processes. *arXiv preprint arXiv:0911.5107*, 2009.

[11] S. Ancelet, M.-P. Etienne, H. Benoît, and E. Parent. Modelling spatial zero-inflated continuous data with an exponentially compound poisson process. *Environmental and Ecological Statistics*, 17(3):347–376, 2010.

[12] K. Balasubramanian and G. Lebanon. The landmark selection method for multiple output prediction. *arXiv preprint arXiv:1206.6479*, 2012.

[13] S. C. Barry and A. H. Welsh. Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2):179–188, 2002.

[14] N. Bernier, K. Thompson, J. Ou, and H. Ritchie. Mapping the return periods of extreme sea levels: allowing for short sea level records, seasonality, and climate change. *Global and Planetary Change*, 57(1):139–150, 2007.

[15] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[16] J. Bjørnar Bremnes. Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review*, 132(1):338–347, 2004.

[17] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *Computer Vision–ECCV 2008*, pages 2–15. Springer, 2008.

[18] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

[19] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[20] D. Böhning, E. Dietz, and P. Schlattmann. Zero-inflated count models and their applications in public health and social science. *Applications of Latent Trait and Latent Class Models in the Social Sciences. Waxman Publishing Co*, 1997.

[21] G. Bontempi. Long term time series prediction with multi-input multi-output local learning. In *Proceedings of the 2nd European Symposium on Time Series Prediction (TSP), ESTSP08*, pages 145–154, 2008.

[22] M. Booij. Extreme daily precipitation in western europe with climate change at appropriate spatial scales. *International Journal of Climatology*, 22(1):69–85, 2002.

[23] G. E. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

[24] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd international conference on Machine learning*, pages 137–144. ACM, 2006.

[25] L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

[26] A. Buja, B. Logan, J. Reeds, and L. Shepp. Inequalities and positive-definite functions arising from a problem in multidimensional scaling. *The Annals of Statistics*, pages 406–438, 1994.

[27] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, 1994.

[28] K. N. Cahill, D. B. Lobell, C. B. Field, C. Bonfils, and K. Hayhoe. Modeling climate and climate change impacts on winegrape yields in california. *American Journal of Enology and Viticulture*, 58(3):414A–414A, 2007.

[29] A. Ceglar and L. Kajfež-Bogataj. Simulation of maize yield in current and changed climatic conditions: addressing modelling uncertainties and the importance of bias correction in climate model simulations. *European Journal of Agronomy*, 37(1):83–95, 2012.

[30] M.-C. Chan, C.-C. Wong, and C.-C. Lam. Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization. In *Computing in Economics and Finance*, volume 61, 2000.

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

[32] P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

[33] 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*, pages 133–141. ACM, 2008.

[34] R. T. Clarke. Estimating time trends in gumbel-distributed data by means of generalized linear models. *Water Resources Research*, 38(7):1111, 2002.

[35] R. T. Clarke. Estimating trends in data from the weibull and a generalized extreme value distribution. *Water resources research*, 38(6):25–1, 2002.

[36] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(12):1553–1566, 2004.

[37] D. Cooley. Extreme value analysis and the study of climate change. *Climatic Change*, 97(1-2):77–83, 2009.

[38] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

[39] C. Cortes and M. Mohri. On transductive regression. *Advances in Neural Information Processing Systems*, 19:305, 2007.

[40] F. G. Cozman, I. Cohen, and M. Cirelo. Unlabeled data can degrade classification performance of generative classifiers. In *FLAIRS Conference*, pages 327–331, 2002.

[41] F. G. Cozman, I. Cohen, M. C. Cirelo, et al. Semi-supervised learning of mixture models. In *ICML*, pages 99–106, 2003.

[42] L. J. D. Erdman and A. Sinko. Zero-inflated poisson and zero-inflated negative binomial models using the countreg procedure. In *SAS Global Forum*, pages 1–11, 2008.

[43] C. Dorland, R. S. Tol, and J. P. Palutikof. Vulnerability of the netherlands and northwest europe to storm damage under climate change. *Climatic change*, 43(3):513–535, 1999.

[44] C. A. Dos Santos, C. M. Neale, T. V. Rao, and B. B. da Silva. Trends in indices for extremes in daily temperature and precipitation over utah, usa. *International Journal of Climatology*, 31(12):1813–1822, 2011.

[45] 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?". *Hydrology and Earth System Sciences Discussions*, 9(4):5355–5387, 2012.

[46] W. Enke and A. Spekat. Downscaling climate model outputs into local and regional weather elements by classification and regression. *Climate Research*, 8(3):195–207, 1997.

[47] A.-C. Favre, S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée. Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(1), 2004.

[48] L. Feudale. *Large scale extreme events in surface temperature during 1950–2003 an observational and modeling study*. PhD thesis, George Mason University.

[49] 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):453–466, 2008.

[50] C. Frei, R. Schöll, S. Fukutome, J. Schmidli, and P. L. Vidale. Future change of precipitation extremes in europe: Intercomparison of scenarios from regional climate models. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 111(D6), 2006.

