Canada Wildfire Analysis and Prediction: Project Report

Problem Definition

Wildfires have a profound impact on Canada's ecosystems, economy, and public safety, with climate change driving an increase in their frequency and severity. By analyzing wildfire data from NASA's Fire Information for Resource Management System (FIRMS), which provides insights into fire brightness and locations, and combining it with historical weather data from Open-Meteo, we aim to explore patterns and build predictive models. These models will help understand and forecast wildfire occurrences based on factors such as location, weather conditions, and time.

The datasets from FIRMS and Open-Meteo are substantial, often exceeding the capacity of a single processor. To effectively analyze this big data, we will leverage distributed computing using Python and Apache Spark, enabling scalable and efficient data processing.

Integrating satellite data from NASA FIRMS with weather data from Open-Meteo posed significant challenges. The datasets differed in temporal and spatial resolutions, requiring meticulous alignment and preprocessing to ensure accuracy. Temporal variations, such as mismatched timestamps, and spatial discrepancies, like differing grid sizes, had to be resolved to create a unified dataset suitable for analysis. Another challenge was understanding the multivariate dependencies between environmental factors such as dryness, temperature, and precipitation. These variables interact in complex ways, and capturing their interdependencies required advanced analysis techniques. Identifying meaningful patterns while accounting for the influence of multiple factors added further complexity. Finally, optimizing predictive models was critical when working with high-dimensional and noisy data. Selecting algorithms that could balance computational efficiency and accuracy, while minimizing the risk of overfitting, demanded careful experimentation and tuning. Achieving robust predictions required iterative testing and validation to ensure generalization to unseen data.

Methodology

Data Sources:

For access to historical data, Open Meteo's api was used. Instead of focusing on hundreds of cities across Canada, our team decided on the cities that usually have the most wildfires. The cities which we chose were: Fort McMurray, Kelowna, Kamloops, Prince George, Vancouver Island, Lytton, Penticton, Williams Lake, Grande Prairie, and Edson. The dataset included features such as date, city, latitude, longitude, variations of temperature, precipitation types and amounts, and wind speed.

In order to access historical wildfire data, we sourced it from NASA FIRMS (MODIS). This dataset includes wildfire-specific data such as brightness, frp, confidence, and bright_t31. After requesting data for Canada, we received the data and joined it with weather data from Open Meteo. This join occurred on longitude, latitude, and date. This join resulted in being able to view weather statistics from days with an active wildfire and non-active wildfires.

Data Prep:

While we have significant and important data for wildfire analysis, adding more meaningful features would result in a more thorough analysis. There are many factors which play a big part in wildfire occurrence, such as soil moisture. When soil moisture is low, we can assume that the vegetation is also dry, which leads to stronger wildfires as dry vegetation acts as a fuel for fire. If the value for soil moisture is higher, we can make estimations for precipitation levels, seasonal changes, and temperature approximations.

Another important feature for wildfire analysis is the intensity of a wildfire. By combining MODIS data: brightness, confidence, and frp, the result yields intensity. Brightness tells us how intense the light is coming from the wildfire. Confidence delivers how certain the satellite is for the occurrence of the wildfire. FRP is a measure of how much heat is radiating from the fire itself. Then, the result, intensity, gives us the combined assumption about the fire itself. A higher intensity value means the fire is burning hotter and producing very high heat. A lower value could possibly reflect a weak or dying wildfire.

By adding all the varieties of precipitation and their respective amounts, we designed a new feature called cumulative precipitation. This assists in many aspects of a wildfire such as intensity and soil moisture. Higher levels of precipitation often dictate lower chances of a wildfire happening. Lower levels contribute to low soil moisture, which in return can result in more wildfire occurrences and higher intensity fires.

While the summary above does not include all of the new features which were added during feature engineering, they dictate the thought behind our analysis. Other features which were not mentioned include: humidity, dryness, risk, and precipitation-sun ratio. All these features were constructed for more meaningful analysis and each plays its importance in wildfire occurrence.

Data Analysis:

Our data analysis was split into 3 sections: feature analysis, analysis per city, and yearly trends. Feature analysis focused on which features correlated to more intense fires or a wildfire event. During this analysis, finding out which feature was the best for wildfire prediction was also calculated. Majority of this analysis occurred in Spark using dataframes aggregations. Plots were constructed using matplotlib and seaborn. Each data frame was transformed into a pandas dataframe before plotting. Another aspect which was mentioned before was feature importance. Our team conducted 2 feature selections using PCA and a RandomForestRegressor.

Analysis by city was also one of our main focuses. Our approach for this analysis was to compare cities and see how their wildfires differed. While wildfire occurrence was our main topic for this, we also wanted to compare a variety of features and their importance to each city's fire risk. Another important aspect to our analysis was yearly trends. By viewing the different changes over the years per each city, we can visualize how the overall changes in climate contribute to either stronger or weaker wildfire seasons.

Machine Learning:

Machine learning was also split into 3 sections: Regression, Classification, and Neural Networks. For classification, both RandomForestClassifier and GBTClassifier were used in a Spark setting. Before any optimization, both models held an accuracy of 87% for predicting a wildfire event. After hyperparameter tuning for both models, they both yield close to 95% in wildfire predictions. The hyperparameter tuning for the GBTClassifier includes learning rate alterations and for RandomForestClassifier: changing values for the number of trees and the maximum size of each tree.

The neural network regressor was exceptionally effective, achieving an impressive coefficient of determination (R^2) of 0.991 and a Root Mean Squared Error (RMSE) of 15.382, highlighting its precision in predicting brightness levels. XGBoost also performed robustly with an RMSE of 16.14 and an R^2 of 0.99, demonstrating high predictive accuracy. In contrast, LightGBM showed decent performance, though slightly less effective, with an RMSE of 15.61 and an R^2 of 0.837. Other regression models, including Linear Regression and Random Forest, did not perform as well, reflecting higher error values in their predictions. Additionally, our neural network classifier, a Multilayer Perceptron (MLP), after hyperparameter tuning, reached an accuracy of 0.999, showing excellent performance without any signs of overfitting, underscoring the strength of our predictive modeling approach.

Repository:

Our GitHub repository includes our entire data pipeline. This includes scripts for feature engineering and the resulting dataset from the merging of all relevant data. It also holds all analysis code including their respective visualizations. Each machine learning model is also included for further predictions and tuning.

Problems Encountered

At the outset of the project, identifying reliable and relevant data sources was a significant challenge. Early attempts with various datasets revealed issues such as incorrect or missing values, limited support for large downloads, and insufficient documentation. These shortcomings hindered initial progress, requiring a shift toward well-documented and robust datasets like those from NASA FIRMS and Open-Meteo, which aligned better with the project goals.

