Exploring Nonlinear Atmospheric Influences on Temperature Using Polynomial Regression at Vancouver International Airport*

Pressure and Wind as Key Drivers of Temperature Variability, with Limited Impact from Precipitation and Gust Speed

Lexun Yu

November 27, 2024

This study investigates how atmospheric factors influence temperature at Vancouver International Airport using a polynomial regression model. The analysis shows that total precipitation has the strongest effect, significantly reducing mean temperature, while wind speed and atmospheric pressure exhibit non-linear and consistent negative impacts, respectively. By explaining 61% of the variability in temperature, the model demonstrates that precipitation, pressure, gust speed and wind speed are key drivers of regional climate patterns. Understanding these relationships helps inform climate adaptation strategies for important infrastructure, including airports, in a warming world.

Table of contents

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

1 Introduction

Climate change is a global challenge 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 contributors to temperature variations.

Some scholars have examined the changing climate. Xu et al. (2009) analyzes 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 caused slight cooling in the eastern equatorial Pacific. However, as Zhang, Zhang, and Chen (2017) argues, 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, 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 separate 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.

The estimand in this study is the effect of atmospheric variables, such as wind speed, atmospheric pressure, total precipitation, and gust speed, on the mean temperature at Vancouver International Airport. The goal is to quantify these relationships and predict temperature variations using a polynomial regression model. To account for the non-linear and skewed nature of the data, transformations such as logarithmic and Box-Cox methods were applied to both predictors and the response variable. The estimand reflects how changes in these atmospheric conditions influence log-transformed mean temperature over time.

The results indicate that total precipitation is the most influential predictor, showing a strong inverse relationship with mean temperature. Wind speed and gust speed demonstrate non-linear effects, with moderate increases associated with temperature rises and higher values contributing to cooling. Pressure consistently exhibits a negative association with temperature, with amplified effects at extreme levels. The model explains 61% of the variance in mean temperature and performs well in validation, showing minimal overfitting and low prediction error.

These results highlight the role of atmospheric variables in regional temperature changes and the effectiveness of the polynomial regression approach in capturing complex relationships. Our findings contribute to a deeper understanding of regional climate changes, which is important for decision-making and resilience planning in an important nodes in global transportation and logistics, where temperature fluctuations directly influence operational efficiency, safety, and infrastructure planning.

The remainder of this paper is structured as follows: Section 2 provides an overview of the data. Section 3 provides the modeling approach of multiple linear regression with polynomial transformation. We then present our results in Section 4 and discuss the implications, limitations, and future research directions in Section 5.

The data gathering and analysis is done in R (R Core Team 2023) with the following packages: knitr (Xie 2014), tidyverse (Wickham et al. 2019), arrow (Richardson et al. 2024), here (Müller 2020), corrplot (Wei and Simko 2024), kableExtra (Zhu 2024), car (Fox and Weisberg 2019), modelsummary (Arel-Bundock 2022), brms (Bürkner 2021), Metrics (Hamner and Frasco 2018).

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). We choose the datasets from the Government of Canada for their coverage and quality control processes.

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 are 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 gauge, 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 has the 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 D. 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 D.

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 obtained from Canadian Centre for Climate Services (2022) and Meteorological Service of Canada (2023).

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)

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

Table 2: Column Headers of Raw AHCCD Data

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

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: historical weather data (raw_data_climate) and AHCCD weather data (raw_data_ahccd). The analysis spans data collected monthly between August 1959 and August 2010 for training and testing purposes.

The cleaned dataset has a range of weather variables providing detailed monthly observations. The `date` variable is 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

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

. Moreover, 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 and Training/Testing Split

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

The summary statistics of the combined dataset are displayed in Table 10.

The cleaned and combined dataset is split into a training group and a testing group randomly. The training dataset contains 70% of the cleaned dataset and the testing dataset contains 30% of the cleaned dataset. The training dataset is used to fit the model. The test dataset is used to evaluate the model's performance on unseen data.

2.4 Characteristics of Cleaned Data

All variables in the dataset are numeric, the histograms are plotted in Figure 1 and Figure 9. The following section explains the characteristics of these variables.

2.4.1 Skewness in Variables

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.

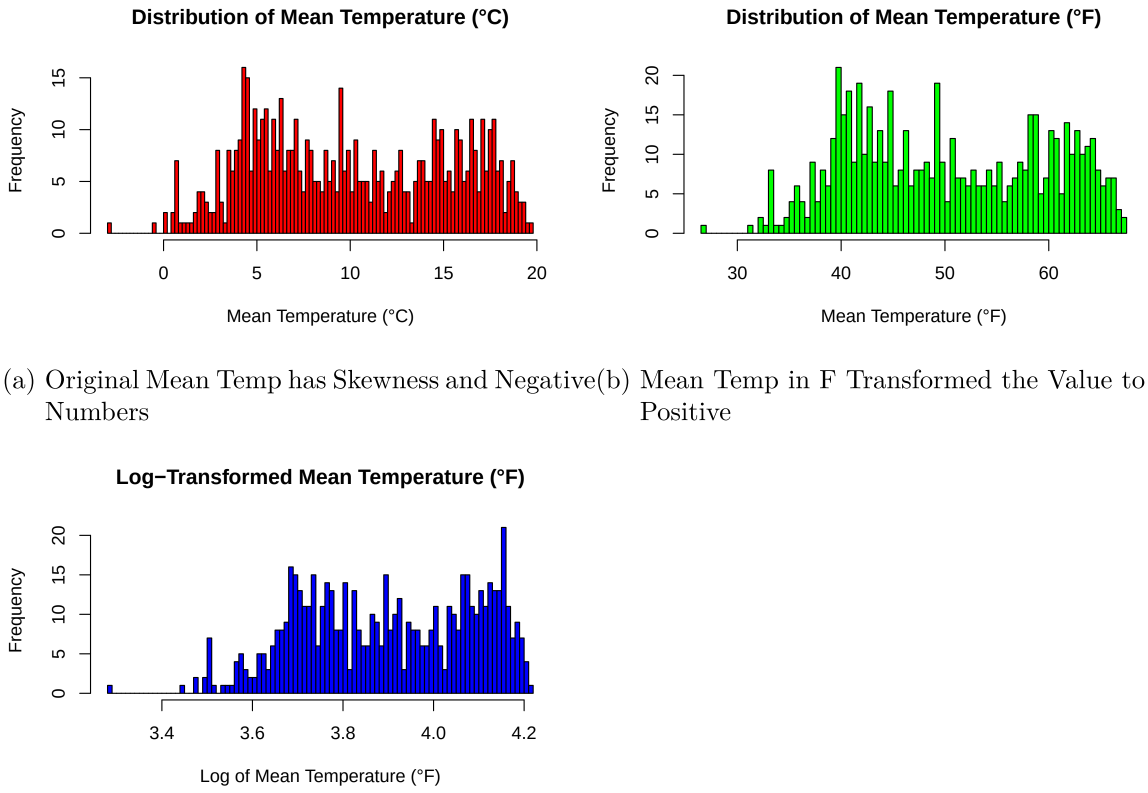
In Figure 9, Transformations are also applied to Wind Speed, Pressure, Total Precipitation and Gust Speed. Total Precipitation has a strong right skew, with most values low and a few extremely high values. We apply log transformation to predictors including wind speed, pressure, and precipitation. For Gust Speed, a Box-Cox transformation is applied to adjust