[51] P. Friederichs and A. Hense. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review*, 135(6):2365–2378, 2007.

[52] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression hidden markov model. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 513–516. IEEE, 2001.

[53] C. Gaetan and M. Grigoletto. A hierarchical model for the analysis of spatial rainfall extremes. *Journal of agricultural, biological, and environmental statistics*, 12(4):434–449, 2007.

[54] S. Ghosh and B. K. Mallick. A hierarchical bayesian spatio-temporal model for extreme precipitation events. *Environmetrics*, 22(2):192–204, 2011.

[55] C. L. Giles, S. Lawrence, and A. C. Tsoi. Noisy time series prediction using recurrent neural networks and grammatical inference. *Machine learning*, 44(1-2):161–183, 2001.

[56] H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of applied meteorology*, 11(8):1203–1211, 1972.

[57] A. M. Greene, A. W. Robertson, P. Smyth, and S. Triglia. Downscaling projections of indian monsoon rainfall using a non-homogeneous hidden markov model. *Quarterly Journal of the Royal Meteorological Society*, 137(655):347–359, 2011.

[58] T. Hastie, R. Tibshirani, and J. Friedman. *Linear Methods for Regression.* Springer, 2009.

[59] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

[60] K. Hayhoe, S. Sheridan, L. Kalkstein, and S. Greene. Climate change, heat waves, and mortality projections for chicago. *Journal of Great Lakes Research*, 36:65–73, 2010.

[61] X. He, Y. Yang, and J. Zhang. Bivariate downscaling with asynchronous measurements. *Journal of agricultural, biological, and environmental statistics*, 17(3):476–489, 2012.

[62] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[63] W.-C. Hong, P.-F. Pai, S.-L. Yang, and R. Theng. Highway traffic forecasting by support vector regression model with tabu search algorithms. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 1617–1621. IEEE, 2006.

[64] G. Hudson and H. Wackernagel. Mapping temperature using kriging with external drift: theory and an example from scotland. *International journal of Climatology*, 14(1):77–91, 1994.

[65] G. Huerta and B. Sansó. Time-varying models for extreme values. *Environmental and Ecological Statistics*, 14(3):285–299, 2007.

[66] Y. Hundecha, A. St-Hilaire, T. Ouarda, S. El Adlouni, and P. Gachon. A nonstationary extreme value analysis for the assessment of changes in extreme annual wind speed over the gulf of st. lawrence, canada. *Journal of Applied Meteorology and Climatology*, 47(11):2745–2759, 2008.

[67] T. Iizumi, M. Nishimori, K. Dairaku, S. A. Adachi, and M. Yokozawa. Evaluation and intercomparison of downscaled daily precipitation indices over japan in present-day climate: Strengths and weaknesses of dynamical and bias correction-type statistical downscaling methods. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 116(D1), 2011.

[68] A. V. Ines and J. W. Hansen. Bias correction of daily gcm rainfall for crop simulation studies. *Agricultural and forest meteorology*, 138(1):44–53, 2006.

[69] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[70] T. H. Jagger, J. B. Elsner, and M. A. Saunders. Forecasting us insured hurricane losses. *Climate extremes and society*, pages 189–208, 2008.

[71] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

[72] R. W. Katz. Statistics of extremes in climate change. *Climatic Change*, 100(1):71–76, 2010.

[73] B. Kedem and K. Fokianos. *Regression models for time series analysis*, volume 488. John Wiley & Sons, 2005.

[74] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. 2010.

[75] P. Kirshen, K. Knee, and M. Ruth. Climate change and coastal flooding in metro boston: impacts and adaptation strategies. *Climatic Change*, 90(4):453–473, 2008.

[76] R. Koenker. Quantile Regression Software. http://www.econ.uiuc.edu/ roger/research/rq/rq.html.

[77] R. Koenker. *Quantile Regresssion.* Wiley Online Library, 2005.

[78] Z. W. Kundzewicz, L. J. Mata, N. Arnell, P. Doll, P. Kabat, B. Jimenez, K. Miller, T. Oki, S. Zekai, I. Shiklomanov, et al. Freshwater resources and their management. 2007.

[79] K. E. Kunkel, K. Andsager, and D. R. Easterling. Long-term trends in extreme precipitation events over the conterminous united states and canada. *Journal of Climate*, 12(8):2515–2527, 1999.

[80] S. Laxman and P. Sastry. A survey of temporal data mining. *Sadhana*, 31(2):173–198, 2006.

[81] Y.-A. Le Borgne, S. Santini, and G. Bontempi. Adaptive model selection for time series prediction in wireless sensor networks. *Signal Processing*, 87(12):3010–3020, 2007.

[82] Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

[83] E. C. Mannshardt-Shamseldin, R. L. Smith, S. R. Sain, L. O. Mearns, and D. Cooley. Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *The Annals of Applied Statistics*, 4(1):484–502, 2010.

[84] J. I. Marden. Positions and qq plots. *Statistical Science*, 19(4):606–614, 2004.