Data quality also posed hurdles, as missing and noisy data disrupted early modeling efforts. Weather data gaps and outliers in fire brightness values had to be addressed to improve data integrity. Interpolation techniques were implemented for missing weather values, while outlier detection and removal ensured the reliability of brightness data. These preprocessing steps enhanced the foundation for analysis and modeling.

Additionally, not all derived features proved useful for prediction. For example, variables like wind gust showed minimal contribution to model performance. To address this, feature importance rankings were employed to identify the most relevant inputs, and dimensionality reduction techniques streamlined the feature set. Computational challenges further compounded these issues, particularly when training neural networks, which required significant resources. Leveraging distributed environments like PySpark and optimizing batch processing allowed for efficient model training and scalability.

Results

Our analysis revealed that brightness is a strong predictor of wildfire occurrence, particularly under dry and warm conditions. High brightness values, particularly under dry and warm conditions, strongly correlated with extreme weather patterns, indicating that brightness levels can serve as a valuable indicator for fire risk, especially in regions experiencing heightened temperatures and dry spells.

Geographically, cities such as Kamloops, Penticton, and Kelowna exhibited consistently high brightness values and elevated wildfire risk. These findings emphasize the need for targeted interventions and preventative measures in these areas, where environmental and climatic conditions exacerbate the potential for wildfire outbreaks. The combination of high brightness , persistent dryness, and temperature spikes make s these cities particularly vulnerable to wildfires.

From an environmental perspective, increasing temperatures and decreasing precipitation were found to directly influence wildfire patterns. The rising temperatures coupled with reduced rainfall are accelerating the frequency and intensity of wildfires, underscoring the importance of understanding these environmental trends for more effective wildfire management strategies.

For our machine learning model performance, we employed a variety of machine learning models to predict wildfire occurrences based on brightness data. Neural networks proved the most effective, achieving a Root Mean Squared Error (RMSE) of 15.38 and a coefficient of determination (R^2) of 0.991 for brightness prediction. This outperformed other methods in handling the dataset's complexity, offering the most reliable predictions.

For classification tasks, the neural network (multilayer perceptron) has the highest accuracy of 99.1 % followed by the Random Forest model demonstrated the highest accuracy at 94.9%, outperforming logistic regression and support vector machines

(SVM), which performed well but were slightly less precise. In regression tasks, the best performance was multilayer perceptron (neural networks) with r2 score of 0.99 and rmse score of 15.38, followed by XGBoost showed strong performance with an RMSE of 16.14 and an R^2 of 0.99 for brightness prediction, while LightGBM delivered an RMSE of 15.61 and an R^2 of 0.837, making it a good alternative but not quite as effective as XGBoost.

Project Summary

★ **Getting the Data (3 points)**
Data was sourced from NASA FIRMS and Open Meteo APIs. We submitted requests to NASA FIRMS for Canadian data and applied logic to extract data for the top ten cities based on latitude and longitude values. APIs were utilized to obtain weather data using the libraries `openmeteo-requests`, `requests-cache`, `retry-requests`, `numpy`, and `pandas`. We acquired detailed satellite and weather data, including brightness and climate variables.

★ **ETL: Extract-Transform-Load (2 points)**
We performed extensive cleaning and merging of datasets, addressing missing values and aligning spatial-temporal dimensions. New features such as dryness index, precipitation ratios, and temperature range were engineered to enhance model inputs.

★ **Problem Definition (3 points)**
We clearly defined wildfire prediction as the problem, focusing on brightness as the target metric and the classification of wildfire occurrences (whether a fire happens or not). This work was motivated by the growing need for wildfire risk management in the context of climate change.

★ **Algorithmic Work (4 points)**

We implemented the following machine learning models for regression to predict brightness (fire intensity) in MODIS data:

- Random Forest

- Linear Regression
- XGBoost
- Multilayer Perceptron

For binary classification of the column `in_modis` (indicating whether a fire occurs), we applied:

- Random Forest Classifier
- Gradient-Boosted Tree Classifier
- Logistic Regression
- Multilayer Perceptron

Machine learning algorithms were implemented with an emphasis on neural networks for brightness prediction and classification.
We tuned models and validated their performance using metrics such as RMSE and accuracy.

★ **Bigness/Parallelization (3 points)**
We leveraged PySpark for distributed data processing and model training.
Workflows were optimized for scalability to handle larger datasets.

★ **UI: User Interface (0 points)**
This project did not include a user interface or interactive front end.

★ **Visualization (3 points)**
We developed insightful visualizations to analyze trends in temperature, precipitation, and wildfire activity.Heat maps and temporal graphs were used to illustrate correlations between features and wildfire risk.
An attached report provides an in-depth explanation of our analysis.

★ **Technologies (2 point)**
We learned and applied PySpark ML for distributed machine learning workflows, employing algorithms such as SVM, logistic regression, linear regression, and neural network perceptron. Additionally, we gained experience with feature engineering

techniques, including PCA and PySpark tree importances, as well as advanced regression methods.

Total: 20 points

This summary captures our project's emphasis on ETL, algorithmic development, and visual analysis, with moderate attention to data acquisition and scalability. While we did not focus on UI development, the strengths in data processing, modeling, and visualization form the core contributions of our project.

Optional in-depth analysis/visualization(also in GitHub repo as data visualization report)

Wildfire Prediction

Data Visualization

MPCS 2024

Vikramjeet singh, Yiming, Harman, Wenhao

Introduction to this project

This project explores wildfires in Canada from 2006 to 2023, utilizing weather data and NASA FIRMS satellite scan data. The dataset includes information on wind, temperature, precipitation, sunlight, location, wildfire dates, and more. Our work involved data analysis and machine learning techniques. In this report, we present the results and insights from the analysis phase.

