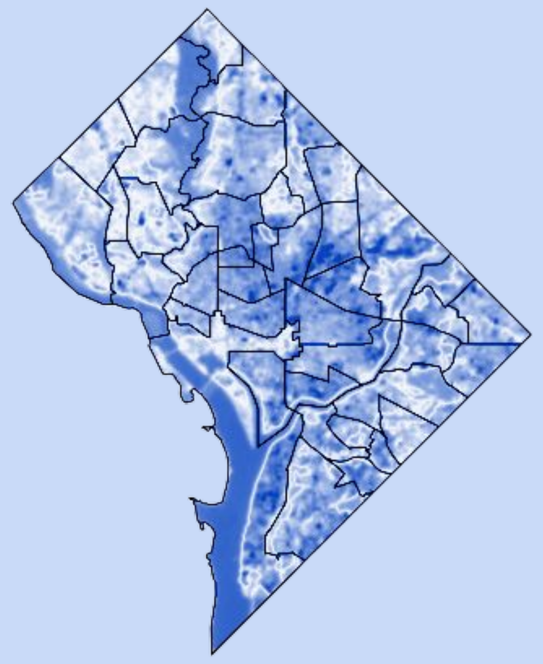


# Predicting Temperature Changes Over Time to Indicate Climate Trends


## Gracie Abrahams, Jeffrey Mann, Luke Carignan

American University, STAT-427


## Introduction

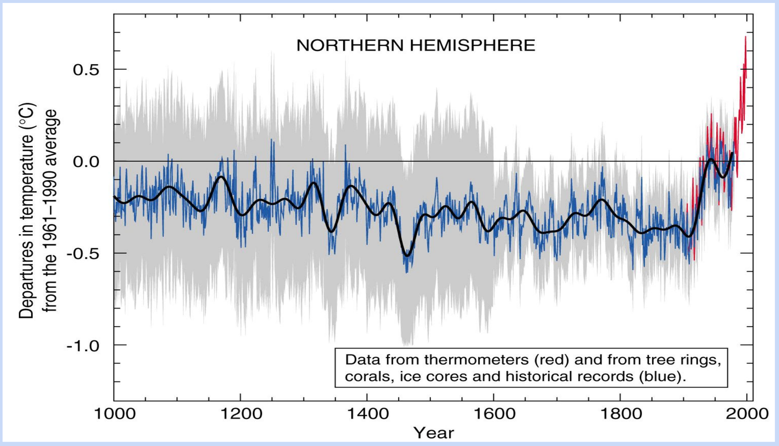



*Figure 1. "Hockey Stick graph" that displays temperature over time in the Northern Hemisphere, illustrating dramatic increase in temperature beginning during the Industrial Revolution.*

Washington, D.C. is experiencing increasing impacts from climate change, including more frequent heatwaves, intense rainstorms, and rising sea levels.

The city has seen a rise in dangerously hot days-from 29 baseline days to projections of 70–80 by 2050 and up to 105 by 2080. These changes threaten vulnerable communities and critical infrastructure.

Predicting future temperature patterns and classifying extreme weather events are essential for supporting D.C.'s climate adaptation and mitigation goals, such as those outlined in Sustainable DC 2.0 and Carbon Free DC.

Using hourly ERA5 climate data from 1940 to present, this project applies machine learning algorithms to:

- Classify extreme weather events affecting the region
- Forecast temperature trends to inform resilience planning

Machine learning-driven insights can help policymakers and communities prepare for a hotter, more volatile climate while advancing equitable climate action across Washington, D.C.

## Objective

This project aims to answer two key questions using ML and historical climate data from D.C.:

- **Can we classify extreme weather events?**
  Using classification algorithms such as KNN, LDA/QDA, and logistic regression, we seek to identify and categorize extreme events like heat waves, storms, and heavy precipitation based on climate data.
- **Can we predict future temperature patterns based on past weather data?**
  Applying regression models, we aim to forecast temperature trends over time to support climate adaptation and mitigation efforts.