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

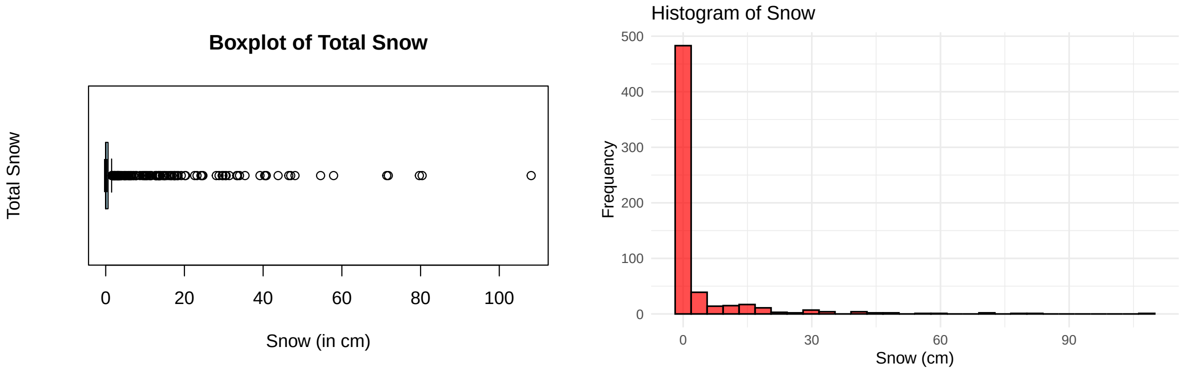
(c) Log-Transformed Data Shows a More Symmetric and Less Skewed Distribution

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

its moderate 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 many observations concentrated at zero and a few extreme outliers far above most of the data. This distribution suggests that the variable snow contains excessive structural zeros, representing instances where no snowfall occurred.



(a) A Large Proportion of Observations With No Snowfall

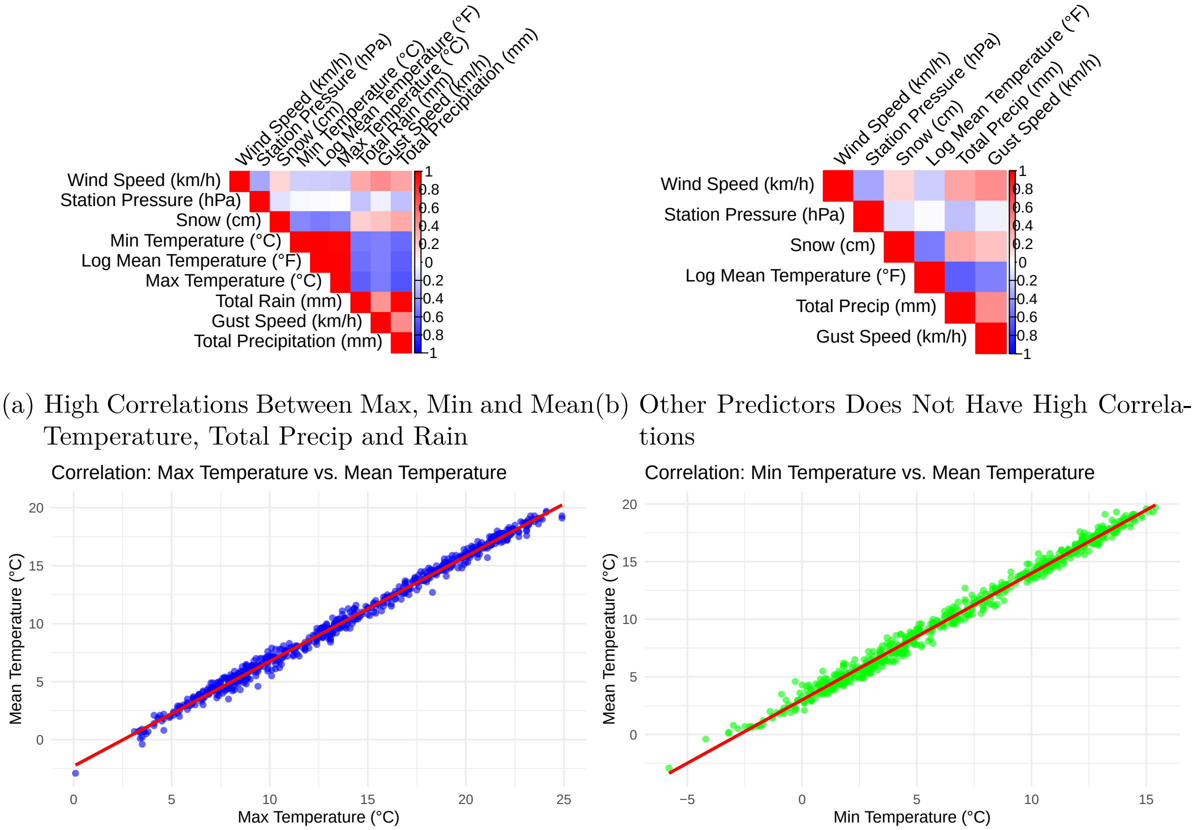
(b) Prevalence of Zero Values and a Skewed Pattern in the Non-zero Snowfall Measurements

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

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

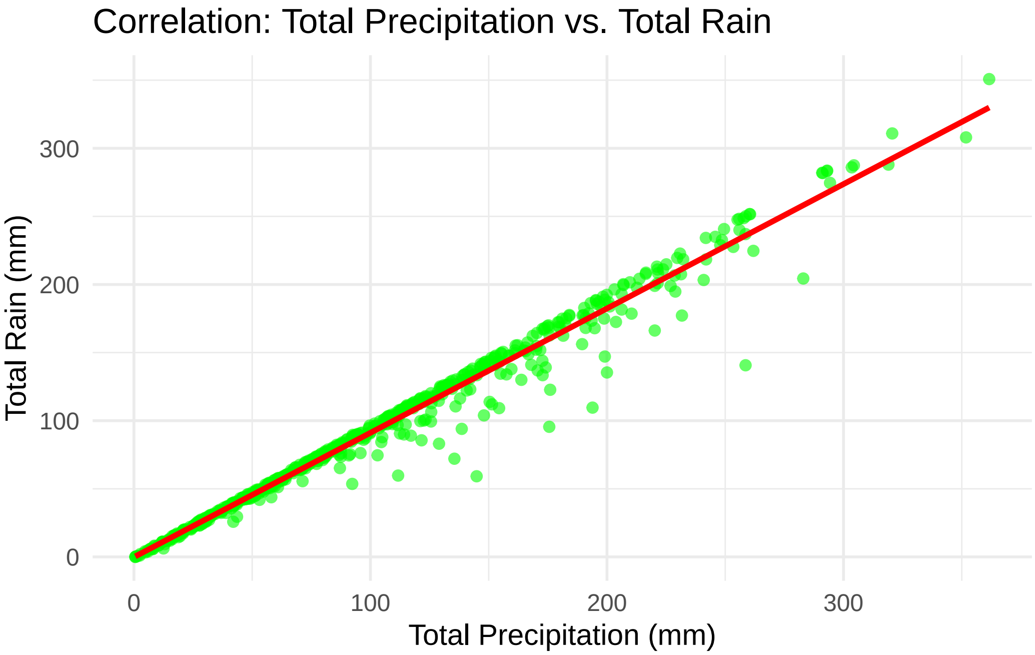


Figure 4: Precipitation and Rain Have High Correlations

3 Model

The goal of our modelling strategy is to find a model that can predict temperature changes with other weather data as predictors such as wind speed, pressure, precipitation, and gust speed.