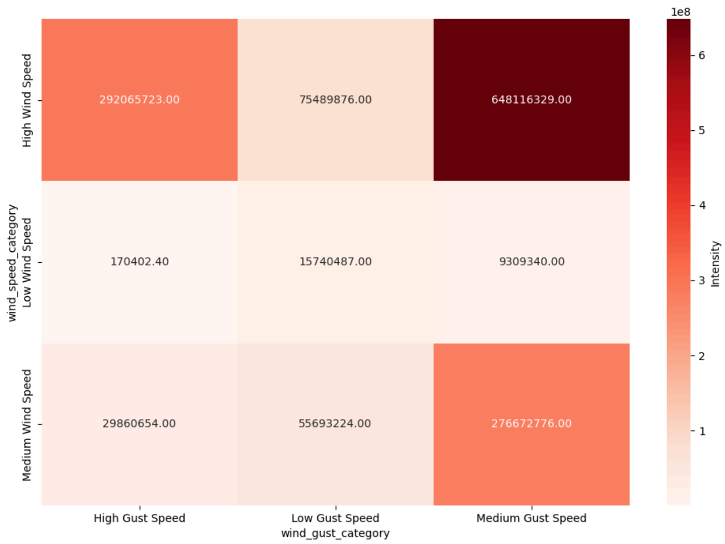
Data Schema

By combining weather data from Open Meteo with NASA FIRMS data, we obtained detailed information about daily weather conditions, wildfire occurrences, their locations, and intensity. We performed a left join on the Open Meteo data using NASA FIRMS data, matching by date and location. Furthermore, we derived new features from the existing ones using relevant theories and formulas.

The dataset includes a variety of features, such as date, city, and geographical coordinates (latitude and longitude). Weather-related attributes include maximum, minimum, and mean temperatures, apparent temperatures, daylight duration, sunshine duration, precipitation (rain, snowfall, and cumulative), precipitation hours, and wind-related metrics such as maximum speed, gusts, and dominant direction. Additional attributes cover shortwave radiation, evapotranspiration, and satellite-specific information like brightness, scan, track, satellite type, instrument, confidence, version, brightness temperature, fire radiative power (FRP), and day/night classification. Derived features include temperature range, relative humidity, cumulative precipitation, dryness index, fire intensity, daylight fraction, wind vector components (x and y), fire risk, soil moisture, and the precipitation-to-radiation ratio.

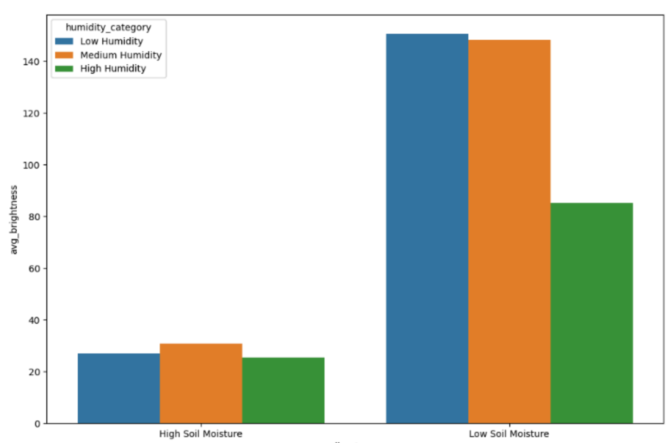
Analysis

Wind Analysis



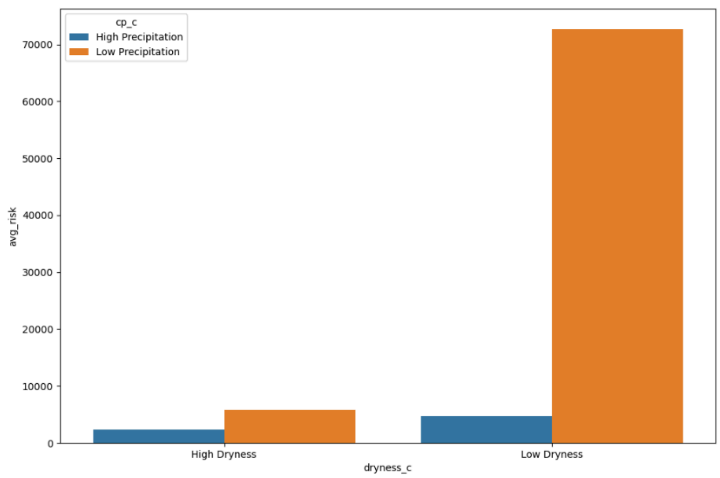
The analysis above shows how gust speed and wind speed contribute to a higher fire intensity using a heatmap. A higher intensity is more likely with medium gust speed and high wind speed and the lowest intensity is correlated with high gust speed and low wind speed. This indicates that wind speed has more of an effect on fire intensity than wind gust.

Soil Analysis



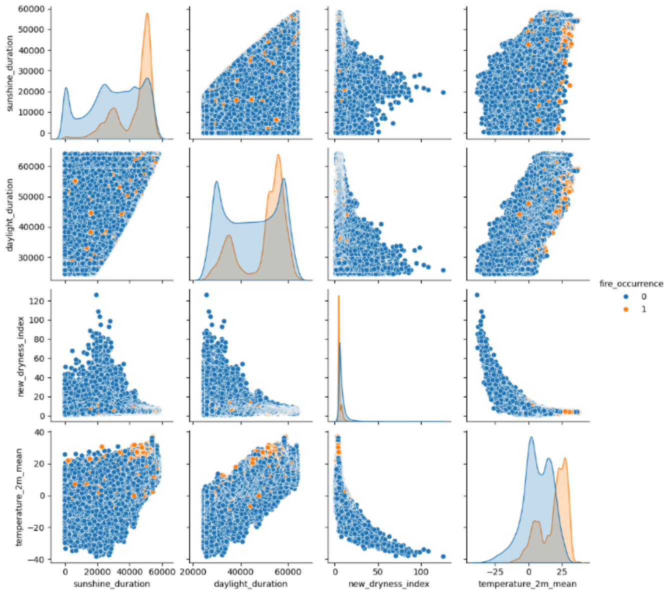
The soil analysis focuses on how soil moisture and humidity levels contribute to average brightness. As the graph indicates, low soil moisture dictates more brightness but when paired with low and medium humidity, the average brightness increases.

Fire Risk (precipitation and dryness)