[85] L. Mearns, W. Gutowski, R. Jones, L. Leung, S. McGinnis, A. Nunes, and Y. Qian. The north american regional climate change assessment program dataset. *National Center for Atmospheric Research Earth System Grid Data Portal, Boulder, CO*, 2007.

[86] L. O. Mearns, W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes, and Y. Qian. A regional climate change assessment program for north america. *Eos, Transactions American Geophysical Union*, 90(36):311, 2009.

[87] N. Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006.

[88] C. Monteleoni, G. A. Schmidt, S. Saroha, and E. Asplund. Tracking climate models. *Statistical Analysis and Data Mining*, 4(4):372–392, 2011.

[89] S. Nadarajah. Extremes of daily rainfall in west central florida. *Climatic change*, 69(2-3):325–342, 2005.

[90] M. Nogaj, P. Yiou, S. Parey, F. Malek, and P. Naveau. Amplitude and frequency of temperature extremes over the north atlantic region. *Geophysical Research Letters*, 33(10), 2006.

[91] A. Ober-Sundermeier and H. Zackor. Prediction of congestion due to road works on freeways. In *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, pages 240–244. IEEE, 2001.

[92] T. O'Brien, D. Sornette, and R. McPherron. Statistical asynchronous regression: Determining the relationship between two quantities that are not measured simultaneously. *Journal of geophysical research*, 106(A7):13247–13, 2001.

[93] S. Ollinger, C. Goodale, K. Hayhoe, and J. Jenkins. Potential effects of climate change and rising co2 on ecosystem processes in northeastern us forests. *Mitigation and Adaptation Strategies for Global Change*, 13(5-6):467–485, 2008.

[94] J. Ospina Norena, C. Gay Garcia, A. Conde, V. Magaña, and G. SÁNCHEZ TORRES ESQUEDA. Vulnerability of water resources in the face of potential climate change: generation of hydroelectric power in colombia. *Atmósfera*, 22(3):229–252, 2009.

[95] M. Pérez-Enciso and M. Tenenhaus. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach. *Human genetics*, 112(5-6):581–592, 2003.

[96] C. Piani, J. Haerter, and E. Coppola. Statistical bias correction for daily precipitation in regional climate models over europe. *Theoretical and Applied Climatology*, 99(1-2):187–192, 2010.

[97] 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. *Journal of Hydrology*, 395(3):199–215, 2010.

[98] M. Rivington, D. Miller, K. Matthews, G. Russell, G. Bellocchi, and K. Buchan. Downscaling regional climate model estimates of daily precipitation, temperature and solar radiation data. *Climate Research*, 35(3):181, 2007.

[99] E. A. Rosenberg, P. W. Keys, D. B. Booth, D. Hartley, J. Burkey, A. C. Steinemann, and D. P. Lettenmaier. Precipitation extremes and the impacts of climate change on stormwater infrastructure in washington state. *Climatic Change*, 102(1-2):319–349, 2010.

[100] M. Rummukainen and S. meteorologiska och hydrologiska institut. *Methods for Statistical Downscaling of GCM Simulations*. SMHI RMK. Swedish Meteorological and Hydrological Institute, 1997.

[101] M. Rusticucci and B. Tencer. Observed changes in return values of annual temperature extremes over argentina. *Journal of Climate*, 21(21):5455–5467, 2008.

[102] R. Samuels, A. Rimmer, A. Hartmann, S. Krichak, and P. Alpert. Climate change impacts on jordan river flow: downscaling application from a regional climate model. *Journal of Hydrometeorology*, 11(4):860–879, 2010.

[103] H. Sang and A. E. Gelfand. Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426, 2009.

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

[105] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[106] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264, 2006.

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

[108] J. M. Themeßl, A. Gobiet, and A. Leuprecht. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10):1530–1544, 2011.

[109] Q. Tian, J. Yu, Q. Xue, and N. Sebe. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 1019–1022. IEEE, 2004.

[110] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[111] E. Towler, B. Rajagopalan, E. Gilleland, R. S. Summers, D. Yates, and R. W. Katz. Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, 46(11), 2010.

[112] E. Vazquez and E. Walter. Multi-output support vector regression. In *13th IFAC Symposium on System Identification*, pages 1820–1825. Citeseer, 2003.

[113] I. Watterson and M. Dix. Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution. *Journal of geophysical research*, 108(D13):4379, 2003.

[114] A. H. Welsh, R. B. Cunningham, C. Donnelly, and D. B. Lindenmayer. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88(1):297–308, 1996.

[115] R. Wilby. Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices. *Climate Research*, 10(3):163–178, 1998.

[116] 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.

[117] R. L. Wilby and T. Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548, 1997.

[118] T. Zhang and F. Oles. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, pages 1191–1198. Citeseer, 2000.

[119] Z.-H. Zhou and M. Li. Semi-supervised regression with co-training. In *IJCAI*, pages 908–916, 2005.

[120] X. Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2:3, 2006.

[121] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.

[122] X. Zhu and A. B. Goldberg. Kernel regression with order preferences. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 681. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.