# Rocky Mountain River Flow Analysis

*Carly Lundgreen and Jared Clark*

*February 9, 2021*

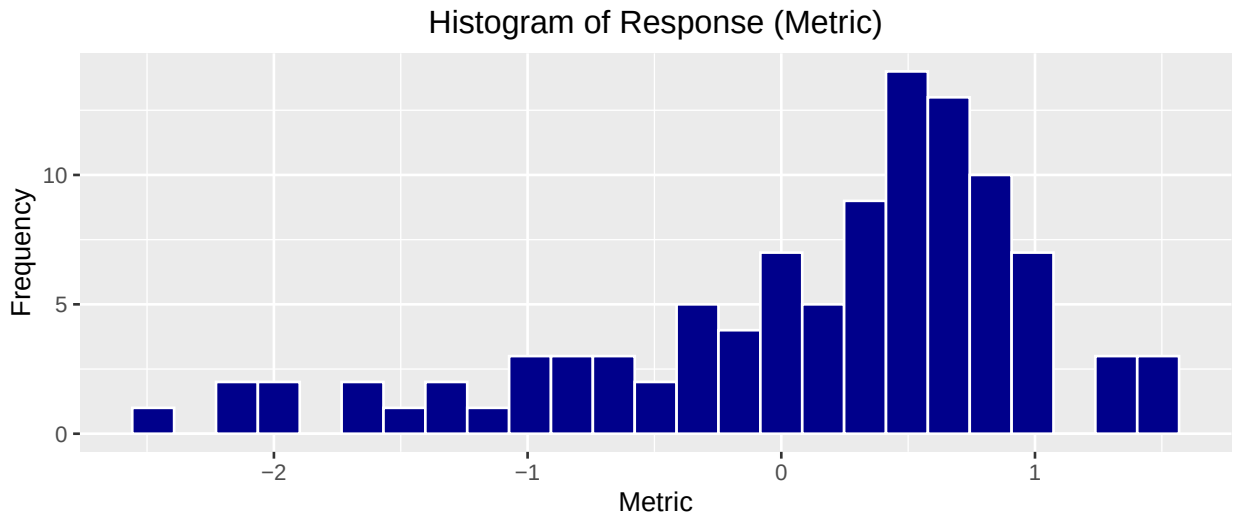
## Introduction

Rivers are extremely important to the vibrancy of nearly all ecosystems on Earth. The presence of adequate river flow is essential in order for the river to provide the maximum benefit to surrounding wilderness, crops, animals, and humans. In the Rocky Mountain region of the United States, rivers play an important part in soil fertility and animal habitats– along with being an important source of water for humans. Irrigation ditches and reservoirs from nearby rivers in this region also allow farmers to irrigate their crops. We aim to analyze and understand how various factors impact the water flow of rivers in the Rocky Mountain region. Discovering the factors with the largest effects on river flow and reporting how predictive these factors are of this overall river flow–as well as understanding how well the factors actually explain it–could influence potential policy adjustments or changes to farming techniques that may help increase (or decrease) river flow based on the needs of the region.

