

## **Forecasting Climate Disaster Frequency and Costs in the United States**

Justin Hay (300236040), Mohamed Reda Jahouri (300262323),

Tijn Schmiehusen (300455529)

Telfer School of Management, University of Ottawa

ADM4307

Dr. Rafid Mahmood

December 5, 2024





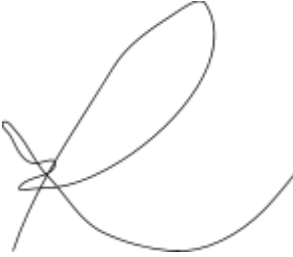
## Statement of Academic Integrity

### Group Assignment Checklist & Disclosure

Please read the disclosure below following the completion of your group assignment. Once all team members have verified these points, hand in this signed disclosure with your group assignment.

1. All team members acknowledge to have read and understood their responsibilities for maintaining academic integrity, as defined by the University of Ottawa's policies and regulations. Furthermore, all members understand that any violation of academic integrity may result in strict disciplinary action as outlined in the regulations.
2. All team members have referenced and/or footnoted all ideas, words, or other intellectual property from other sources used in completing this assignment.
3. A proper bibliography is included, which includes acknowledgment of all sources used to complete this assignment.
4. This is the first time that any member of the group has submitted this assignment or essay (either partially or entirely) for academic evaluation.
5. No member of the team has utilized unauthorized assistance or aids including but not limited to outsourcing assignment solutions, and unethical use of online services such as artificial intelligence tools and course-sharing websites.
6. Each member of the group has read the full content of the submission and is assured that the content is free of violations of academic integrity. Group discussions regarding the importance of academic integrity have taken place.
7. All team members have identified their individual contributions to the work submitted such that if violations of academic integrity are suspected, then the student(s) primarily responsible for the violations may be identified. Note that the remainder of the team may also be subject to disciplinary action.

Course Code:	ADM 4307
Assignment No. / Title:	Assignment 2
Use of Plagiarism Detection Tools (e.g., Original):	<input type="checkbox"/> Yes (Required by Course / Professor) <input type="checkbox"/> Yes (Self-Conducted) <input type="checkbox"/> No (Not Applicable for Type of Assignment) <input checked="" type="checkbox"/> No (Not Conducted)
Date of Submission:	November 25, 2024

Name	Signature	Student Number
Tijn Schmiehusen		300455529
Justin Hay		300236040
Mohamed Reda Jahouri		300262323

### Group Member Contributions

Justin Hay	Tijn Schmiehusen	Mohamed Reda Jahouri
<ul style="list-style-type: none"> <li>- Preprocessing</li> <li>- Intro, business questions, monthly analysis, conclusion</li> <li>- Slides formatting</li> <li>- Paper formatting</li> </ul>	<ul style="list-style-type: none"> <li>- Yearly counts models</li> <li>- Out of scope model for yearly counts</li> </ul>	<ul style="list-style-type: none"> <li>- Data preprocessing</li> <li>- Yearly cost models</li> <li>- Improvement of models</li> <li>- Yearly Business Questions</li> <li>- Yearly cost slides analysis</li> </ul>

We acknowledge that there was a relatively equivalent amount of work done throughout the semester.

## Table of Contents

<b>Abstract.....</b>	<b>6</b>
<b>1.0 Introduction.....</b>	<b>7</b>
1.1 The Relevance of Exploring Climate Disasters.....	7
1.2 Business Questions.....	7
<b>2.0 Dataset.....</b>	<b>8</b>
2.1 Data Pre-Processing.....	9
2.1.1 Extraction.....	9
2.1.2 Transformations in R.....	10
2.2 Monthly Overview.....	11
2.2.1 Monthly Counts.....	11
2.2.2 Monthly Costs.....	12
2.3 Yearly Overview.....	13
2.3.1 Yearly Counts.....	13
2.3.2 Yearly Costs.....	14
<b>3.0 Analysis.....</b>	<b>15</b>
3.1 Monthly Analysis.....	15
3.1.1 Stationary Seasonality.....	15
3.1.2 Variable Seasonality.....	19
3.2 Yearly Analysis.....	21
3.2.1 Yearly Counts.....	21
3.2.2 Yearly Costs.....	24
3.2.2.1 Decomposition.....	24
3.2.2.2 Model Analysis.....	25
3.2.2.3 Model Improvement.....	27
<b>4.0 Conclusion.....</b>	<b>30</b>
4.1 Revisiting Business Questions.....	30
4.2 Limitations and Next Steps.....	31
<b>References.....</b>	<b>33</b>

**Abstract**

Each year, climate /disasters such as droughts and hurricanes affect millions of people in the United States. Understanding the nature of climate disasters and potentially predicting their frequency and cost would directly benefit key stakeholders such as locally impacted economies and insurance companies. The analysis conducted in the report is broken down into monthly and yearly timeframes, which allowed us to examine both local seasonality and forecast multi-year trends. In addition, both frequency and costs of disasters over both time periods were examined, with the intention of gaining a better understanding of the patterns belonging to both variables as well as their correlation to each other. Overall, the monthly analysis found that there is a strong seasonality in the summer months that is increasing over time, and there is a large spike of costs in the late summer due to tropical cyclones. The yearly analysis found that both counts and costs are increasing steadily over time, although it will take further analysis and data to be able to properly analyze costs specifically.

## 1.0 Introduction

### 1.1 *The Relevance of Exploring Climate Disasters*

According to a variety of sources across the internet, a climate/weather disaster is a natural disaster that pertains specifically to weather- and climate-related occurrences (United Nations, 2023). Examples of these disasters include droughts, flooding, hail, severe weather, tornadoes, tropical cyclones, wildfires, and winter storms. These disasters accounted for over \$2.75 trillion dollars in the United States alone since the National Centers for Environmental Information (NCEI) began tallying costs in 1980 (NCEI, 2024). These disasters have incurred tremendous costs on various key stakeholders across the United States, most notably the communities and local economies impacted by the damages. Costs incurred by climate disasters also influence government policy at various levels as well as corporate decision-making for companies in various sectors, but most notably in insurance. In addition to examining costs, disaster counts over a time span are an extremely important metric to consider. Analyzing the number of disasters that have occurred during the same time period is important for two reasons. Firstly, counts have a clear upwards trend over time, which makes it useful as explanatory variable for predicting costs. Secondly, examining counts helps to quantify a phenomenon known as compound climate disasters, which can be defined as an exponential increase in damages due to multiple disasters happening in overlapping time periods (Crimmins, 2023). Predicting the trends and seasonality of both counts and costs of climate disasters may be able to help local stakeholders, governments, and corporations better prepare for the severity of these events as well as mitigate losses.

### 1.2 *Business Questions*

When looking at analyzing trends and patterns in climate disaster data to answer business questions, we chose to separate our data into two timeframes: monthly and yearly. Below we can outline certain specific questions we wish to examine throughout our analysis.

Monthly Questions	Business Purposes
Is there seasonality?	As the broadest, yet likely most important question, understanding seasonality will help stakeholders adjust their preparation for various times of year depending on how intense that time is.

What times of year do specific disasters occur?	Understanding specific disaster timings can especially help local economies and government to plan and prepare and educate communities.
Is there a correlation between disaster types and cost?	Understanding which disaster costs are the most expensive can help stakeholders such as insurance companies plan pricing policies and set aside resources prior to the events occurring.
Is seasonality changing over time?	Understanding how seasonality is changing over time is extremely valuable for long term planning purposes such as government policy planning.

**Figure 1: Table Outlining Monthly Questions and Business Purposes**

<b>Yearly Questions</b>	<b>Business Purposes</b>
Is there seasonality/trend?	Understanding a multi-year seasonality / trend will directly aid stakeholders in conducting multi-year planning initiatives and resource allocation.
Can we forecast yearly costs/counts?	Similarly to above, forecasting these patterns will help stakeholders such as local economies better prepare for future years.
Are there any additional explanatory variables that may help with forecasting?	Understanding variables that are correlated with costs and counts may help stakeholders create connections between disasters and factors such as climate change.
Is there any correlation between disaster costs and counts?	Understanding how costs are correlated to counts can help stakeholders understand how they are related and potentially prepare for the fluctuation in costs based on changes in counts.

**Figure 2: Table Outlining Yearly Questions and Business Purposes**

## 2.0 Dataset

The dataset chosen as a foundation for our analysis is entitled “Weather and Climate Billion-Dollar Disasters to Affect the U.S. from 1980 - 2024”. The main purpose of this dataset is to track each disaster that has cost over \$1 billion in the United States and to categorize them by type. The threshold of \$1 billion allows us to define each disaster that has occurred as an extreme weather event, which in turn provides us with a baseline for



conducting analysis over various time periods. In addition, the cost for each disaster is calculated from a wide variety of sources, including the Insurance Services Office, Property Claim Services, etc., which provides a relatively comprehensive understanding of the total costs for each disaster. Although confidence intervals were not included in certain cost analysis figures, it is important to keep in mind that these costs are still estimates and are not 100% accurate.

## *2.1 Data Pre-Processing*

The overall pre-processing of the dataset ended up requiring a considerable amount of work since we were extracting data for both costs and counts on a monthly and yearly basis.

