Climate Change Impacts on Watershed's Streamflow

Kristo Krugger, Jiayu Peng, Te Jung Chen

April 24, 2020

1 Abstract

In order to understand the effects of climate change, it is important to study the relationship between attributes that are believed to be significant and a meaningful measure of changes in climate. One such measure is the change in streamflow of a given watershed. The goal of this analysis is to demonstrate to what extent the hydrologic behaviour (streamflow) of a given watershed is sensitive to changes in climatic attributes such as temperature, total precipitation, and the ratio of snowmelt to total precipitation. In particular, we use parametric methods to explore the relationship between streamflow and each of the climatic attributes of interest. We conclude that for the vast majority of the watershed stations under consideration, changes in streamflow are most strongly associated with changes in total precipitation.

2 Introduction

For the past century, relevant data have been collected at various watershed stations across Canada. From these data, the daily changes in a watershed's streamflow levels can be obtained. This measure serves as a quantifiable measure of interest that can be modeled as a function of climatic attributes using statistical methods. The main question to be answered is whether temporal changes in temperature and snow-to-precipitation ratio have a greater impact on changes in a watershed's streamflow than changes in total precipitation. It is also of interest to quantify this impact and rank the climatic attributes according to their association with changes in streamflow. To address these objectives, we explore the data and perform linear regression analysis.

3 Data Description

The dataset includes yearly observations for 1425 natural watershed (not affected by human) stations within Canada between the years of 1909 and 2019. The data points consist of daily measurements for streamflow, temperature, snowmelt and total precipitation, aggregated (averaged) over a year. The data contain many observations in which at least one value of the variables of interest is missing; this is mainly because only years with at least 90% of the observations (330 days in a year) for each of the variables of interest are considered meaningful, which is a standard approach in hydrology. For the purposes of our analyses, we only consider watershed stations containing at least 10 complete observations. A complete observation is defined to be a row in the dataset that contains no missing values for all of streamflow, temperature, snowmelt, and total precipitation. More details about each variable and an illustration of the dataset can be found in Table 1 and Table 2 respectively.

Table 1: Description of the variables in the dataset.

Name	**Description**	**Units**
Year	The year corresponding to the data point. A year has been specified to start in October and end in September.	None
Streamflow	Annual average daily streamflow. This means that in a given year, on average X amount of streamflow occurred during each day.	mm/day
T_{min}	Annual average daily minimum temperature.	°C
T_{max}	Annual average daily maximum temperature.	°C
Snowmelt	Annual average daily snowmelt This means that in a given year, on average X amount of snowmelt occurred during each day.	mm/day
Precipitation	Annual average daily total precipitation. This means that in a given year, on average X amount of total precipitation occurred during each day.	mm/day
Station	The station (location) corresponding to the watershed at which the measurements were taken.	None
Snow-Precip-Ratio	The ratio of Snowmelt to total precipitation.	None

Table 2: A small subset of the complete data.

Year	Streamflow	T_min	T_max	Snowmelt	Precipitation	Station	Snow_Precip_Ratio
1987	0.43	-3.30	9.58	0.34	1.21	station_1068	0.28
1988	0.22	-2.16	10.20	0.35	1.21	station_1068	0.29
1989	0.43	-3.87	8.97	0.37	1.25	station_1068	0.30
1990	0.73	-3.27	9.89	0.32	1.14	station_1068	0.28
1991	0.41	-1.94	11.17	0.39	1.18	station_1068	0.33

4 Methodology

In order to quantify effects of total precipitation, temperature, and snow-to-precipitation ratio on streamflow, a linear regression model is fit to the data. In this model, precipitation, T_{min}, and snow-to-precipitation ratio are used as explanatory variables and streamflow is used as the response variable. From this model, we extract the type III sum of squares from an ANOVA table and calculate the percentage that each of the climatic factors contributes to the total variability in streamflow that is explained by the regression model. This metric is used to rank the importance of the climatic factors. This analysis and computation is done for each station in the dataset using the car (Fox and Weisberg 2019) package in R.

5 Exploratory Analysis

In order to understand the data and formulate an appropriate model, we first explore various aspects of the dataset.

5.1 Missing Values

In this section we explore the distribution of missing values such as the distribution of missing observations (observations with at least one field missing) and the distribution of complete observations (observations with no missing fields). We also explore the time intervals for which data are missing. Figure 1 below illustrates the distributions of missing observations and complete observations.

