My title*

My subtitle if needed

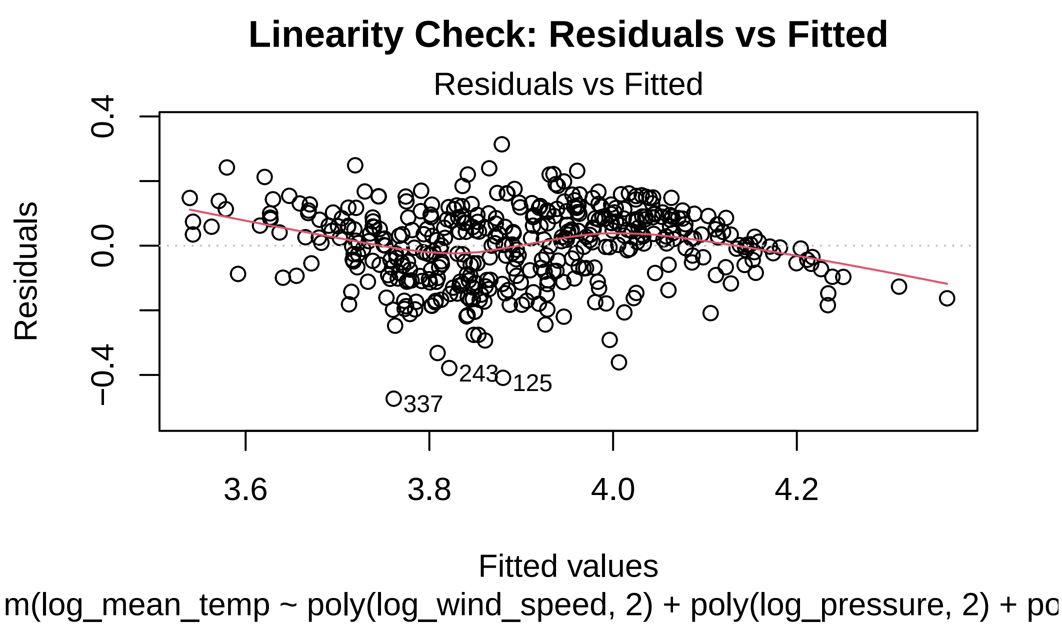
Lexun Yu

November 23, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

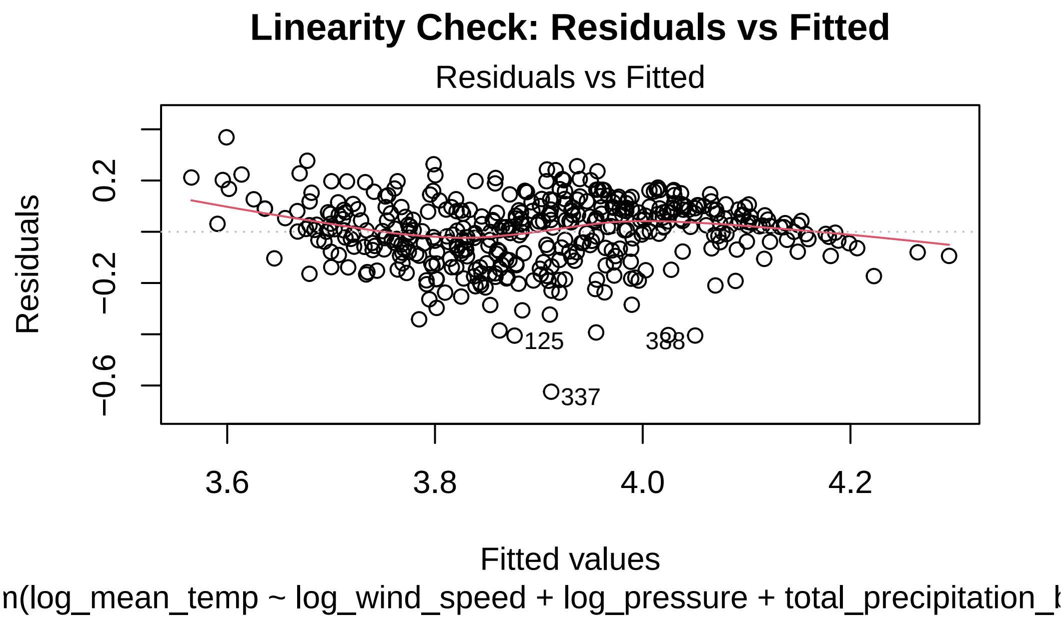
```
# Calculate diagnostic measures
glm_model <- glm
# Assuming `glm_model` is your fitted model

# 1. Linearity: Residuals vs Fitted Plot
plot(glm_model, which = 1, main = "Linearity Check: Residuals vs Fitted")
```



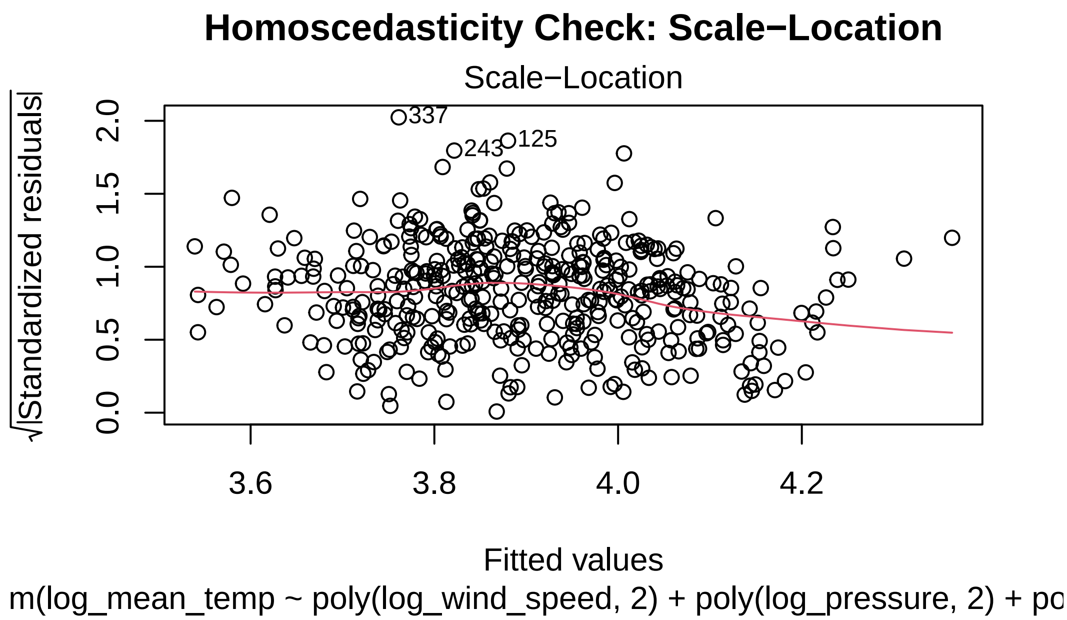
*Code and data are available at: https://github.com/yulexun/ClimateChangeYVR.

```r
plot(glm_log, which = 1, main = "Linearity Check: Residuals vs Fitted")
```

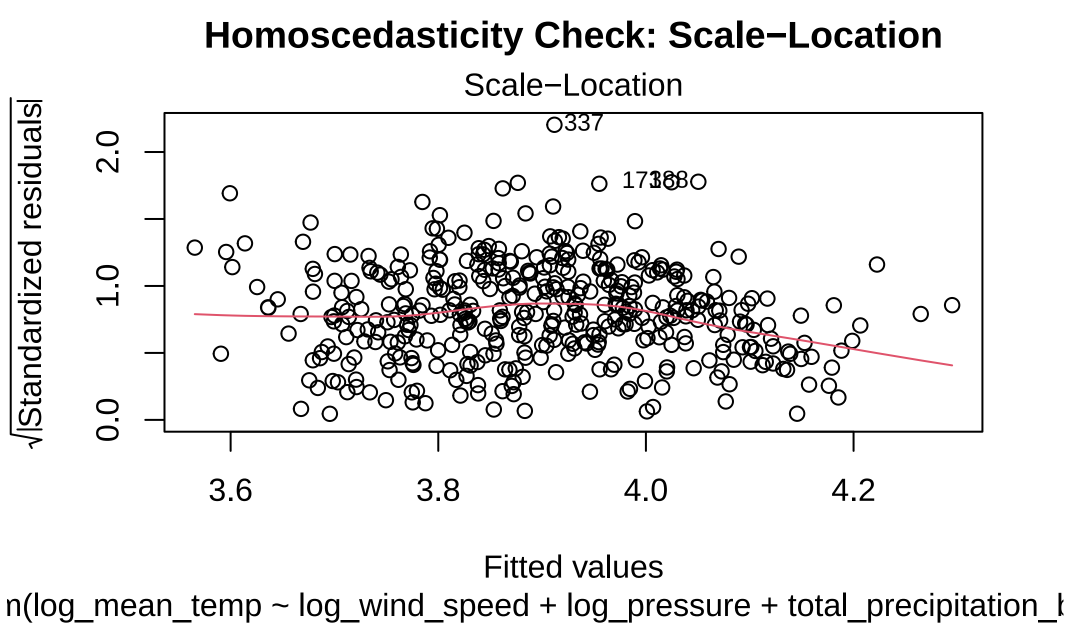



```r
# Look for a random scatter of points. Patterns indicate non-linearity.

# 2. Homoscedasticity: Scale-Location Plot
plot(glm_model, which = 3, main = "Homoscedasticity Check: Scale-Location")
```