We build a linear model, a Bayesian model, and a linear model with 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. We choose linear regression instead of the general model because all the variables are numeric. According to Kumar (2023), linear regression is applicable when the response is continuous and normally distributed, which is more applicable to our dataset.

3.1 Model Set-up

The final model we chose is the linear model with polynomial transformation.

This polynomial linear regression model predicts the log-transformed mean temperature (`log_mean_temp`) based on quadratic polynomial transformations of four predictors:

- 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`).

The model introduces non-linear relationships by including polynomial terms up to the second degree (quadratic) for each predictor.

The model is fitted with R (R Core Team 2023), the Bayesian Model in Section B is built using brms (Bürkner 2021) and rstanarm (Brilleman et al. 2018).

3.2 MLR Polynomial Model

The Model LP is mathematically expressed as:

$$\begin{aligned}
\text{Log Mean Temperature} &= \beta_0 + \beta_1 \cdot \text{Log Wind Speed} + \beta_2 \cdot (\text{Log Wind Speed})^2 && (1)\\
&+ \beta_3 \cdot \text{Log Pressure} + \beta_4 \cdot (\text{Log Pressure})^2 && (2)\\
&+ \beta_5 \cdot \text{Box-Cox Total Precipitation} + \beta_6 \cdot (\text{Box-Cox Total Precipitation})^2 && (3)\\
&+ \beta_7 \cdot \text{Log Gust Speed} + \beta_8 \cdot (\text{Log Gust Speed})^2 && (4)\\
&+ \epsilon && (5)
\end{aligned}$$

Where:

1. **Intercept (β_0):**
   - Represents the baseline log-transformed mean temperature when all predictors are zero.
2. **Predictors and Their Polynomial Terms:**
   - **Wind Speed (Log Wind Speed):**
     - $\beta_1 \cdot$ Log Wind Speed models the linear effect of wind speed.
     - $\beta_2 \cdot (\text{Log Wind Speed})^2$ captures the quadratic (non-linear) effect of wind speed.
   - **Atmospheric Pressure (Log Pressure):**

  - $\beta_3 \cdot$ Log Pressure models the linear effect of atmospheric pressure.
  - $\beta_4 \cdot$ (Log Pressure)^2 captures the quadratic effect of atmospheric pressure.
- **Total Precipitation (Box-Cox Total Precipitation):**
  - $\beta_5 \cdot$ Box-Cox Total Precipitation models the linear effect of total precipitation.
  - $\beta_6 \cdot$ (Box-Cox Total Precipitation)^2 captures the quadratic effect of total precipitation.
- **Gust Speed (Log Gust Speed):**
  - $\beta_7 \cdot$ Log Gust Speed models the linear effect of gust speed.
  - $\beta_8 \cdot$ (Log Gust Speed)^2 captures the quadratic effect of gust speed.

3. **Residual Error (ϵ):**
   - Represents the variation in the log-transformed mean temperature that is not explained by the predictors.

The purpose of including polynomial terms in this model is to capture non-linear relationships between the predictors and the response variable, allowing the model to fit more complex, curved patterns that a purely linear model cannot accommodate. This enhances the model's predictive performance, making it capable of explaining variance that would otherwise remain unaccounted for in a simple linear regression, while still maintaining the interpretability and simplicity of a linear regression framework.

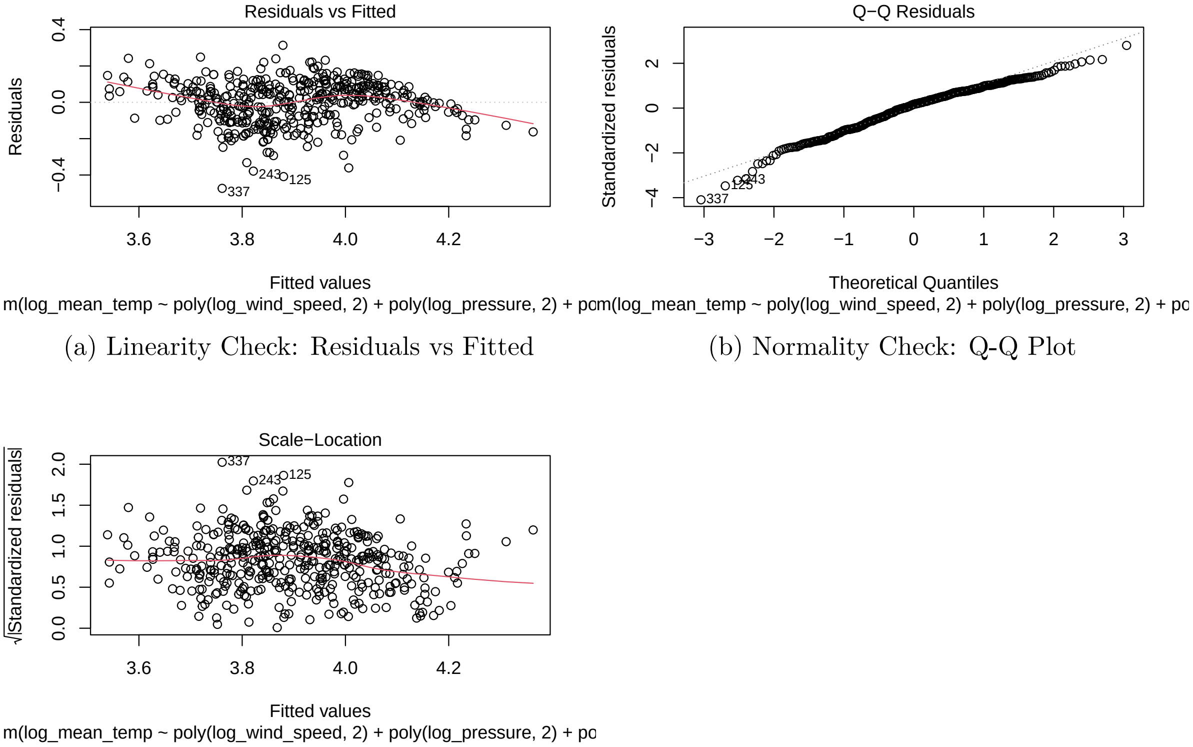
3.3 Model Validation

The linear regression model was chosen to analyze the relationship between wind speed, atmospheric pressure, total precipitation, and gust speed with the log-transformed mean temperature. This model's main goal is to quantify each predictor's effect on the response variable and to make predictions. The linear model is suitable because the relationship between the predictors and the response variable was found to be linear after applying logarithmic and Box-Cox transformations to address non-linearity and skewness.