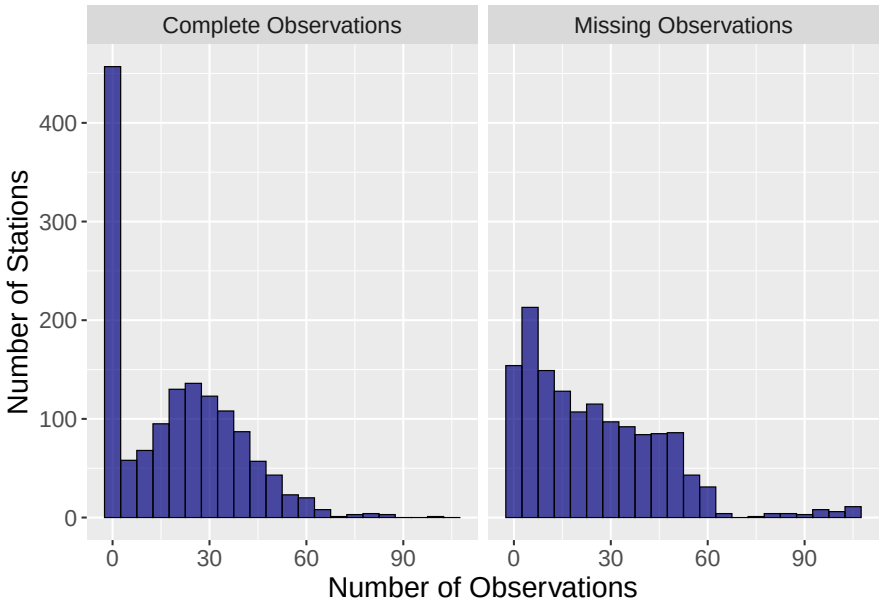

Figure 1: Distribution of the number of complete and incomplete observations.

From the figure, we can observe that there are many (just over 450 stations) stations with 0 complete observations.

Figure 2 below illustrates the distribution (over time) of complete observations for a subset of 30 randomly selected stations.

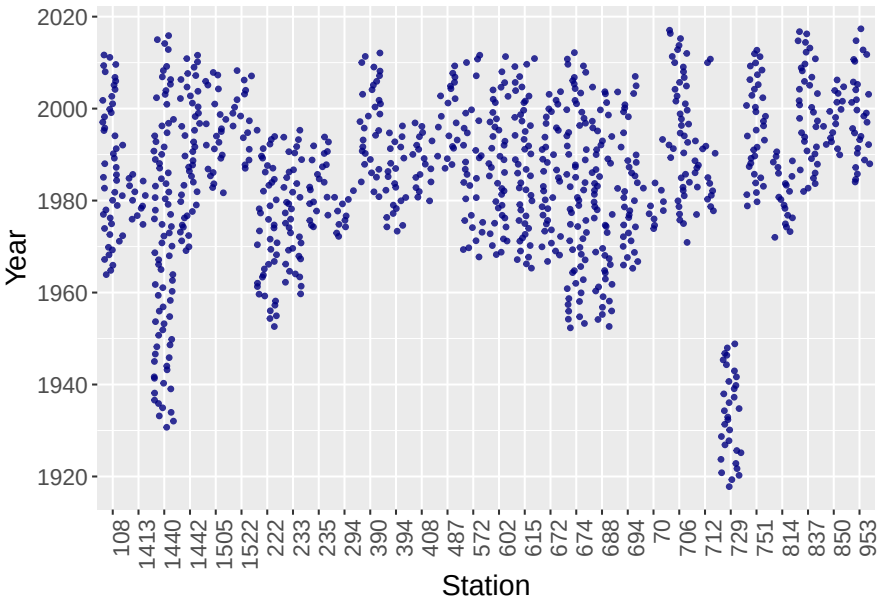

Figure 2: A jitter plot of the observations over time.

We can observe that there are some stations with missing data over a large period of time and that there are some stations that contain data for a different time interval than other stations.

5.2 Variable Relationships

Figure 3 below illustrates the pairwise correlations between each of the variables in the dataset. The figure shows a random subset of 4 stations for which these correlation matrices is calculated.