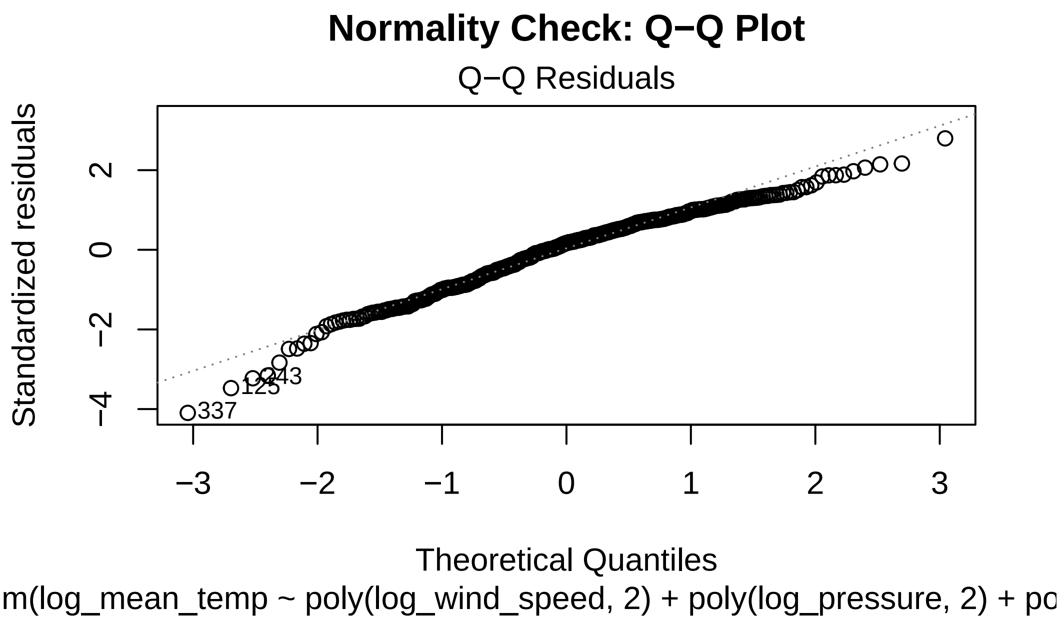



```r
plot(glm_log, which = 3, main = "Homoscedasticity Check: Scale-Location")
```

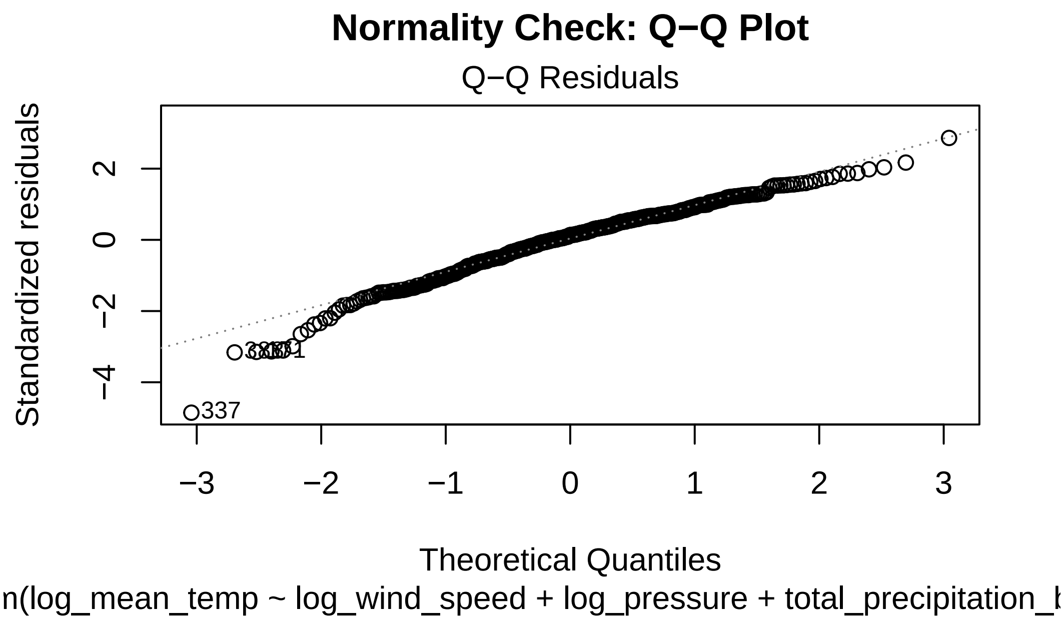


```r
# Points should be evenly spread. A funnel shape indicates heteroscedasticity.

# 3. Normality of Residuals: Q-Q Plot
plot(glm_model, which = 2, main = "Normality Check: Q-Q Plot")
```

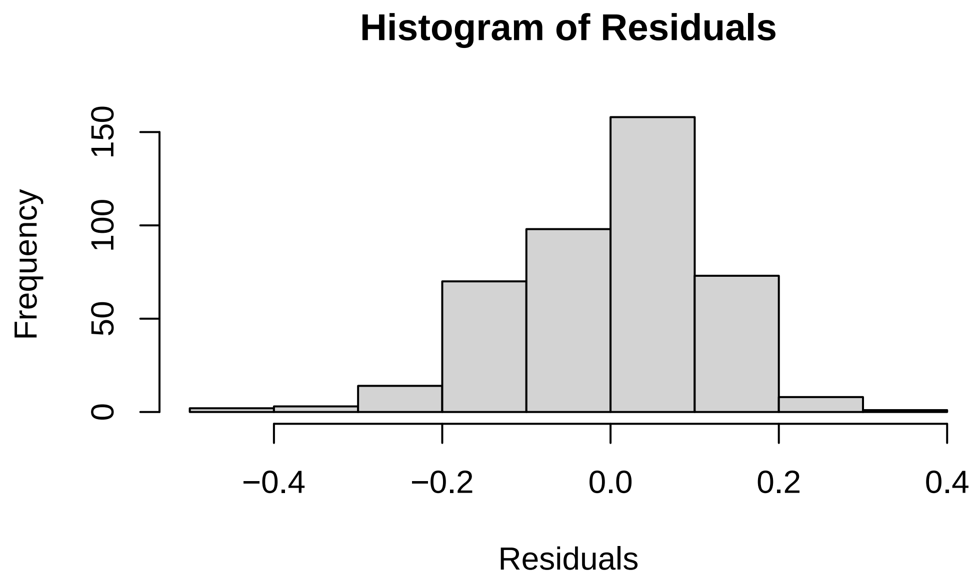



```r
plot(glm_log, which = 2, main = "Normality Check: Q-Q Plot")
```

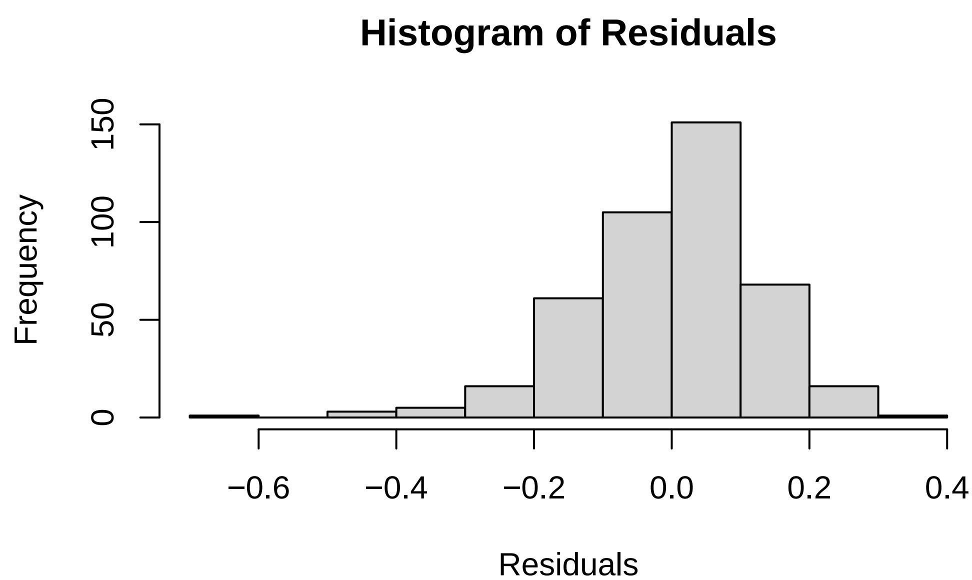



```
# Points should lie approximately along the diagonal line.

# 4. Residual Histogram: Another Normality Check
residuals <- residuals(glm_model)
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals")
```




```r
residuals <- residuals(glm_log)
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals")
```



```r
# Check for symmetry. A skewed histogram suggests non-normal residuals.

# 5. Multicollinearity: Variance Inflation Factor (VIF)
vif_values <- vif(glm_model)
print(vif_values)
```