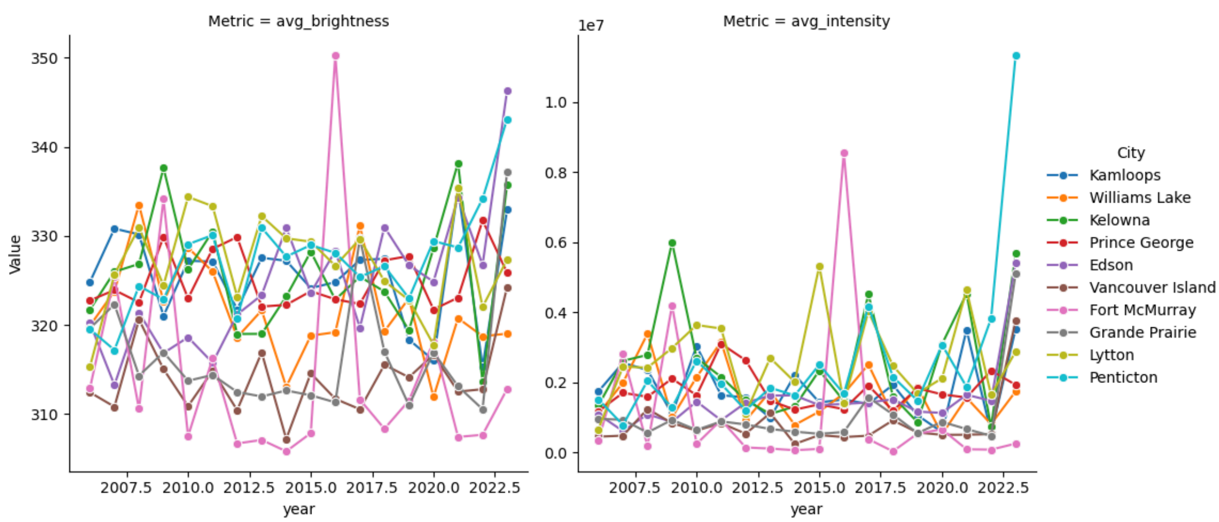
The graph above implies how dryness and precipitation relate to average fire risk. Low precipitation levels and low dryness imply a higher average fire risk. The graph also shows that dryness seems to be the dictating feature as high dryness with low precipitation has a lower fire risk.

Feature Correlation (sunshine_duration, daylight_duration, dryness, temperature)



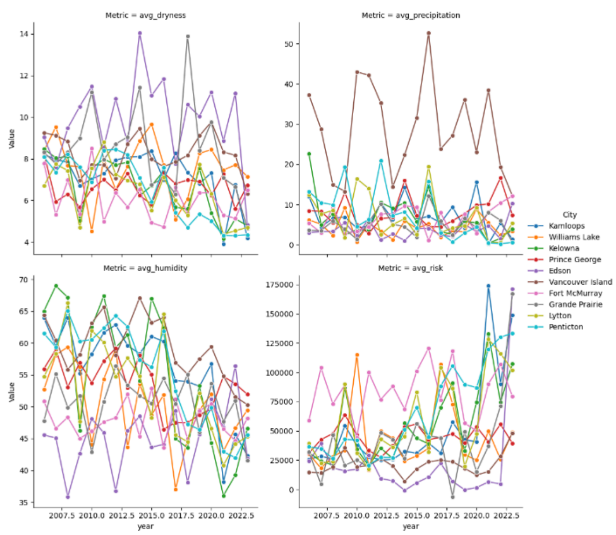
The graph above explains how the values for the following features: (sunshine_duration, daylight_duration, dryness, temperature), relate to a fire occurrence. So as the graph shows, higher temperatures and higher daylight and sunshine duration imply a fire occurrence. Meaning, higher values for all these features paired with higher temperature values imply a fire is present.

Average intensity and brightness per city by year



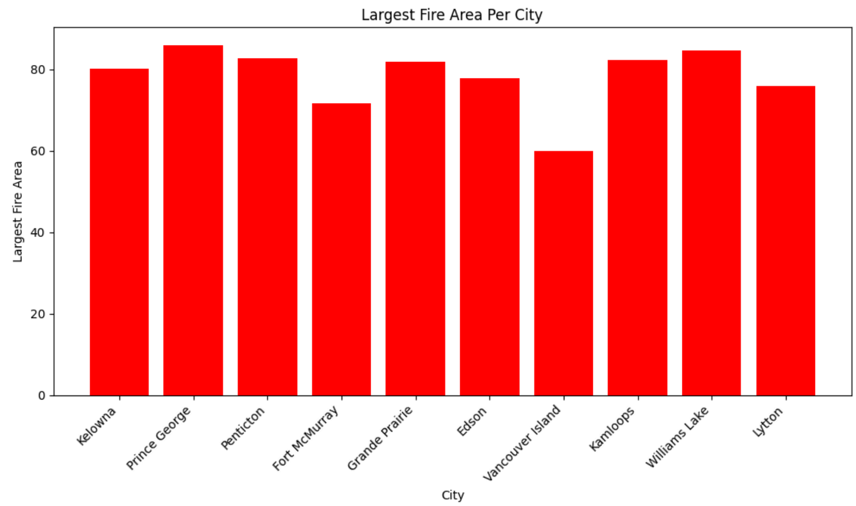
The graph above shows how average intensity and brightness changes over the years per city. Intensity: intensity seems to be normal for the most part, but some cities show a spike in certain years such as Penticton (2024), Fort McMurray (2016), and Kelowna (2010). Brightness: brightness also has the same trend as intensity, the spikes include Fort McMurray (2016), Edson (2024), and Penticton (2024). From this graph, we can see that there were significant wildfire activities in 2024 for Penticton and 2016 for Fort McMurray.

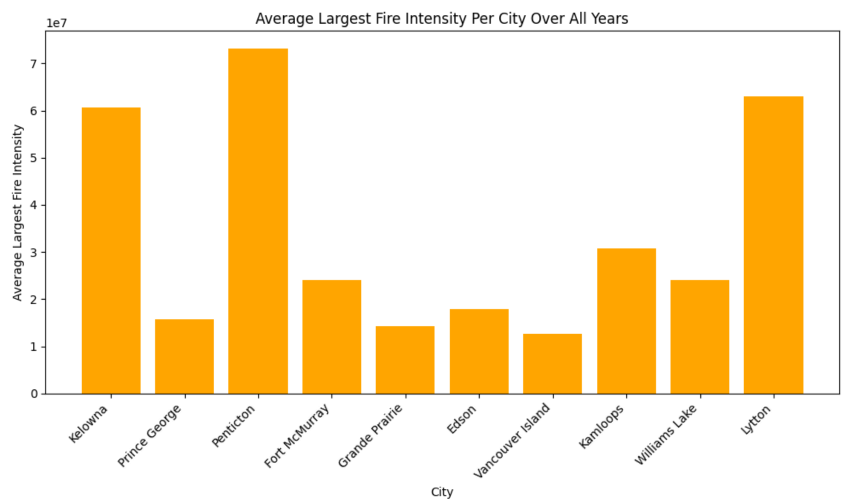
Yearly City Trends



The graph above displays the trends for features: dryness, humidity, risk, and precipitation per city. Excluding risk, we can see how the 3 other features influence fire risk. As average humidity lowers over the years, the fire risk increases for each city. Cities with higher precipitation values also show lower levels of fire risk.

