Extreme Value Prediction for Zero-Inflated Data

Fan Xin[1] and Zubin Abraham[2]

[1] Department of Statistics, Michigan State University
[2] Department of Computer Science & Engineering, Michigan State University
{fanxin,abraha84}@msu.edu

Abstract. Depending on the domain, there may be significant ramifications associated with the occurrence of an extreme event (for e.g., the occurrence of a flood from a climatological perspective). However, due to the relative low occurrence rate of extreme events, the accurate prediction of extreme values is a challenging endeavor. When it comes to zero-inflated time series, standard regression methods such as multiple linear regression and generalized linear models, which emphasize estimating the conditional expected value, are not best suited for inferring extreme values. And so is the case when the the conditional distribution of the data does not conform to the parametric distribution assumed by the regression model. This paper presents a coupled classification and regression framework that focuses on reliable prediction of extreme value events in a zero-inflated time series. The framework was evaluated by applying it on a real-world problem of statistical downscaling of precipitation for the purpose of climate impact assessment studies. The results suggest that the proposed framework is capable of detecting the timing and magnitude of extreme precipitation events effectively compared with several baseline methods.

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.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 318–329, 2012.

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 [13], 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 paper 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 our proposed framework outperforms the baselines for majority of the weather stations investigated in this study.

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

- We compare and analyze 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.
- We present a approach optimized for modeling zero-inflated data that outperforms the baseline methods in predicting the value of extreme data points.
- We successfully demonstrated the proposed approach to the real-world problem of downscaling precipitation climate data with application to climate impact assessment studies.

2 Related Work

The motivation behind the presented model is accurately predicting extreme values in the presence of zero-inflated data. Previous studies have shown that

additional precautions must be taken to ensure that the excess zeros do not lead to poor fits [2] 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).

Generally, simple modeling of zero values may not be sufficient, especially in the case of zero-inflated climate data such as that of precipitation where extreme value observations, (that could indicate floods, droughts, etc) need to be accurately modeled. Due to the significance of extreme values in climatology and the increasing trend in extreme precipitation events over the past few decades, a lot of work needs to be done in analysing the trends in precipitation, temperature, etc., for regions in United states, Canada, among others [3]. Katz et al. introduces the common approaches used in climate change research, especially with regard to extreme values[4].

The common approaches to modeling extreme events are based on general extreme value theory [5], Pareto distribution [10], generalized linear modeling [6], hierarchical Bayesian approaches [9], etc. Gumbel [8] and Weibull [12] are the more common variants of general extreme value distribution used. There are also Bayesian models [11] that try augmenting the model with spatial information. Watterson et al. propose a model that also deals with the skewness of non-zero data/intermittency of precipitation using gamma distribution to interpret changes in precipitation extremes [7]. In contrast, the framework presented in this paper handles the intermittency of the data by coupling a logistic regression classifier to the quantile regression part of the model.

3 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 $\tau \in \{n+1, \cdots, n+m\}$, our 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.

3.1 Generalized Linear Model(GLM) 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 our 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 us to use Poisson distribution or an exponential distribution to describe the data. Hence, in our experiments we always choose $\log(\cdot)$ as the link function and choose to use Poisson distribution. We scale the Y used in the regression model to be $10 \times Y$:

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

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

The histogram in Figure 1 is a representation of the data belonging to station-1. It is clear that the number of zero is too large. The second histogram which is without zero looks similar to a kind of Poisson or exponential distribution.

Considering the large number of zeros, one is motivated to perform classification first to eliminate the zero values before any regression. There are many

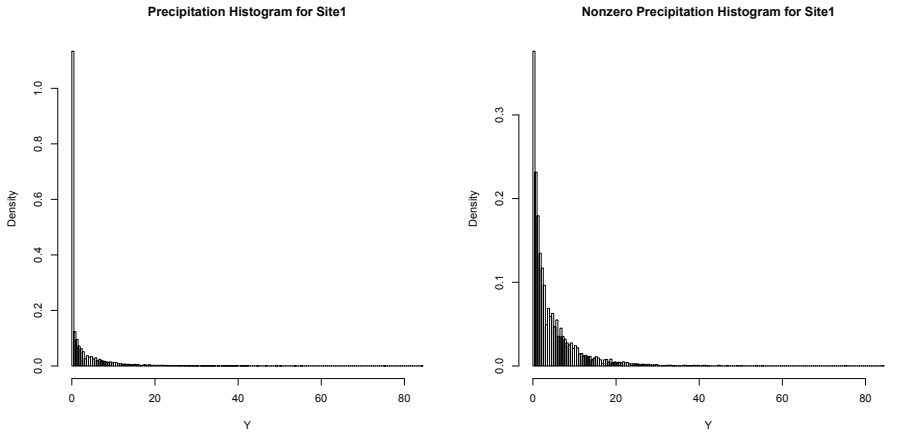

Fig. 1. Comparison of the histogram of the original distribution of data at Station-1 with its truncated counterpart

classification methods available. But for the purpose of our experiments, we use logistic regression (which is also a variation of GLM) 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 we derive the fitted values, 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, we got fitted-value $\mathbf{f'}$ for all X_i
Finally, we report the product of those two fitted-values $\hat{\mathbf{Y}} = \mathbf{f}^* \times \mathbf{f'}$

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

3.2 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, we used the R package 'pscl' to fit the model.