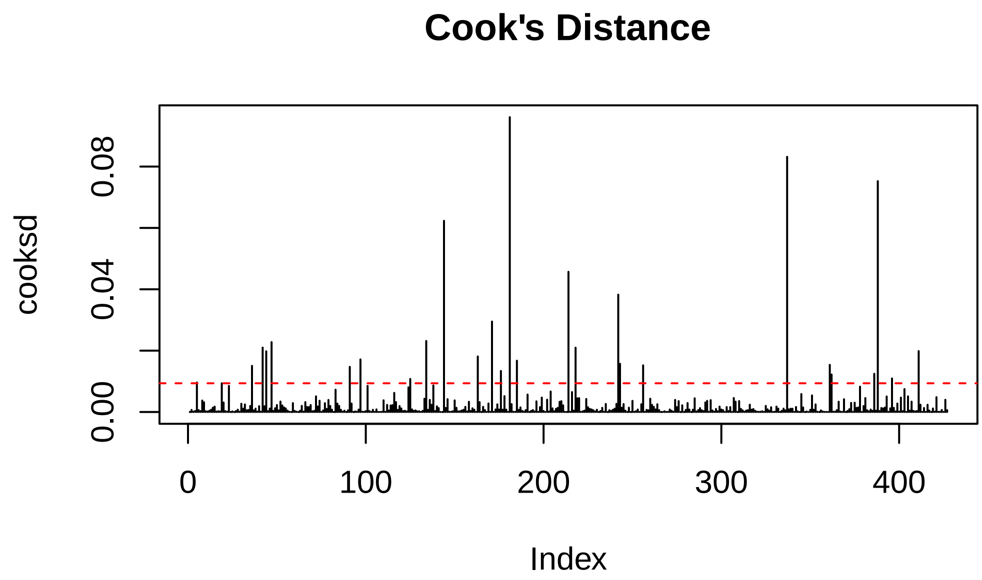
```
                                     GVIF Df GVIF^(1/(2*Df))
poly(log_wind_speed, 2)             1.685596  2        1.139432
poly(log_pressure, 2)               1.385884  2        1.085005
poly(total_precipitation_boxcox, 2) 1.478386  2        1.102674
poly(log_gust_speed, 2)             1.628509  2        1.129659
```

```r
vif_values <- vif(glm_log)
print(vif_values)
```

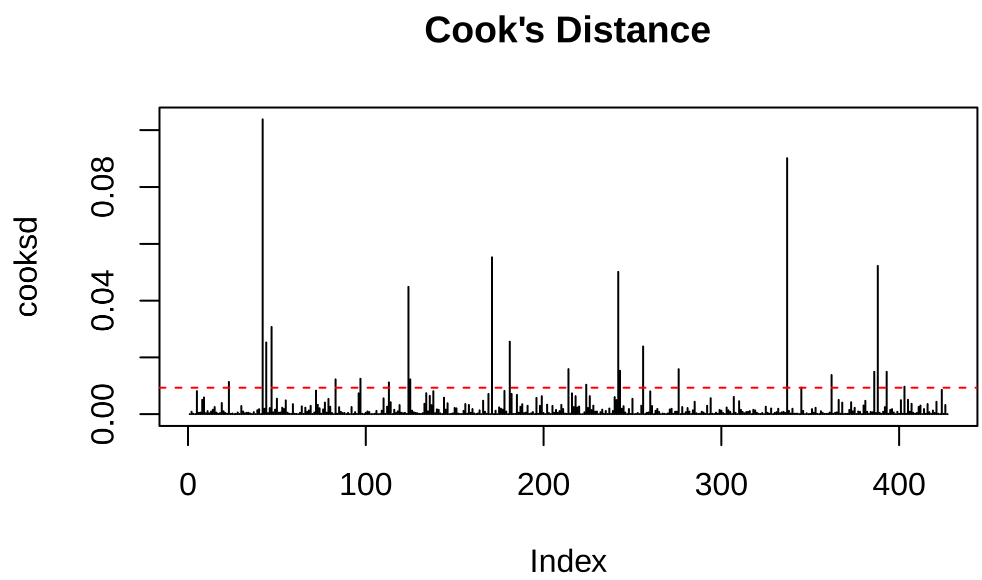
```
            log_wind_speed               log_pressure
                  1.453373                   1.217738
total_precipitation_boxcox             log_gust_speed
                  1.350802                   1.477064
```

```r
# VIF > 5 indicates high multicollinearity. Consider removing highly correlated predictors.

# 6. Cook's Distance: Influence of Observations
cooksd <- cooks.distance(glm_model)
plot(cooksd, main = "Cook's Distance", type = "h")
abline(h = 4 / nrow(model.frame(glm_model)), col = "red", lty = 2)
```

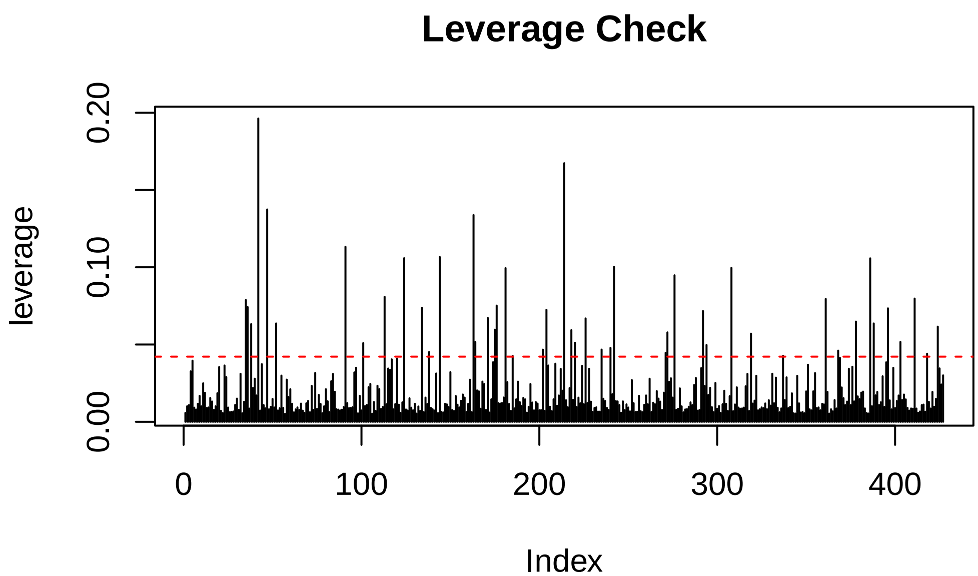



```
cooksd <- cooks.distance(glm_log)
plot(cooksd, main = "Cook's Distance", type = "h")
abline(h = 4 / nrow(model.frame(glm_log)), col = "red", lty = 2)
```

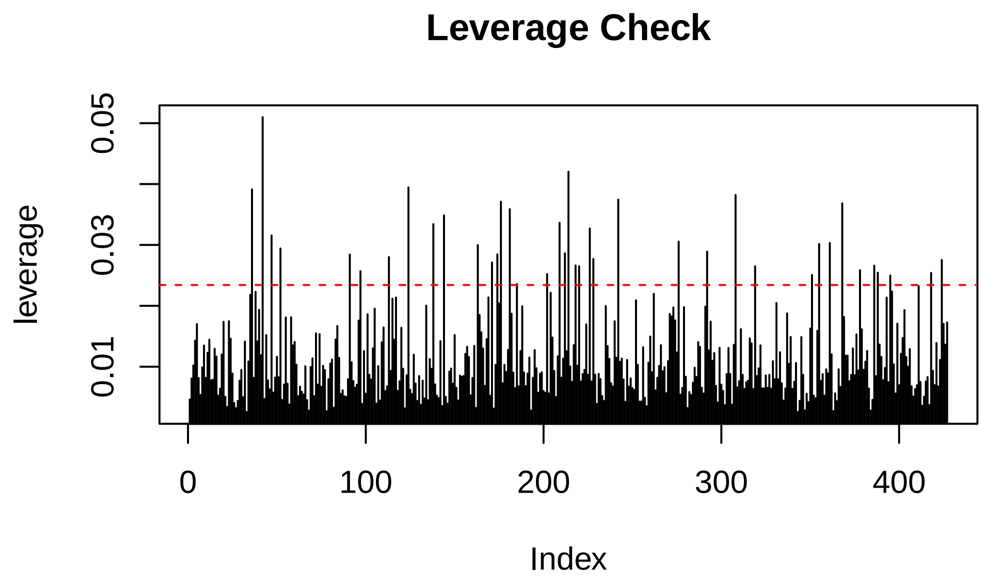


```r
# Points above the red line are highly influential.

# 7. Leverage: Identify High Leverage Points
leverage <- hatvalues(glm_model)
plot(leverage, main = "Leverage Check", type = "h")
abline(h = 2 * mean(leverage), col = "red", lty = 2)
```




```r
leverage <- hatvalues(glm_log)
plot(leverage, main = "Leverage Check", type = "h")
abline(h = 2 * mean(leverage), col = "red", lty = 2)
```




```r
# Points above the red line have high leverage.


AIC(glm_model)
```

```
[1] -601.1772
```

```r
AIC(glm_log)
```

```
[1] -524.8584
```

```r
BIC(glm_model)
```

```
[1] -560.6093
```

```r
BIC(glm_log)
```

```
[1] -500.5177
```

1 Introduction

Climate change is a global challenges today. Patterns such as rising temperatures, shifting weather systems, and increased frequency of severe weather events. In 2021, floods swept through streets in Japanese cities, displacing millions, while extreme heat fueled wildfires in Siberia (Greenpeace East Asia 2021). Climate change impacts human health, ecosystems, food security, water supplies, and economic stability. Understanding the factors driving temperature changes is necessary for designing effective mitigation strategies. This requires examining the various contributors to temperature variations.