Largest Fire Intensity Per City

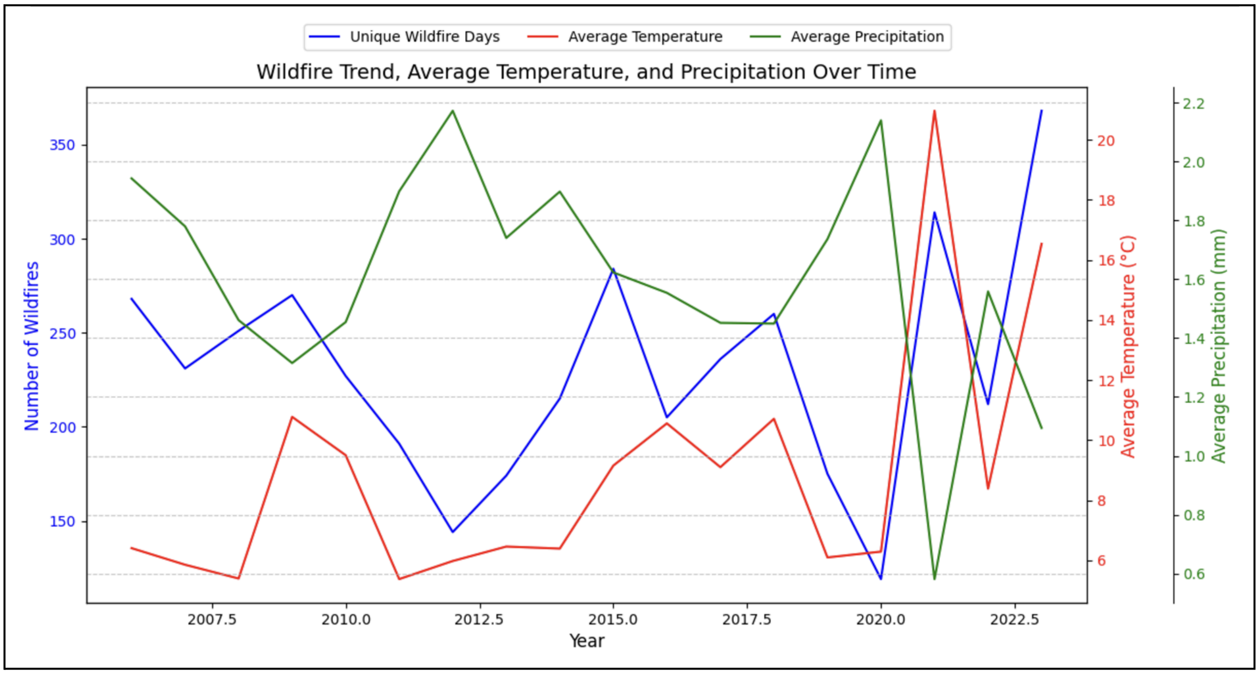




The graph above displays the largest fire happening in each city in terms of intensity and area. The area is calculated as the area of the smallest matrix covering all the locations in one city. For area, all the cities perform similarly excluding Vancouver Island and Fort Mcmurray. For intensity, **Kelowna**, **Penticton** and **Lytton** are the top three cities with the highest forest fire intensity. With VIII and X analysis combined, we can conclude that people living in these three cities need to pay more attention to the possibility of forest fires happening in summer times.

Wildfire Day Counts, Temperature And Dryness Index Trend Analysis

One of the hypotheses is that the number of wildfires is associated with temperature and dryness, as wildfires are likely to occur in an environment that is hot and dry. The following graph described those three factors over the time period from 2006 to 2023.

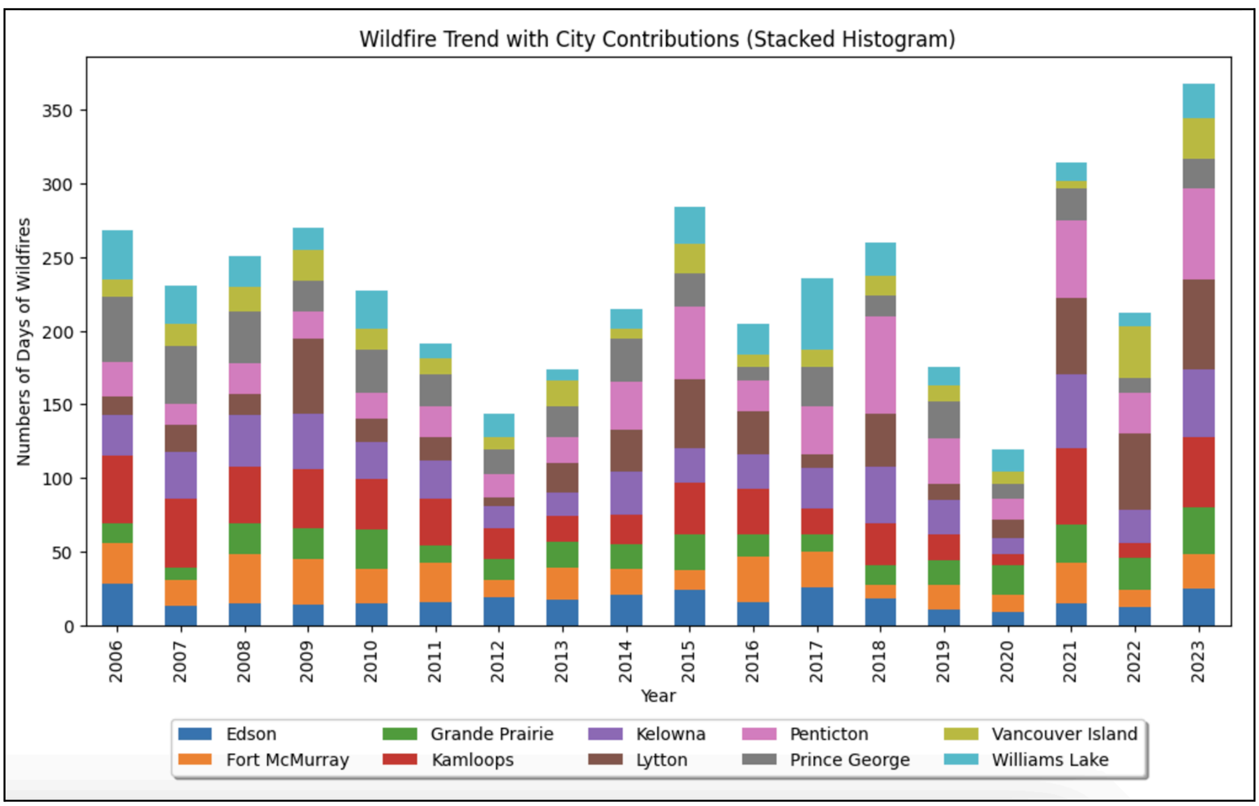


Key observations include the correlation between temperature, precipitation, and wildfire activity. The blue line represents wildfire count, the red line indicates average temperature, and the green line corresponds to average precipitation.