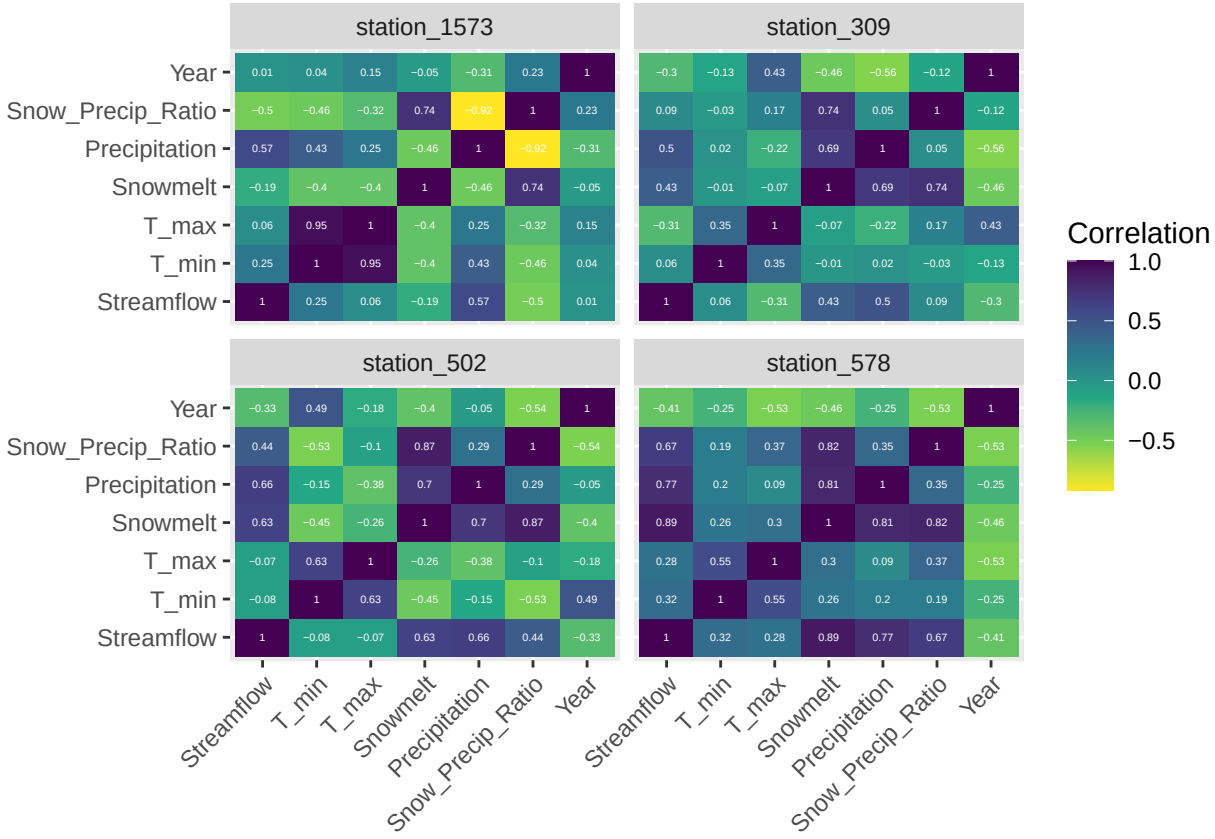

Figure 3: A heatmap for the pairwise corralations between variables.

There seems to be an inconsistency between these correlations across stations, which might indicate that the relationships between the variables vary from station to station.

Figure 4 below shows the variability in streamflow across a subset of 30 randomly selected stations.

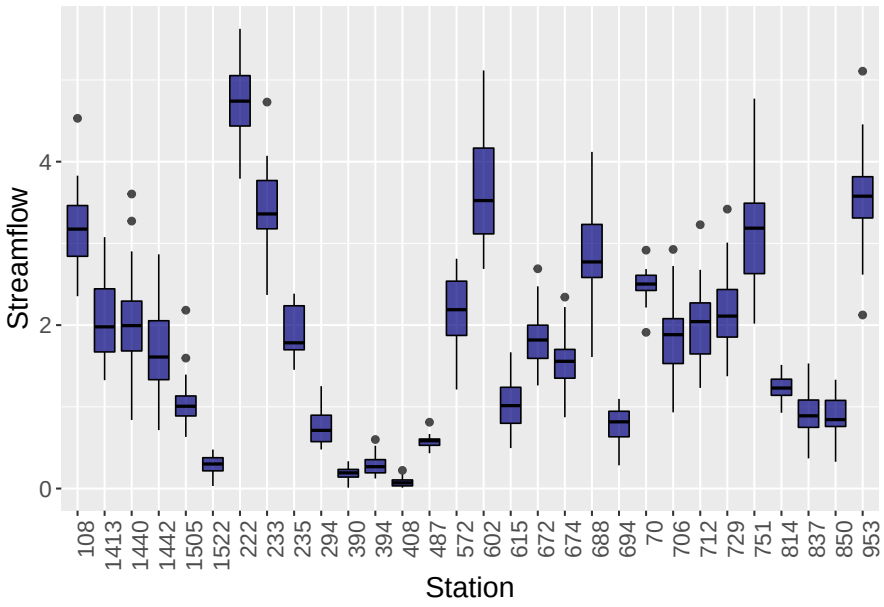

Figure 4: Streamflow variability between stations.

It is apparent that the level of streamflow seems quite different across the stations. This variability suggests that including the station effect in the linear regression model or fitting a linear regression model to each station might be a suitable approach for this problem.

Figure 5 below shows the patterns in the variables for 4 randomly selected stations.