Some scholars have examined the changing climate. Xu et al. (2009) analyze the effects of rising temperatures in the Himalayas, highlighting increased frequency and duration of extreme events and shifts in ecosystems. These changes pose challenges to water supply, agriculture, and human populations. Visser et al. (2021) investigates the relationship between precipitation and temperature using data from the Australian Bureau of Meteorology. Visser's regression model indicates that average precipitation intensities increase with temperature, suggesting more intense rainfall in a warmer climate. The role of sea level pressure is also significant. Wills et al. (2022) note that observed trends in sea level pressure have intensified warming in the Indo-Pacific Warm Pool and caused slight cooling in the eastern equatorial Pacific. However, as Zhang, Zhang, and Chen (2017) argue, much of the research has focused on temperature and precipitation. Zhang, Zhang, and Chen (2017) expands on this by incorporating additional predictors—relative humidity and wind speed—and concludes, using data from the Ministry of Agriculture of China, that these variables are important in understanding climate dynamics.

Temperatures significantly impact airport operations. Rising temperatures significantly affect aircraft performance, potentially leading to take-off weight restrictions and the need for longer runways. This directly impacts airport capacity and operations (Coffel and Horton 2015). Temperature forecast models vary in different locations, and different regions have unique climate characteristics that models may not fully capture (American Meteorological Society n.d.).

This research paper aims to identify the factors influencing temperature at Vancouver International Airport and build a model for temperature prediction with the data obtained from Canadian Centre for Climate Services (2022) and Meteorological Service of Canada (2023). Located on the west coast of Richmond, the airport sits on Sea Island, surrounded by water. As a transportation hub for passengers and freight, it is important to assess the location's safety in a warming climate.

Estimand paragraph

Results paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

The data gathering and analysis is done in R (R Core Team 2023) with the following packages: knitr (**knitr?**), tidyverse (Wickham et al. 2019), ggplot2 (**ggplot2?**), dplyr (**dplyr?**), arrow (**arrow?**), here (**here?**), and lubridate (Grolemund and Wickham 2011).

2 Data

2.1 Measurement

The measurement of Canadian weather data involves a network of weather stations and data collection methods managed by Environment and Climate Change Canada (ECCC). These stations continuously measure meteorological parameters such as temperature, precipitation, wind speed, and pressure (Meteorological Service of Canada 2023).

According to the glossary published by Meteorological Service of Canada (2023), Each day, measurement of temperature, rain, snow, precipitation, and gust speed are recorded. The wind and gust speed is measured in km/h with anemometer dials at a standard height of 10 meters above the ground. Rain and precipitation are measured in millimeter using the standard Canadian rain guage, a cylindrical container 40 cm high and 11.3 cm in diameter. Snow is measured in centimeters at several points that appear representative of the immediate area and then averaged. These raw data are combined to one entry and added to the historical climate database with a generated climate id and the station's location and id. Each row also have month and year of the data measured.

For climate research, including climate change studies, Environment and Climate Change Canada (2021a) has developed the Adjusted and Homogenized Canadian Climate Data (AHCCD) dataset. This dataset undergoes rigorous quality control and homogenization processes to address non-climatic factors that can affect long-term data consistency, such as station relocations or changes in instrumentation. The AHCCD ensures that observed trends reflect actual climate changes rather than artificial shifts in the data. In the AHCCD dataset, the precipitation, rain, pressure, snow and wind speed are adjusted with models to account for missing data and other non-climate factors. The detailed adjustments and corrections are documented in Section E. For example, precipitation measurements, which are often underestimated, are adjusted to ensure accuracy, especially in regions like the Arctic (Environment and Climate Change Canada 2021b). In the AHCCD dataset, parameters measured are recorded with the units, date, station ids, location and unique identifiers. The AHCCD data maintains a one-to-one correspondence with the historical weather dataset by a matching station id system, ensuring that each entry in the AHCCD aligns directly with a specific observation in the historical dataset.

The limitations are documented in Section E.

Table 1: Column Headers of Raw Climate Data

Longitude (x)	Latitude (y)	Station Name
Climate ID	Date/Time	Year
Month	Mean Max Temp (°C)	Mean Max Temp Flag
Mean Min Temp (°C)	Mean Min Temp Flag	Mean Temp (°C)
Mean Temp Flag	Extr Max Temp (°C)	Extr Max Temp Flag
Extr Min Temp (°C)	Extr Min Temp Flag	Total Rain (mm)
Total Rain Flag	Total Snow (cm)	Total Snow Flag
Total Precip (mm)	Total Precip Flag	Snow Grnd Last Day (cm)
Snow Grnd Last Day Flag	Dir of Max Gust (10's deg)	Dir of Max Gust Flag
Spd of Max Gust (km/h)	Spd of Max Gust Flag	Longitude (x)

2.2 Raw Data

In this project, we focus on weather data from YVR Airport, extracting only the datasets containing measurements taken at this specific location from the database. In both datasets, each row corresponds to a single averaged observation for a specific month and year. Each entry includes climate information such as temperature and wind speed, with their respective units recorded alongside the values. Additionally, a unique station ID and geographic coordinates (x, y) are included at the beginning of each entry for reference. The column headers of the raw historical weather dataset is displayed in Table 1. The column headers of the AHCCD dataset is displayed in Table 2.

The variables in the two datasets contains the following:

- Geographical Information: Longitude (x) and Latitude (y), with corresponding identifiers for location (Station Name in Table 1, station_id and province in Table 2).
- Temperature Metrics: Mean, maximum, and minimum temperatures (Mean Temp, Mean Max Temp, Mean Min Temp, Extr Max Temp, Extr Min Temp) and associated flags for data validity in Table 1. Similar metrics (temp_mean, temp_max, temp_min) in Table 2, with additional units included.
- Precipitation and Snowfall: Total precipitation (Total Precip) and total snow (Total Snow), with flags for data quality in Table 1. Equivalent precipitation and snow variables (total_precip, snow) in Table 2, with units explicitly defined.
- Wind and Gust Metrics: Direction and speed of maximum gusts (Dir of Max Gust, Spd of Max Gust) in Table 1, with units and flags. Wind speed (wind_speed) and related metrics in Table 2, with units included.
- Pressure Information: Sea level and station pressure variables in Table 2 (pressure_sea_level, pressure_station) with units.
- Temporal Information: Date and time variables (Date/Time in Table 1, date, period_value in Table 2) to track observations across time periods.

Table 2: Column Headers of Raw AHCCD Data

x	y
temp_mean_units___temp_moyenne_unites	temp_max_units___temp_max_unites
total_precip___precip_totale	temp_min___temp_min
rain___pluie	total_precip_units___precip_totale_unites
pressure_sea_level_units___pression_niveau_mer_unite	snow_units___neige_unites
temp_max___temp_max	lat___lat
identifier___identifiant	pressure_station___pression_station
lon___long	wind_speed___vitesse_vent
period_value___valeur_periode	period_group___groupe_periode
wind_speed_units___vitesse_vent_unites	temp_mean___temp_moyenne
pressure_station_units___pression_station_unites	province___province
station_id___id_station	temp_min_units___temp_min_unites
pressure_sea_level___pression_niveau_mer	snow___neige
date	rain_units___pluie_unites

- Flags and Identifiers: Flags for data validity in both tables, such as precipitation flags, temperature flags, and identifiers like Climate ID or identifier.

2.3 Data Cleaning

The data cleaning process consists of two steps. First, we standardize and clean the column headers. Second, we merge the two datasets into a single combined dataset. The dataset used in this analysis combines information from two distinct sources: climate data (raw_data_climate) and historical weather data (raw_data_ahccd). The analysis spans data collected between 1959 and 2010 for training and testing purposes.

The cleaned dataset contains a range of weather variables providing detailed monthly observations. The **`date`** variable represents the observation month, standardized to the first day of each month. **`wind_speed`** (km/h) captures average monthly wind speeds, while **`total_precipitation`** (mm) measures the total monthly precipitation, including rain and snow. **`snow`** (mm) records total snowfall, and **`pressure_station`** (kPa) indicates atmospheric pressure at the observation station. **`max_temp`** (°C), **`min_temp`** (°C), and **`mean_temp`** (°C) represent the monthly averages of maximum, minimum, and overall temperatures, respectively. **`total_rain`** (mm) focuses solely on rainfall amounts, distinct from snowfall. **`gust_speed_km_h`** (km/h) records the monthly average of maximum gust speeds. Additionally, constructed variables include **`mean_temp_F`**, the mean temperature was converted to Fahrenheit using

$$(\text{mean\_temp} \times 1.8) + 32$$

, and `log_mean_temp`, the log-transformed Fahrenheit temperature, was calculated as

$$\log(\text{mean\_temp\_F})$$

. Additionally, a Box-Cox transformation was applied to the `total_precipitation` variable to address skewness and stabilize variance, resulting in the new variable `total_precipitation_boxcox`. For `gust_speed_km_h`, `wind_speed` and `pressure_station`, a log transformation was used to stabilize variance and reduce right-skewness in their distribution, creating the new variable `log_gust_speed`, `log_wind_speed` and `log_pressure`.

All column names were cleaned and standardized using `janitor` in tidyverse (Wickham et al. 2019) to ensure consistency and readability. Dates were parsed into a unified format (`yyyy-mm-dd`) and aligned with monthly observations using the lubridate package (Grolemund and Wickham 2011). The datasets were merged into a single combined dataset using the `date_time` variable as the common key. Finally, constructed variables, including `mean_temp_F`, `total_precipitation_boxcox`, `log_gust_speed`, `log_mean_temp`, `log_wind_speed` and `log_pressure` were added to the cleaned data.

2.3.1 Cleaned Data

The top 6 rows of the cleaned data is displayed in Table 3.

The summary statistics of the combined dataset is displayed in Table 5.