### 2.1.1 Extraction

The first step in order to prepare the data was to import it from its source file in the National Oceanic and Atmospheric Administration OneStop Database, which contains thousands of datasets regarding various climate trends all around the world (NOAA, n.d.). The most recent data for our dataset (Q2 2024) was extracted from the website using a Zip file extractor, which provided us with raw uncleaned data. After some transformations in Excel, we were able to organize all the data to show the following columns in figure 3 below. One thing to note is that the “Counts” column in the figure below was not included in the original dataset. Each value in this column contains the number “1” since it represents the event occurring a total of 1 time.

Name	Type of Disaster	Begin Date	End Date	Adjusted Cost	Unadjusted Cost	Deaths	Counts
------	------------------	------------	----------	---------------	-----------------	--------	--------

**Figure 3: Table showcasing the Dataset Columns**

Once this data was organized, it was transformed into two CSVs, one for both monthly and yearly purposes. In the monthly CSV, only the unadjusted costs and deaths were removed since they were not relevant to the purposes of our analysis. In the yearly CSV, the begin date, unadjusted cost, and deaths were removed since we were only looking to analyze specific points in time. Once the CSVs were cleaned, they were imported into separate R files in order to be transformed into time series objects.

### 2.1.2 Transformations in R

When conducting our preliminary analysis, we first agreed to not remove any outliers from our monthly or yearly data. The reasoning behind this was that if there were indeed any patterns to the outliers, we did not want to remove them so that we could not analyze them. There was, however, one preprocessing task that ended up being necessary over all time frames. When importing our data into R, it did not recognize time periods that had costs equal to 0. When plotting, it would skip over these values and take an average when moving to the next value. In order to remedy this issue, we used the “fill\_gaps” function in order to capture the data and then used the “is.na” function in order to set that value to 0. We intentionally chose to make those time periods equal to 0 since it more accurately reflects the data.

```
ts_cost <- ts_cost %>% tsibble::fill_gaps()
ts_cost[is.na(ts_cost)] <- 0
```

#### **Figure 4: The Code for Imputing Missing Values**

Looking specifically at monthly data, multiple transformations were made in order to more accurately represent the frequency of events. As mentioned in section 2.1.1, both start dates and end dates were intentionally in the CSV file. These dates were kept so that the data could count the frequency of the events happening over each month that it happened rather than in just one month of the year. The data was transformed in R to duplicate the values of each disaster depending on each month that it occurred between its start date and end date. When this transformation was applied, the cost of each disaster was modified to be spread evenly throughout each month that it occurred by dividing total cost by number of months. This methodology may not be completely accurate since it does not account for the fact that certain costs may be incurred even after the disaster ends. For this reason, costs are examined more thoroughly in the yearly analysis but are still analyzed as an auxiliary variable throughout our monthly analysis.

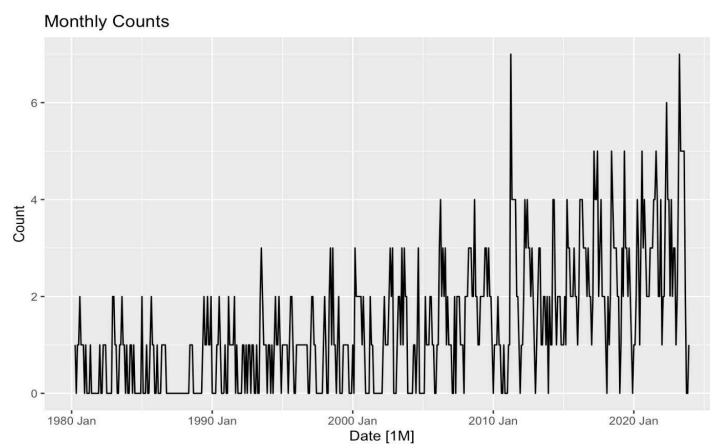
To obtain yearly data, we had to conduct various transformations to obtain counts and costs data. After loading our CSV data in R, we had to perform some data aggregation and grouping to obtain the yearly values as our initial data set is an event based dataset. First, we created a “Counts” helper column with values of “1” as mentioned in section 2.1.1 to be able to calculate disaster counts. For our grouping, we chose to use the end date based on the assumption that disaster costs are incurred after the disaster has ended. We extracted the year

value from the end date, and used it as our grouping field and summed the adjusted costs and the counts to obtain the yearly adjusted costs and yearly disaster counts respectively. These two fields are going to serve as the main data for our yearly counts and costs analysis as well as forecast..

## 2.2 Monthly Overview

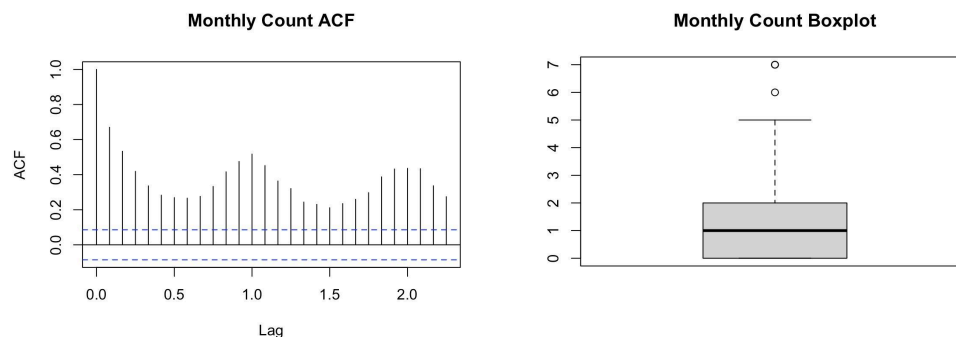
### 2.2.1 Monthly Counts

To start, the monthly count data was plotted. The time series plot below represents the total count of disasters each month from 1980 - 2023.



**Figure 5: Monthly Count Time Series**

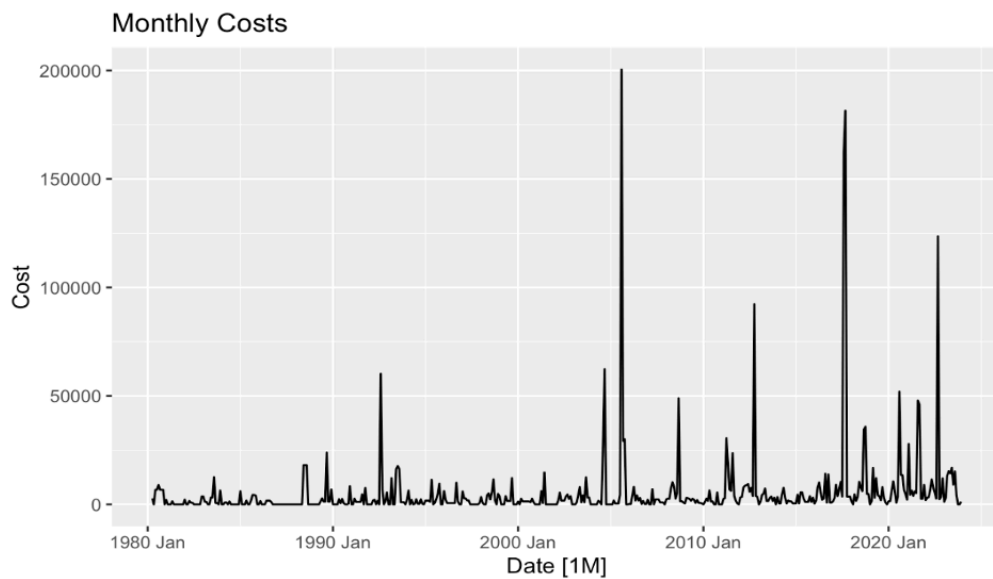
Looking at the figure above, there is most likely some upward trend, but seasonality is hard to tell without further examining the data. Next, an ACF plot and boxplot were modeled to better understand the data's autocorrelation and outliers.



**Figure 6: Monthly Count ACF**

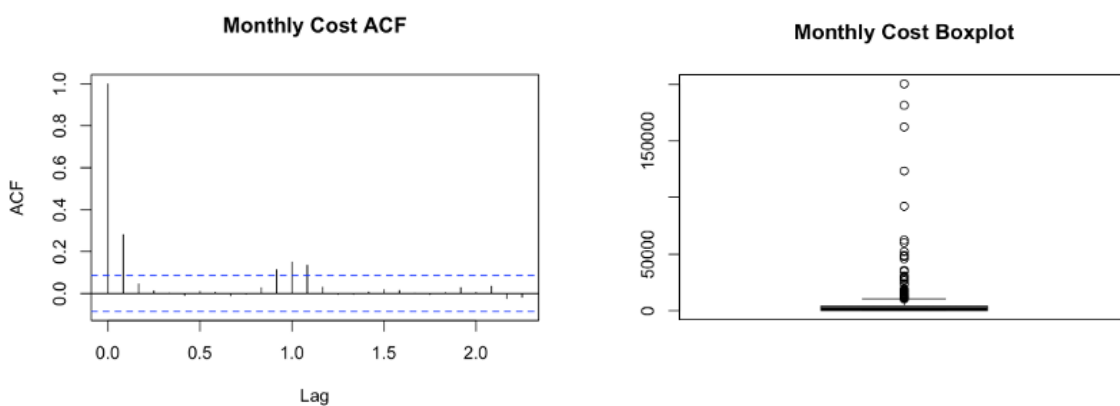
The ACF plot not only confirms some upwards trend with a high initial autocorrelation, it also suggests some seasonality due to the peaks and troughs in autocorrelation throughout the lags. Looking next at a boxplot for the data, there is a definite right skew with 2 clear outliers, but the data looks promising for further analysis. Next, the monthly cost data was examined. The time series below represents the total costs each month from 1980 - 2023.