3.3.1 Assumptions Fit the Data

The model, which uses linear regression, has four assumptions: linearity, homoscedasticity, and approximate normality of residuals.

The linearity assumption states that the relationship between each predictor and the response variable is linear. This assumption was evaluated by examining scatterplots of the predictors against the response variable and by analyzing residuals versus fitted values. As shown in Figure 5a, The red smooth line indicates only a slight curvature, and the residuals scatter randomly around the 0 line. The resulting diagnostic plots showed no obvious systematic patterns, indicating that the linearity assumption is satisfied.



(a) Linearity Check: Residuals vs Fitted

(b) Normality Check: Q-Q Plot

(c) Homoscedasticity Check: Scale-Location

Figure 5: Model LP meets all assumtions of linear regression

Homoscedasticity requires that the variance of the residuals remains constant across all levels of the predictors. In Figure 5c, the points (standardized residuals) are evenly spread across the range of fitted values. There is no clear fan or funnel shape, suggesting the variance of the residuals is constant, which supports the assumption of homoscedasticity.

Linear regression assumes that the residuals follow a normal distribution. This assumption was assessed using Q-Q plots in Figure 5b, where the residuals were compared to a theoretical normal distribution. Most points aligned closely with the diagonal line, indicating that the residuals are normally distributed.

Independence assumes that each observation in the dataset is unrelated to others. This was ensured during data collection or preparation according to ECCC's methodology as discussed in Section D.

Multicollinearity occurs when predictors are highly correlated, making it difficult to isolate their individual effects. Variance Inflation Factor (VIF) quantifies how much multicollinearity inflates the variance of the estimated regression coefficients. The Generalized VIF (GVIF) is used because each polynomial term represents more than one degree of freedom. The result in Table 4 are below the commonly accepted threshold of 5, indicating low multicollinearity in the predictors.

Table 4: VIF Value Indicates Low Multicollinearity

	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	1.129660

3.3.2 Validation on Test Data

We validate our Model LP with test data to evaluate the model's generalizability and performance on unseen data.

The R2 value calculated on the test data ($R^2 = 0.59$) is very close to the R2 value of the model on the training data ($R^2 = 0.61$), as discussed in Section 4.2.6 This similarity indicates that the model generalizes well to unseen data, as its performance on the test set is consistent with its performance during training. The training MSE of 0.0137 and test MSE of 0.0158 indicate the model performs well on both datasets with minimal overfitting. The small difference between the two R2 values and the small difference between the errors suggests good generalization to unseen data.

Table 5: R2 and MSE value on test data is close to train data

Metric	Value
SSR	2.8960166
TSS	7.0693079
R2	0.5903394
R2-Train	0.6104151
MSE-Train	0.0136690
MSE-Test	0.0158252

3.4 Model Justification

The decision to use a linear regression model with polynomial terms for this analysis is grounded in the need to capture non-linear relationships between the predictors and the log-transformed mean temperature (log_mean_temp). Initial exploratory data analysis and diagnostic plots indicated that the relationships between the predictors, such as log-transformed wind speed, pressure, total precipitation, and gust speed, and the response variable were not strictly linear. Applying second-degree polynomial terms allows the model to account for curvature and interactions within the data while maintaining interpretability.

The polynomial linear regression model was selected over alternative non-linear methods due to its balance between complexity and simplicity. Polynomial terms extend the linear model's capacity to capture non-linear patterns while preserving the interpretability of regression coefficients. The inclusion of transformations, such as logarithmic and Box-Cox, further improves the model by addressing issues of skewness and heteroscedasticity observed in the raw data. By incorporating quadratic terms for each predictor, the model gains flexibility in representing real-world data relationships, reducing bias, and addressing systematic patterns in residuals that may indicate non-linearity.

Finally, the model assumptions were evaluated in Section 3.3. Residual plots confirmed that the linearity assumption was met, while the Scale-Location plot and diagnostic tests supported the homoscedasticity of residuals. Variance Inflation Factor (VIF) values for the predictors were all well below the threshold, indicating no significant multicollinearity. These results ensure that the model remains valid and reliable for inference and prediction.

4 Results

Our results are summarized in Table 6, which displays the estimate, standard error, and p-value for each coefficient.

Table 6: Summary of Model LP

	(1)
(Intercept)	3.899
	(0.006)
	(<0.001)
poly(log_wind_speed, 2)1	0.260
	(0.144)
	(0.072)
poly(log_wind_speed, 2)2	−0.449
	(0.126)
	(<0.001)
poly(log_pressure, 2)1	−0.388
	(0.131)
	(0.003)
poly(log_pressure, 2)2	−0.973
	(0.125)
	(<0.001)
poly(total_precipitation_boxcox, 2)1	−2.016
	(0.141)
	(<0.001)
poly(total_precipitation_boxcox, 2)2	0.218
	(0.121)
	(0.072)
poly(log_gust_speed, 2)1	−0.980
	(0.145)
	(<0.001)
poly(log_gust_speed, 2)2	0.347
	(0.123)
	(0.005)
Num.Obs.	427
R2	0.610
R2 Adj.	0.603
AIC	−601.2
BIC	−560.6
Log.Lik.	310.589
RMSE	0.12

4.1 Temperature Conversion

In the analysis, the mean temperature was log-transformed to stabilize variance and improve model interpretability. However, the predicted values from the model are in the log-transformed scale and need to be transformed back to their original scale for practical interpretation. The back-transformation involves exponentiating the predictions to return them to the temperature scale. It is converted with:

$$\text{Temperature (Fahrenheit)} = e^{\text{log\_mean\_temperature}} \tag{6}$$

Since the original temperature data is in Fahrenheit, the results must first be converted from the logarithmic scale to Fahrenheit, and then subsequently to Celsius using the standard conversion formula:

$$\text{Temperature (Celsius)} = (\text{Temperature (Fahrenheit)} - 32) \times \frac{5}{9} \tag{7}$$

In this paper, we interpret the result in the model based on the log transformed mean temperature. For practice applications, the temperature predicted must be converted.

4.2 Coefficients

Figure 6 illustrates the estimated coefficients for the polynomial regression model, along with their 95% confidence intervals. Positive coefficients indicate predictors that increase the log-mean temperature, while negative coefficients represent predictors associated with a decrease, with confidence intervals crossing zero suggesting lower statistical significance.