2.4 Characteristics of Cleaned Data

All of the variables in the dataset are numeric, the histograms are plotted in Figure 1 and Figure 6. The following section explains the characteristics of these variables.

2.4.1 Skewness in Mean Temperature, Total Precipitation and Gust Speed

Figure 1 displays the histogram of the response variable Mean Temperature. Figure 1a shows the original mean temperature, which is skewed and includes negative values, making it unsuitable for direct modeling. To address this, We first transformed the data to Fahrenheit in Figure 1b, shifting all values to be positive. However, to further normalize the distribution and reduce skewness, we applied a logarithmic transformation in Figure 1c. The log transformation stabilizes variance, improves symmetry, and addresses non-linearity in the data, making it more appropriate for modelling.

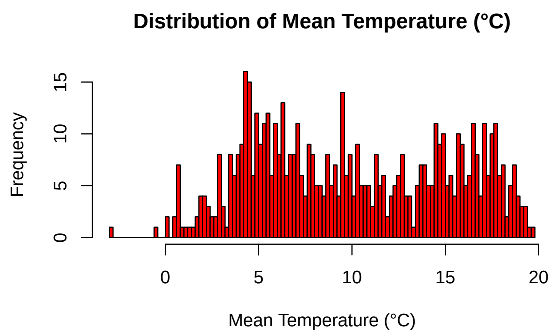
In Figure 6, Transformations are also applied to Wind Speed, Pressure, Total Precipitation and Gust Speed. Total Precipitation has a strong right skew, with most values low and a few extreme high values. We apply log transformation to predictors including wind speed, pressure and precipitation. For Gust Speed, a Box-Cox transformation is applied to adjust its moderate

Table 3: Sample of Cleaned Weather Data

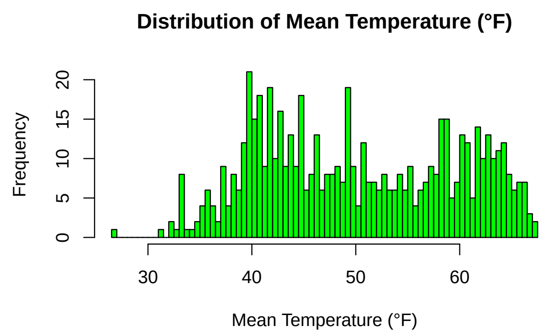
Wind Speed	Total Precip.	Snow	Pressure	Max Temp	Min Temp	Mean Temp	Rain
14.1	51.1	0.0	1015.9	20.7	12.1	16.4	44.7
13.9	153.9	0.0	1013.7	17.3	10.0	13.7	143.5
14.0	95.4	0.0	1016.9	13.4	7.2	10.3	87.1
14.8	166.8	7.8	1022.0	8.4	2.2	5.3	148.8
14.4	153.1	0.2	1019.0	7.2	1.3	4.3	142.2
13.7	172.7	18.3	1016.8	5.3	0.1	2.7	144.0

Log of Gust Speed	Log of Wind Speed	Log of Pressire
3.85	2.65	6.92
4.34	2.63	6.92
4.22	2.64	6.92
4.61	2.69	6.93
4.26	2.67	6.93
4.43	2.62	6.92

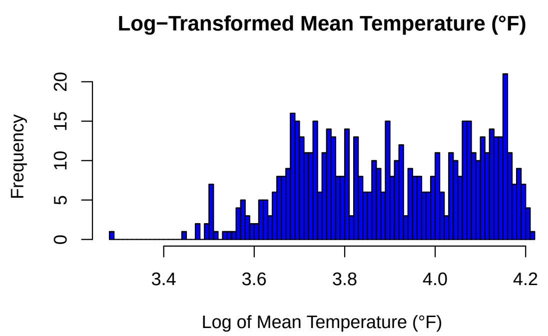
Gust Speed	Log of Mean Temp	Box-Cox Total Precip.
47	4.12	10.15
77	4.04	17.61
68	3.92	13.94
100	3.73	18.30
71	3.68	17.57
84	3.61	18.61



(a) Original Mean Temp has Skewness and Negative Numbers



(b) Mean Temp in F Transformed the Value to All Positive



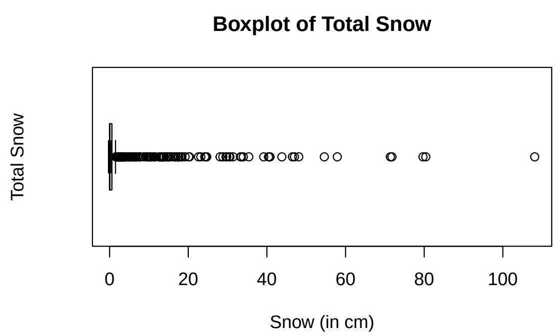
(c) Log-Transformed Data Shows a More Symmetric and Less Skewed Distribution

Figure 1: Mean Temp Shows More Normality and Less Skewness After Adjustment

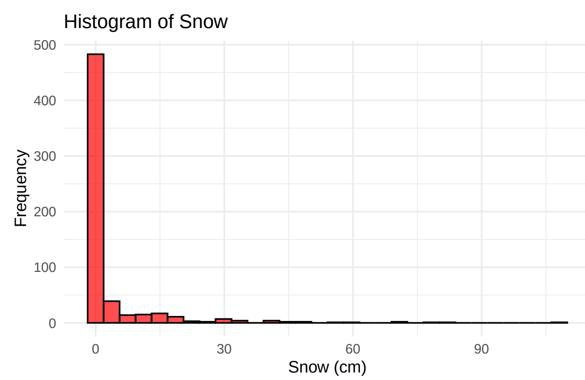
skewness. This transformation reshaped the data to better approximate a normal distribution. These adjustments improve the suitability of these variables for statistical analyses that assume normality.

2.4.2 Total Snow Is Zero-Inflated

Figure 2 clearly shows significant zero inflation, with a large number of observations concentrated at zero and a few extreme outliers far above the majority of the data. This distribution suggests that the variable snow contains excessive structural zeros, likely representing instances where no snowfall occurred.



(a) A Large Proportion of Observations With No Snowfall



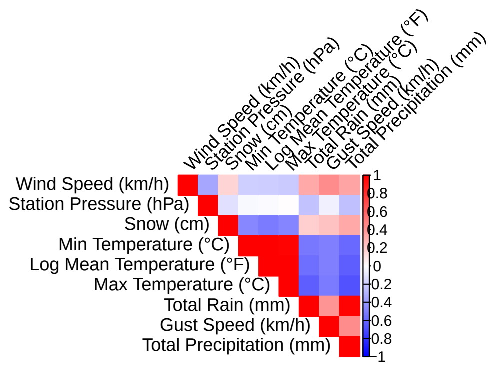
(b) Prevalence of Zero Values and a Skewed Pattern in the Non-zero Snowfall Measurements

Figure 2: Total Snow shows Zero Inflation

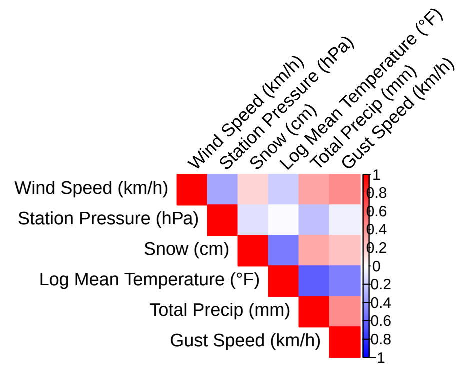
2.4.3 Variables with Strong Linear Relationships

2.4.3.1 Maximum Temperature, Minimum Temperature and Mean Temperature

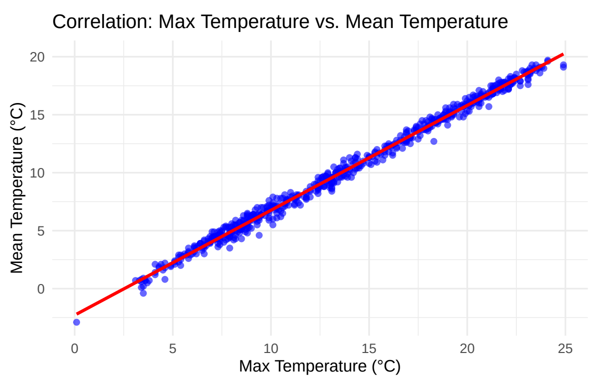
Figure 3a highlights strong relationships between temperature variables, showing strong positive correlations between Max Temperature (°C), Min Temperature (°C), and Mean Temperature (°C). Scatter plots in Figure 3c and Figure 3d show near-perfect linear relationships, indicating that Max Temperature (°C) and Min Temperature (°C) are highly collinear with Mean Temperature (°C). In contrast, other predictors shown in Figure 3b, such as Wind Speed (km/h), Station Pressure (hPa), and Total Rain (mm), show weaker correlations with the temperature variables and with each other, suggesting they contribute unique and independent information to the model.



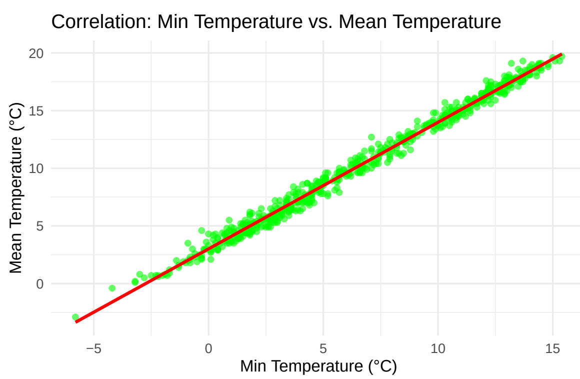
(a) High Correlations Between Max, Min and Mean Temperature, Total Precip and Rain



(b) Other Predictors Does Not Have High Correlations



(c) Linear Relationship Between Max and Mean Temperature



(d) Linear Relationship Between Min and Mean Temperature

Figure 3: Temperature Values Have High Correlations

2.4.3.2 Total Precipitation and Total Rain

Similar to temperature, precipitation and total rain also have a relatively strong linear relationship as illustrated in Figure 4.