In years with higher average temperatures, such as 2016 and 2021, there is a noticeable increase in wildfire count, suggesting a potential link between temperature and wildfire activity. Conversely, years with higher average precipitation generally show lower wildfire counts, indicating precipitation may act as a mitigating factor for wildfires. The interplay between these variables highlights how climatic factors like temperature and precipitation contribute to wildfire dynamics, underscoring the importance of understanding environmental influences for better wildfire management.

Geographical Wildfire Day Counts Analysis

This stacked histogram visualizes the annual trend of wildfires across various cities from 2006 to 2023, with contributions from specific cities represented by different colors. The y-axis shows the total number of wildfires, while the x-axis represents the years. Each segment in a bar corresponds to the number of wildfires reported in a particular city, allowing for a clear comparison of contributions by location over time.

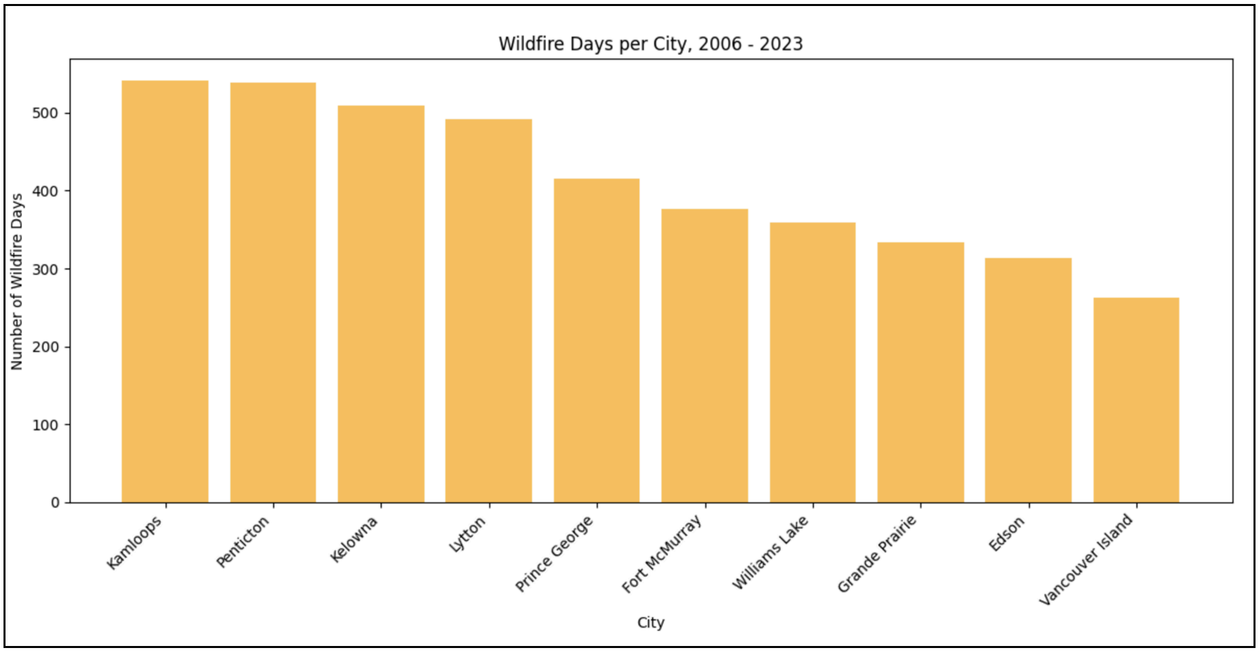


Key observations include a noticeable spike in wildfire occurrences in 2015 and

2023, indicating potentially extreme conditions during those years. Conversely, years like 2012 and 2020 exhibit relatively lower wildfire counts. Among the cities, locations such as Prince George, Kamloops, and Penticton seem to consistently contribute a significant portion to the total wildfire count, reflecting areas of high fire activity. This trend suggests that wildfire management efforts might need to prioritize these cities due to their recurrent vulnerability.

From 2020 to 2021, there is a significant leap in the total number of wildfires, as illustrated by the marked increase in the bar height. This sharp rise suggests a drastic shift in environmental conditions or other contributing factors, such as prolonged droughts, higher temperatures, or increased human activities during this period. The cities of **Kamloops, Penticton, Kelowna** and **Lytton** appear to play a prominent role in this increase, with larger segments in 2021 compared to 2020. This jump highlights the need to investigate and address the underlying causes of such spikes to mitigate future wildfire risks.

This bar chart shows the total number of wildfire days recorded for each city. Kamloops and Penticton lead with the highest numbers of wildfire days, both exceeding 500 days, indicating that these cities experience prolonged and frequent wildfire activity compared to others. This trend highlights them as high-risk areas requiring more focused wildfire prevention and mitigation strategies.