3.3 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:

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

where

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

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 $E\rho_\tau(y - \hat{y})$ 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, we always used $\tau = 0.95$ 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.

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

For the same reason as mentioned in the Section 3.1, a classification method should be incorporated along with the regression model. We used logistic regression for classification, and quantile regression on those nonzero Y. Finally, we report the product of those two fitted values. Quantile regression may return a negative value, which we force to 0. We do this because precipitation is always non-negative.

4 Framework for Integrated Classification and Regression

Now that we have introduced quantile regression, which is an integral part of our objective function we will elaborate the motivation behind the various components of the proposed objective function. 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, our framework consists of both components. For the classifier component we use least square support vector machine and for the regression component, we use the intuition of quantile regression 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.

4.1 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:

$$\begin{aligned} \arg\min_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} L(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) = & \frac{1}{n}\sum_{i=1}^{n}(1-(2y_i-1)f_i)^2 \\ & + \frac{1}{n^*}\sum_{i=1}^{n} y_i \rho_\tau(y_i' - f_i' \times (f_i+1)/2) + \lambda(||\boldsymbol{\omega}_1||^2 + ||\boldsymbol{\omega}_2||^2) \end{aligned} \tag{1}$$

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

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

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 $\boldsymbol{\omega}_2$.

We consider each data point 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. (1995). 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 our model includes a quantile regression component, which is not differentiable, this method of optimization is well suited to our objective function.

To solve the objective function, we used the inverse logistic function of $\mathbf{x}_i^T \boldsymbol{\omega}_2$ instead of $\text{sign}((\mathbf{x}_i^T \boldsymbol{\omega}_2 + 1)/2))$. 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, we use the binary-fitted values from the SVM component.

5 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, we analyzed the behavior of baseline models and contrasted them with ICRE in terms of relative performance of the various models when applied to this real world dataset to forecast future values of precipitation.

5.1 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.

5.2 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, we used simulated values for the corresponding predictor variables obtained from H3A2a climate scenario as $\mathbf{X}_U$, while $\mathbf{X}_L$ are values obtained from NCEP. All the experiments were run for 37 stations.

5.3 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.

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 $E(Y|X) = \lambda$

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.

Quantile Regression (QR). The baseline QR refers to the regular quantile regression described earlier in the preliminary section 3

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.

Zero Inflated Poisson(ZIP). Zero Inflation Poisson model used as a baseline and is similar to the ZIP model described in Section 3.

5.4 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, which we use 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 paper, 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 will be 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.

5.5 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 .

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 paper is on extreme data points within zero-inflated data, the impact of the classifier is unclear. In this section, we compare the impact of including the classifier in modeling extreme values of zero-inflated data. We compared QR with QR-C and GCM with GCM-C and show the results in Table 1. Note that the percentage of wins for F-measure, recall, precision may not total to 100 in the case of a tie.

Table 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 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.

Comparison of ICRE to Baseline Methods. Table 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 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

Table 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

identifying extreme events, 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-estimate 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 Conclusions

This paper compare and analyze 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. For future work, we plan to extend the framework to a semi-supervised setting.

References

1. Canadian Climate Change Scenarios Network, Environment Canada, `http://www.ccsn.ca/`
2. Ancelet, S., Etienne, M.-P., Benot, H., Parent, E.: Modelling spatial zero-inflated continuous data with an exponentially compound poisson process. Environmental and Ecological Statistics (April 2009), doi:10.1007/s10651-009-0111-6
3. Kunkel, E.K., Andsager, K., Easterling, D.: Long-Term Trends in Extreme Precipitation Events over the Conterminous United States and Canada. J. Climate, 2515–2527 (1999)
4. Katz, R.: Statistics of extremes in climate change. Climatic Change, 71–76 (2010)
5. Gaetan, C., Grigoletto, M.: A hierarchical model for the analysis of spatial rainfall extremes. Journal of Agricultural, Biological, and Environmental Statistics (2007)
6. Clarke, R.T.: Estimating trends in data from the Weibull and a generalized extreme value distribution. Water Resources Research (2002)
7. Watterson, I.G., Dix, M.R.: Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution. Journal of Geophysical Research (2003)
8. Booij, M.J.: Extreme daily precipitation in Western Europe with climate change at appropriate spatial scales. International Journal of Climatology (2002)
9. Ghosh, S., Mallick, B.: A hierarchical Bayesian spatio-temporal model for extreme precipitation events. Environmetrics (2010)
10. Dorland, C., Tol, R.S.J., Palutikof, J.P.: Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change. Climatic Change, 513–535 (1999)
11. Cooley, D., Nychka, D., Naveau, P.: Bayesian spatial modeling of extreme proecipitation return levels. Journal of the American Statistical Association, 824–840 (2007)
12. Clarke, R.T.: Estimating trends in data from the Weibull and a generalized extreme value distribution. Water Resources Research (2002)
13. Wilby, R.L.: Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection. Climate Research 10, 163–178 (1998)