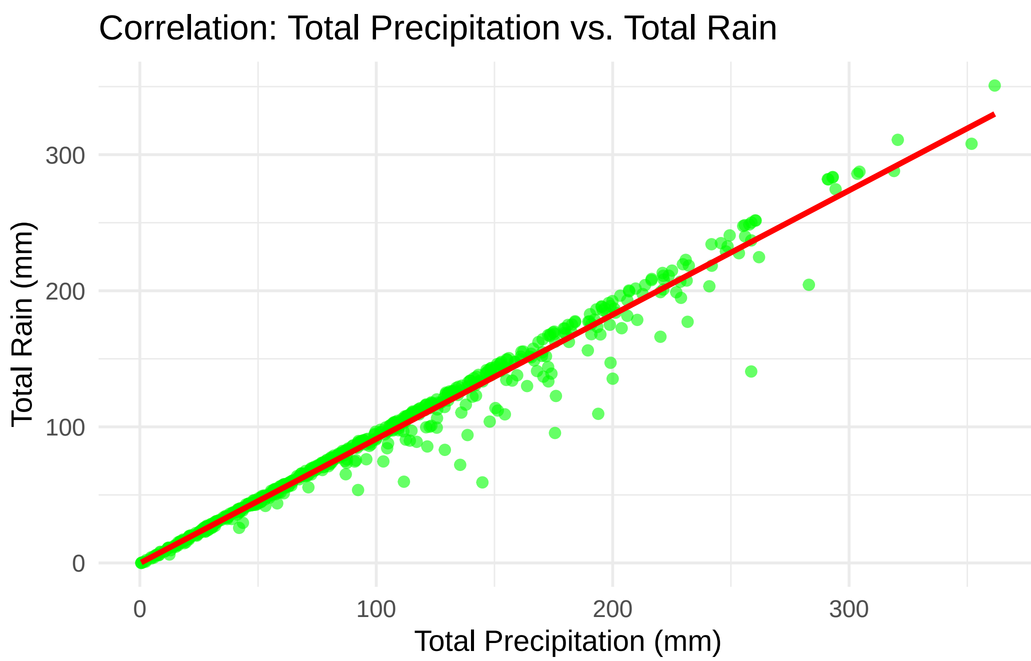


Figure 4: Precipitation and Rain Have High Correlations

3 Model

The goal of our modelling strategy is to find a model that can predict temperature changes with other weather data.

We build a bayesian model, a general linear model and a general linear model with a 2 degrees of polynomial transformation. We determine the best model is the linear model with polynomial transformation. The detailed steps are recorded in Section B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