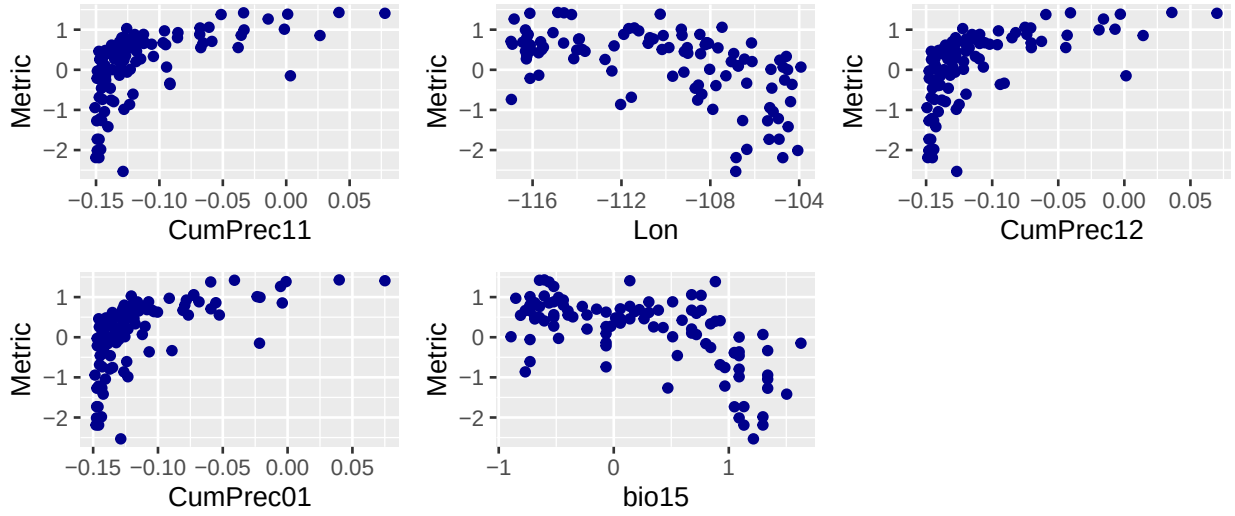
The dataset we utilized as part of this analysis contains 102 river flow measurements at various stations along rivers in the Rocky Mountain region. 98 covariates, such as latitude and longitude, average population densities for various years, landcover types, soil drainage quality, and monthly and annual precipitation measures, were also recorded. The response variable, `Metric`, is a standardized and unitless measure of river flow, with low values indicating lower overall flow and high values indicating the opposite. We recognize that there are likely spatial correlations between river stations, but we will disregard this for the purposes of this analysis.

A particular challenge with this dataset is the high dimensionality resulting from the large number of covariates. Because of this, we must consider the 'Curse of Dimensionality'–which means that as the dimension (i.e. number of columns in the data) increases, the data becomes increasingly sparse/scattered. This lack of local information in these sparse areas can lead to overfitting statistical models, which may increase the the variability of the model. This would lead to a high predictive mean square error (MSE), indicating that a model has poor predictive ability. This sparseness resulting from the large number of dimensions may also lead to 'false positives', where associations between variables are falsely assumed. We will attempt to reduce the dimension of these data without losing important covariate information. By devising a low-dimensional representation of the data, we can avoid the aforementioned problems. The two statistical methods we will utilize are Ridge regression (a form of penalized least squares) and principle component analysis. These methods will be outlined in detail in following sections.

A histogram of the response variable, `Metric`, is displayed below. The clear left-skewness indicates a larger spread among very low measurements of river flow. It appears that the majority of observations are between 0 and 1, which likely indicate 'normal' levels of flow (not too high or too low).




We have included scatterplots of five selected covariates against the response variable `Metric` below.`CumPrec01`,`CumpPrec11` and `CumPrec12` indicate, respectively, the cumulative January, November, and December precipitations for the watershed upstream of grdc station in millimeters. Longitude, along with `bio15`, which is a measure of precipation seasonality (coefficient of variation) also seem to have strong relationships with `Metric`.




We chose to display the five variables that had the strongest correlations with metric. It does not appear that the relationships between these variables and our response are very linear in nature, however, so Pearson correlations are not appropriate to report. Because of this, we have reported the five strongest distance correlations between `Metric` and all covariates in Table 2, included below. These distance correlations indicate which variables have a strong relationship with the response, whether linear or not. Interestingly, it appears that nearly all of the strongest distance correlations are the same as those given by Pearson correlations. However, the distance correlations paint a more accurate picture of the covariates' relationship with the response.

We also note that due to the large number of covariates, it would be computationally intensive to generate pairs plots for all variables, however the patterns in the plots below lead us to assume that the curvature likely appears in many other relationships between metric and the other variables. Linearity and model details will be discussed further in the following sections.

Table 1: Strongest Distance Correlations

| | Metric |
|---|---|
| CumPrec02 | 0.582 |
| CumPrec11 | 0.587 |
| CumPrec12 | 0.597 |
| bio15 | 0.614 |
| CumPrec01 | 0.617 |

## Model Specification

In our analysis of the river data, we used both Ridge regression and principle component analysis. These methods were chosen since there is a large number of explanatory variables contained in the data set. Using the multiple linear regression model would likely lead to large variances for the parameters. Essentially, the parameter estimates would be overly sensitive to the observations.