### Why is this important?

- Washington, D.C. faces increasing extreme weather events that threaten vulnerable communities and infrastructure.
- Accurate predictions enable better preparation and targeted climate adaptation strategies.
- Decrease health risks from extreme weather
- Supports the city's goals to reduce emissions and build resilience by 2050.

## Methodology


**Data Preprocessing**
included verification of longitude and latitude values for the District of Columbia, filtering to only include observations between 2000 and 2025, and checking for null values in each column.

**Feature Selection**
Entailed choosing variables relevant to extreme weather events, including temperature, wind speed, precipitation totals, and surface air pressure. To predict future temperature patterns in the District of Columbia, air temperature data was converted from Kelvin to Fahrenheit and coupled with time variables including month and year. This allowed for future value prediction to represent both projected annual temperature changes, and also monthly, as both time intervals are common in climate modeling and resiliency planning.

**Exploratory Data Analysis**
Was done on temperature data to later predict future temperature patterns. The EDA gave insight into the variability of temperature year-over-year and by month between 2000 and 2025, as well as extreme temperature events and the average range of surface temperature observed. Further EDA was conducted on precipitation totals and wind speed using histogram outputs, which aided in understanding normality in each predictor.

**Modeling and Algorithm Comparison**
Gave insights for both questions of interest. To classify extreme weather events using predictors such as precipitation, pressure, and wind speed, KKN Modeling was initially used to test accuracy of our classification. After completion, multiple models were employed including Logistic Regression, Linear and Polynomial Regression, LDA/QDA, and Lasso. LDA and QDA were used to assess class separation and classification boundaries under different assumptions, while lass Regression was implemented to perform variable selection and reduce overfitting in our models.

## Results

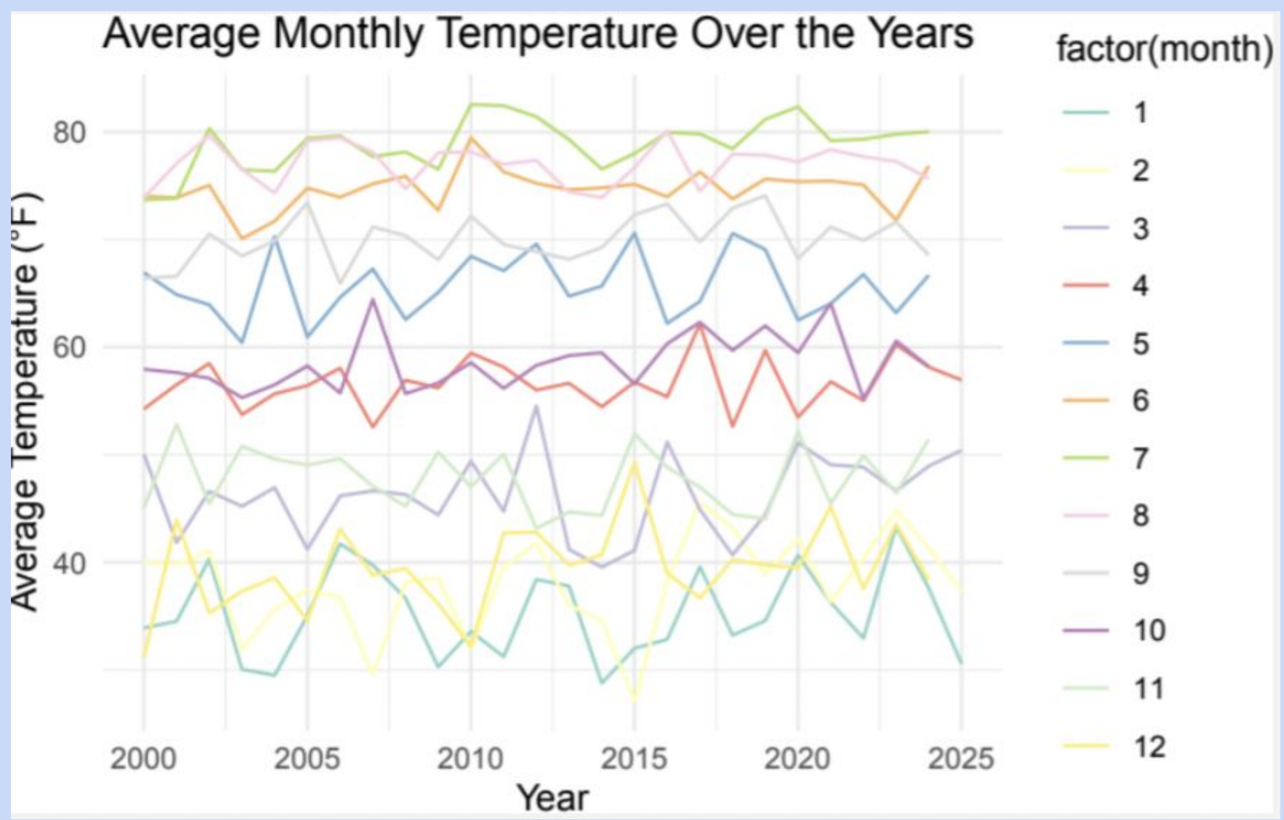
**Classification: Can we classify extreme weather events?**