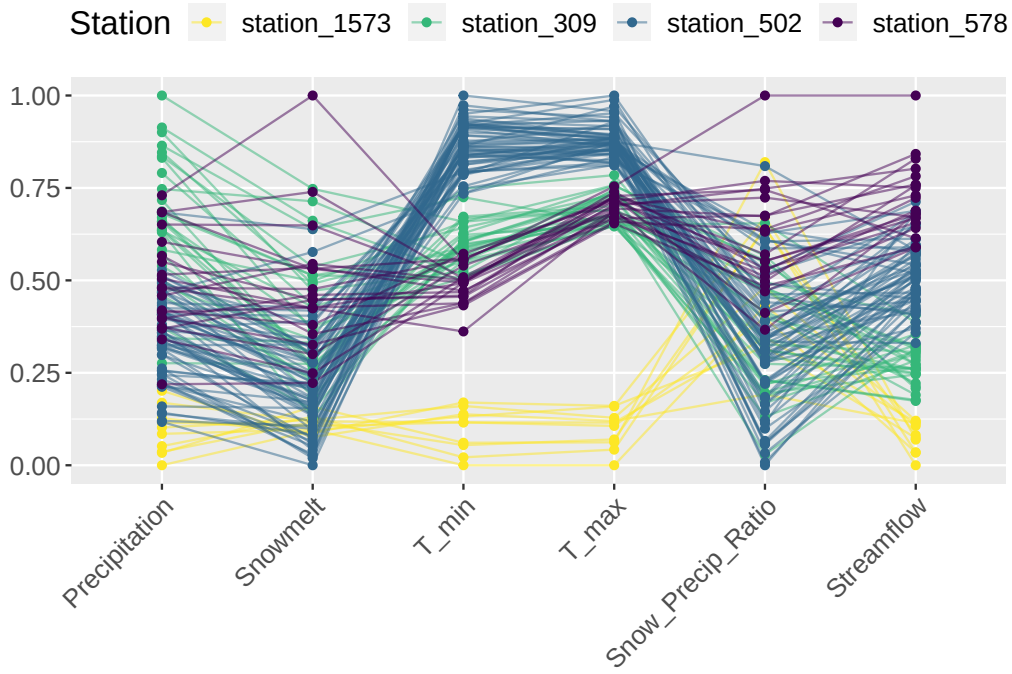

Figure 5: Parallel coordinate plot for a random subset of stations.

The plot above suggests that the variable patterns vary from station to station and that modeling the interaction between station and some of the explanatory variables or fitting a linear regression model to each

station might be appropriate.

6 Results

The comparisons of the effect of each climatic attribute is carried out using linear regression analysis. A linear regression model is fit to the data corresponding to each individual station using the lm function in R. From these models, variable importance metrics are calculated to rank the climatic attributes. The results are presented below using a single station with complete observation and are summarized for all stations using various visualizations.

Table 3: A table of the 95 percent confidence intervals for the effects of the climatic attributes.

	2.5 %	Estimate	97.5 %
T_min	-0.22	0.19	0.60
Precipitation	1.65	2.04	2.43
Snow_Precip_Ratio	-6.85	-1.96	2.94

Table 3 displays the 95% confidence intervals of the true association between Streamflow and each of Precipitation, T_{min}, and Snow_Precip_Ratio for a single station. There are no missing values in the dataset corresponding to the single station that is used to fit the linear regression model. This table indicates that we can be 95% confident that a rise of 1mm/day is associated with a rise between the range of $[1.65, 2.43]$mm/day when all other attributes holds constant. The same logic applies to the other climatic attributes.

Figures 6, 7, and 8 provide a visualization of the 95% confidence intervals of the true association between Streamflow each of Precipitation, T_{min}, and Snow_Precip_Ratio.

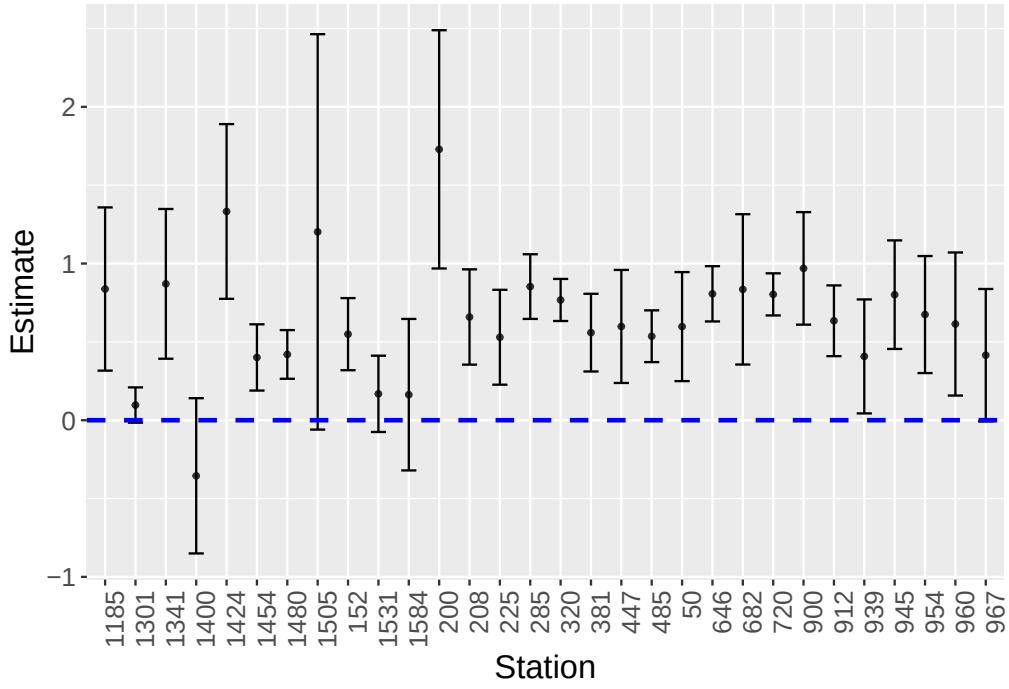

Figure 6: The 95 percent confidence interval of the effect of precipitation for 30 randomly selected stations.