4.2.1 Intercept Indicates a Strong Baseline In Temperature

The intercept of the model is estimated at **3.899**, representing the predicted value of the log-transformed mean temperature when all predictors are at their mean (after centering and scaling in the polynomial terms). This value serves as the baseline against which the effects of all predictors are measured. A significant intercept indicates the model reliably predicts the baseline log-mean temperature.

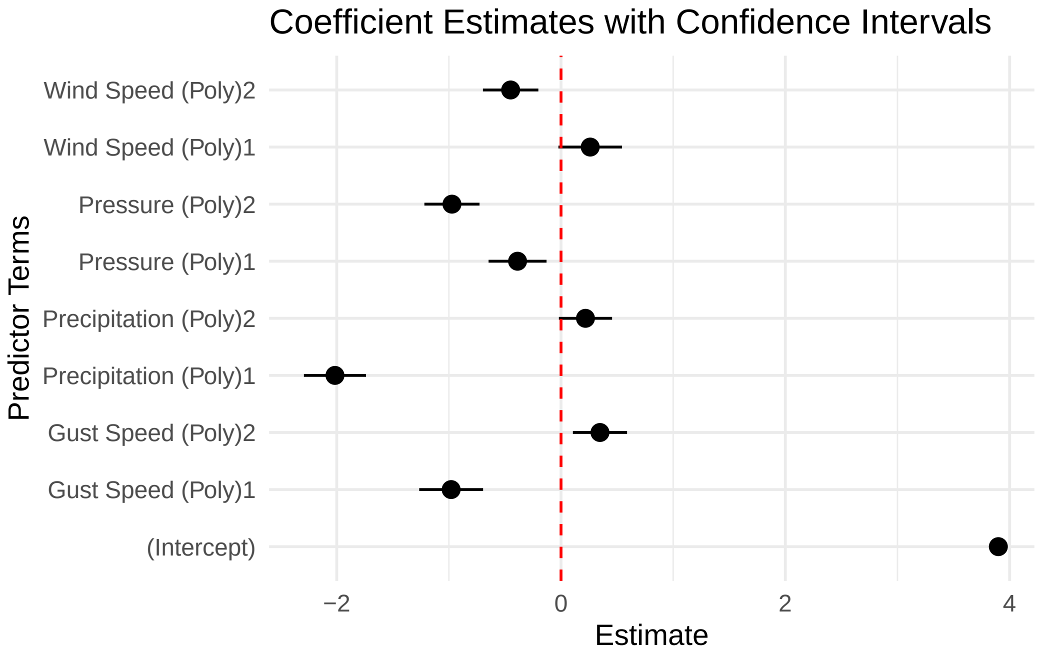


Figure 6: Intercept Indicates a Strong Baseline (95% CI)

4.2.2 Wind Speed: Non-Linear Increase and Decline in Temperature

The first-degree polynomial term for log-transformed wind speed (poly(log_wind_speed, 2)1) has a positive coefficient of **0.260**, suggesting that increasing wind speed is associated with a slight increase in log(mean_temp). However, the second-degree term (poly(log_wind_speed, 2)2) has a negative coefficient of **-0.449**, indicating a reversing effect at higher wind speeds. This quadratic relationship reflects non-linear relationships between wind speed and temperature.

4.2.3 Atmospheric Pressure: Consistent Negative Impact on Temperature

Log-transformed atmospheric pressure shows a consistently negative relationship with log(mean_temp). The first-degree term (poly(log_pressure, 2)1) has a coefficient of **-0.388**, and the second-degree term (poly(log_pressure, 2)2) is more negative at **-0.973**. This indicates that higher atmospheric pressure is associated with a reduction in log(mean_temp), and the quadratic term amplifies this effect, particularly at extreme values of pressure.

4.2.4 Total Precipitation: The Strongest Negative Effect

The Box-Cox transformed total precipitation has the strongest impact on log(mean_temp) among all predictors. The first-degree term (poly(total_precipitation_boxcox, 2)1) has a significant negative coefficient of **-2.016**, implying that higher precipitation reduces log(mean_temp) substantially. The second-degree term (poly(total_precipitation_boxcox, 2)2) is positive at **0.218**, indicating a slight mitigation of this negative effect at extreme precipitation levels.

4.2.5 Gust Speed: Decreases Temperature but Shows Curvature

The first-degree term for log-transformed gust speed (poly(log_gust_speed, 2)1) has a negative coefficient of **-0.980**, showing that increasing gust speed reduces log(mean_temp). However, the second-degree term (poly(log_gust_speed, 2)2) has a positive coefficient of **0.347**, suggesting a minor curvature in the relationship where the negative effect is less pronounced at higher gust speeds.

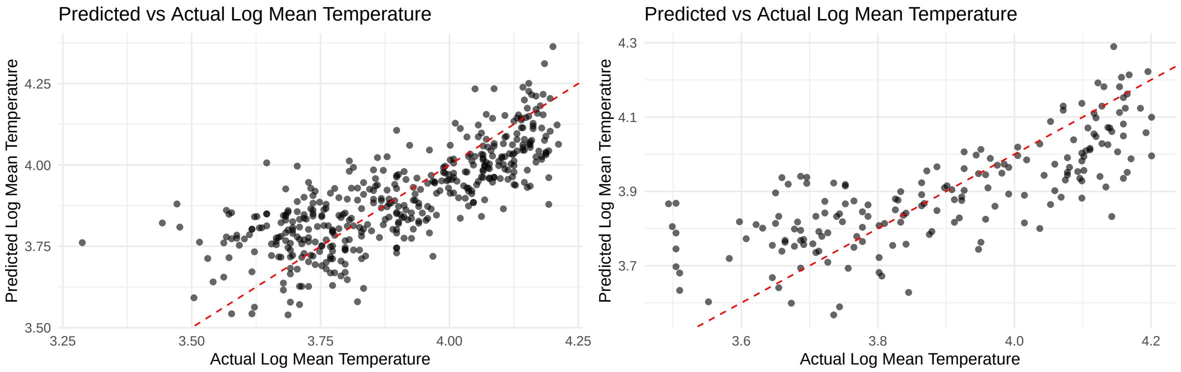
4.2.6 Model Metrics: Explains 61% Variance with Low RMSE

The model explains a significant portion of the variance in log(mean_temp), as indicated by an R^2 value of **0.610** and an adjusted R^2 of **0.603**. These values demonstrate that approximately 61% of the variability in log(mean_temp) is explained by the predictors, even after accounting for the complexity of the model.

The model has a low RMSE of **0.12**, indicating accurate predictions with minimal average error. The AIC (**-601.2**) and BIC (**-560.6**) values suggest a strong fit compared to alternative models.

4.3 Predicted vs. Actual Plot Shows Consistent Performance of the Model

As displayed in Figure 7, the predicted vs. actual plots for both the training and test datasets demonstrate a strong alignment along the diagonal. In Figure 7a, The points closely align along the red diagonal line, indicating that the model predictions are highly accurate for the training dataset. In Figure 7b, the test dataset shows a comparable trend, with predictions aligning well with actual values, suggesting the model generalizes effectively to unseen data. These results confirm that the model provides a good fit for the response variable in both datasets.