### 2.2.2 Monthly Costs



**Figure 7: Monthly Cost Time Series**

Looking at the figure above, it is difficult to determine whether any seasonality or trend exists. Modeling an ACF plot and boxplot will be helpful in further understanding this data.



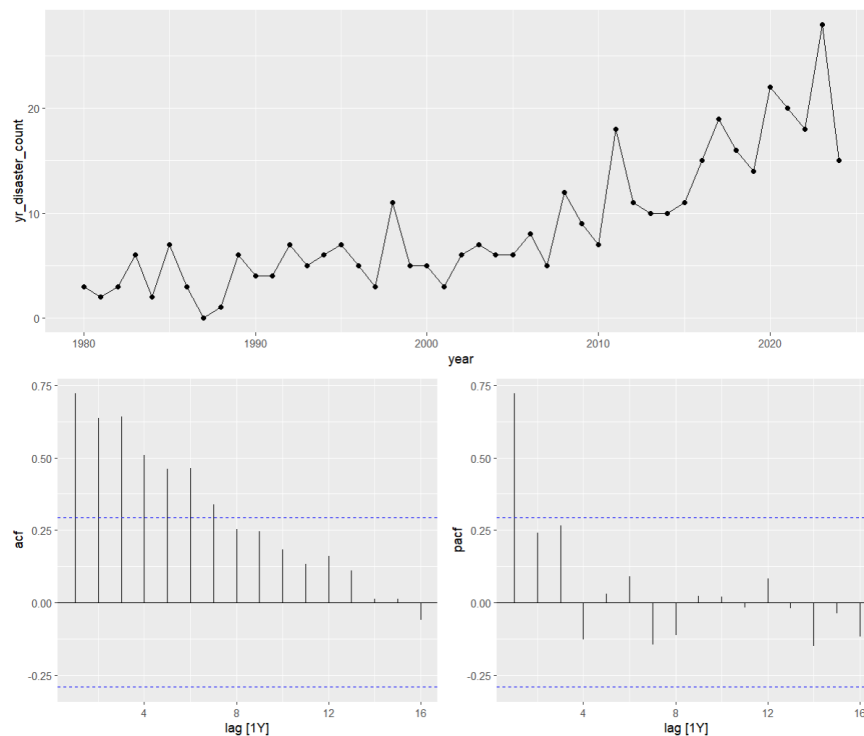
**Figure 8: Monthly Cost ACF**

Looking at the ACF plot above, we see some trend, but little to no meaningful seasonality. In addition, the boxplot is extremely right skewed, with plenty of outliers. This dataset is going to be difficult to work with, so alongside the reasoning given in section 2.1.2, monthly cost data will be used more so for auxiliary analysis to support findings for counts but not as primary analysis.

## 2.3 Yearly Overview

### 2.3.1 Yearly Counts

First of all we plotted some graphs of the counts to get a feel for the data. The data is from 1980-2024 but we choose to filter out the data from 2024 since it was not yet complete. Let's first take a look at the graphs:



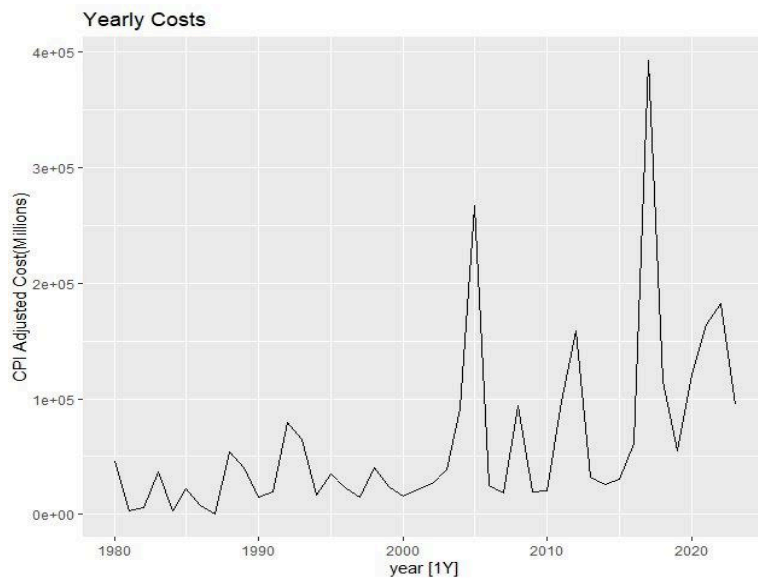
**Figure 9: Yearly Count Time Series and ACF**

In the top graph, which plots the counts over time, it is visible that there is a clear upwards trend in the data. It is also visible that there is a lot of variability in the data. Next we turn our attention to the bottom left graph, where a high autocorrelation for the lower lags can be observed. This means that the past (although more recent) values influence the future values.

Finally we direct our attention to the bottom right graph, here a big drop in the PACF, immediately after lag 1 can be observed. This indicates that a model where the future values mostly depend on the most recent past values, like an AR(1) model would be good.

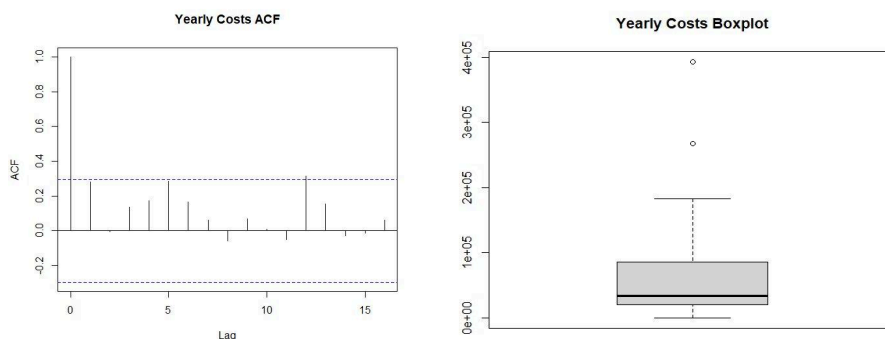
### 2.3.2 Yearly Costs

Yearly Cost data was plotted similarly to the other datasets between the years 1980 and 2023. Looking at the plots we can see some interesting observations.



**Figure 10: Yearly Cost Time Series**

By looking at the graph, we can see that the data has a slight upwards trend that we are going to confirm that later on in our analysis and we can distinguish that there is not a lot of seasonality but to solidify our observation we can plot an ACF plot.



**Figure 11: Yearly Cost ACF and Boxplot**

Looking at the ACF, we can see a peak in lag 1 then remotely no significant values after that indicating that there is no strong autocorrelation between the cost values from year to year, therefore no apparent seasonality. Additionally, we can see that the boxplot reflects two outliers, based on our information, these two values fall in the year 2005 and 2017. We will continue exploring this dataset in the yearly cost analysis portion of the report.

### **3.0 Analysis**

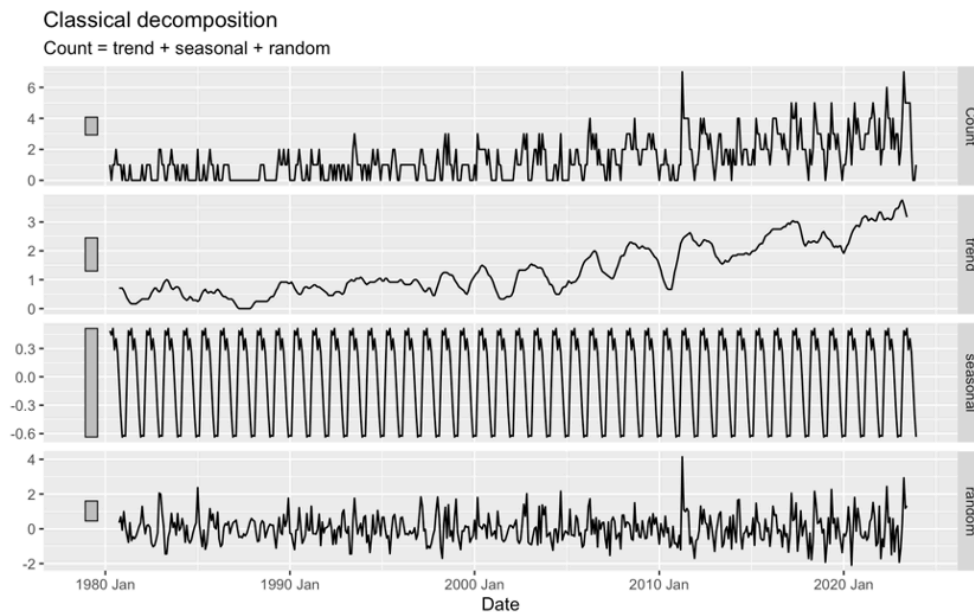
As mentioned in the introduction, the analysis section is split into two main timeframes: monthly and yearly. Broadly, the monthly analysis will provide insights pertaining to monthly seasonality, whereas the yearly analysis will provide insights to trend through comparing multiple forecasting models.