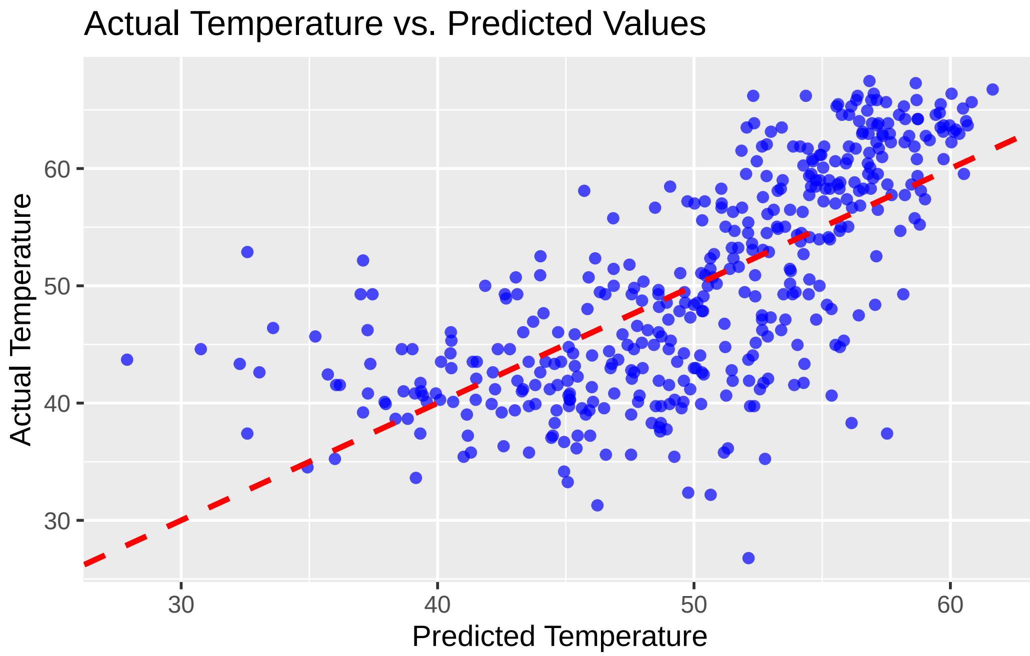
$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.2 MLR Model

```
fitted_values <- fitted(m4)
ggplot(data = train_data, aes(x = fitted_values, y = mean_temp_F)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed", size = 1) +
  labs(x = "Predicted Temperature", y = "Actual Temperature", title = "Actual Temperature vs
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



3.2.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A License

Contains information licensed under the Open Government Licence – Canada

B Model Details

B.1 MLR Model With Every Predictor in Cleaned Data

The first model predicts mean temperature (mean_temp_F) based on multiple predictors: wind speed (wind_speed), total precipitation (total_precipitation), snow (snow), station pressure (pressure_station), maximum temperature (max_temp), minimum temperature (min_temp), total rainfall (total_rain), and gust speed (gust_speed_km_h).

The fitted model is:

$$
\begin{aligned}
\text{mean\_temp\_F} = \beta_0 + \beta_1 \cdot \text{wind\_speed} + \beta_2 \cdot \text{total\_precipitation} + \beta_3 \cdot \text{snow} \quad (7)\\
+ \beta_4 \cdot \text{pressure\_station} + \beta_5 \cdot \text{max\_temp} + \beta_6 \cdot \text{min\_temp} \quad (8)\\
+ \beta_7 \cdot \text{total\_rain} + \beta_8 \cdot \text{gust\_speed\_km\_h} + \epsilon \quad (9)
\end{aligned}
$$

- β_0: Intercept
- $\beta_1, \beta_2, \dots, \beta_8$: Coefficients representing the change in mean_temp_F for a one-unit increase in the respective predictor, holding other variables constant.
- ϵ: Residual error, assumed to be normally distributed with mean 0.

This model has the following summary statistics in Table 4.

The model's coefficients suggest an issue of multicollinearity, particularly due to the inclusion of highly correlated predictors such as maximum temperature, minimum temperature, and mean temperature, as discussed in Section 2.4.3.1. Multicollinearity inflates the standard errors of the coefficients, making it difficult to determine the individual contribution of these variables to the response variable. Despite the model showing a perfect R2 and adjusted R2 , these metrics are misleading because the presence of highly correlated predictors often leads to overfitting. This is evident from the small coefficient magnitudes and nearly zero p-values, which do not reflect the true independent influence of the predictors. Such multicollinearity can undermine the model's interpretability and generalizability to new data.

Table 4: Summary Statistics Shows a Large R2 in Model 1, Potential Variability in Model 2, Model L fits performs than Model 2

	Model 1	Model 2	Model L
(Intercept)	33.333	479.165	55.192
	(1.148)	(133.508)	(17.928)
wind_speed	−0.001	0.406	
	(0.002)	(0.188)	
total_precipitation	−0.001	−0.077	
	(0.001)	(0.005)	
snow	0.000		
	(0.001)		
pressure_station	−0.001	−0.409	
	(0.001)	(0.131)	
max_temp	0.900		
	(0.004)		
min_temp	0.899		
	(0.004)		
total_rain	0.001		
	(0.002)		
gust_speed_km_h	0.000	−0.181	
	(0.000)	(0.026)	
log_wind_speed			0.150
			(0.050)
log_pressure			−7.283
			(2.586)
total_precipitation_boxcox			−0.022
			(0.001)
log_gust_speed			−0.238
			(0.033)
Num.Obs.	427	427	427
R2	1.000	0.491	0.525
R2 Adj.	1.000	0.486	0.521
AIC	−1247.6	2835.7	−524.9
BIC	−1207.0	2860.0	−500.5
Log.Lik.	633.787	−1411.843	268.429
RMSE	0.05	6.60	0.13

B.2 MLR Model Without Multicollinearity Variables

We then build our second model.

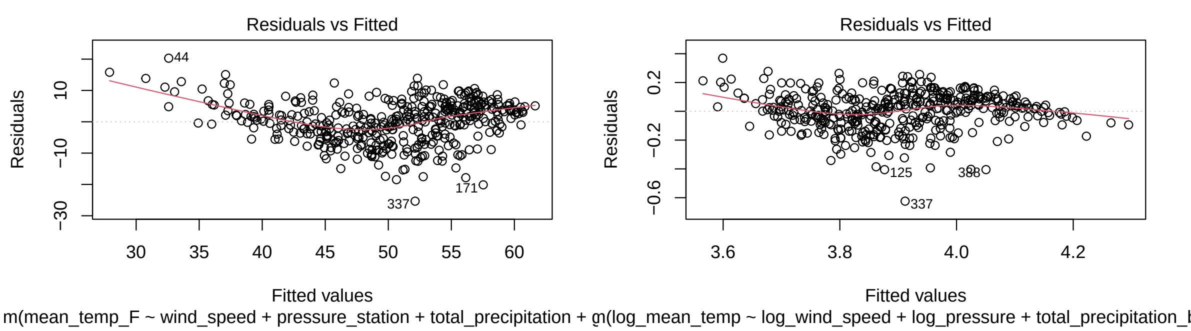
This model predicts mean temperature (mean_temp_F) based on a subset of predictors: wind speed (wind_speed), station pressure (pressure_station), total precipitation (total_precipitation), and gust speed (gust_speed_km_h).

$$\begin{aligned} \text{mean\_temp\_F} = \beta_0 + \beta_1 \cdot \text{wind\_speed} + \beta_2 \cdot \text{pressure\_station} \quad (10) \\ + \beta_3 \cdot \text{total\_precipitation} + \beta_4 \cdot \text{gust\_speed\_km\_h} + \epsilon \quad (11) \end{aligned}$$

- β_0: Intercept.
- $\beta_1, \beta_2, \beta_3, \beta_4$: Coefficients representing the change in mean_temp_F for a one-unit increase in each respective predictor, holding others constant.
- ϵ: Residual error, assumed to be normally distributed with mean 0.

This simplified model excludes highly correlated predictors, such as maximum and minimum temperatures, to reduce multicollinearity and improve interpretability.

In the summary of our second model as shown in Table 4, all predictors have relatively small coefficients, suggesting incremental effects on the response variable. The relatively large standard errors of some coefficients, such as the intercept, indicate potential variability or noise in the data. For instalce, from Figure 1 and Figure 6, we observe skewness and non-normal distribution in both predictor and the response. According to Figure 5a, The model does not sufficiently explain the variability in the response variable, due to non-linearity or unaddressed skewness in the data. This plot suggests that the model's assumptions of linearity and homoscedasticity (constant variance of residuals) are violated.



(a) In Model 2, The residuals are not evenly distributed

(b) In Model L, The residual is more evenly distributed

Figure 5: Residual vs Fitted Plot of Model 2 and L

B.3 MLR Model with transformed variables

In our third model L, we use log and Box-Cox transformation to ensure linearity and homoscedasticity in all predictors and the response. The detailed steps are documented in Section 2.4.3. This linear model predicts the log-transformed mean temperature (`log_mean_temp`) based on log-transformed wind speed (`log_wind_speed`), log-transformed pressure (`log_pressure`), Box-Cox-transformed total precipitation (`total_precipitation_boxcox`), and log-transformed gust speed (`log_gust_speed`).

We build our Model L as the following:

$$\begin{aligned} \text{log\_mean\_temp} = \beta_0 + \beta_1 \cdot \text{log\_wind\_speed} + \beta_2 \cdot \text{log\_pressure} \quad (12) \\ + \beta_3 \cdot \text{total\_precipitation\_boxcox} + \beta_4 \cdot \text{log\_gust\_speed} + \epsilon \quad (13) \end{aligned}$$

- β_0: Intercept.
- $\beta_1, \beta_2, \beta_3, \beta_4$: Coefficients representing the change in `log_mean_temp` for a one-unit increase in each predictor, holding other variables constant.
- ϵ: Residual error, assumed to follow a Gaussian (normal) distribution.

The inclusion of the Box-Cox-transformed total precipitation further refines the model by accommodating non-linearity in precipitation data. The Gaussian family ensures that the residuals of the response variable follow a normal distribution after the transformations. As shown in Figure 5b, this model reduces heteroscedasticity, minimizes non-linear patterns in residuals, and improves overall interpretability and fit. Each coefficient indicates the multiplicative effect of a one-unit change in the respective predictor on the mean temperature after applying the logarithmic transformations. This model fits better than Model 2, as indicated in Table 4, as the R2 and adjusted R2 are higher, AIC, BIC are smaller.

B.4 Bayesian Model

After fitting the linear regression model using log and Box-Cox transformations, We extend the analysis by testing a Bayesian regression model (Model B). This model also predicts the log-transformed mean temperature (`log_mean_temp`) but incorporates prior and Bayesian inference to evaluate the uncertainty of parameter estimates. The predictors remain the same: wind speed (`wind_speed`), station pressure (`pressure_station`), Box-Cox-transformed total precipitation (`total_precipitation_boxcox`), and log-transformed gust speed (`log_gust_speed`).

The Bayesian model is defined as:

$$
\begin{aligned}
\text{log\_mean\_temp} &\sim \mathcal{N}(\mu, \sigma^2), && (14)\\
\mu &= \beta_0 + \beta_1 \cdot \text{wind\_speed} + \beta_2 \cdot \text{pressure\_station} && (15)\\
&\quad + \beta_3 \cdot \text{total\_precipitation\_boxcox} + \beta_4 \cdot \text{log\_gust\_speed}. && (16)
\end{aligned}
$$

The prior distributions for the parameters are:

- Coefficients $(\beta_1, \beta_2, \beta_3, \beta_4)$:

  $$\beta_i \sim \mathcal{N}(0, 10), \quad \text{for } i = 1, 2, 3, 4,$$

  reflecting moderate uncertainty centered around zero.

- Intercept (β_0):

  $$\beta_0 \sim \mathcal{N}(0, 10),$$

  indicating prior uncertainty about the baseline log-mean temperature.

The model was fit using **Hamiltonian Monte Carlo (HMC)** via the `brms` package (**brm?**). It uses: - **4 chains** for convergence, - **2000 iterations** per chain to ensure stability, - **4 cores** for parallel computation, enabling efficient sampling.

B.5 Polynomial Linear Model

The residual plot in Figure 5b shows a non-linear pattern, as indicated by the curved trend in the residuals. This suggests that the relationship between the predictors and the response variable is not fully captured by a linear model. Adding polynomial terms could help address this non-linearity by allowing the model to fit curved relationships.

C Summary Statistic of cleaned dataset

Table 5: Summary Statistics of Raw Climate Data

Wind Speed	Total Precipitation	Snow	Pres.	Max Temp	Min Temp
Min. : 8.4	Min. : 0.60	Min. : 0.00	Min. :1006	Min. : 0.100	Min. :-5.800
1st Qu.:12.7	1st Qu.: 47.98	1st Qu.: 0.00	1st Qu.:1015	1st Qu.: 8.625	1st Qu.: 2.600
Median :13.9	Median : 88.30	Median : 0.00	Median :1016	Median :13.200	Median : 6.050
Mean :14.0	Mean :103.09	Mean : 3.65	Mean :1016	Mean :13.722	Mean : 6.484
3rd Qu.:15.2	3rd Qu.:146.05	3rd Qu.: 0.60	3rd Qu.:1018	3rd Qu.:19.000	3rd Qu.:10.800
Max. :22.5	Max. :361.60	Max. :108.10	Max. :1025	Max. :24.900	Max. :15.400

Mean Temp	Rain	Max Gust Speed	Mean Temp in F	Log of Mean Temp
Min. :-2.90	Min. : 0.00	Min. : 33.00	Min. :26.78	Min. :3.288