(a) The model is accurate in the train data

(b) The model is accurate in the test data

Figure 7: Predicted vs. Actual plot shows good fit

4.4 P-value

In Table 6, the p-value in regression results show that several coefficients are statistically significant at the 0.05 level, and all coefficients are significant at the 0.1 level. The intercept and terms for the second-degree polynomials of log_pressure and total_precipitation_boxcox are highly significant ($p < 0.001$). Similarly, the second-degree term of log_gust_speed ($p = 0.00496$) and the first-degree term of log_pressure ($p = 0.00337$) are significant at the 0.01 level. Marginal significance is observed for the first-degree terms of log_wind_speed ($p = 0.0721$) and total_precipitation_boxcox ($p = 0.0716$).

5 Discussion

This paper examines the factors influencing mean temperature at Vancouver International Airport using data collected from Environment and Climate Change Canada. A polynomial regression model was developed to predict log-transformed mean temperature based on atmospheric predictors, including wind speed, atmospheric pressure, total precipitation, and gust speed. These variables were transformed using logarithmic or Box-Cox methods to stabilize variance and address non-linearity in their relationships with temperature.

The analysis included detailed data preparation, combining raw observational datasets with adjusted and homogenized climate data. The dataset was cleaned, standardized, and split into training and test groups to ensure robust model validation. Exploratory data analysis found skewness and zero inflation in several predictors, needing transformations for improved model suitability.

The study focused on building a polynomial regression model. By incorporating quadratic terms for all predictors, the model captured non-linear relationships and explained 61% of the variance in mean temperature. Model performance was evaluated through diagnostic tests, residual analysis, and comparisons of predicted versus actual values. Metrics such as RMSE confirmed the model's accuracy and generalizability to unseen data.

5.1 Summary of Results

5.1.1 Precipitation Leads, While Wind and Pressure Add Complexity

This study identified total precipitation as the strongest predictor of temperature variation, demonstrating a robust inverse relationship with mean temperature. The analysis showed that wind speed and gust speed have non-linear effects, with the direction and magnitude of their impact dependent on the quadratic terms included in the model. Atmospheric pressure consistently reduced temperature, with the quadratic term further emphasizing this effect at extreme levels of pressure. The inclusion of polynomial terms allowed the model to address curvatures in relationships that linear terms alone could not capture. The model performed well, explaining 61% of the variance and maintaining its accuracy across training and test datasets. The low RMSE confirms the model's ability to approximate observed temperatures effectively.

These results reflect the complexity of meteorological interactions influencing temperature. While total precipitation emerged as the dominant factor, the important roles of wind and pressure underscore the importance of considering multiple predictors simultaneously. Together, these variables reflect the dynamics influencing temperature at Vancouver International Airport (YVR).

5.1.2 Non-Linear Effects of Atmospheric Predictors on Temperature

The study highlights a non-linear relationship between predictors, such as wind speed, and mean temperature. For instance, at moderate wind speeds, the findings suggest a slight increase in temperature, potentially due to enhanced mixing of atmospheric layers that can bring warmer air from aloft. However, as wind speed increases further, its effect reverses, leading to a decrease in temperature. This could be attributed to the cooling influence of stronger winds dispersing heat more effectively or increasing evaporation rates. These findings suggest the importance of accounting for non-linear effects in predictive climate models, as linear assumptions would overlook such complex dynamics. Further research could inspect how wind speed interacts with other variables, such as humidity or surface characteristics, to influence temperature.

5.1.3 Dominance of Precipitation in Predicting Temperature Variation

The analysis identifies precipitation as the most influential factor in predicting temperature changes, with a strong inverse relationship observed. This result is consistent with the idea that precipitation events often coincide with increased cloud cover, which limits solar radiation and leads to cooler surface temperatures. Additionally, evaporative cooling during and after precipitation events may further contribute to this effect. The quadratic nature of the precipitation term suggests diminishing effects at extreme levels, which could reflect saturation in cooling mechanisms or localized anomalies in precipitation patterns. These findings emphasize the role of hydrological processes in modulating regional climate and provide a basis for refining climate prediction models to incorporate precipitation variability more effectively.

5.2 Comparing with Previous Studies

5.2.1 Our finding does not aligh with Visser

The findings diverge from the work of Visser et al. (2021), who identified a positive association between precipitation intensity and temperature in Australia. This contrast highlights the importance of regional climatic differences and suggests that factors such as geographic features, prevailing weather patterns, and local topography may alter these relationships. For example, Vancouver's coastal climate likely moderates precipitation's impact on temperature compared to Australia's arid and semi-arid conditions.

5.2.2 Our finding align with Wills

Our results align with Wills et al. (2022), who identified atmospheric pressure as a cooling factor in equatorial Pacific regions. This consistency reinforces the broader applicability of pressure-temperature relationships in diverse climatic contexts. The agreement with Wills suggests that pressure effects may operate at larger spatial scales, transcending localized conditions.

5.3 Weaknesses and Future Directions

5.3.1 Limitations on the Model LP

While the polynomial linear regression model captures much of the relationship between the predictors and the log-transformed mean temperature, some limitations remain. Figure 5a indicates that the relationship is not completely linear, with minor deviations suggesting that the model may not fully account for certain non-linearities. Additionally, Figure 5c shows slight evidence of heteroscedasticity at the extreme fitted values, indicating that the variance of residuals is not entirely constant. These limitations suggest that the model could benefit from further refinement, such as applying additional transformations, inspecting interaction terms, or using alternative techniques like weighted least squares or robust regression to address potential heteroscedasticity.

5.3.2 Limitation on the Response Log Transformed Mean Temperature

The temperature in the model is transformed, which adds complexity for real world applications. Moreover, the use of mean temperature as the sole response variable, though practical, limits the information gained. Mean temperature smooths out extremes, thereby overlooking information about temperature variability. Extremes, such as maximum and minimum temperatures, play a significant role in understanding climate impacts, particularly in terms of heatwaves or freezing conditions.

Future research could expand the scope of the analysis to include a range of temperature metrics. Analyzing temperature extremes or variability measures, such as diurnal temperature range or seasonal deviations, would provide a richer understanding of atmospheric influences on climate. Incorporating additional response variables could also show interactions between predictors that are not apparent when analyzing mean temperature alone.

5.3.3 Limitation on Using Temperature as the only Response

As Coffel and Horton (2015) discussed, using mean temperature as the sole response variable, while good for general climate modeling, presents limitations when assessing airport safety. Aviation safety is influenced not just by average temperature but also by its variability, particularly extremes that can significantly impact operations. For example, high maximum temperatures reduce air density, affecting aircraft lift during takeoff. Similarly, low minimum temperatures can lead to ice formation on runways and aircraft surfaces, posing safety hazards. Atmospheric conditions that affect aviation, such as wind speed and sea level pressure, are not captured by mean temperature alone. These variables are directly tied to operational challenges and safety considerations.