In comparing both Ridge regression and principle component analysis, we chose to focus this report on the results from the Ridge regression, since this model produced a smaller MSE and one of the major goals of this analysis is to be able to predict river flow. Ridge was chosen over LASSO since the main goal was prediction and not variable reduction. In a preliminary comparison, Ridge appeared to introduce less bias than LASSO and produced a smaller mean square error (MSE).

Principle component analysis (PCA), another method for the analysis of many predictors, has the advantage of not losing much information from the original data set, while also reducing the variability of the parameters. This is done by reducing the original predictors into a smaller set of new predictors. These new predictors are linear combinations of the original ones and are defined to be orthogonal (which avoids issues with collinearity). Just like Ridge and LASSO though, PCA introduces some level of bias.

Ridge regression is a form of penalized least squares. The response is still assumed to have a linear relationship with the predictors, however extreme values for coefficients incur a large penalty. The idea is to balance the penalty and the level of model fit in order to decrease the variance of the parameters.

The Ridge model is provided below:

$$\vec{Y} = X\ \vec{\beta} + \vec{\epsilon}$$

In this model, $\vec{Y}$ is a column vector containing the 102 river flow responses, referred to `Metric`. The X matrix is of size 102 by 96 and contains a column of ones to denote the intercept. Note that the other columns contain the explanatory information given in the data set, however, note that everything has been centered and scaled. We have also dropped two variables from the data, which will be justified in the following section.

The vector $\vec{\beta}$ contains the 96 coefficient parameters. Lastly, $\vec{\epsilon}$ is a random vector that accounts for the variability in the responses not explained by our predictors.

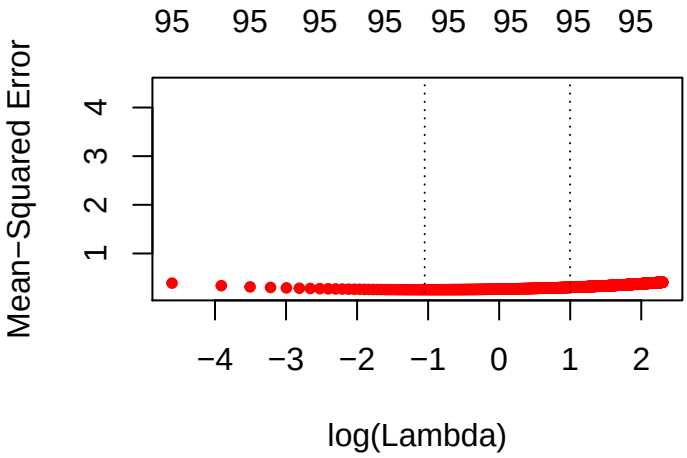
We are not making any assumptions about the distribution of $\vec{\epsilon}$. Except for ignoring the distributional assumptions on the error terms, this is similar to a multiple linear regression model, however the estimation technique is different. In estimating $\vec{\beta}$, the following algorithm is used:

$$\min_{\vec{\beta}} \sum_{i=1}^{102} (y_i - x_i'\ \vec{\beta}) + \lambda \sum_{p=1}^{95} \beta_p^2$$

Recognize that $y_i$ is the ith river flow metric response while $x_i$ is the ith row of the X matrix defined above. The shrinkage parameter $\lambda$ defines the penalty associated with the size of the coefficients. A closed form solution is provided below:

$$\vec{\beta}_{est} = (X'X + \lambda I)^{-1} X'(\vec{Y} - \bar{Y})$$

Note that 0.35 is the chosen value for the shrinkage parameter, $\lambda$, which was determined through cross validation. The following figure shows the mean-squared error (MSE) for different values of $log(\lambda)$. Note that the figure plots $log(\lambda)$ values against MSE, but we calculated $e^{log(\lambda)} = \lambda$ and reported this value. The MSE was found to be minimized at the value given above.