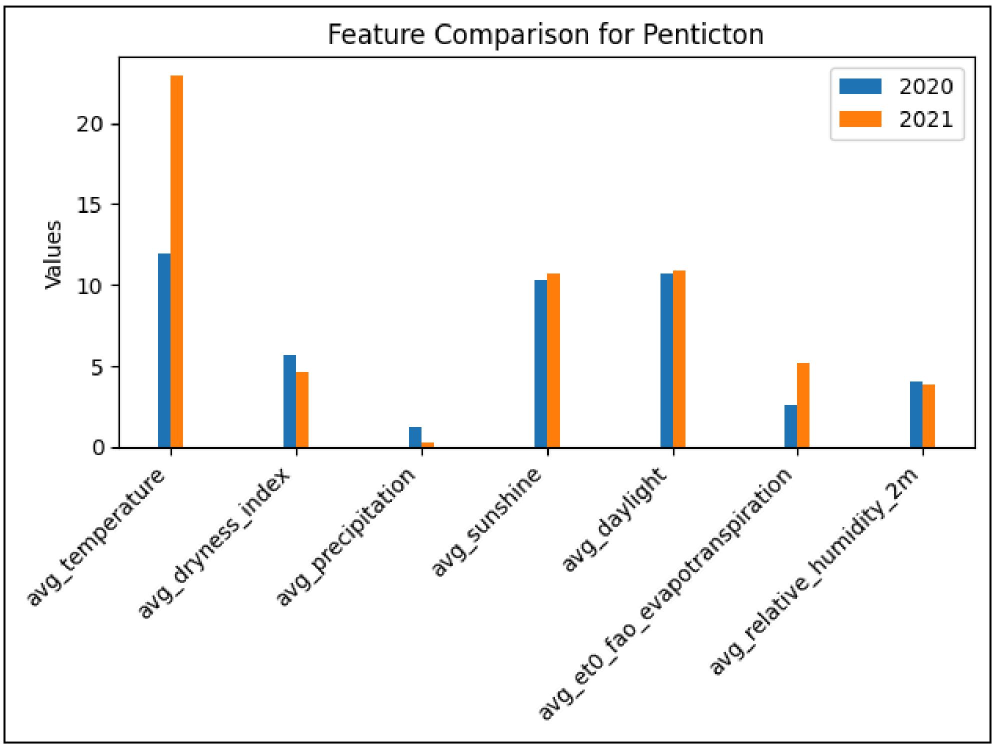


Cities like Vancouver Island and Edson are at the lower end of the spectrum, with fewer wildfire days, suggesting they face relatively less frequent or severe wildfire conditions. However, even these cities are not immune to wildfires, emphasizing the need for consistent monitoring and preparedness across all regions. Overall, the distribution suggests significant geographic variability in wildfire patterns, influenced by local climate, vegetation, and possibly human activities.

2020 - 2021, A Gigantic Increase in the Number of Wildfire Days

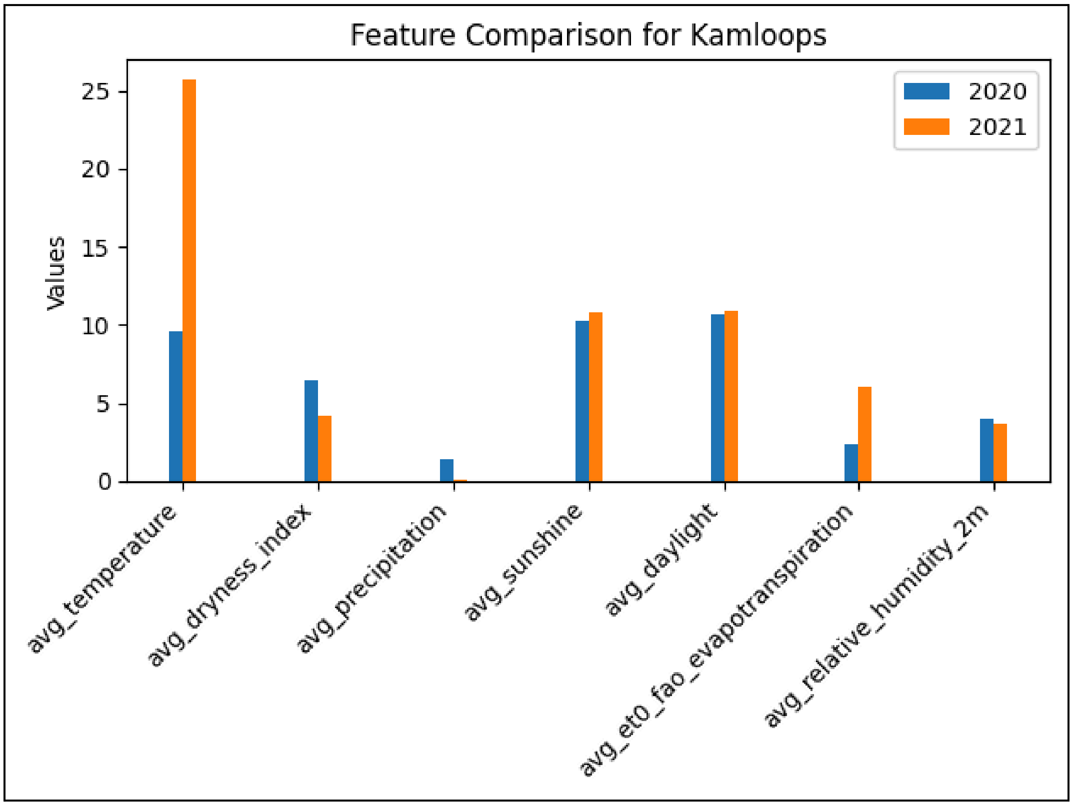
As mentioned in the previous analysis, we observe a large increase in the number of days of wildfire occurring during 2020 - 2021 mainly in the following cities: Kamloops, Penticton, Kelowna and Lytton. In this section, we will take a look at those four cities in terms of their average values during this time period.

Penticton



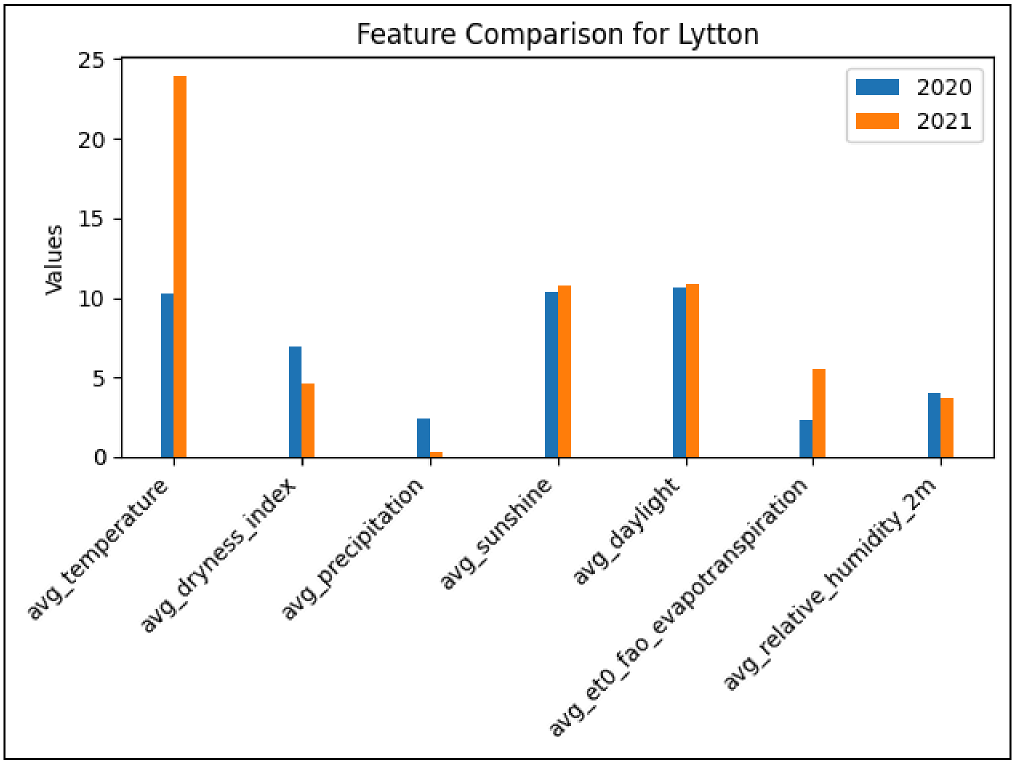
Penticton saw an average temperature increase from 11.99°C in 2020 to 22.93°C in 2021. The dryness index declined from 5.72 to 4.67, while precipitation dropped from 1.29 mm to 0.32 mm. Sunshine duration rose modestly from ~31,632 hours to ~44,301 hours, and daylight increased from ~44,237 hours to ~52,502 hours. Evapotranspiration also increased, rising from 2.61 to 5.25. Relative humidity decreased from 54.77% to 45.33%, contributing to the overall arid conditions.