For instance, wind speed is a major factor in airport operations, influencing takeoff, landing, and overall flight stability. Sudden changes in wind speed or direction, including gusts and crosswinds, can compromise safety and require adjustments in scheduling and runway usage. Analyzing wind speed alongside temperature would provide a better understanding of the conditions that lead to unsafe situations.

Sea level pressure is another important variable, affecting aircraft performance and weather systems. Variations in pressure can lead to storms, precipitation, or clear but frigid conditions that impact runway conditions and aircraft operation. Low-pressure systems often bring intense winds and poor visibility, while high-pressure systems can lead to temperature extremes that influence infrastructure and fuel performance.

Including wind speed and sea level pressure in future analyses would enhance the ability to assess and predict conditions relevant to airport safety. These variables are fundamental to the aviation sector and provide a more detailed view of the atmospheric dynamics that affect operations. Addressing these factors would align the study with the needs of airport safety management and planning.

5.3.4 Limitation on ECCC's methodology

The reliance on Environment and Climate Change Canada (ECCC) adjusted and homogenized climate data introduces methodological constraints. While these datasets are extensively quality-controlled, they are subject to adjustments for station relocations, instrument changes, and missing data imputation as discussed in Section D. Such adjustments, while necessary, could obscure trends or introduce biases that might not reflect current climate conditions. This is especially relevant when studying localized phenomena where raw observational data might show small scale of changes.

To address these concerns, future research could compare adjusted datasets with raw, unprocessed data. Integrating alternative data sources, such as satellite-derived measurements or high-resolution reanalysis products, might complement the ECCC data. Furthermore, temporal limitations of the dataset, such as monthly aggregation, restrict the analysis of short-term weather events. Employing higher temporal resolution data, such as hourly or daily measurements, would allow for a detailed examination of precipitation and temperature interactions over shorter time scales.

5.3.5 Potential Directions for Expanded Analysis

Expanding this research to include airports in different regions around the world would provide a broader understanding of how variables influence airport safety. Airports operate in different climatic and geographic conditions such as coastal locations and high-altitude terrains, have unique challenges influenced by local weather patterns. Investigating these variations could show patterns and trends that are not observable when focusing on Vancouver's airport.

For instance, airports in tropical regions might experience high wind speeds and pressure fluctuations due to frequent storms, whereas those in colder regions are more likely to face challenges related to low temperatures and ice. Conducting similar analyses at airports in different environments could show how predictors like wind speed, sea level pressure, and temperature interact under different climatic regimes.

Additionally, comparing airports with varying altitudes and proximity to water would allow researchers to inspect the role of geographical features in modifying atmospheric conditions. For example, high-altitude airports may be more affected by low atmospheric pressure, while coastal airports might experience more significant impacts from wind and precipitation variability.

Appendix

A License

Contains information licensed under the Open Government Licence – Canada.

B Additional 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:

$$\begin{aligned}
\text{mean\_temp\_F} = {} & \beta_0 + \beta_1 \cdot \text{wind\_speed} + \beta_2 \cdot \text{total\_precipitation} + \beta_3 \cdot \text{snow} && (8)\\
& + \beta_4 \cdot \text{pressure\_station} + \beta_5 \cdot \text{max\_temp} + \beta_6 \cdot \text{min\_temp} && (9)\\
& + \beta_7 \cdot \text{total\_rain} + \beta_8 \cdot \text{gust\_speed\_km\_h} + \epsilon && (10)
\end{aligned}$$

- 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 7.

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 7: 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.

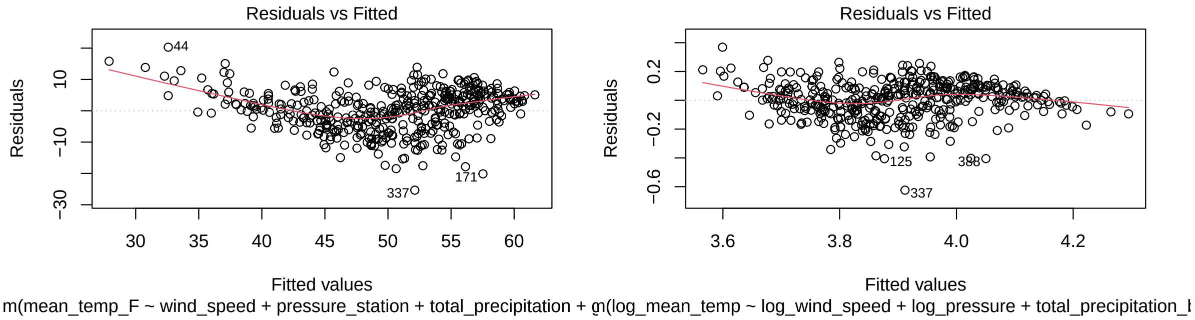
This model predicts mean temperature (mean_temp_F) based on a subset of predictors: wind speed (wind_speed), station pressure (pressure_station), total precipitation (total_precipitation), and gust speed (gust_speed_km_h).

$$\begin{aligned} \text{mean\_temp\_F} &= \beta_0 + \beta_1 \cdot \text{wind\_speed} + \beta_2 \cdot \text{pressure\_station} && (11) \\ &\quad + \beta_3 \cdot \text{total\_precipitation} + \beta_4 \cdot \text{gust\_speed\_km\_h} + \epsilon && (12) \end{aligned}$$

- 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 7, all predictors have small coefficients, suggesting incremental effects on the response variable. The large standard errors of some coefficients, such as the intercept, indicate potential variability or noise in the data. For instance, from Figure 1 and Figure 9, we observe skewness and non-normal distribution in both predictor and the response. According to Figure 8a, 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 8: 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:

$$\begin{aligned}\text{log\_mean\_temp} &= \beta_0 + \beta_1 \cdot \text{log\_wind\_speed} + \beta_2 \cdot \text{log\_pressure} \qquad (13)\\ &+ \beta_3 \cdot \text{total\_precipitation\_boxcox} + \beta_4 \cdot \text{log\_gust\_speed} + \epsilon \qquad (14)\end{aligned}$$

- 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 8b, 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 7, 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:

$$\begin{aligned}
\text{log\_mean\_temp} &\sim \mathcal{N}(\mu, \sigma^2), && (15)\\
\mu &= \beta_0 + \beta_1 \cdot \text{log\_wind\_speed} + \beta_2 \cdot \text{log\_pressure\_station} && (16)\\
&\quad + \beta_3 \cdot \text{total\_precipitation\_boxcox} + \beta_4 \cdot \text{log\_gust\_speed}. && (17)
\end{aligned}$$

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 the `brms` package (Bürkner 2021). It uses:

- 4 chains for convergence,
- 2000 iterations per chain to ensure stability,
- 4 cores for parallel computation, enabling efficient sampling.

Unlike linear regression, which provides point estimates and assumes fixed parameter values, Bayesian regression incorporates prior knowledge and generates posterior distributions, giving a probabilistic framework that quantifies uncertainty in parameter estimates.