#### *3.1 Monthly Analysis*

To begin, the monthly analysis will precede the yearly analysis since we are conducting it as a form of descriptive analytics. There will be no forecasting done in this section, since the main focus is purely to understand past seasonality and how it has changed from 1980 until now. In order to understand both those questions, we decided to use two decompositions that were capable of isolating seasonality: a classical decomposition and an STL decomposition. First, the classical decomposition was chosen to measure the overall seasonality of the model since its seasonality component does not change over time. This is tremendously helpful in getting a high-level view of which disasters are the most common as well as the correlation between disasters and cost. Next, the STL decomposition since it changes over time, thus allowing us to gain a better understanding of how seasonality is changing over time. We will be able to understand whether different times of year are becoming more common, as well as whether the seasonality is increasing overall.

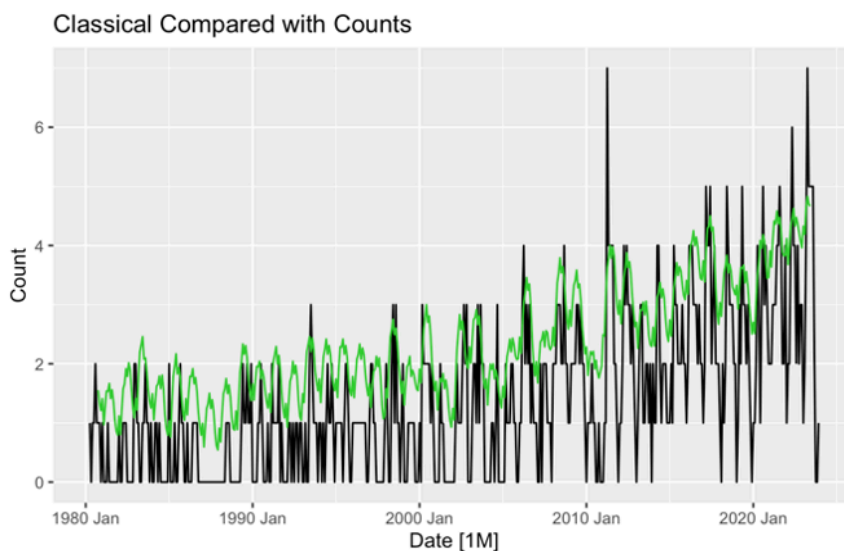
##### 3.1.1 Stationary Seasonality

To start, a classical decomposition was applied to the data. We tried applying a multiplicative classical decomposition, but it ended up having a higher overall error, so we ended up using an additive model. Found below are all the components of the decomposition.



**Figure 12: Classical Decomposition of Monthly Counts**

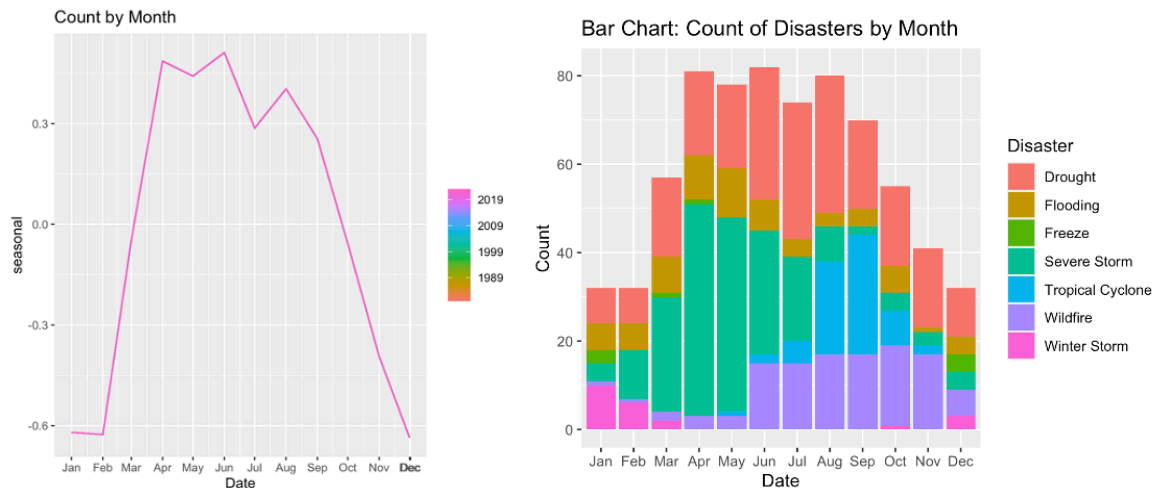
Although the trend is not the main focus of the decomposition, the decomposition shows a relatively steady upwards trend. In addition, it also shows a clear seasonality component that isn't visible when looking solely at the time series. Looking at the residuals, there is no clear pattern, however there seems to be a few outliers that skew the data, as well as a non-zero mean. To further understand whether the classical decomposition is capturing the seasonality properly, it was plotted next to the original time series.



**Figure 13: Monthly Count Time Series Layered with Classical Decomposition**

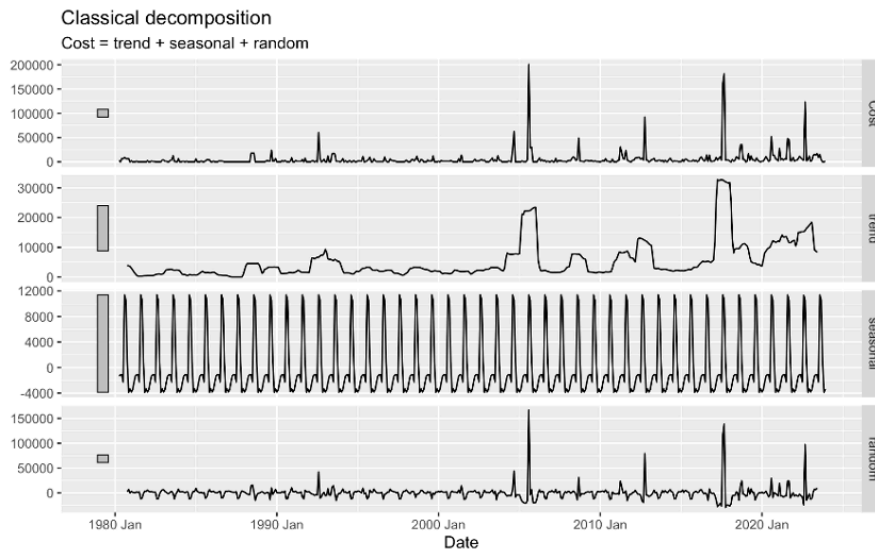


As we can see, the monthly seasonality seems to follow the trends of the original time series quite accurately but doesn't seem to capture the increasing intensity of the seasonality over time. Although the classical decomposition doesn't explain how the seasonality changes over time, it is still extremely valuable in understanding the overall seasonality of climate disasters, especially as it pertains to specific disasters.



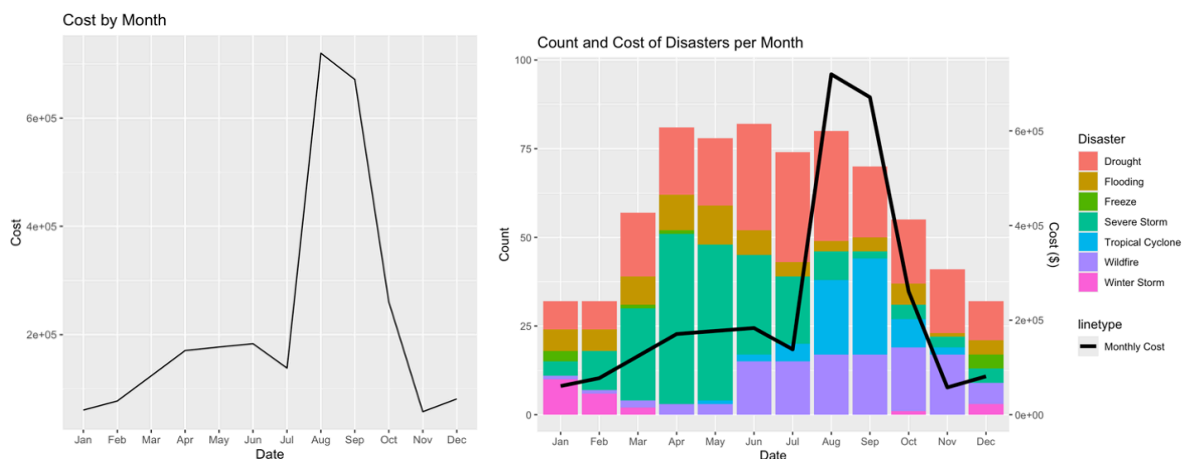
**Figure 14: Monthly Count Seasonality and Bar Chart**

Looking closer at the seasonal pattern provided by the classical decomposition, we can see that there are generally two peaks in the year, with one occurring in the early summer and one occurring in the late summer. Once we compare this seasonality with a bar chart outlining each type of disaster and their overall counts, we can see that the summer months not only have higher counts in terms of types of disasters, but those disasters also occur more frequently than those in the winter months. The most frequently occurring disasters seem to be severe storms, droughts, and tropical cyclones. Although these events occur more frequently, a crucial next step in understanding them is to determine whether they are correlated to seasonal costs. In order to better understand this, a classical decomposition of seasonal costs was plotted.