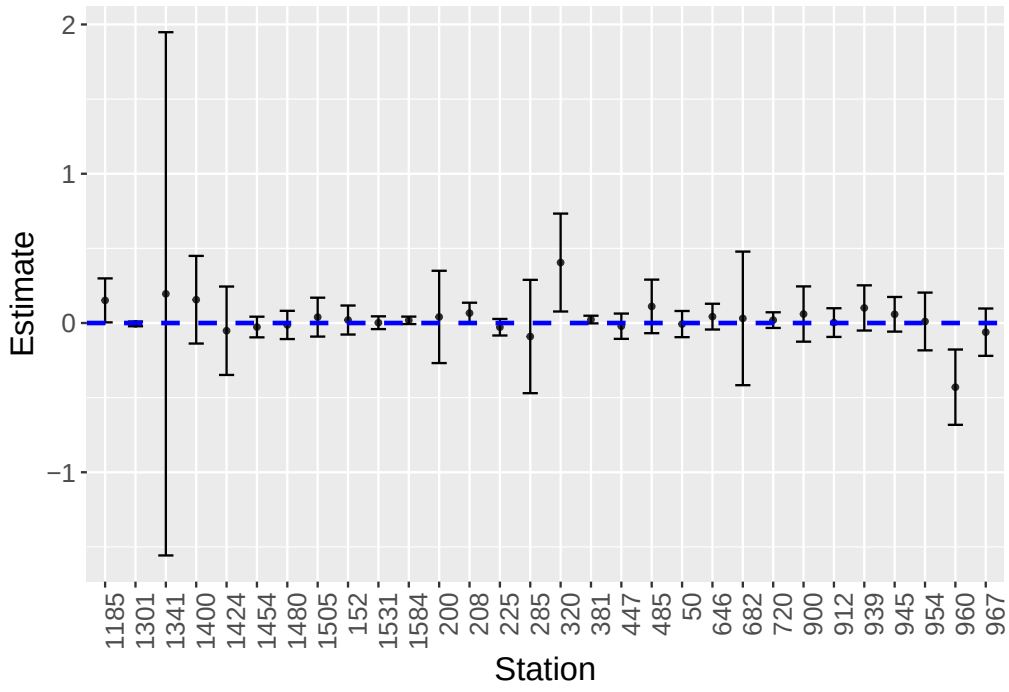

Figure 7: The 95 percent confidence interval of the effect of temperature for 30 randomly selected stations.

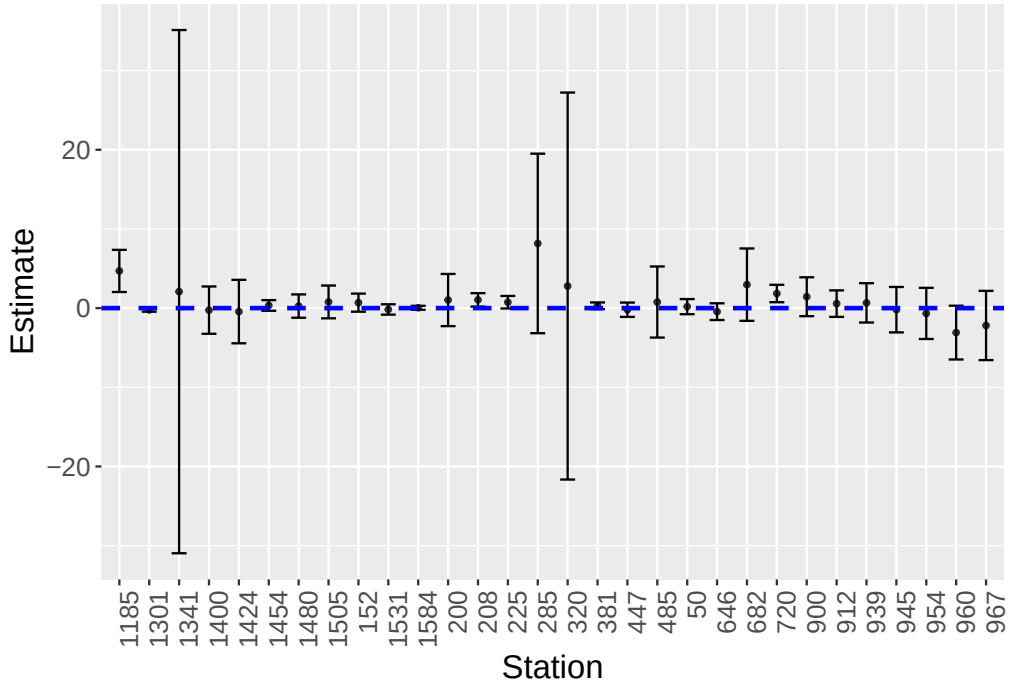

Figure 8: The 95 percent confidence interval of the effect of snow-to-precipitation ratio for 30 randomly selected stations.

The intervals above correspond to a subset of 30 randomly selected stations. From these figures, we can observe that 0 is included in the 95% confidence intervals of T_{min} and Snow_Precip_Ratio for the majority of the stations. This indicates that for the majority of the stations, it can be stated with 95% confidence

that there is no association between Streamflow and each T_{min} as well as Snow_Precip_Ratio.