Using three algorithms, we can classify extreme weather based on temperature, wind speed, and precipitation rates.

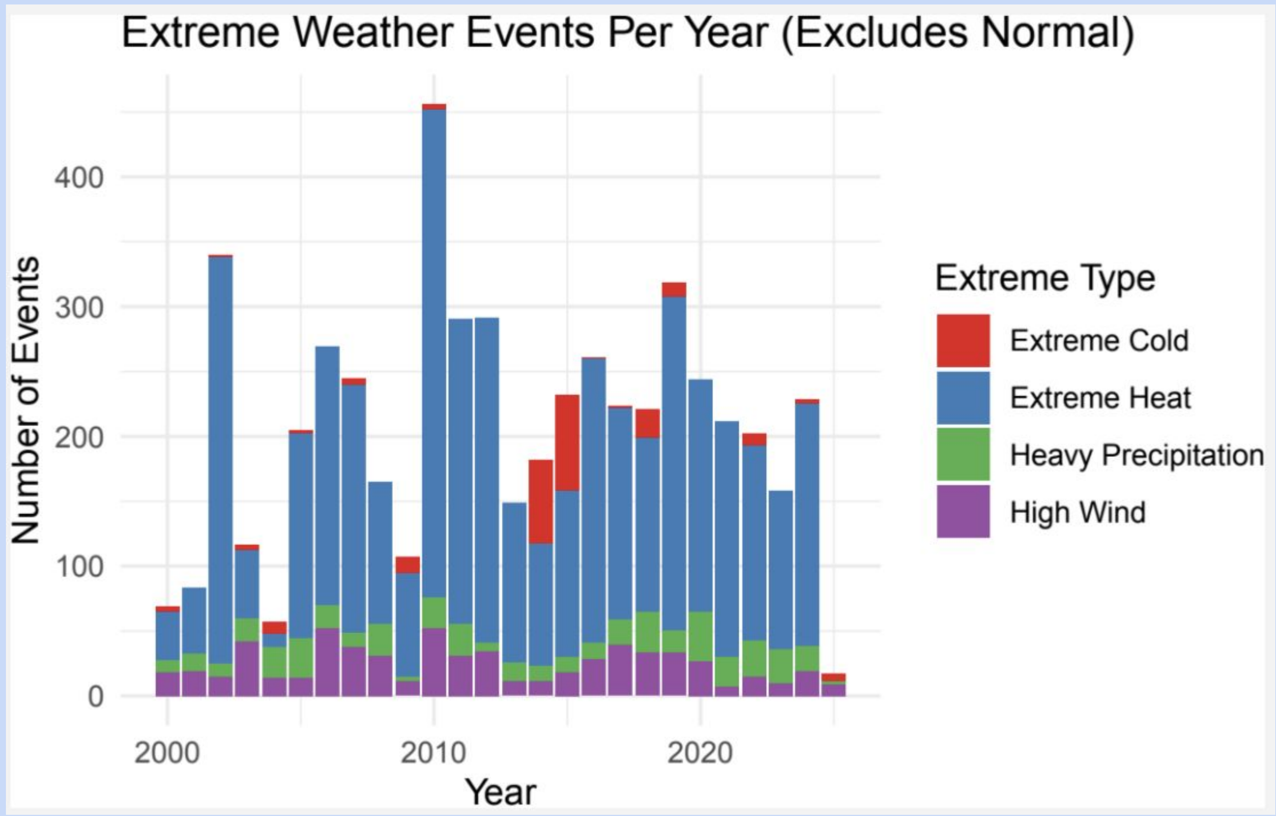
The following models were used and evaluated:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| KNN | 88% | 89% | 87% | 88% |
| Logistic Reg | 6% | 4% | 4% | 4% |
| LDA | 91.9% | 95% | 88% | 92% |
| QDA | 93.9% | 95% | 93% | 94% |