This model assumes that the average `Metric` measurement is a linear combination of the predictors. If this condition is not met, any results will be misleading as they are subject to unrealistic constraints. For example, prediction will not make sense as all predictions will fall on the estimated line. There is also a somewhat loose assumption of independence. Each observation is treated separately in the estimation technique. If the observations are not independent, it would make more sense to use a model that accounts for this. There is no distributional assumption on the residuals. In this way Ridge regression is extremely flexible.

## Model Justification

Using Ridge as our approach to the analysis, the model doesn't exclude any variables from the data set by necessity. The built-in penalization was our method for reducing the variance of our parameters. However, two variables associated with soil drainage were removed as they did not vary across the observations. This standard deviation of zero caused issues with our comparison model.

Any independence assumption will be assumed. There could certainly be some level of spatial correlation, but we will not explore this as part of our analysis. In the context of Ridge regression, where distributional assumptions have not been made, the possible spatial correlation does not influence the least squares estimates.

The other assumption with this model is that of linearity. Since there are still many predictors, it doesn't make sense to check for linearity in all cases. In fact with so many factors, we don't necessarily believe that we have the "correct" model. Rather, it seems reasonable to assume that the constraint of linearity will provide a useful model.

The $R^2$ term for this model is 0.78 meaning that the model is able to explain 78% of the variability in the river flow metrics. This suggests that the model fits the data well. In comparison, the $R^2$ for the PCA is 0.7378.

The model also does an adequate job with prediction. In 500 cross-validation studies, the mean root predicted mean square error was found to be 0.124. Note that the metric measurements have a range of approximately three, with an interquartile range (IQR) of about one. This indicates that the model is able to do a decent job at predicting river flow levels. As a point of comparison, the mean RPMSE for the PCA approach was about 0.59. In line with expectations, the predictions produced by the Ridge model are slightly biased. These model metrics are discussed further in the Results section.

# Results

The fitted Ridge regression model (which utilized a cross-validated shrinkage parameter of 0.35 from a training set of 85% of the data) resulted in estimates of the effects of each of the 96 factors. The factors with the five strongest effects are shown in the table below. 95% confidence interval bounds, obtained through bootstrapping, are included for each of these effects. It appears that `cls2`, or the percent of broadleaf evergreen landcover in the region, has a very large effect on river flow. Based on the results in Table 2, a 1% increase in broadleaf evergreen landcover will lead to an increase in river flow of between 37 and 165, on average. Realistically, based on the extremely small range of `cls2` values in the data, we will likely never see an increase of 1% broadleaf evergreen landcover. Therefore a 1% increase of `cls2` is an extreme amount, and would thus lead to an extreme (and likely unrealistic) increase in river flow. Note that the covariates were centered and scaled previously, but the coefficient estimates have been unscaled for better interpretation.

Other factors that strongly affect river flow include `cls10` (the percentage of snow/ice), `HydroLakes_Area_sqkm` (lake area in square km), and `cls8` (percentage of regularly flooded vegetation). `CumPrec03` and `CumPrec04` (Cumulative March and April precipitation for the watershed upstream of grdc station in millimeters) are also in the top five factors with the strongest effects.

We compared the effects from the Ridge regression model to effects (and bootstrapped confidence intervals) calculated from a PCA using 12 principle components. This number was selected after cross-validating using a training set of about 85% of the data and comparing the MSEs associated with each component. Note that the most influential factors were different betweent the two models. The only overlapping variable is the percentage of broadleaf evergreens, which was by far the variable with the strongest effect in the Ridge results.

Table 2: Ridge Results

| Variable | Effect | 2.5% | 97.5% |
|---|---|---|---|
| CumPrec03 | 0.5594 | 0.200 | 1.001 |
| CumPrec04 | 0.6534 | 0.170 | 1.038 |
| cls8 | 1.3318 | 0.358 | 2.072 |
| HydroLakes_Area_sqkm | 1.6471 | 0.052 | 3.933 |
| cls10 | 2.8047 | -4.992 | 22.164 |
| cls2 | 80.8638 | 37.634 | 165.123 |