1st Qu.: 5.50	1st Qu.: 43.65	1st Qu.: 51.00	1st Qu.:41.90	1st Qu.:3.735
Median : 9.60	Median : 80.45	Median : 59.00	Median :49.28	Median :3.898
Mean :10.13	Mean : 93.94	Mean : 61.11	Mean :50.23	Mean :3.899
3rd Qu.:14.90	3rd Qu.:133.25	3rd Qu.: 70.00	3rd Qu.:58.82	3rd Qu.:4.074
Max. :19.70	Max. :350.80	Max. :126.00	Max. :67.46	Max. :4.212

Box Cox of Precipitation	Log of Gust Speed	Log of Wind Speed	Log of Pressure
Min. :-0.4593	Min. :3.497	Min. :2.128	Min. :6.914
1st Qu.: 9.8199	1st Qu.:3.932	1st Qu.:2.542	1st Qu.:6.923
Median :13.4170	Median :4.078	Median :2.632	Median :6.924
Mean :13.4494	Mean :4.085	Mean :2.628	Mean :6.924
3rd Qu.:17.1709	3rd Qu.:4.248	3rd Qu.:2.721	3rd Qu.:6.925
Max. :26.3305	Max. :4.836	Max. :3.114	Max. :6.933

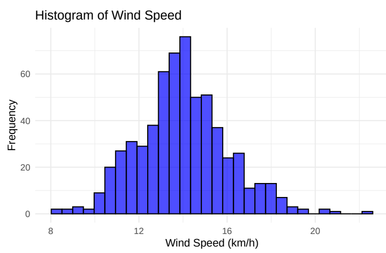
D Additional Model Detail

E Methodology of ECCC

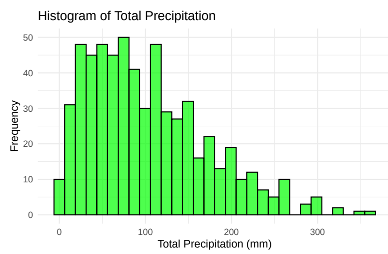
The Adjusted and Homogenized Canadian Climate Data (AHCCD) is a collection of climate datasets developed by Environment and Climate Change Canada (2021a). These datasets provide long-term, quality-controlled data that have been adjusted to correct for non-climatic influences.

E.1 Population, Frame, and Sample

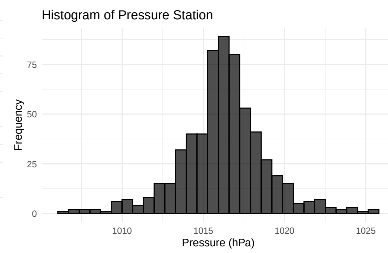
The population of interest in the AHCCD is the entirety of Canada's climate data, representing diverse geographical regions and climate conditions. The frame of the dataset are the climato-



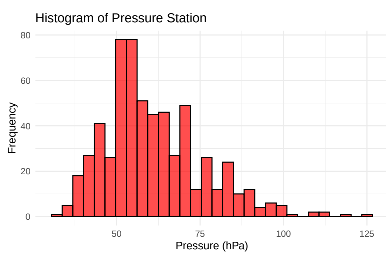
(a) Wind Speed Shows Symmetric Distribution



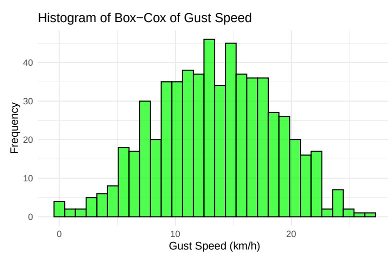
(b) Total Precipitation Shows Normal Distribution but Skewed Towards Low Values



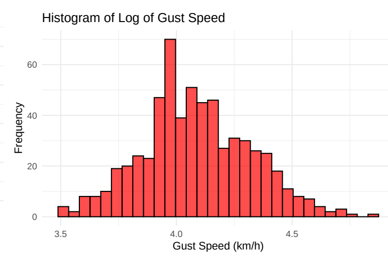
(c) Pressure Station Has a Nearly Normal Distribution



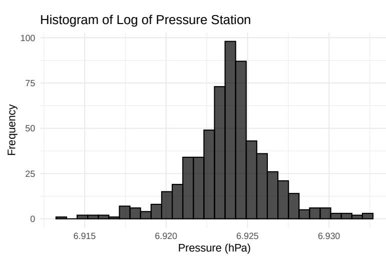
(d) Gust Speed Shows Normal Distribution and Moderate Skewness



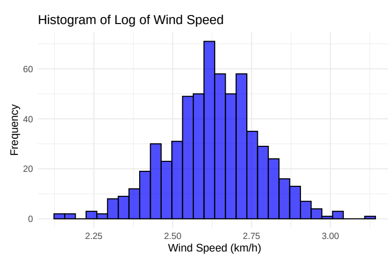
(e) Box-Cox Transformed Precipitation shows Reduces Skewness



(f) Log Transformed Gust Speed shows Reduces Skewness



(g) Log Transformed Pressure shows Reduces Skewness



(h) Log Transformed Wind Speed shows Reduces Skewness

Figure 6: Other Variables Show Normal Distribution

logical stations maintained by the ECCC that span across the countries in important locations such as airports, and banks of lakes or rivers. These stations record data on climate elements such as temperature, precipitation, surface pressure, and wind speed over extended periods. The sample is the selected stations across Canada, with adjustments applied to address inconsistencies. The datasets cover periods extending back to 1895 for precipitation, while other variables like wind speed and surface pressure start from 1953 or later. The recorded sample consists of monthly, seasonal, and annual data about surface air temperature, precipitation, pressure, and wind speed, according to Environment and Climate Change Canada (2021a).

E.2 Sample Corrections and Adjustments

The original data for AHCCD are extracted from the National Climate Data Archive of Environment Canada. These data include daily observations, such as maximum and minimum temperatures, precipitation, surface pressure, and wind speed. Observations are quality-controlled and adjusted to correct for biases due to changes in instruments, observation procedures, and other factors.

Precipitation data adjustments account for wind undercatch, evaporation, and gauge-specific losses. According to Environment and Climate Change Canada (2021b), corrections to account for wind undercatch, evaporation, and gauge specific wetting losses were implemented, especially in snowy conditions where snowfall is not fully captured by standard gauges. Corrections are made with the study by Devine and Mekis.

Surface air temperature adjustments apply Quantile-Matching techniques to remove inhomogeneities. According to Environment and Climate Change Canada (2021c), With Vincent and Wang's third generation homogenized temperature, Quantile-Matching ensures that the temperature data remain consistent across different periods, even when observation practices change.

Surface pressure and wind speed data undergo adjustments based on metadata and statistical tests for systematic shifts. According to Environment and Climate Change Canada (2021e), wind speed is first adjusted with a logarithmic wind profile, then tested for homogeneity using a technique based on regression models. It involves the identification of variation due to changes in anemometer and location change. The pressure data is corrected due to systematic shifts of non-updated station elevation and relocation, as stated by Environment and Climate Change Canada (2021d).

E.3 Sampling Approach and Trade-offs

According to the published methodology and the webpage by Dunbar (2020), they employ a systematic sampling approach by selecting specific climatological stations with long-term, consistent data records. In some cases, observations from neighboring or overlapping stations are merged to extend time series. The AHCCD dataset may also contain missing values, which can vary depending on the variable, station, and time. Additionally, the AHCCD dataset is site-specific, meaning it provides data specific to individual observation stations.

E.4 Missing Data Handling

Non-response, such as gaps in the data due to missing records, is managed by employing statistical and physical methods to homogenize the data. For instance, the AHCCD adjusts for shifts detected through historical evidence and metadata analysis. For large amount of missing data, ECCC mark the data as NA in the dataset (Canadian Centre for Climate Services 2022).

E.5 Strengths and Weaknesses

The AHCCD by Dunbar (2020) provides long-term, high-quality climate records adjusted for non-climatic factors such as changes in instrumentation, observation procedures, and station relocations, ensuring consistency and reliability for trend analysis in climate change.

The documentation acknowledges the possibility of missing values, which naturally arise in long-term observational datasets due to factors such as station interruptions, relocation, or equipment malfunctions (Environment and Climate Change Canada 2021a). Moreover, the dataset's coverage in Arctic regions is limited to the restricted to the mid-1940s to present, as this limitation reflects the historical absence of earlier systematic observations in these remote regions.

F Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

References

American Meteorological Society. n.d. "Weather Analysis and Forecasting." *American Meteorological Society*. Accessed November 21, 2024. https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/weather-analysis-and-forecasting2/.

Canadian Centre for Climate Services. 2022. "Adjusted and Homogenized Canadian Climate Data." https://climate-change.canada.ca/climate-data/#/adjusted-station-data.

Coffel, E., and R. Horton. 2015. "Climate Change and the Impact of Extreme Temperatures on Aviation." *Weather, Climate, and Society* 7 (1): 94–102. https://doi.org/10.1175/WCAS-D-14-00026.1.

Dunbar, Alyssa. 2020. "Adjusted and Homogenized Canadian Climate Data (AHCCD) - Data Collection Methodology." https://open.canada.ca/data/en/dataset/9c4ebc00-3ea4-4fe0-8bf2-66cfe1cddd1d/resource/26545adf-e689-4d83-8f2d-9aad3dfa6f57.

Environment and Climate Change Canada. 2021a. "Adjusted and Homogenized Canadian Climate Data." Datasets. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data.html.

———. 2021b. "Climate Data: Adjusted Precipitation Data." Research;program results. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data/precipitation.html.

———. 2021c. "Climate Data: Homogenized Surface Air Temperature Data." Program descriptions;program results. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data/surface-air-temperature.html.

———. 2021d. "Climate Data: Homogenized Surface Pressure Data." Research. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data/surface-pressure.html.

———. 2021e. "Climate Data: Homogenized Surface Wind Speed Data." Research;datasets. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data/surface-wind-speed.html.

Greenpeace East Asia. 2021. "5 Ways the Climate Crisis Will Change Asia." *Greenpeace East Asia*. https://www.greenpeace.org/eastasia/blog/6802/5-ways-the-climate-crisis-will-change-asia/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Meteorological Service of Canada. 2023. "Past Weather and Climate." https://www.canada.ca/en/services/environment/weather.html.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Visser, Johan B., Conrad Wasko, Ashish Sharma, and Rory Nathan. 2021. "Eliminating the 'Hook' in Precipitation-Temperature Scaling." *Journal of Climate*, September, 1–42. https://doi.org/10.1175/JCLI-D-21-0292.1.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wills, Robert C. J., Yue Dong, Cristian Proistosecu, Kyle C. Armour, and David S. Battisti. 2022. "Systematic Climate Model Biases in the Large-Scale Patterns of Recent Sea-Surface Temperature and Sea-Level Pressure Change." *Geophysical Research Letters* 49 (17): e2022GL100011. https://doi.org/10.1029/2022GL100011.

Xu, Jianchu, R. Edward Grumbine, Arun Shrestha, Mats Eriksson, Xuefei Yang, Yun Wang, and Andreas Wilkes. 2009. "The Melting Himalayas: Cascading Effects of Climate Change on Water, Biodiversity, and Livelihoods." *Conservation Biology* 23 (3): 520–30. https://doi.org/10.1111/j.1523-1739.2009.01237.x.

Zhang, Peng, Junjie Zhang, and Minpeng Chen. 2017. "Economic Impacts of Climate Change on Agriculture: The Importance of Additional Climatic Variables Other Than Temperature and Precipitation." *Journal of Environmental Economics and Management* 83 (May): 8–31. https://doi.org/10.1016/j.jeem.2016.12.001.