*Table 1. Evaluation Metrics for Classification Models*




*Fugure 2. Average Temperature by Month between 2000 and 2025*




*Fugure 3. Extreme Weather Event Frequency Per Year Excluding Normal Events*

## Conclusion

We evaluated multiple classification methods to predict extreme weather events, including KNN, logistic regression, LDA, and QDA. KNN initially appeared highly accurate (>98%), but deeper inspection revealed this was due to severe class imbalance. After balancing the data, KNN's performance became more realistic, achieving ~88% accuracy and a balanced accuracy of 0.5. Logistic regression struggled even after rebalancing, with only 6.4% accuracy, indicating it could not capture complex decision boundaries. LDA and QDA, by contrast, performed significantly better, with classification accuracies of 91.97% and 93.99% respectively, showing that distribution-based models can better handle the underlying structure of extreme weather classification. Cross-validation was used to select the optimal number of neighbors in KNN and to ensure fair model comparison. These results highlight the importance of both class balance and cross-validation in evaluating classifier performance in imbalanced real-world datasets.

To evaluate model performance, we used RMSE (Root Mean Squared Error) and adjusted $R^2$ across four regression approaches: linear, polynomial, ridge, and lasso. The standard linear model performed best overall, with an adjusted $R^2$ of 0.9165 and a low residual error, indicating a strong linear relationship between temperature and the selected predictors. Polynomial regression, despite modeling seasonal trends with month, showed weaker performance due to its limited feature scope. Ridge regression used cross-validation to select the optimal penalty parameter ($\lambda$), but still underperformed with an RMSE of 4.65. Lasso regression also used cross-validation for tuning and achieved a much lower RMSE of 2.94 by automatically selecting relevant features. These results underscore the effectiveness of regularized models when handling collinearity or when model simplicity is preferred. Cross-validation played a key role in objectively comparing flexibility and prediction error, helping avoid overfitting.

## Limitations

Despite some promising results from our prediction modeling, several limitations exist, including:

**Data Constraints:** Our modeling use a limited dataset from a single location and a short time span relative to typical climate modeling, which restricts this study's applicability to broader spatial or temporal contexts.

**Model Complexity:** KNN may not capture complex atmospheric interactions that impact both temperature anomalies and overall patterns, as well as other climate mechanisms across the broader Northern Hemisphere and Atlatnic Ocean which can also impact localized weather patterns. These realities do present possible underfitting of our models in some scenarios.

**Feature Limitations:** While notable weather variables were used in this study, the study could benefit from a wider range of weather variables to predict future temperature and/or classify extreme weather events. Examples of additional useful variables include solar radiation, wind direction, humidity, cloud cover, or visibility distance.

**Short-Term Applicability:** climate modeling generally suffers low accuracy rates because of the complexity and variability of observed climate conditions year over year. Oftentimes, studies just a few years old that concluded then-current climate conditions are questionable in relevance to today, as regional and global climate conditions can change rapidly year to year.