**Figure 15: Monthly Costs Classical Decomposition**

Looking at the classical decomposition, it is clear to see that the model is extremely influenced by outliers. Both the trend and residual components point towards the fact that the model is not usable in any circumstances, however the seasonality component does provide us with a value that we can compare to the counts data.



**Figure 16: Monthly Costs Seasonality and Correlation with Counts**

Isolating the seasonal component of the decomposition, we can see a clear spike in cost during the late summer. When plotted on top of the bar chart, we can see that it seems to be most commonly associated with tropical cyclone season. In order to confirm this hypothesis, a quick multiple linear regression model was run using cost as dependent variable and types of disasters as the independent variables.

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1562.2      868.6   1.799  0.0726 .
DisasterFlooding    1579.1    1864.8   0.847  0.3974
DisasterFreeze     2564.0    4485.4   0.572  0.5678
DisasterSevere Storm 1882.5    1430.2   1.316  0.1886
DisasterTropical Cyclone 24576.3 1995.5 12.316 <2e-16 ***
DisasterWildfire   -293.6    1506.6  -0.195  0.8456
DisasterWinter Storm 3205.0    3008.9   1.065  0.2872
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13200 on 622 degrees of freedom
Multiple R-squared:  0.2095,    Adjusted R-squared:  0.2019
F-statistic: 27.48 on 6 and 622 DF,  p-value: < 2.2e-16

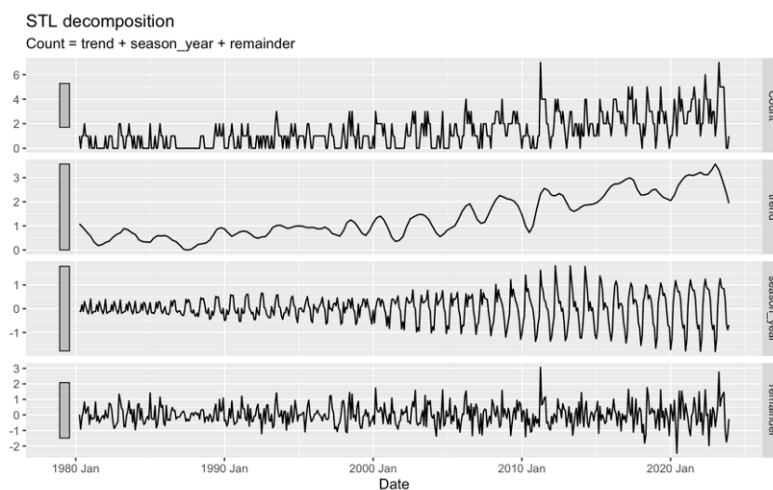
```

**Figure 17: Code Output for Linear Regression Model**

Based on the outputs above, we can clearly see that tropical cyclones have a strong correlation ( $p \leq 0.001$ ) with cost.

### 3.1.2 Variable Seasonality

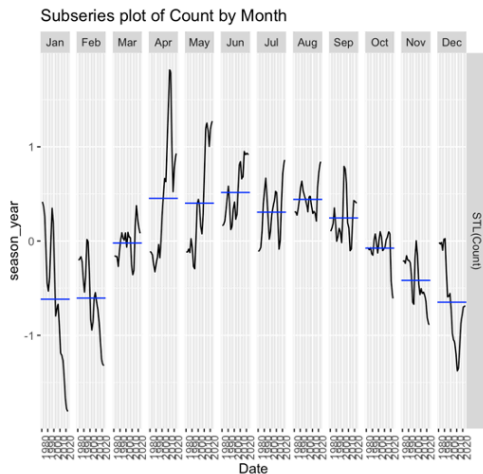
After determining that there is indeed some change in intensity of seasonality over time for counts, the next intuitive step is to measure how exactly the seasonality is changing. In order to do so, an STL decomposition was fitted to the data. The main advantage of using an STL decomposition as opposed to the classical method is that it can actually measure the change in seasonality over time, which proves extremely helpful in our case.



**Figure 18: STL Decomposition of Counts**

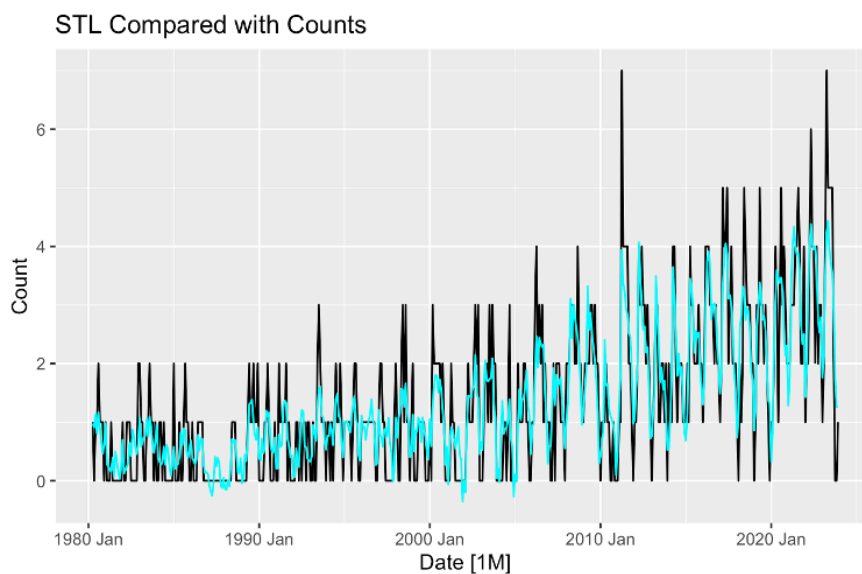
Starting with the trend component, the STL provides a similar value to the classical decomposition, although more smoothed. One thing to note is that if we wanted to forecast this data, we might want to apply a box cox transformation to the trend component since it

looks somewhat non-linear. The residuals also seem to be smaller and closer to 0, although In addition, there seems to be a clear increase in seasonality over time. Although this may be due to the outlier occurring in 2011, it looks to be a shift from two peaks in the summer to one larger peak occurring in the early summer months. We can examine these changes more closely through the usage of a subseries plot.



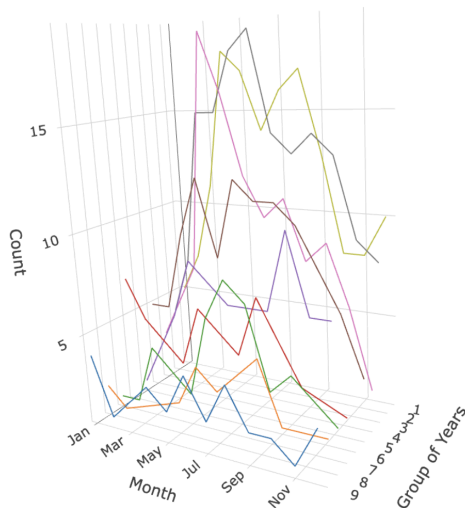
**Figure 19: Subseries Plot of STL Decomposition**

As mentioned above, the subseries plot indeed shows the early summer months seasonality component increasing over time, while the later some months tend to stay relatively similar. With this in mind, we can plot the decomposition against the original time series.



**Figure 20: Monthly Count Time Series Layered with STL Decomposition**

As we can see from the figure above, the STL decomposition fits the data much better than the classical decomposition does. It is able to effectively capture the increase in seasonality over time, however, it does begin to underestimate the seasonal peaks of the data as seasonality increases. As mentioned earlier, applying a box-cox transformation might be useful if the STL decomposition were to be chosen for forecasting purposes. The final chart created for the monthly analysis includes a 3 dimensional representation of how the peaks in seasonality have changed as well as how the counts have changed overall. This wasn't a necessary chart as all the relevant business questions have been answered but rather just a novel way of representing the data and further solidifying the conclusions found above. There is also an interactive website that can be accessed which showcases the dataset (<https://sites.google.com/view/adm4307/home>).



**Figure 21: Three-Dimensional Plot Measuring Seasonality by Grouping of Years**

### *3.2 Yearly Analysis*

#### 3.2.1 Yearly Counts

Now we will move on to some prediction models for the yearly counts. We tried 5 models to find the best one, the results are as follows: (please note that the ARIMA is an  $\text{arima}(0,1,1)$  model).