Figure 9 below illustrates the most important climatic attribute for each station, based on the importance metric described in the methodology section.

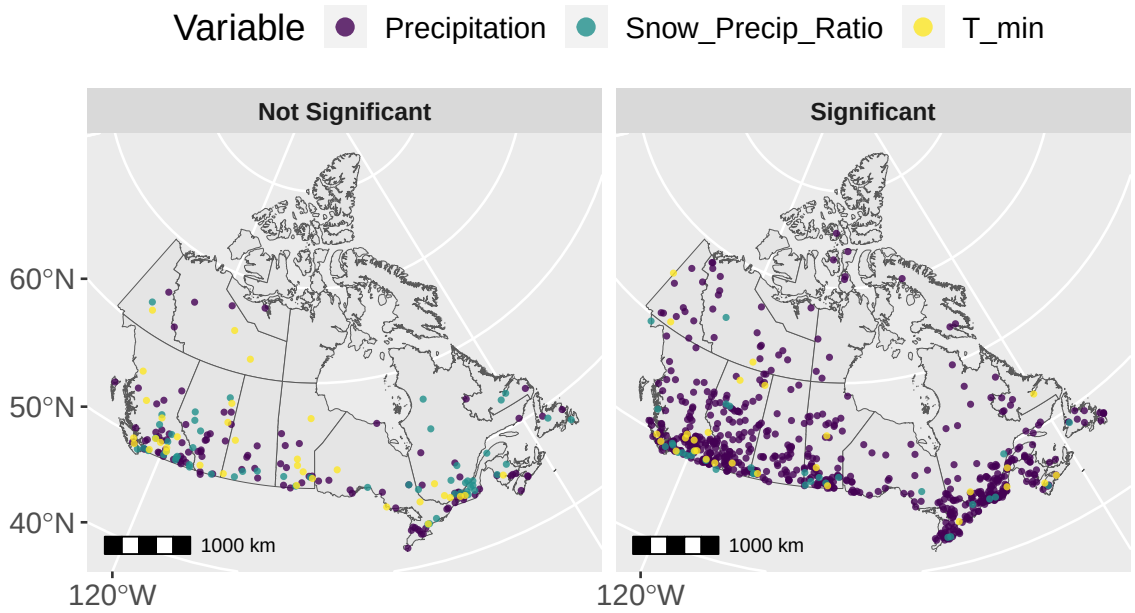

Figure 9: A map showing the most important variable for each station. Left: most important and not significant. Right: most important and significant.

The colour of the points in the plot above is used to distinguish the most important climatic attribute at each station. Both maps illustrate the climatic factor that each station considers the most important climatic factor. However, the map on the left illustrates the case where the most important factor is not considered statistically significantly different from 0 at a 5% significance level. The map on the right illustrates the case where the most important factor is considered statistically significantly different from 0 at a 5% significance level. The figure suggests that Precipitation seems to be considered the most dominant climatic attribute among the majority of the stations.

Figure 10 below depicts the number of stations that consider each of the climatic factors statistically significantly different from 0 at a 5% significance level and the number of stations that consider each of the climatic factors as the most important factor.

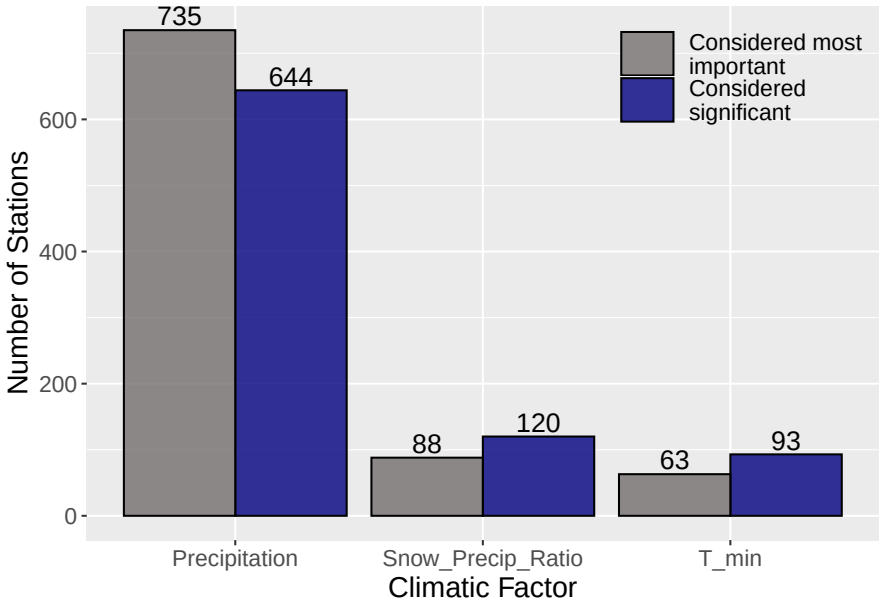

Figure 10: A bar chart of the number of stations that consider each factor most important and the number of stations that consider each factor significant.

We can observe from the bar chart above that precipitation is considered significant by the most number of stations and it is also considered the most important factor by the most number of stations.