B.5 The Linear Model Has Better Fit than Bayesian Model

We calculate the RMSE value and MAE value of the Bayesian model, and the Linear model based on the test dataset, because it evaluates how well the model performs on data it has never seen before, providing a realistic measure of predictive accuracy.

The RMSE measures the average squared difference between the observed (y_i) and predicted $(\hat{y}_i)$ values. It is calculated as:

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

Where:

Table 8: Summary of Bayesian Model and Comparison of RMSE / MAE on Test Data

	(1)
b_Intercept	12.942
b_wind_speed	0.009
b_pressure_station	−0.008
b_total_precipitation_boxcox	−0.022
b_log_gust_speed	−0.232
sigma	0.130
Num.Obs.	427
R2	0.522
R2 Adj.	0.511
ELPD	259.9
ELPD s.e.	19.4
LOOIC	−519.9
LOOIC s.e.	38.7
WAIC	−519.9
RMSE	0.13

Model	RMSE	MAE
Linear	3.548929	2.766038
Bayesian	3.556934	2.777702

- y_i: The actual value of the i-th observation.
- $\hat{y}_i$: The predicted value for the i-th observation.
- n: The total number of observations.

The MAE measures the average absolute difference between the observed (y_i) and predicted ($\hat{y}_i$) values. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Where:

- y_i: The actual value of the i-th observation.
- $\hat{y}_i$: The predicted value for the i-th observation.
- n: The total number of observations.

By comparing the RMSE and MAE results in Table 8, the Linear Model L has slightly lower RMSE and MAE compared to the Bayesian model, suggesting it has a marginally better fit for the given data. As a result, we chose the Linear Model over the Bayesian model.

B.6 Polynomial Linear Model

Since we choose the Linear Model L over Bayesian Model, we inspect the residual plot of the linear model. The residual plot in Figure 8b 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. The detail of the new model is discussed in Section 3.

Table 10: 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

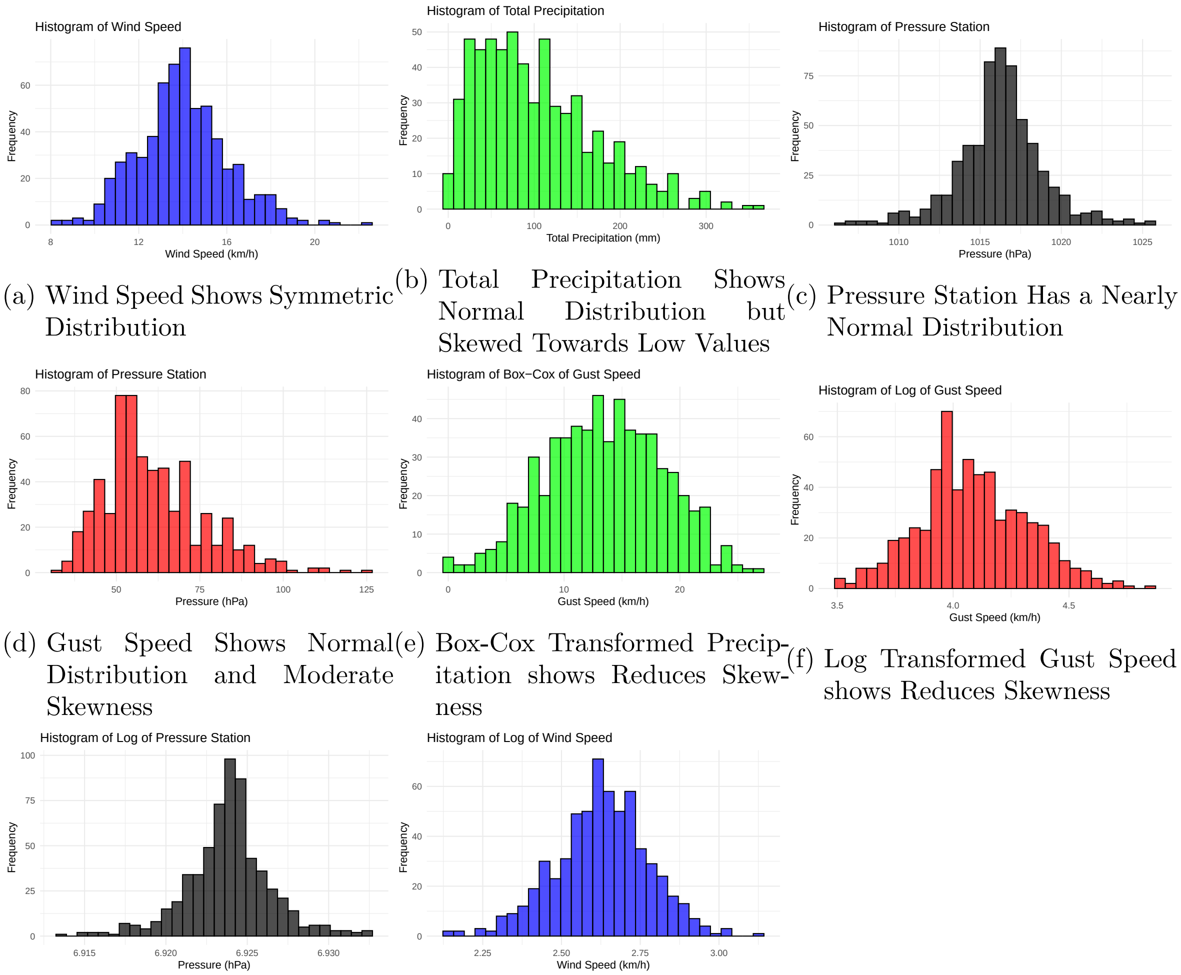
Mean Temp	Rain	Max Gust Speed	Mean Temp in F	Log of Mean Temp
Min. :-2.90	Min. : 0.00	Min. : 33.00	Min. :26.78	Min. :3.288
1st Qu.: 5.50	1st Qu.: 43.65	1st Qu.: 51.00	1st Qu.:41.90	1st Qu.:3.735
Median : 9.60	Median : 80.45	Median : 59.00	Median :49.28	Median :3.898
Mean :10.13	Mean : 93.94	Mean : 61.11	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

C Additional Data Details

The summary statistics of the cleaned data are shown in Table 10.

As discussed in Section 2, the histogram of variables before and after transformation is displayed in Figure 9.



(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 9: Other Variables Show Normal Distribution

D 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.

D.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 climatological 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).

D.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).

D.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.

D.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).

D.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.

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/.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Bürkner, Paul-Christian. 2021. "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software* 100 (5): 1–54. https://doi.org/10.18637/jss.v100.i05.

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.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third.

Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

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/.

Hamner, Ben, and Michael Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. https://CRAN.R-project.org/package=Metrics.

Kumar, Ajitesh. 2023. "GLM Vs Linear Regression: Difference, Examples." *Analytics Yogi*. https://vitalflux.com/glm-vs-linear-regression-difference-examples/.

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

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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.

Wei, Taiyun, and Viliam Simko. 2024. *R package 'corrplot': Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

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.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

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.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.