```
# A tibble: 5 × 10
```

	.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	ARIMA	Test	7.84	9.56	7.95	43.1	44.8	NaN	NaN	0.265
2	ETS	Test	6.79	8.48	6.89	33.7	34.7	NaN	NaN	0.253
3	Mean	Test	11.3	12.4	11.3	62.3	62.3	NaN	NaN	0.253
4	Naive	Test	7.3	8.89	7.3	36.9	36.9	NaN	NaN	0.253
5	TSLM	Test	6.05	7.55	6.08	30.1	30.5	NaN	NaN	0.166

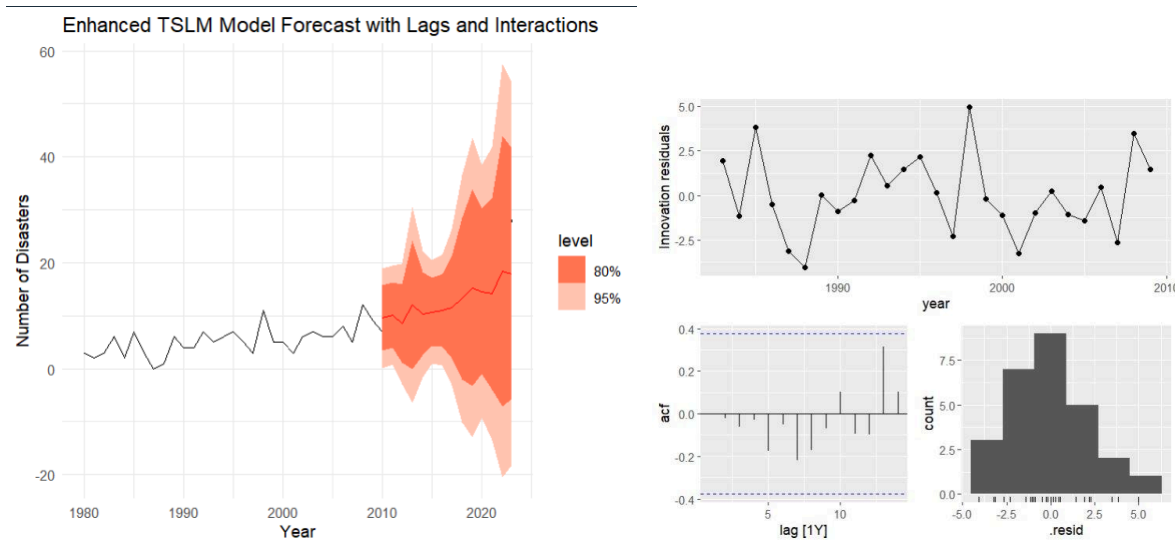
**Figure 22: Accuracy Metrics of the Yearly Counts Forecasts**

From the table, it can be observed that the TSLM model is the best, since the ME, RMSE and the MAE are the lowest. To try and make this forecast more accurate, we tried incorporating a lagged variable in the TSLM, as was suggested in the previous paragraph, however this yielded less accurate results. We also tried applying a much larger TSLM model that includes multiple lagged variables as well as interaction terms, the formula of which is shown below this paragraph, however since this is a bit beyond the scope of the course we will not go into too much detail about this.

$$\text{Disaster Count at time } t = \beta_0 + \beta_1 \cdot (\text{Trend at time } t) + \beta_2 \cdot (\text{Lag 1 at time } t) + \beta_3 \cdot (\text{Lag 2 at time } t) + \beta_4 \cdot (\text{Lag 3 at time } t) + \beta_5 \cdot (\text{Lag 1 at time } t \times \text{Trend at time } t) + \beta_6 \cdot (\text{Lag 2 at time } t \times \text{Trend at time } t) + \epsilon_t$$

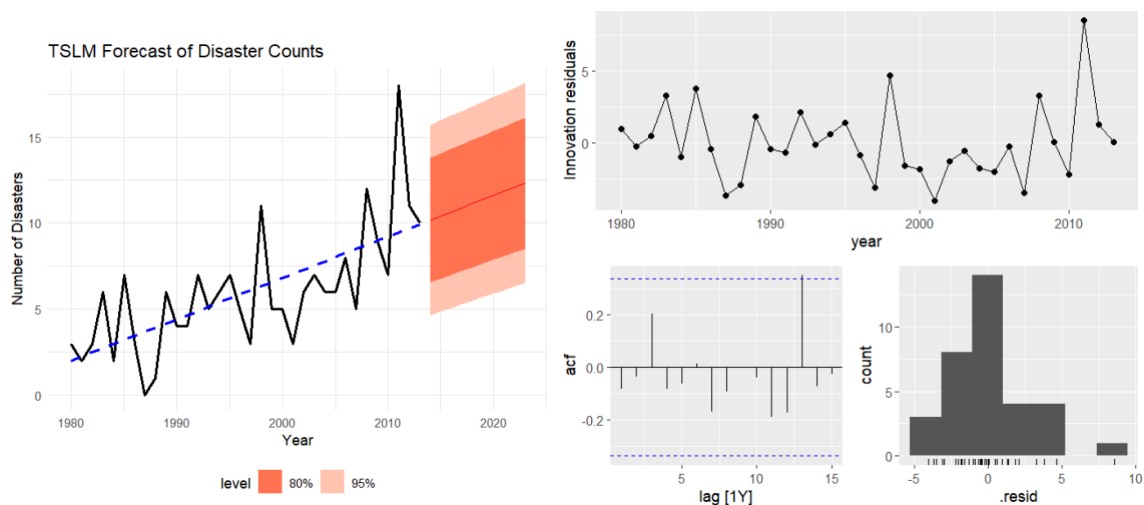
	RMSE	MAE	MAPE	SMAPE	ME
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	5.00	3.92	23.1	26.1	-2.95

**Figure 23: Equation and Accuracy Metrics of Enhanced TSLM forecast**



**Figure 24: Enhanced TSLM Forecast of Yearly Counts with ACF and Residual Plot**

We can however see that this model yields more accurate results which shows that with more advanced techniques the results can improve. The standard TSLM model, as mentioned before, yielded the best results and hence we choose this for our predictions.



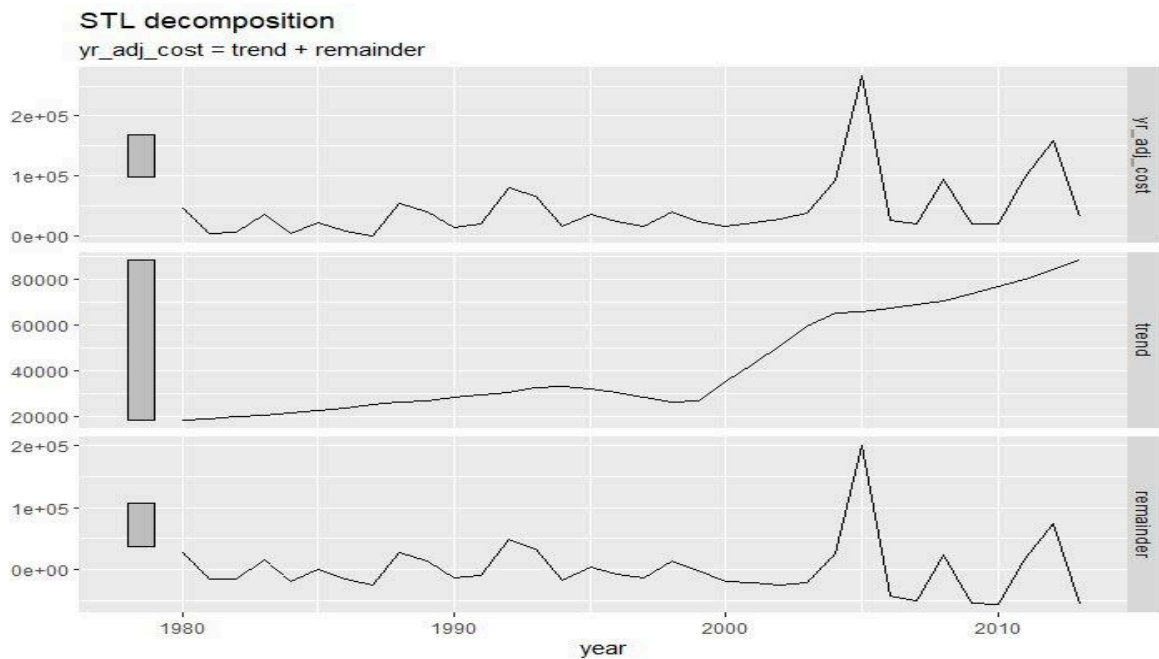
**Figure 25: TSLM forecast of Yearly Counts with ACF and Residual Plot**

Above we can see the graphs of the TSLM model, the graph on the left shows the prediction and the graph on the right shows the ACF and the plot of the residuals. If we look at the ACF, we can see that everything stays between the blue dotted lines, except for lag 13, where it slightly exceeds the line. However, this is more likely a coincidence rather than a trend that we could adjust our forecasts on, since it is very far into the past.

### 3.2.2 Yearly Costs

#### 3.2.2.1 Decomposition

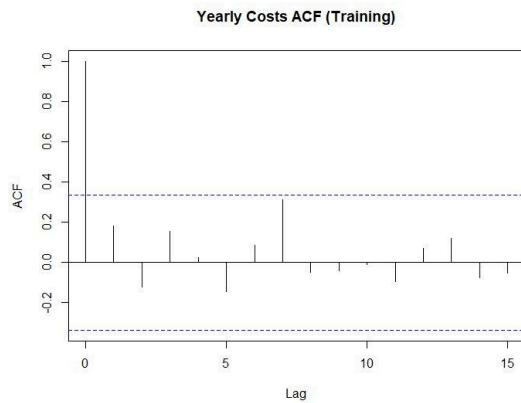
Similar to the yearly counts analysis, the data was allocated following an 80/20 training and testing split, so data points from 1980-2013 were used to train our models while points from 2014-2023 were used in testing. First, to have a better look at our training split, we perform an STL decomposition along with plotting the ACF.



**Figure 26: STL Decomposition of Yearly Cost Data (Training)**

Looking at the decomposition, we can see that our training data have a slight upwards trend with the slope increasing at around the year 1999. Additionally, we see that our remainder data mimics the peaks of the main data, indicating that our data does not have any specific predictable pattern.





**Figure 27: ACF Plot of Yearly Cost Data (Training)**

From the ACF plot of the training data, we can confidently say that our training data does not exhibit any seasonal pattern through the years. This can help us determine which models will most likely fit our data.

### 3.2.2.2 Model Analysis

Following the previous decomposition analysis, we chose to fit our training data using three models as well as mean and naive models as benchmarks:

- ARIMA(0,1,1): this is equivalent to a Moving Average model with first degree of differencing. This model was chosen since our data was not stationary (upwards trend), and did not contain a significant amount of autocorrelation.