Kamloops



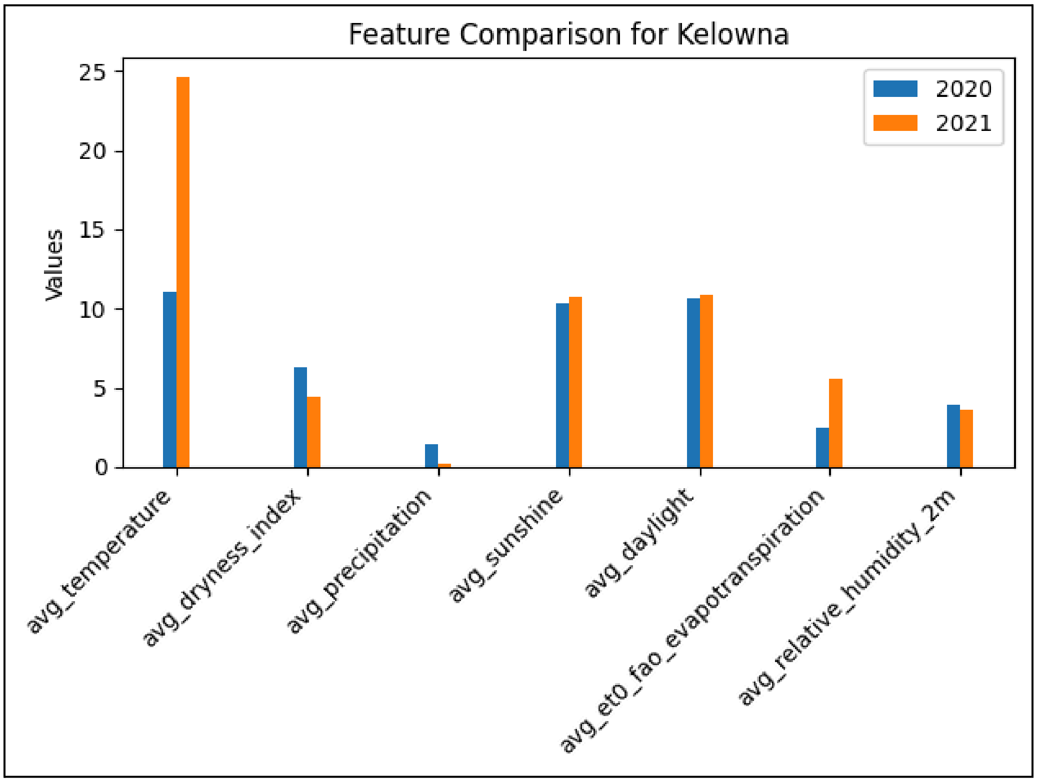
In 2021, Kamloops experienced a dramatic rise in average temperature from 9.65°C to 25.71°C, coupled with a significant drop in the dryness index from 6.44 to 4.20. Precipitation decreased drastically, from 1.39 mm to just 0.15 mm, highlighting much drier conditions. Sunshine and daylight durations increased substantially, with sunshine rising from ~30,060 hours to ~47,062 hours and daylight from ~43,975 hours to ~54,161 hours. Evapotranspiration more than doubled from 2.35 to 6.11, indicating higher water loss due to increased heat and sunshine. Relative humidity dropped from 51.59% to 39.33%, further emphasizing the arid conditions.

Lytton



In Lytton, the average temperature rose significantly from 10.25°C to 23.91°C in 2021. The dryness index saw a marked decrease from 6.96 to 4.64, while average precipitation dropped from 2.42 mm to 0.34 mm, reflecting notably drier conditions. Sunshine duration increased dramatically from ~31,427 hours to ~47,262 hours, with daylight rising from ~44,271 hours to ~54,490 hours. Evapotranspiration rose from 2.33 to 5.55, indicating greater water loss, and relative humidity decreased from 52.20% to 41.87%, further emphasizing the dry conditions.

Kelowna



Kelowna experienced a significant temperature increase in 2021, with the average rising from 11.04°C to 24.59°C. The dryness index fell from 6.27 to 4.45, indicating drier conditions, and precipitation levels dropped from 1.40 mm to 0.18 mm. Sunshine duration rose sharply from ~30,837 hours to ~45,928 hours, while daylight duration increased from ~44,325 hours to ~52,429 hours. Evapotranspiration also more than doubled, increasing from 2.52 to 5.60. Relative humidity dropped from 49.11% to 37.59%, making the atmosphere much drier overall.

In Summary

In 2021, all four cities—Kamloops, Kelowna, Lytton, and Penticton—experienced significant increases in temperature and dryness compared to 2020. Average temperatures rose sharply, with increases of around 14°C to 15°C across the board, accompanied by a notable drop in the dryness index. For instance, Kamloops' dryness index dropped from 6.44 to 4.20, while similar patterns were observed in the other cities. Precipitation levels plummeted, with most cities experiencing reductions of over 80%. Kamloops and Kelowna saw their average precipitation fall from around 1.4 mm to below 0.2 mm, while Lytton and Penticton faced similarly steep declines. These reductions in precipitation were paired with significant increases in sunshine and daylight hours, which rose by roughly 15,000 and 10,000 hours, respectively, across all locations.

Evapotranspiration levels more than doubled, reflecting heightened water loss from the soil due to elevated temperatures and prolonged sunlight exposure. Kamloops saw a rise from 2.35 to 6.11, with other cities following similar trends. At the same time, relative humidity dropped significantly, indicating much drier atmospheric conditions. For example, Kamloops' relative humidity decreased by over 12%, while other cities saw comparable declines. These combined changes suggest that 2021 was considerably hotter, drier, and more prone to extreme conditions, raising concerns about the increasing risk of wildfires and the broader impacts of climate change in these regions.