7 Discussion

A linear regression model is fit to the data for each station with Streamflow level as the response variable and total precipitation, minimum temperature, and snow-to-precipitation ratio as input variables. Although the assumptions underlying the linear regression model are checked (see Appendix) for some of the stations, it would be extremely time consuming to check for the validity of these assumptions for every station, given the large number of stations. The main concern is the violation of independence between the observations. Although this assumption is satisfied by most of the stations that are checked, it is also violated by some of the stations. Thus the significance of some of the explanatory variables must be interpreted with some skepticism in the presence of correlation between observations.

It is also worth noting that due to the privacy policies of the data, the correct locations of the stations are not used for this analysis. For this reason, no inference is made related to the spatial location of the stations. For illustrational purposes, coordinates of hydrometric stations within Canada ("Hydrometric Station Metadata Index" 2020) are randomly sampled and assigned to watershed stations. These coordinates should be replaced with the real coordinates for further analysis (if of interest).

8 Conclusion

From the analysis of the data with respect to the inferential nature of the question in hand, it can be stated with confidence that, in general, total precipitation is the most dominant climatic attribute associated with changes in the hydrologic behaviour of watersheds across Canada. Since the analysis is done individually for each station, it can be observed from the results that this is not true for every station. However, the result seems to hold for the vast majority of the stations under consideration.

9 References

Fox, John, and Sanford Weisberg. 2019. "An R Companion to Applied Regression." Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

"Hydrometric Station Metadata Index." 2020. *Canada.ca*, February. Government of Canada. https://wateroffice.ec.gc.ca/station_metadata/station_characteristics_e.html.

Schwarz, Carl James. 2019. "Detecting Trends over Time," November. http://people.stat.sfu.ca/~cschwarz/CourseNotes/PDFbigbook-ALL/R-chapter-16.pdf.

10 Appendix

10.1 Checking Model Assumptions

10.1.1 Independent Observations

It is suspected that since the data corresponds to repeated measurements taken over time for single watersheds, the observations might not be independent. One of the main assumptions of the linear regression model is that the observations are independent from each other. If this assumption is violated, it is possible to mistakenly declare variables as significant when in fact they are not. This assumption can often be informally detected by looking at a time plot of the residuals and a lag plot of the residuals (Schwarz 2019). The lag residual plot shows the residual at time t versus the residual at time $t-1$. If there is a strong relationship between the two residuals, this indicates that autocorrelation may be present. A time plot shows the residuals over time, if the time plot shows a consistent pattern of runs of residuals above the 0 line and then below the 0 line, this is an indication that autocorrelation may be present.

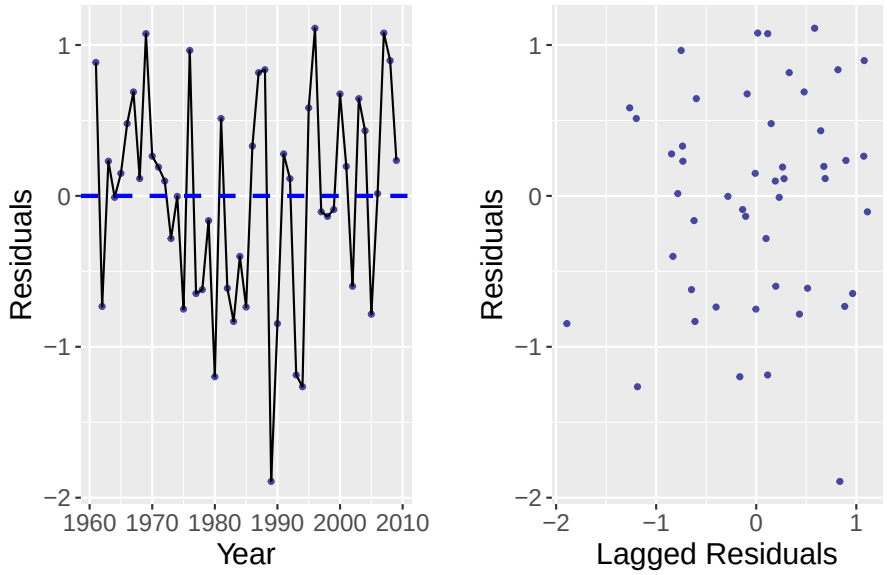

Figure 11: Diagnostic plots for autocorrelation. Left: a time plot of the residuals. Right: a lag plot of the residuals.

The time plot on the left side in figure 11 shows somewhat of a pattern in the residuals above and then below 0 line; however, this pattern does not seem too consistent. The lag plot on the right side of the figure suggests that there is not a strong relationship between the residuals and the lagged residuals.

The Durbin Watson Test can also be used to test for autocorrelation. This test tests for the absence (null hypothesis) vs the presence of autocorrelation (alternative hypothesis). A small p-value (< 0.05) indicates strong evidence of autocorrelation.

Table 4: Durbin-Watson test: `lin.model`

Test statistic	P value	Alternative hypothesis
1.791	0.4081	true autocorrelation is not 0

Table 4 suggests that there is not enough evidence to indicate the presence of autocorrelation.