```
Series: yr_adj_cost
Model: ARIMA(0,1,1)

Coefficients:
    ma1
    -0.8651
s.e.    0.0884
```

**Figure 28: ARIMA(0,1,1) Report**

- ETS(A,N,N): Also referred to, as a Simple Exponential Smoothing model. This model was chosen since our data does not have any form of seasonality and a weak upwards trend. In our analysis, we compared this model to a Holt Linear (ETS(A,A,N)) that factors the trend but the SES was a better fit.

```

Series: yr_adj_cost
Model: ETS(A,N,N)
Smoothing parameters:
  alpha = 0.1129713

Initial states:
  l[0]
22961.46

sigma^2: 2781362613

```

**Figure 29: ETS(A,N,N) Report**

- TSLM (with trend): The third model is a linear regression using the trend as an explanatory variable. This model was chosen as our data did not have any autocorrelation that can be captured in the previous models so relying on an explanatory variable would potentially yield better outcomes in terms of fit.

```

Series: yr_adj_cost
Model: TSLM

Residuals:
    Min       1Q   Median       3Q      Max
-50939 -26791 -13529  23606 206318

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7853.4    17014.7   0.462   0.6475
trend()       2037.6     848.1    2.403   0.0223 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48520 on 32 degrees of freedom
Multiple R-squared:  0.1528,    Adjusted R-squared:  0.1263
F-statistic: 5.773 on 1 and 32 DF, p-value: 0.022252

```

**Figure 30: TSLM Report**

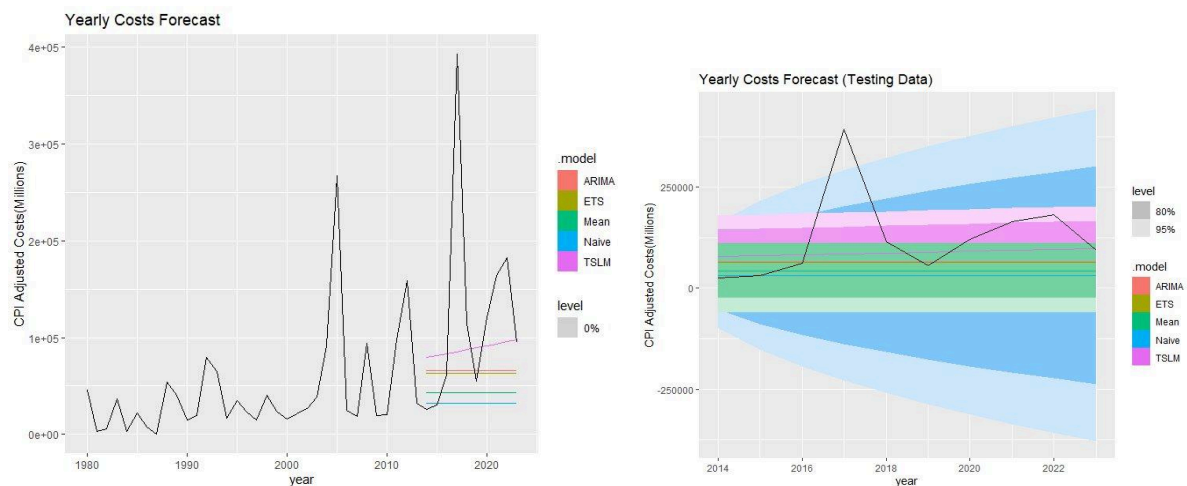
After fitting these models to our yearly costs, we obtain the following results:

	.model	sigma2	log_lik	AIC	AICc	BIC	ar_roots	ma_roots	MSE
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<list>	<list>	<dbl>
1	ARIMA	2786713756.	-406.	816.	816.	819.	<cpl>	<cpl>	NA
2	ETS	2781362613.	-429.	863.	864.	868.	<NULL>	<NULL>	2.62e9
3	TSLM	2353748048.	-414.	738.	738.	742.	<NULL>	<NULL>	NA
4	Naïve	4541354698.	NA	NA	NA	NA	<NULL>	<NULL>	NA
5	Mean	2694150587.	NA	NA	NA	NA	<NULL>	<NULL>	NA

**Figure 31: Summary of fitted Models**

From the model summary, we can see that the linear regression model with trend fit the data the best followed by the MA model (ARIMA(0,1,1)) then SES(ETS(A,N,N)), as it has the lowest adjusted AIC and BIC.

Taking this further, we forecast using these models on the testing split, and we obtain the following results:



**Figure 32: Forecast of fitted models over testing split**

From the presented graphs we can see that the most accurate model is the Linear Regression model followed by MA, then SES. These results reflect exactly what we were able to conclude using the fit analysis in the previous step. Yet to solidify our answer, we can conduct an accuracy analysis over the testing data.

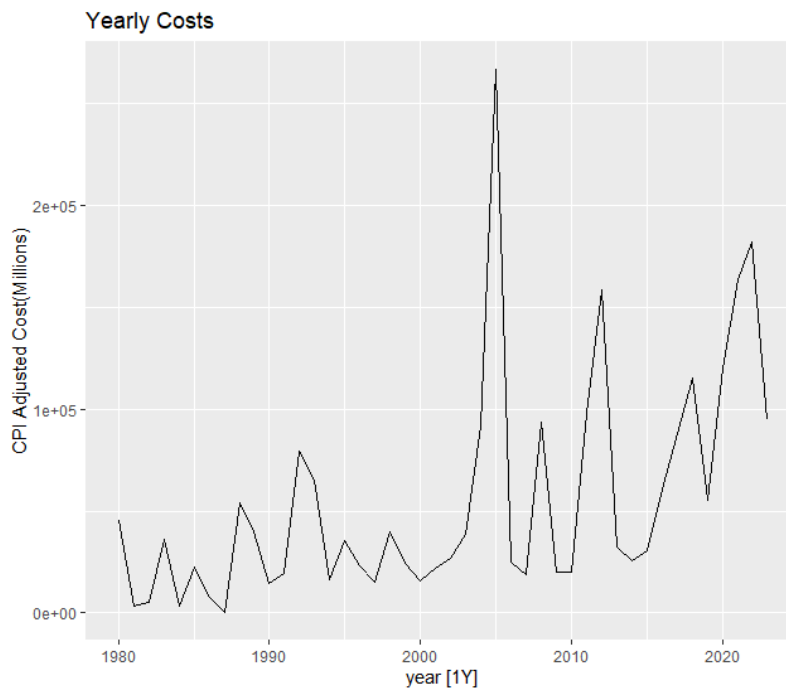
	.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	ARIMA	Test	58440.	118098.	76554.	2.66	62.5	NaN	NaN	-0.0275
2	ETS	Test	61306.	119543.	77127.	6.90	60.7	NaN	NaN	-0.0275
3	Mean	Test	80706.	130558.	86874.	35.6	58.2	NaN	NaN	-0.0275
4	Naive	Test	92084.	137882.	93701.	52.4	58.6	NaN	NaN	-0.0275
5	TSLM	Test	35878.	107629.	68515.	-25.9	69.4	NaN	NaN	-0.0444

**Figure 33: Accuracy Test over testing split**

From our accuracy test, we can clearly see that the most accurate model is the Linear Regression model with the lowest RMSE of 107629. Yet, it has a high MAPE, indicating that this model is not fit for forecasting. One of the reasons these results might not have the best accuracy might be the presence of outliers in our testing split. In the next section, we are going to attempt to improve this linear regression model.

### 3.2.2.3 Model Improvement

To improve our linear regression models, we took multiple steps. First, we decided to impute our outlier values in our yearly adjusted costs dataset as outliers might be affecting the accuracy of our models. Second, we experimented with various explanatory variables for our regression models to find the best fit and accuracy.



**Figure 34: Yearly Cost Plot with Imputed Outliers**

In this dataset, outlier values were identified and imputed using moving average method then we fitted multiple linear regression models on its training split (1980-2013):