Table 3: PCA Results

| Variable | Estimate | 2.5% | 97.5% |
|---|---|---|---|
| meanPercentDC_ModeratelyWell | -0.043 | -0.087 | -0.004 |
| cls2 | 0.055 | 0.022 | 0.091 |
| meanPercentDC_SomewhatExcessive | 0.061 | 0.028 | 0.098 |
| bio15 | -0.073 | -0.111 | -0.034 |
| strmOrder | 0.054 | 0.015 | 0.095 |
| cls7 | -0.061 | -0.100 | -0.016 |

The Ridge model resulted in an $R^2$ of 0.784, which can be interpreted as the proportion of variability in river flow that is explained by the covariates. The variables in the data do a decent job of explaining changes in river flow. The $R^2$ from the PCA, in comparison, is 0.738, which means the PCA model's covariates do a poorer job of explaining changes in river flow.

Finally, the root predictive mean square error (RPMSE) from the Ridge model, which was cross-validated on 500 studies using a training set of 85% of the data, is 0.124, which leads us to conclude that predictions from this model were off by this amount on average. The RPMSE was much smaller than the interquartile range of river flow observations, which is 0.99. This is excellent. The RPMSE for the PCA was found to be 0.59, which is a much larger proportion of the IQR. We selected the Ridge model due to the lower RPMSE.

We also note that the bias from the Ridge model is 0.472, which is greater than the bias of 0.009 from the PCA. RPMSE is more indicative of a model's predictive ability because our method, with a least squares penalization, leads to a decrease in the variance. This decrease in variance will usually compensate for any increase in bias due to the penalized least squares method. The same logic holds for dimension reduction through PCA. Therefore, we are not worried about the bias coming from the Ridge model. The Ridge model has a higher predictive ability than PCA in a direct comparison of the MSEs, so we select this model over a principle component analysis.

# Conclusion

By fitting a Ridge regression model, which was selected after comparison to a principle component analysis, we calculated the five biggest climate / river network / human factors that impact overall river flow and these effects and associated bootstrapped intervals were reported in the Results section, where the variables and their descriptions were also discussed. The factors that we reported seem like logical variables to be used in the explanation of river flow changes. For example, it makes sense that a larger lake area is associated with higher river flow levels. Similarly, the amount of rain in Spring, when rain is most likely to fall, is another logical explanatory variable for river flow levels. Larger percentages of flooded vegetation are also likely to be positively associated with river flow. It is interesting that the strongest effect on river flow comes from the percentage of broadleaf evergreen landcover in the area–these trees must be very dependent upon the flow of rivers in the Rocky Mountains.

The Ridge regression model had a relatively high $R^2$ of about 0.78, which means that the factors included in the model do relatively well in predicting river flow. Note that in our Ridge regression setting, all the covariates were included in the model, but many of them had extremely small effects, signalling tht they have little to no association with river flow. We reported the factors that had the largest effects, and therefore contributed the most to the explanation of variability in river flow measurements. We also discussed the RPMSE of the Ridge model, which was about 1/10th of our response variable's IQR–which means that the model has relatively high predictive ability.

The biased nature of the coefficient estimates in the Ridge regression model are perhaps a small concern when considering using the model for prediction. However, the reduction in variance due to the penalized least squares method likely compensates for the increase in bias that comes from the penalization. Because the Ridge model does not "zero out" coefficients, there may be a chance of the model overfitting to our data, or "overreacting" to changes in the data, more than a LASSO model would, for example. The LASSO model zeroes out coefficients that have little effect on the response, which would mean a simpler model with fewer coefficients. We did fit a LASSO model, however, and the RPMSE was a bit higher than the Ridge, which led us to select the Ridge model for the purposes of this analysis.

Further analysis of these data may include a look at collinearity between explanatory variables. A principle component analysis effectively removes any collinearity, but it may be interesting to see if variables can be excluded before even attempting to fit a model. Also, penalized least squares models may have higher variability due to the effect of collinearity. It may also be interesting to attempt a prediction of river flow at a certain location with specific characteristics to "put the model to work" and assist farmers and/or

researchers in their attempt to understand how various factors affect the river flow that is so important to the Rocky Mountain region.

**Teamwork**

Carly wrote the introduction, results, and conclusion. She also did the formatting and wrote the PCA code. Jared wrote the other sections, and did the lasso and ridge code.