10.1.2 Variance Homogeneity and Normality of Errors

Violations in the variance homogeneity assumption might result in mistakenly declaring variables as significant when in fact they are not. We can test this assumption by plotting the residuals vs the fitted value yielded by the model in question. This plot must show no clear pattern or relationship between the residuals and the fitted values.

Violations in the assumption of normality of the error terms could result in unreliable estimates of the linear regression coefficients. We can test the validity of this assumption by plotting a histogram that displays the distribution of the residuals. If the residuals seem to be approximately normally distributed around 0, the model does not violate this assumption.

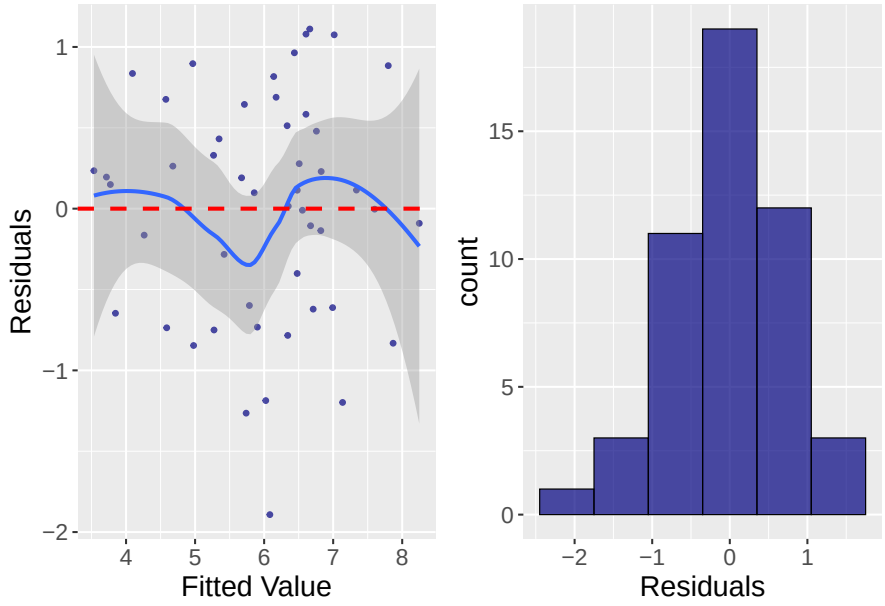

Figure 12: Left: a residual vs fitted value plot. Right: a histogram of the residuals.

10.1.3 No Multicollinearity

The presence of multicollinearity might result in mistakenly declaring variables as significant when in fact they are not. We can detect the presence of multicollinearity by calculating the VIF values for each of the explanatory variables. A value between 1 and 5 indicates that there is no presence or mild presence of multicollinearity.

Table 5 shows no signs of severe multicollinearity; all of the VIF values seem to be less than 5.

10.1.4 No Outliers

The presence of influential points could have a big impact on the slopes of the regression hyperplane, which could affect the estimated effects of the explanatory variables. Influential points can be detected using

Table 5: VIF table for linear regression model.

	VIF
T_min	1.76
Precipitation	1.00
Snow_Precip_Ratio	1.76

Cook's distance. A point with Cook's distance greater than $\frac{4}{n}$ is considered an influential point, where n is the number of observations used to fit the regression model.

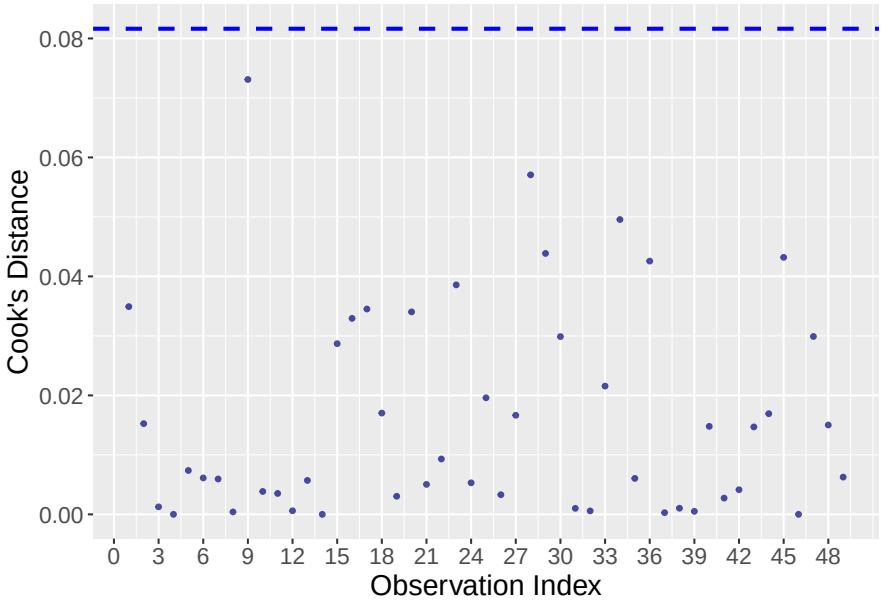

Figure 13: A plot of the Cook's distance for every observation.

Figure 13 suggests that there are no signs of influential points.

To see the entire process of this analysis, including all the code, click here to see the accompanying notebook.