```
tslm_model <- training_data_av %>% model(TSLM1 = TSLM(yr_adj_cost ~ trend()),
  TSLM2 = TSLM(yr_adj_cost ~ trend(knots = 1999)),
  TSLM3 = TSLM(yr_adj_cost ~ yr_disaster_count),
  TSLM4 = TSLM(yr_adj_cost ~ trend() + yr_disaster_count),
  TSLM5 = TSLM(yr_adj_cost ~ trend(knots = 1999) + yr_disaster_count),
  TSLM6 = TSLM(yr_adj_cost ~ yr_disaster_count + Average_Fahrenheit_Temperature))
```

**Figure 35: Testing Linear Regression Model Improvement**

- TSLM1: Our benchmark model from our previous analysis. Trend as an explanatory variable.
- TSLM2: Piecewise regression factoring the knot in data trend at the year 1999 as mentioned in section 3.2.2.1.
- TSLM3: Linear Regression relying on disaster counts as an explanatory variable

- TSLM4: Linear Regression model combining both trend and disaster counts
- TSLM5: Piecewise Regression model combining both the trend (with knot at 1999) and disaster counts
- TSLM6: Linear Regression combining disaster counts as well as US Temperature Data as the second explanatory variable.(U.S. climate normals, 2024)

After fitting these models, we obtain the following results:

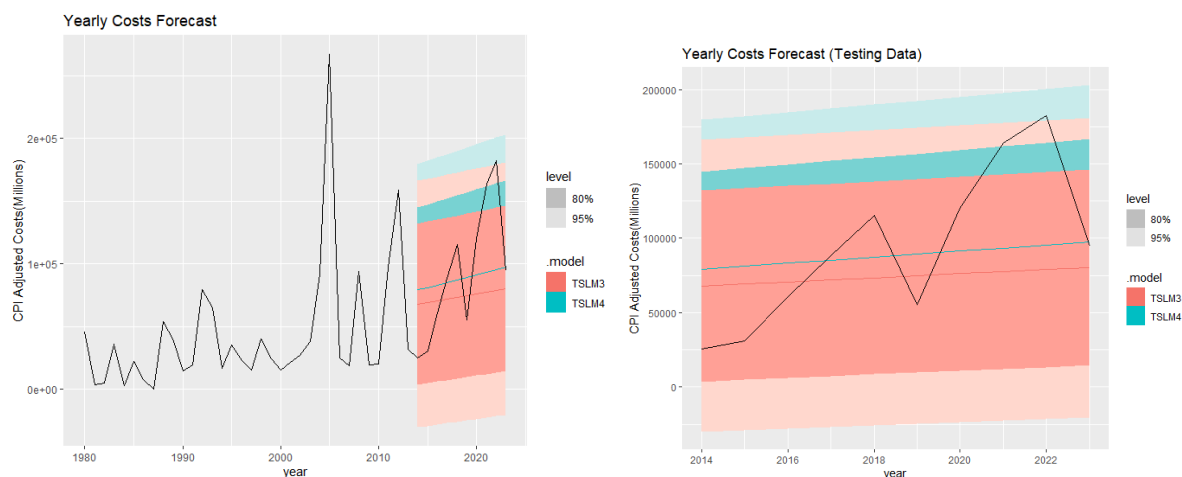
```

.model r_squared adj_r_squared sigma2 statistic p_value df log_lik
<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl>
1 TSLM1 0.153 0.126 2.35e9 5.77 0.0223 2 -414.
2 TSLM2 0.165 0.111 2.39e9 3.06 0.0610 3 -414.
3 TSLM3 0.157 0.131 2.34e9 5.96 0.0203 2 -414.
4 TSLM4 0.185 0.133 2.34e9 3.52 0.0418 3 -413.
5 TSLM5 0.189 0.108 2.40e9 2.33 0.0942 4 -413.
6 TSLM6 0.167 0.114 2.39e9 3.12 0.0584 3 -414.
# i 7 more variables: AIC <dbl>, AICC <dbl>, BIC <dbl>, CV <dbl>,
# deviance <dbl>, df.residual <int>, rank <int>

```

**Figure 36: Fitness test of Linear Improved Linear Regression Models**

By comparing the adjusted R2, we find that TSLM3 and TSLM4 have the best results, yet given their low scores we can infer that they are not the strongest models. We can still proceed with forecasting using these models to determine accuracy and forecasting power of these two models.



**Figure 37: Forecast of Selected Linear Regression Models over Testing Split**

Looking at the forecast, we can see that there is definitely an improvement over the previous models in terms of accuracy yet the prediction intervals are still considerably wide which hints towards low forecasting power.

```
# A tibble: 2 × 10
  .model .type    ME    RMSE    MAE    MPE    MAPE    MASE    RMSSE    ACF1
  <chr>   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 TSLM3 Test 19605. 51037. 41629. -12.8  55.9    NaN    NaN  0.463
2 TSLM4 Test  5412. 46084. 38049. -33.4  61.9    NaN    NaN  0.439
```

**Figure 38: Accuracy Test of selected linear regression models over testing split**

Looking at the accuracy test, we can see that the RMSE has improved dramatically from our previous linear regression model, decreasing from a score of 107629 to 5412 during the best case scenario (TSLM4). Nevertheless, we do not see any considerable improvement in the MAPE score indicating low forecasting power. Although the TSLM4 model has an improved RMSE score, and is relatively accurate over the testing split, this model cannot be used for accurate forecasting. In the conclusion, we will discuss why this may be the case and what are the necessary improvements that can potentially solve this issue.

## 4.0 Conclusion

### 4.1 Revisiting Business Questions

To conclude, it first makes sense to revisit the business questions introduced in section 1.2 in order to see what both the monthly and yearly analysis has accomplished.

Monthly Questions	Results
Is there seasonality?	We found there to be a clear seasonality for disaster counts in the summer months. Specifically, there are two peaks: one in the early summer and one in the late summer.
What times of year do specific disasters occur?	It generally depends on the disaster, and the location that the stakeholders are living in.
Is there a correlation between disaster types and cost?	There is a strong correlation between tropical storms and costs. Because of the disparity between tropical storm costs and other disasters, no other disasters were seen as statistically significant.
Is seasonality changing over time?	Seasonality is changing over time since it is becoming more intense as time progresses. The actual times of year where disasters are common do not change significantly, however.

**Figure 39: Conclusion Summary of Monthly Analysis**

Yearly Questions	Results
Is there seasonality/trend?	Based on our analysis, yearly disaster data follows a slight upwards trend and some kind of seasonality. As for cost data, we observe a slight upwards trend with no seasonality between the yearly values.
Can we forecast yearly costs/counts?	To be able to forecast both costs and counts, we will require more data to be able to accurately forecast both values. Specifically for yearly costs, incorporating location data would potentially yield better results as costs might be higher in urban areas instead of rural areas.
Are there any additional explanatory variables that may help with forecasting?	Additional explanatory variables that would prove useful would be count of disaster per type, location data (urban, semi-urban, rural), also we can potentially have some correlation between intensity of the event (Saffir-Simpson Hurricane Wind Scale etc.).
Is there any correlation between disaster costs and counts?	From our analysis, we can see that there is a relationship between yearly costs and yearly counts of disasters. This relationship is positive, in the case that disaster counts increase per year, yearly disaster costs increase.

**Figure 40: Conclusion Summary of Yearly Analysis**

#### *4.2 Limitations and Next Steps*

While writing this report, we encountered limitations that are well-documented above, specifically when looking at yearly data. Firstly, the data itself presented a challenge for two main reasons. The first reason is that there just simply wasn't enough data to warrant a full yearly analysis considering the intensity of the outliers. The second reason is that the data is missing a crucial element: location data. Location data would be able to make our forecasts much more accurate overall since it would provide us with not only a greater quantity of data but also relevant data that can be used to make more accurate decisions. The omission of location data also negatively impacted our results in regards to our business questions. For example, monthly seasonality across the United States differs widely depending on the region, so to say that there is only one peak in seasonality is a somewhat inaccurate statement

when various stakeholders are taken into account. Secondly, we found that splitting the work over both months and years based on the data we had ended up making the analysis more shallow and less focused on answering a strong business question. If we focused specifically on one of the two timeframes, we might have been able to deliver more pertinent analysis specifically regarding the stakeholders involved.

In the future, we would likely implement two things in order to make our analysis stronger. First, we would most likely find a way of incorporating location data into our forecasts, potentially using machine learning techniques to process the large quantity of data in order to create better insights. Next, we would choose to focus on a specific time period for further research, most likely monthly data since there is a clear seasonality and would be easier to forecast overall.



## References

- Crimmins, A. R., & Singh, D. (2023, November 14). *Focus on Compound Events*. Fifth National Climate Assessment. <https://nca2023.globalchange.gov/chapter/focus-on-1/>
- National Centers for Environmental Information. (2024). *Billion-Dollar weather and climate disasters | national centers for environmental information (NCEI)*. [Www.ncei.noaa.gov. https://www.ncei.noaa.gov/access/billions/](https://www.ncei.noaa.gov/access/billions/)
- National Oceanic and Atmospheric Association. (n.d.). *Webpack App*. Data.noaa.gov. <https://data.noaa.gov/onestop/>
- United nations. (2023). *What is climate change?* United Nations; United Nations. <https://www.un.org/en/climatechange/what-is-climate-change>
- U.S. climate normals. (2024, November 19). National Centers for Environmental Information (NCEI). <https://www.ncei.noaa.gov/products/land-based-station/us-climate-normals>